Chapter 2
Frequency Distributions

2.1 Introduction

One principal aim of any statistical enquiry is to be able to understand and describe the population of interest. For example, a farm survey is aimed at estimating current crop output and evaluating the impact of various government policies; a consumer survey will be interested in assessing how much of its product is being consumed and what is the chance of increasing production if some action is taken. Thus, the first task of a statistical staff is that of organizing the data in the form that salient characteristics can be easily seen.

Suppose in your enumeration area, 35 farming households were sampled, and the weights of heads of households in kilograms (to nearest whole number) as obtained from the field are shown in Table 2.1:

Table 2.1 Weights of heads of households in kilograms

70	66	60	55	61	63	72
68	60	60	63	60	75	68
59	71	53	76	64	64	52
64	64	68	64	66	67	63
64	70	69	68	63	59	57

These data are what we call *raw data*, that is, data as obtained from the field. With the data in this form, very little information can be obtained about the population. The first possible thing that we can do is to put the data in what we call an *array*. An array is the arrangement of the values in ascending or descending order of magnitude. For example, if we put the data in an ascending array we have the following results:

52	53	55	57	59	59	60
60	60	60	61	63	63	63
63	64	64	64	64	64	64
66	66	67	68	68	68	68
69	70	70	71	72	75	76

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_2,

```
NAME C1 'WEIGHTS'
SET C1
DATA>70 66 60 55 61..........57
DATA>END
SORT C1 C2
PRINT C1 C2
```

2.2 Frequency Distributions

2.2.1 Ungrouped Distribution

The above initial analysis can be improved by finding out how many farmers have specific weights.

Sample No.	Weights	No. of farmers having such weights
1	52	1
2	53	1
3	55	1
4	57	1
5	59	2
6	60	4
7	61	1
8	63	4
9	64	6
10	66	2
11	67	1
12	68	4
13	69	1
14	70	2
15	71	1
16	72	1
17	75	1
18	76	1

Note that the total should be equal to the number of households. This classification tells us more about the sample; for example, we could see that:

(i) most farmers have different weights
(ii) the most popular (or common) weight of household head is 64 kg.

This is an example of ungrouped frequency distribution. The display is called a *frequency table.*

Definition

The number of farmers having a certain weight is called its *frequency.* In general, the number of times a particular variable/individual occurs is called its frequency. This is represented by "f." For example, the frequency of 67 is 1, that of 68 is 4, etc.

2.2.2 Grouped Distribution

One serious disadvantage of the classification above is that the table may be too long. Take an example when we consider the weights of a sample of 200 households. The analysis in the form of the preceding section becomes too cumbersome and uninformative.

A more convenient way of summarizing a large mass of raw data is to group the observations/variables (in this case) weights into categories and find out how many household heads belong to each category, for example, how many household heads have weights?

- 52 kg to a weight less than 54 kg
- 54 kg to a weight less than 56 kg
- 56 kg to a weight less than 58 kg, etc.

We write the above in a more shortened form:

- 53 kg - under 54 kg
- 54 kg - under 56 kg
- 56 kg - under 58 kg

Each of these categories is called a class interval. A simple procedure we use is what we call *Tally Score Method.* This method consists of making a stroke in the proper class for each observation and summing these for each class to obtain the frequency. It is customary for convenience in counting to place each fifth stroke through the preceding four as shown below.

Weights in kg	Tally	No. of farmers (f)
52 - under 56	111	3
56 - under 60	111	3
60 - under 64	11111 1111	9
64 - under 68	11111 1111	9
68 - under 72	11111 111	8
72 - under 76	11	2
76 - under 80	1	1
	Total	35

Descriptive Analysis

(i) No household head has weight that is less than 52 kg and more than 80 kg.

(ii) The most common weight is somewhere between 60 and 68 kg.

(iii) Most of the farmers have weights from 56 to 72 kg, that is, $3+9+9+8=29$ or 83 % of the farmers.

This is an example of a grouped frequency distribution.

Definitions

- Class Interval: Each category is called a class interval or simply a class.
- Class Limits: These are the end numbers of each class, e.g., 52, 56, 58, etc.
- Upper Class Limit: This is the larger number of the class intervals, e.g., 56.
- Lower Class Limit: This is the smaller number of the class intervals, e.g., 52.
- Size or Width of a class interval: This is the difference between the upper and lower class limits, e.g., $56 - 52 = 4$.
- Class Mark: This is the midpoint of the class interval and is defined as

$$\frac{\text{Upper Class Limit} + \text{Lower Class Limit}}{2}, \quad \text{e.g.,} \quad \frac{56 + 52}{2} = 54, \quad \text{etc.}$$

- Class Boundary: When the upper limit of each class is the same as the lower limit of the next class, the class limits are called class boundaries (above example).

2.2.3 Constructing a Frequency Distribution

There is no hard and fast rule for the construction of frequency distribution, but the following procedures may be followed:

(i) Try to use equal class interval width. This is useful for comparative purposes and for easier calculations.
(ii) The number of classes should not be too many or too few. A rough guideline for constructing k classes for a sample data is the smallest integer value of k such that $2^k \geq n$, where n is the sample size. In the example above, the sample size is 35 and since $2^5 \leq 35 \leq 2^6$, we would employ $k = 6$ classes. Note that in our example above, we have used seven classes.
(iii) It is advisable to use class interval width of multiples of 2, 5, or 10.

In our example above, if we choose $k = 6$ classes, then, the class width is computed as

$$\frac{\text{Largest value} - \text{Smallest value}}{\text{class size}} = \frac{\text{Range}}{\text{Class size}} = \frac{\text{Range}}{k} = \frac{76 - 52}{6} = 4$$

We would usually increase this class width by a little notch say, to 4.2 or 4.5. Suppose we choose 4.5. We can now start the construction of our classes by starting from a value that is slightly less than the minimum. Our minimum in this case is 52. Suppose we start with 51.5. We then give below the construction of the six classes with a class width of 4.5.

Weights	Midpoints	Tally	Frequency (f)
51.5 - <56.0	53.75	111	3
56.0 - <60.5	58.25	11111 11	7
60.5 - <65.0	62.75	11111 11111 1	11
65.0 - <69.5	67.25	11111 111	8
69.5 - <74.0	71.75	1111	4
74.0 - <78.5	76.25	11	2
Total			35

Note The idea of having equal class interval may be waived in a lot of cases. For example, when we have a lot of classes with very few values, it might be advisable to lump them together.

Another example is the case when some classes are unbounded, that is, when we have the case of *open class intervals.* The table below gives the ages of pupils in a primary school in years.

Age (years)	Frequency (f)
Under 6	3
6 - 7	39
8 - 9	42
10 - 11	40
12 - 13	36
Above 13	7

Note that the classes Under 6 and Above 13 have no lower limit and upper limit, respectively.

2.2.4 Other Forms of Frequency Distribution

Relative Frequency

We may be interested in the proportion of our sample or population that falls in a certain class. In this case, we make use of relative frequency. The result of dividing each class frequency by the total frequency of all classes and multiplying the result by 100 is the relative frequency.

Weights	Frequency	Relative frequency (%)
52 - under 56	3	$\frac{3}{35} \times 100 = 8.6$
56 - under 60	3	$\frac{3}{35} \times 100 = 8.6$
60 - under 64	9	$\frac{9}{35} \times 100 = 25.7$
64 - under 68	9	$\frac{9}{35} \times 100 = 25.7$
68 - under 72	8	$\frac{8}{35} \times 100 = 22.9$
72 - under 76	2	$\frac{2}{35} \times 100 = 5.7$
76 - under 80	1	$\frac{1}{35} \times 100 = 2.9$
Total	35	100.10

The relative frequency is mostly useful for easy comparison of two or more frequency distributions. A biological example for instance is the situation where we wish to compare the number of seeds germinating in two varieties of a plant.

The following data in Table 2.2 are used to illustrate the comparative use of the relative frequency approach.

Table 2.2 Age distribution of grade and pupils in Gabon, 1962.

	Frequency (f)		
Age (years)	Boys	Girls	Total
10 - 11	6	5	11
12 - 13	119	49	168
14 - 15	210	102	312
16 - 17	169	75	244
18 - 19	34	4	38
20 - 21	12	-	12
22 - 23	2	-	2
Total	552	235	787

Source: Fundamentals in Educational Planning, (UNESCO)

One cannot compare these values straightaway because the population of the boys in the school is greater than that those of girls, so expectedly, the figures for boys will be greater than those for girls. However, to compare both results, we would need to convert both frequencies into relative frequencies. The relative frequency is very useful for an easy comparison of two or more frequency distributions. We give an example of such a use with the data below which relate to the age distribution of pupils in Gabon in 1962.

	Relative frequencies		Total relative
Age (years)	Boys	Girls	frequency
10 - 11	1.1	2.1	1.4
12 - 13	21.5	20.9	21.3
14 - 15	38.0	43.4	39.6
16 - 17	30.6	31.9	31.0
18 - 19	6.2	1.7	4.8
20 - 21	2.2	0	1.5
22 - 23	0.4	0	0.3
Total	100	100	100

The results from the above analysis suggest the following:

(i) Gabonese government should encourage more girls to school.
(ii) The proportional distribution of ages by sex is close enough except for age group 14 - 15 (difference = 5.4 %) and 18 - 19 (difference = 4.5 %).
(iii) More boys of older age stay at school.

2.2.5 Cumulative Frequency Distributions

Suppose for the data in Table 2.1, we are interested in answering questions such as:

- How many household heads weigh less than 53 kg?
- How many household heads weigh more than 52 kg?

The answers to these and other similar questions are best answered through *cumulative frequency* distributions.

Weights in kg	Frequency	Cumulative frequency from below	Cumulative frequency from above
52 - under 56	3	3	35
56 - under 60	3	6	32
60 - under 64	9	15	29
64 - under 68	9	24	20
68 - under 72	8	32	11
72 - under 76	2	34	3
76 - under 80	1	35	1
Total	35		

No. of farmers whose weights are less than 52 kg = 0
No. of farmers whose weights are less than 56 kg = 3
No. of farmers whose weights are less than 60 kg = 6
No. of farmers whose weights are less than 64 kg = 15
No. of farmers whose weights are less than 68 kg = 24
No. of farmers whose weights are less than 72 kg = 32
No. of farmers whose weights are less than 76 kg = 4
No. of farmers whose weights are less than 80 kg = 35

The above are obtained from the cumulative frequency distribution from below. Similarly, we have,

No. of farmers whose weights are greater than 52 kg = 35
No. of farmers whose weights are greater than 56 kg = 32
No. of farmers whose weights are greater than 60 kg = 29
No. of farmers whose weights are greater than 64 kg = 20
No. of farmers whose weights are greater than 68 kg = 11
No. of farmers whose weights are greater than 72 kg = 3
No. of farmers whose weights are greater than 76 kg = 1
No. of farmers whose weights are greater than 80 kg = 0

The above are similarly obtained from the cumulative frequency distribution from above.

2.3 Graphical Representation of Data

Many people have a strong aversion for anything having numbers and tables, so it might be useful to represent frequency distribution in a more appealing form. One method is to represent the frequency distribution in graphic form which is more informative to the layman. We consider some cases:

2.3.1 The Dotplot

One very useful and simple graphical way to display data is by the use of the graphical method called the *dotplot*. The plot employs a horizontal line with the appropriate axis mark to reflect the range of the data. Each sample observation is then represented in the graph by a single dot above the horizontal line at the specified value. For instance, a data value of 55 is represented by a single dot in the figure below, while a value of 64 is represented by six dots that are stacked above one another. The figure below is a MINITAB output of the dotplot for the data in Table 2.1. We could see that the interval 59–70 contains most of our data values. Further, the plot provides visual information which otherwise could not be discerned from mere looking at the original data in Table 2.1.

```
MTB > DotPlot 'WGT'.

Dotplot: WGT
                                     :
                             :     : :       :
             . .   .   .   : : .   : :   : . : . : . .     . .
         +---------+---------+---------+---------+---------+-------WGT
      50.0      55.0      60.0      65.0      70.0      75.0
```

A very useful advantage of the dotplot is in comparative analysis of two distributions.

2.3.2 The Stem and Leaf Display

The stem and leaf plot offers a quick way to graphically display the shape of continuous type data while including the actual numerical values in the graph. That is, the plot retains the original values of the data. The stem and leaf works best for small numbers of observations as each item of data must be listed. Below is a MINITAB command to construct a stem and leaf display of the data in Table 2.1 which was stored in column 1, C1.

```
MTB > STEM AND LEAF C1;
SUBC> INCREMENT=5.

Stem-and-leaf of weights   N  = 35
Leaf Unit = 1.0

    2    5 23
    6    5 5799
  (15)   6 000013333444444
   14    6 66788889
    6    7 0012
    2    7 56
```

The first column from the MINITAB output for stem and leaf display gives the cumulative frequencies, both from above and below to the interval in which the *median* is located. Thus the parentheses around 15 indicate that the median is in that class interval. The column also tells us that 6 household heads have weights below 60 and 14 who have weights of at least 70.

To construct the stem and leaf display, we note that the minimum datum here is 52 and the maximum is 76. Thus, we could make this a *one-stemmer* by having as stems the tens digits 5, 6, and 7, while the ones digit would then constitute the leaves. This would only result in only three classes, which would not give a fair pictorial representation of the data. This approach is displayed in the following:

```
MTB > STEM AND LEAF C1;
SUBC> INCREMENT=10.

Stem-and-leaf of weights   N  = 35
Leaf Unit = 1.0

    6    5 235799
  (23)   6 00001333344444466788889
    6    7 001256
```

The stem and leaf display we have in the figure above is an example of a *two-stemmer* display. Here the 5's for instance are broken into two groups; 50 - 54 and 55 - 59. That is, the leaves in both groups are respectively the digits {1, 2, 3, 4} and {5, 6, 7, 8, 9}. The two stemmers can be invoked in MINITAB by using the subcommand *increment = 5* while the one-stemmer can similarly be invoked by using the subcommand *increment = 10.* Other forms of the stem and display are the *five-stemmer* and the *ten-stemmer*. For a five stemmer, we would have for the 5's the following stems.

Stems	Leaves
2*	With unit digits 0 or 1
2t	With unit digits 2 or 3
2f	With unit digits 4 or 5
2s	With unit digits 6 or 7
5•	With unit digits 8 or 9

In this splitting, the symbol t is used for the digits 2 and 3; f for four and five; and s for six and seven. We again give this display for our data in Table 2.1. The display is generated by the MINITAB subcommand *increment = 2.*

```
MTB > stem and leaf c1;
SUBC> increment=2.

Stem-and-leaf of weights   N  = 35
Leaf Unit = 1.0

    2    5 23
    3    5 5
    4    5 7
    6    5 99
   11    6 00001
   15    6 3333
   (6)   6 444444
   14    6 667
   11    6 88889
    6    7 001
    3    7 2
    2    7 5
    1    7 6
```

We note that in all the above MINITAB displays of the stem-and-leaf plots, the MINITAB orders the leaf units. However, one needs to very careful with stem-and-leaf displays because the display itself does not tell you the actual value of the data. The actual value is provided by the leaf *unit* = statement which is given just above the display. For example, if the leaf unit = 1.0 had been leaf unit = 10, then the smallest data element would have been 520. Similarly, if the leaf unit had been leaf unit = 0.001 instead of 1.0, then the smallest data element would have been 0.052. We give an example below where the data in Table 2.1 were multiplied each by 10, and the resulting stem and leaf display below (a five-stemmer) gives the leaf unit = 10, indicating that the minimum data element is 520 and the maximum being 760. Notice that this display is very similar in every respect to the five-stemmer display above, except for the leaf unit value.

```
MTB > LET C3 = C1*10
MTB > STEM AND LEAF C3

Stem-and-leaf of C3        N  = 35
Leaf Unit = 10

    2    5 23
    3    5 5
    4    5 7
    6    5 99
   11    6 00001
   15    6 3333
   (6)   6 444444
   14    6 667
   11    6 88889
    6    7 001
    3    7 2
    2    7 5
    1    7 6
```

Springer

http://www.springer.com/978-3-319-05554-1

Applied Statistical Methods in Agriculture, Health and Life Sciences
Lawal, B.
2014, XXXIII, 799 p. 536 illus., 61 illus. in color.,
Hardcover
ISBN: 978-3-319-05554-1