Frequency Distributions

A frequency distribution is a set of possible values of a variable, together with the associated frequencies. It is a type of table used to organize data. The left column include all possible values of the variable being studied. The right column is a list of the frequencies or number of observations for each value or class. It's an organized tabulation of the number of observations located in each category on the scale of measurement.

Frequency distribution can be classified into:

i) Simple frequency distributions – also called discrete frequency distribution.
ii) Grouped frequency distribution – also called continuous frequency distribution.

Discrete frequency distributions

In a data set where the values are largely repeated, we can prepare a table showing the different values of the variable along with the number of repetitions of each value.

Illustration 1:

The following distribution shows the weekly wages (in thousands) of employees of a certain firm.

10	12	11	15	16	15	14	20	21	19
18	17	16	14	15	13	15	11	21	19
12	14	15	20	19	18	16	16	17	18
18	16	15	14	12	11	17	19	13	12
9	10	17	16	18	19	16	17	9	14

Required:

(i) Construct a discrete frequency distribution.
(ii) Determine the number of employees
(iii) Determine the total weekly wage bill.

Procedure:

1. In the first column of the table, list all the values that are included in the data set without repetition in ascending order i.e from smaller to largest.
2. In the second column, put a tally mark against each value whenever it occurs in the data set. To facilitate counting, tally marks are arranged in blocks of five by crossing the four tally marks already marked by the fifth one.
3. Count the tally marks and write the frequencies in the third column.

Relative and percentage frequency distribution

Sometimes it is desirable to express the frequencies in the form of proportion or percentage of the total frequency. This may be useful for comparing the distributions of frequencies for two or more data sets of unequal size, since the actual frequencies would tend to be greater for each class within the larger group than for a class in the smaller one.

The relative frequency of an observation is obtained by dividing the frequency of the observation by the total frequency. Proportions describe the frequency in relation to total number of scores.

The percentage frequency of an observation is obtained by expressing the frequency of that observation as a percentage of total frequency. The percentage frequency is obtained by dividing

each frequency by the number of observations and multiplying the resulting proportion by 100. A relative or percentage frequency describes the proportion of data values that fall within each category.

Cross tabulations

Cross tables are also called cross tabs or contingency tables. A cross table is used to display a summary of the distribution of two or more variables.

Illustration 2:

The following data represents the monthly wages (in thousands) of employees of a small company.

G	W	G	W	G	W	G	W
F	70	M	75	M	72	M	72
M	72	M	70	F	70	M	70
M	70	F	75	F	70	F	72
F	76	F	60	M	75	F	75
F	70	M	60	M	72	M	72
M	75	F	70	F	72	M	72
F	60	M	75	M	74	M	74
F	60	F	74				

Column headings: G – gender, W-wage
Row headings : M – Male, F – Female

Required:

(i) Prepare the wage frequency distribution table
(ii) Once completed, include the relative frequency and the percentage frequency.
(iii) Prepare a cross-tab with the gender variable in the columns and wage in the rows.
(iv) What percentage of the males earned sh.70,000 per month.
(v) Determine the total monthly wage bill.
(vi) Describe the total month wage bill for female employees.

Illustration 3:

The total monthly energy consumption for each of the 20 manufacturing firms in a certain town are shown below:
587, 350, 584. 292, 424, 726, 512, 705, 765, 528, 730, 682, 586, 630, 489, 655, 634, 666, 629, 660.

Required:

i. A simple frequency distribution
ii. Determine the total monthly energy consumption for all firms.
iii. A grouped frequency distribution starting with class 250-349 and using a class width of 100.

Grouped frequency distributions

A grouped frequency distribution consists of class intervals and their corresponding frequencies. It is a table that divides the data values into classes and shows the number of observed values that fall into each class.

A grouped frequency distribution is used when a set of data covers a wide range of values. The observations are spread out over a wide range such that few observations, if any, are repeated.

Definitions associated with grouped frequency distribution

1. **Class limits**

   These are the values that define the observations to be included in each class. Every class is defined by two values; the lower class limit and the upper class imit. The lower class limit represents the smallest data value that can be included in the class. The upper class limit represents the largest data value that can be included in the class, i.e class limits determine which data values are assigned to each class.

   Table 1.3: Energy consumption

Class	Frequency	Midpoint	Class total
250-349			
350-449			
450-549			
550-649			
650-749			
750-849			

2. **Class midpoints**

   In the process of forming a grouped frequency distribution, the individual observations lose their identity when grouped into classes. We make the assumption that grouped observations are uniformly distributed between the class limits. Based on this assumption, each observation is represented by the value that is situated at the centre of the class. Class midpoint is the value that is situated at the centre of a class and is given by the average of the lower and upper limits.

   $$\text{Class midpoint} = \frac{\text{Lower class limit} + \text{Upper class limit}}{2}$$

3. **Class width**

   Class width is the difference between the lower class limit of one class and the lower class limit of the next class.

4. **True class limits** – also called class boundaries. Class boundaries are used to separate the classes so that there are no gaps in the frequency distribution.

Conversion of class limits to class boundaries

1. Find the difference between the upper limit of a class and the lower limit of the next class interval.

2. Half of the difference is added to the upper limit of each class and the remaining half is deducted from the lower limit of each class interval.

NB: Class width is the difference between the lower and the upper class boundaries of a class.

Steps in the construction of a grouped frequency distribution.

1. Find the smallest and largest observations in the data set and find the range.
2. Decide on the number of classes into which data are to be grouped.

According to Sturge's rule, the approximate number of classes can be determined by the formula

$K = 1 + 3.3 \log n.$

where K is the number of classes and N is the total number of observations.

However, to get convenient class limits, we may choose slightly more or fewer classes i.e the above rule is a guideline, setting number of classes requires judgement.

3. Determine the approximate class width. The class width is obtained by dividing the range by the number of classes, i.e class width $= \frac{\text{Range}}{K}$ - The value obtained from this formula can be rounded off to a more convenient value.

To obtain nice class limits, round the class width up to an appropriate value. The width of a class should be a whole number conveniently divisible by 2, 5, or 10

4. Find the class limits of the lowest class. Set the lower limit for the first class as a multiple of the class width. The lower limit should be less than or equal to the lowest data value.
5. Obtain the class limits for the remaining classes by adding the class width to the limits of the previous class.
6. Count the data value in each class and enter in the frequency column.

Illustration 4:

The following distribution represents the marks of students in an examination.

59	16	20	10	11	38	10	23	13	18	31	16
20	12	11	7	24	11	27	17	42	13	17	14
37	19	22	16	16	14	19	26	26	28	50	9
10	10	18	21	21	15	37	13	8	58	11	23
31	17										

Required:

i. Prepare a grouped frequency distribution.
ii. Determine the actual total marks obtained by all students.
iii. Use the table obtained in (i) above to estimate total marks obtained by all students.

Exclusive and Inclusive Classes

Exclusive Classes

In exclusive classes the upper limit of one class is equal to the lower limit of the next class. This method ensures continuity between successive classes.

Examples

0 – 5 ⟶ means 0 and above but below 5

5 – 10 ⟶ means 5 and above but below 10

10 – 15 ⟶ 10 and under 15

15 – 20 ⟶ 15 and under 20

Observations equal to the upper limit of a class are excluded from that class and are included in the next class e.g. 15 is entered in class 15-20. Exclusive classes exclude all observations that are equal to the upper limit of the class.

Inclusive Class

In inclusive classes both the lower limit and the upper limit of a class are included in that class. The upper limit of a given class is not equal to the lower limit of the next class.

Example

0 – 4
5 – 9
10 – 14
15 – 19

Each class has 2 limits; the lower limit and the upper limit. The usual practice is to let the lower limit of the first class be a convenient number slightly below or equal to the lowest value in the data set.

Cumulative frequency distribution

Sometimes we may be interested to know the number of observations less than or more than a specified value. This is given by a cumulative frequency distribution. Cumulative frequency is the number of observations located below the upper class boundary.

A cumulative frequency distribution is derived from a frequency distribution by merging successive class intervals until all have been cumulated.

There are two types of cumulative frequency distributions:

(i) Less than cumulative frequency distribution

This is obtained by adding the frequencies of all observations less than upper class boundary of a class. **More than cumulative frequency distribution**

This is obtained by adding the frequencies corresponding to more than lower class boundary of a class. For more than cumulative frequency distributions, lower limit of the class is taken as the reference point.

Illustration 5:

A business firm wants to analyse the characteristics of its employees. A sample of 50 employees is selected, and the age of each at nearest birthday is determined. The following are the ages obtained from individual employee records in the Human Resource Department.

32	60	39	23	30	29	26	41	40	22	52	46	35
25	33	33	20	42	34	43	41	36	58	21	24	55
51	18	40	44	38	21	49	36	43	22	23	51	40
50	45	48	17	45	45	43	21	53	19	63		

Required:

a) Organize the data in a grouped frequency distribution table starting with class 15 – 19 and using a class width of 5.
b) Determine the actual sum of ages of all employees.
c) Use the table obtained in (a) above to estimate the sum of ages of all employees.
d) Prepare the following distributions: a less than cumulative frequency, a more than cumulative frequency, a percentage frequency, a less than cumulative percentage frequency and a more than cumulative percentage frequency distribution.
e) The firm decides to give bonuses of Ksh 100, 200, 300, 400, 500, 600,700, 800, 900 and 1000 for individuals in the age groups 15 -19, 20 - 24 and so on up to the highest salary group respectively. Find the total bonus paid by the firm.

Illustration 6:

A biologist who is interested in knowing whether or not a given consignment of seeds is worth planting conducts an experiment as follows: he takes 100 seeds from the consignment and keeps them under carefully standardized conditions favourable for germination. After 14 days, he counts the seeds that have germinated and records their number. He repeats this experiment 100 times and obtains the following data:-

Number of seeds germinating

86	95	92	89	92	92	91	90	88	93	92
95	94	88	93	92	91	87	94	92	92	88
95	91	92	88	92	91	92	90	89	86	93
89	90	90	89	94	89	95	86	93	93	94
93	91	89	95	92	94	90	86	89	93	92
86	86	94	91	90	92	93	94	92	95	88
94	94	91	88	88	93	90	91	88	90	92
91	93	91	94	87	92	94	86	91	91	94
90	88	93	94	85	95	87	93	92	89	93
92										

Required:

(a) Construct a simple frequency distribution.
(b) Use the table to determine the number of seeds that germinated.
(c) What percentage of the seeds germinated?

Illustration 7:

Twenty five army inductees were given a blood test to determine their blood type. The data set is

A	B	B	AB	O
O	O	B	AB	B
B	B	O	A	O
A	O	O	O	AB
AB	A	O	B	A

Construct a frequency distribution for the data.

The primary purpose of a frequency distribution is to provide a description of an entire set of scores. A frequency distribution allows us to see at a glance the entire set of scores. It shows whether the scores are generally high or low, whether they are concentrated in one area or spread out across the entire scale and generally provides an organized picture of the data. In addition it allows us to see the location of any individual score relative to all of the other scores in the set.