1b

Olopoenia, R. A. and D. D. Ajayi (2005) Data Exploration and Description. In: Gbadegesin, A.S; R. Olopoenia and A. Jerome (eds), Statistics for the Social Sciences, Chapter 4, pp. 25-33. Ibadan University Press. –Nigeria (50.0%)

STATISTICS FOR THE SOCIAL SCIENCES

Edited by
Adeniyi Gbadegesin
Razaq Olopoenia
Afeikhena Jerome

Ibadan University Press
2005

CONTENTS

4

DATA EXPLORATION AND DESCRIPTION

Rasaq A. Olopoenia and 'Dickson Dare Ajayi

Usually, after collecting raw data, we are faced with the problem of how to organize such data numerically. We arrange such *numerical data* in ascending or descending order of magnitude. Such an arrangement is called an *array*. The difference between the largest and smallest numbers is called the *range of the data.* For instance, if the largest weight of 100 male students is 74 kg and the smallest weight is 60 kg, the range is 74-60 = 14 kg.

Frequency Distribution

The frequency distribution is one of the major techniques of data presentation. Usually, it is useful to summarize large masses of raw data into *classes,* or *categories,* and to determine the number of individuals belonging to each class. This is called the class frequency. The tabular arrangement of data by classes, together with the corresponding class frequency, is called a *frequency distribution* or *frequency table.* The frequency distribution of the weights of 100 male students is shown in Table I.

Table 1:
Frequency Distribution of Male Students

Weight (kg)	Number of Students
60 -62	5
63 -65	18
66 -68	42
69 -71	27
72 –74	8
Total	100

The first class (or category), for example, consists of weights from 60 to 62 kg, which is indicated by the range symbol 60 – 62. Since five students have weights belonging to this class, the corresponding class frequency is 5. Data organized and summarized in this way are called *grouped data.* Grouping of data generally destroys much of the original detail of the data, but an important element of it is in the *clear "overall" picture* that is obtained, and in the vital relationships that are thereby made evident.

Class Intervals and Class Limits

In our example, a symbol defining a class, such as 60 – 62 in the table, is *called a class interval.* The end numbers are called *class limits.* The smaller number (60) is the *lower class limit* and the larger number (62) is the *upper class limit.* A class interval that has either no upper class limit or lower class limit indicated is called an *open class interval.* For example, referring to age groups of individuals, the class interval "65 years and over" is an open class interval.

Class Boundaries

In our example, the class interval 60 – 62 theoretically includes all measurements from 59.5 to 62.5kg. These numbers, indicated briefly by the exact numbers 59.5 and 62.5, are called *class boundaries,* or true class limits. The smaller number (59.5) is the *lower class boundary,* while the larger number (62.5) is the *upper class boundary.*

Class Interval

Usually, the difference between the lower and upper class boundaries is referred to as the *class width, class size,* or *class length.* When all the class intervals of a frequency distribution have equal width, the common width is denoted by C. This is equal to the differences between two successive lower limits or two successive upper class limits. From our table a = 62.5 – 59.5 = 65.5 – 62.5 = 3.

Class Mark

The class mark is the midpoint of the class interval. It is obtained by adding the lower and upper class limits and dividing by 2. Therefore, the class mark of the interval 60 − 62 is (60 + 62)/2 = 61. The class mark is also called the class midpoint. For further analysis all observations in a given class interval are assumed to coincide with the class mark. All weights in the class interval 60–62 kg are considered to be 61 kg.

Rules for forming a Frequency Distribution

1. Determine the largest and smallest numbers in the raw data and find the range (the difference between the largest and smallest number)
2 Divide the range into a convenient number of class intervals having the same size. When this is not possible, use class intervals of different sizes or open class intervals. The number of class intervals is usually taken between 5 and 2, depending on the data. Class intervals are also chosen so that the class marks (or midpoints) coincide with the actually observed data. However, the class boundaries should not coincide with the actually observed data
3. Determine the number of observations falling into each class interval, that is, find the class frequencies. This is done by counting.

Histograms and Frequency Polygons

There are two graphic representations of frequency distributions. These are *histograms* and *frequency polygons*

1. A histogram consists of a set of rectangles

   (a) with bases on a horizontal axis (the axis), centres at the class marks, lengths equal to the class interval sizes, and
   (b) with areas proportional to the class frequencies.

2. A frequency polygon is a line graph of the class frequency plotted against the class mark. It can be obtained by connecting the midpoints of the tops of the rectangles in

the histogram. For our table, the histogram and frequency polygon are shown in Figure 1.

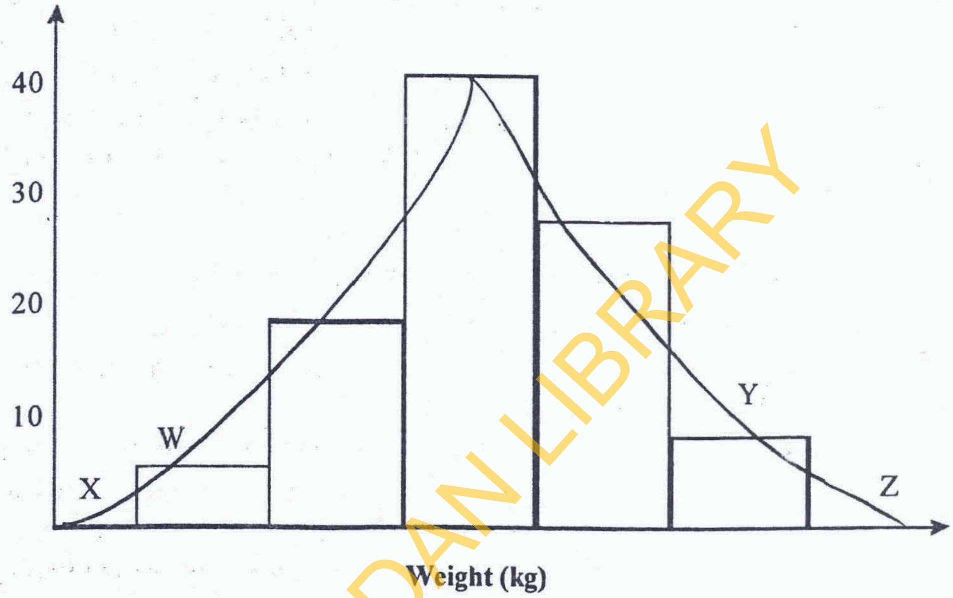


Fig 1: Histogram and Frequency Polygon

It is customary to add the extensions WX and YZ to the next-lower and higher-class marks, which have a corresponding class frequency of zero.

Relative Frequency Distribution

The relative frequency of a class is the frequency of tne class divided by the total frequency of all classes and is generally expressed as a percentage. In our example, the relative frequency of the class 66–68 is 42/100 = 42%. The sum of the relative frequencies of all classes is clearly 1 or 100%

When the frequencies are replaced with the corresponding relative frequencies, the resulting table is called a *relative frequency distribution,* or *relative frequency table.* This can be graphically represented in a *relative frequency histogram* (or percentage histogram) and *relative frequency polygon* (or percentage polygon), respectively.

Cumulative Frequency Distributions and Ogives

The total frequency of all values less than the upper class boundary of a given class interval is called the *cumulative frequency* up to and including that class interval. In our table, the cumulative frequency up to and including the class interval 66–68 is 5 + 18+ 42 = 65, signifying that 65 students have weights less than 68.5 kg. A table presenting such cumulative frequencies is called a *cumulative frequency distribution, a cumulative frequency table* or, briefly, a *cumulative distribution.*

Table 2.
Cumulative Frequency Distribution

Weight (kg)	Number of Students
<59.5	0
<62.5	5
<65.5	23
<68.5	65
<71.5	92
<74.5	100

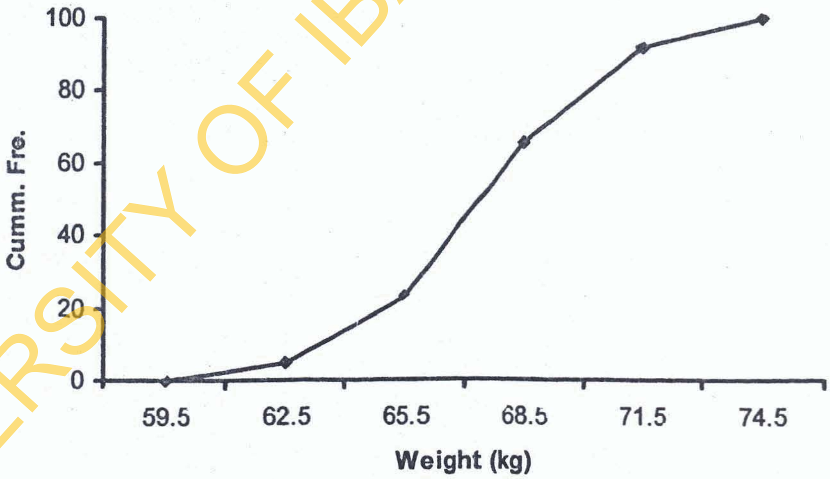


Fig. 2: Cumulative Frequency Polygon

A graph showing the cumulative frequency that is less than any upper class boundary plotted against the upper class boundary is called a *cumulative frequency polygon* or ogive. This is shown in

Fig. 2 for the student weight distribution indicated in Table 2. In some situations, it may be necessary to consider a cumulative frequency distribution of all values greater than or equal to the lower class boundary of each class interval. It should be noted that because in this case we consider weights of 59.5kg or more, 62.5kg or more, etc., this is often called an *"or more "* cumulative distribution, while the one considered above is a *"less than"* cumulative distribution. What is clear, however, is that one is easily obtained from the other.

For example, Table 4 shows the "or more" cumulative frequency for the distribution of wages of 65 employees. To obtain the "or more" cumulative frequency, add successive entries from column 2 of Table 3, starting at the bottom. Thus we have 7 = 2 + 5, 17 = 2+5+10, 31 = 2+5+10+14 etc (see Table 4).

Table 3:
Frequency Distribution of Wages

Wages (₦)	Number of Employees
250.00 - 259.99	8
260.00-269.99	10
270.00 - 279.99	16
280.00-289.99	14
290.00-299.99	10
300.00 - 309.99	5
310.00-319.99	2
Total	65

Table 4:
"Or More" Cumulative Frequency Distribution of Wages

Wages (₦)	"Or More" Cumulative Frequency
250.00 or more	65
260.00 or more	57
270.00 or more	47
280.00 or more	31
290.00 or more	17
300.00 or more	7
310.00 or more	2
320.00 or more	0

The entries can also be obtained in reverse order by subtracting each entry in column 2 of Table 3 from the total; thus 57 = 65-8, 47 = 65-18, etc. The corresponding ogives are then called "or more" and "less than" ogives. The "less than" is implied whenever we refer to cumulative distributions or ogives without qualification.

Relative Cumulative Frequency Distributions and Percentage Ogives

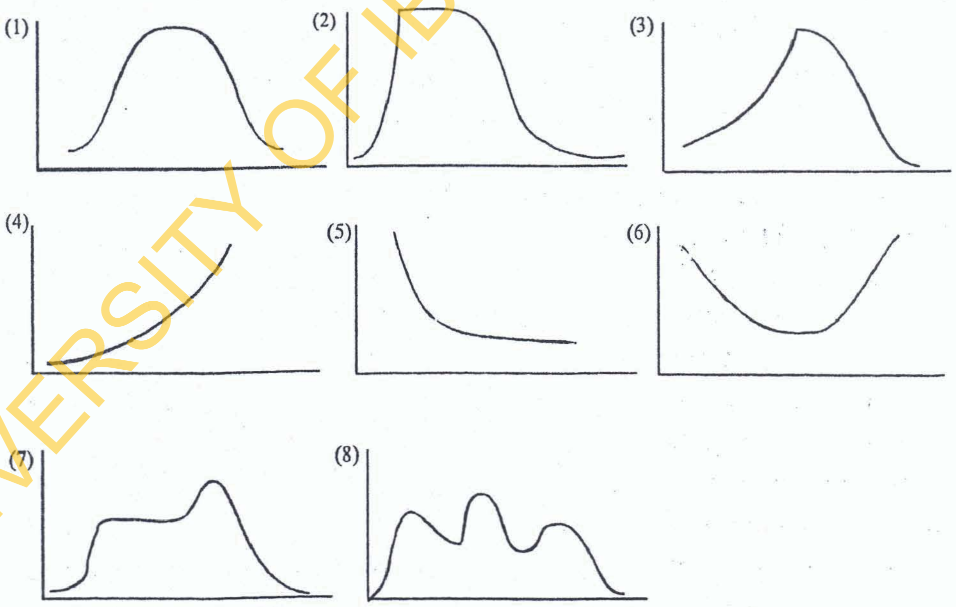
Relative cumulative frequency, or percentage cumulative frequency, is the cumulative frequency divided by the total. For example, the relative cumulative frequency of weights less than 68.5kg in our table is 65/100 = 65%. This signifies that 65% of the students have weights less than 68.5kg. When relative cumulative frequencies are used, either in tables or figures, in place of cumulative frequencies, the results are called *relative cumulative frequency distributions (or percentage cumulative distributions)* and *relative cumulative frequency polygons (or percentage ogives),* respectively.

Frequency Curves and Smoothed Ogives

Often times, the data collected are considered as belonging to a sample drawn from a large population. Since so many observations are available in the population, in theory, it is possible (for continuous data) to choose very small class intervals and still have sizeable numbers of observations falling within each class. One would therefore expect the frequency polygon or relative frequency polygon for a large population to have several small, broken line segments that closely approximate to a curve. Such curves are called *frequency curves* or *relative frequency curves,* respectively. Theoretically, such curves can be approximated by smoothing the frequency polygons or relative frequency polygons of the sample, with the approximation improving as the sample size is increased. For this reason, a frequency curve is sometimes called a *smoothed frequency polygon.* In like manner, *smoothed ogives* are obtained by smoothing the cumulative frequency polygons, or ogives. But an ogive is usually easier to smoothen than a frequency polygon.

Types of Frequency Curves

Diagrams



1. Symmetrical
2. Positive Skewness
3. Negative Skewness
4. J-Shaped
5. Reverse J-Shaped
6. U-shaped
7. Bimodal
8. Multimodal

The *symmetrical or bell-shaped* frequency curves are characterized by equidistant observations from the centre location. This may be likened to the normal curve.

- In the *moderately asymmetrical,* or *skewed,* frequency curves, the tail of the curve to one side is longer than that to the other. The skewness, therefore, is either to the left or right.
- In a *J-shaped* or reverse J-shaped curve a maximum occurs at one end.
- A *U-shaped* frequency curve has maxima at both ends
- A *bimodal* frequency curve has two maxima
- A *multimodal* frequency curve has more than two maxima.

REFERENCE

Spiegel, M.R. 1992. Frequency Distributions. In Schaum's *Outline Series of Theory and Problems of Statistics*. Chapter 2, pp. 36- 57.