Chapter 2

Frequency Distributions and Graphs

1. Introduction

When conducting a statistical study, the researcher must gather data for the particular variable under study.

To describe situations, draw conclusions, or make inferences about events, the researcher must organize the data in some meaningful way. The most convenient method of organizing data is to construct a *frequency distribution*

After organizing the data, the researcher must present them so they can be understood by those who will benefit from reading the study. The most useful method of presenting the data is by constructing *statistical charts* and *graphs.* There are many different types of charts and graphs, and each one has a specific purpose.

2. Organizing Data

Suppose a researcher wished to do a study on the number of miles that the employees of a large department store traveled to work each day. The researcher first would have to collect the data by asking each employee the approximate distance the store is from his or her home. When data are collected in original form, they are called **raw data.** In this case, the data are

1	2	6	7	12	13	2	6	9	5
18	7	3	15	15	4	17	1	14	5
4	16	4	5	8	6	5	18	5	2
9	11	12	1	9	2	10	11	4	10
9	18	8	8	4	14	7	3	2	6

Since little information can be obtained from looking at raw data, the researcher organizes the data by constructing a frequency distribution. The **frequency** is the number of values in a specific class of the distribution. For this data set, a frequency distribution is shown.

Class Limits (in miles)	Frequency
1-3	10
4-6	14
7-9	10
10-12	6
13-15	5
16-18	5
	Total 50

Now some general observations can be obtained from looking at the data in the form of a frequency distribution.

A **frequency distribution** is the organization of raw data in table from, using classes and frequency.

The classes in this distribution are 1-3, 1-4, etc. These values are called **class limits**. Two types of frequency distributions that are most often used are the *categorical frequency distribution* and the *grouped frequency distribution.* The procedures for constructing this distribution are shown now.

2.1 Categorical Frequency Distributions

The categorical frequency distribution is used for data that can be placed in specific categories, such as nominal or ordinal level data. For example, data such as political affiliation, religious affiliation, or major field of study would use categorical frequency distributions.

Example 1

20 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.

Solution

There are four blood types, which will be used as the classes in the distribution.

Step1: Make a table as shown

A	**B**	**C**	**D**
Class	**Tally**	**Frequency**	**Percent**
A			
B			
O			
AB			

Step2: Tally the data and place the results in column **B**

Step3: Cont the tallies and place the results in column **C**

Step4: Find the percentage of values in each class by using the formula

$$\% = \frac{f}{n} \cdot 100\%$$

where f = frequency of the class and n = total number of values. For example, in the class of type A blood, the percentage is

$$\% = \frac{5}{25} \cdot 100\% = 20\%$$

Note: Percentage is not normally a part of frequency distribution, but they can be added since they are used in certain types of graphical representation, such as pie chart.

Step5: Find the totals for column **C** and **D**

A Class	B Tally	C Frequency	D Percent
A	~~////~~	5	20
B	~~////~~ //	7	28
O	~~////~~ ////	9	36
AB	////	4	16
		25	100

2.2 Grouped Frequency distributions

When the range of the data is large, the data must be grouped into classes that are more than one unit in width. For example, a distribution of the number of hours that boat batteries lasted is

Class Limits	Class Boundaries	Frequency	Cumulative Frequency
24-30	23.5-30.5	3	3
31-37	30.5-37.5	1	4
38-44	37.5-44.5	5	9
45-51	44.5-51.5	9	18
52-58	51.5-58.5	6	24
59-65	58.5-65.5	1	25

In this distribution, the value 24 and 30 of the first class are called *class limits.* The **lower class limit** is 24; it represents the smallest data value that can be included in the class. The **upper class limit** is 30; it represents the largest data value that can be included in the class. The numbers in the second column are called **class boundaries.** These numbers are used to separate the classes so that there are no gaps in the frequency distribution. The gaps are due to the limits; for example, there is a gap between 30 and 31. Finally, the **class width** for a class in a frequency distribution is found by subtracting the lower (or upper) class limit of one class from the lower (or upper) class limit of the next class. For example the class width in the preceding distribution on the duration of boat batteries is 7, found from 31-24=7. The class width can also be found by subtracting the lower boundary from the upper boundary for any given class. In this case, 30.5-23.5 = 7

The researcher must decide how many classes to use and the width of each class. To construct a frequency distribution, follow these rules:

(i). *There should be between* 5 and 20 *classes.* Although there is no hard-and –fast rule for the number of classes contained in a frequency distribution, it is of the utmost importance to have enough classes to present a clear description of the collected data.

(ii). T*he class width should be an odd number*. This ensures that the midpoint of each class has the same place value as the data. The **class midpoint** X_m is obtained by adding the lower and upper boundaries and dividing by 2, or adding the lower and upper limits and dividing by 2:

$$X_m = \frac{\text{lower boundary + upper boundary}}{2}$$

or $$X_m = \frac{\text{lower limit + upper limit}}{2}$$

For example, the midpoint of the first class in the example with boat batteries is

$$\frac{24+30}{2} = 27 \text{ or } \frac{23.5+30.5}{2} = 27$$

Rule 2 is only a suggestion, and it is not rigorously followed.

(iii). *The classes must be mutually exclusive.* Mutually exclusive classes have non-overlapping class limits so that data cannot be placed into two classes.

Age
10-20
20-30
30-40
40-50

Not mutually exclusive

Age
10-20
21-31
32-42
43-53

Mutually exclusive

(iv). *The classes must be continuous.* Even if there are no values in a class, the class must be included in the frequency distribution. There should be no gaps in a frequency distribution. The only exception occurs when the class with a zero frequency is the first or last class. A class with a zero frequency at either end can be omitted without affecting the distribution.

(v). *The class must be exhaustive.* There should be enough classes to accommodate all the data.

(vi). *The classes must be equal in width.* This avoids a distorted view of the data.

One exception occurs when there is an **open-ended distribution-**i.e., it has no specific beginning value or no specific ending value. Here are the class limits for two open-ended distributions.

Age
10-20
21-31
32-42
43-53
54-64
65 and above

Minutes
Below 110
110-114
115-119
120-124
125-129
130-134

Example

These data represent the record high temperatures for each of the 50 states. Construct a grouped frequency distribution for the data using 7 classes.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

The procedure for constructing a grouped frequency distribution for numerical data follows

Step1: Determine the classes

Find the highest value and lowest value: $H = 134$ and $L = 100$

Find the range: $R = H - L = 134 - 100 = 34$

Select the number of classes desired. In this case, 7 is arbitrarily chosen. Find the the class with by dividing the range by the number of classes.

$$\text{width} = \frac{R}{\text{number of classes}} = \frac{34}{7} = 4.9$$

Round the answer up to the nearest whole number: $4.9 \approx 5$

Select a starting point for the lowest class limit. This can be the smallest data value or any convenient number less than the smallest data value. In this case 100 is used. Add the width to the lowest score taken as the starting point to get the lower limit of the next class. Keep adding until there are 7 classes, as shown 100, 105, 110, etc.

Subtract 1 unit from the lower limit of the second class to get the upper limit of the first class. Then add the width to each upper limit to get all the upper limits.

$$105 - 1 = 104$$

The first class is100 - 104, the second is 105 – 109, etc.

Find the class boundaries by subtracting 0.5 from each lower class limit and adding 0.5 to each upper class limit: 99.5-104.5, 104.5-109.5, etc.

Step2: Tally a data

Step 3: Find the numerical frequencies from the tallies

Step4: Find the cumulative frequencies.

The completed frequency distribution is

Class limits	**Class boundaries**	**Tally**	**Frequency**	**Cumulative frequency**
100-104	99.5-104.5		2	2
105-109	104.5-109.5		8	10
110-114	109.5-114.5		18	28
115-119	114.5-119.5		13	41
120-124	119.5-124.5		7	48
125-129	124.5-129.5		1	49
130-134	129.5-134.5		1	50

When the range of the data values is relatively small, a frequency distribution can be constructed using single data values for each class. This type of distribution is called an **ungrouped frequency distribution** and shown next.

The steps for constructing a grouped frequency distribution are summarized in the following Procedure.

(i). Determine the classes: Find the highest and lowest value, find the range, select the number of classes desired, find the width by dividing the rage by the number of classes and rounding up, select a starting point (usually the lowest value or any convenient number less than the lowest value), add width to get the lower limits, find the upper class limits, and the boundaries.

(ii). Tally the data.

(iii) Find the numerical frequencies from the tallies.

(iv). Find the cumulative frequencies.

3. Histograms, Frequency Polygons, and Ogives

3.1 Histogram

A **histogram** is a graph that displays the data by using contiguous vertical bars (unless the frequency of a class is 0) of various heights to represent the frequencies of the classes.

Example

Construct a histogram to represent the data shown below for the record high temperature:

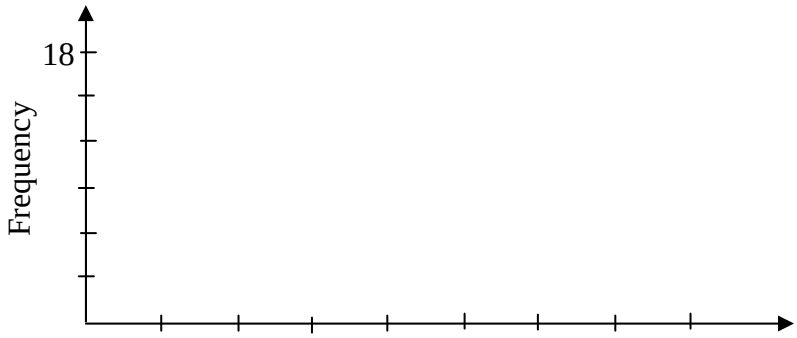
Class boundaries	**Frequency**
99.5-104.5	2
104.5-109.5	8
109.5-114.5	18
114.5-119.5	13
119.5-124.5	7
124.5-129.5	1
129.5-134.5	1

Solution

(i). Draw and label the *x* and *y* axis. The *x* axis is always the horizontal axis, and the *y* axis is always the vertical axis.

(ii). Represent frequency on *y* axis and the class boundaries on the *x* axis.

(iii). Using the frequencies as the heights, draw vertical bars for each class.



3.2 The Frequency Polygon

The **frequency polygon** is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoint (class mark) of the classes. The frequencies are represented by the heights of the points.

Example

Use the above data to construct a frequency polygon.

Solution

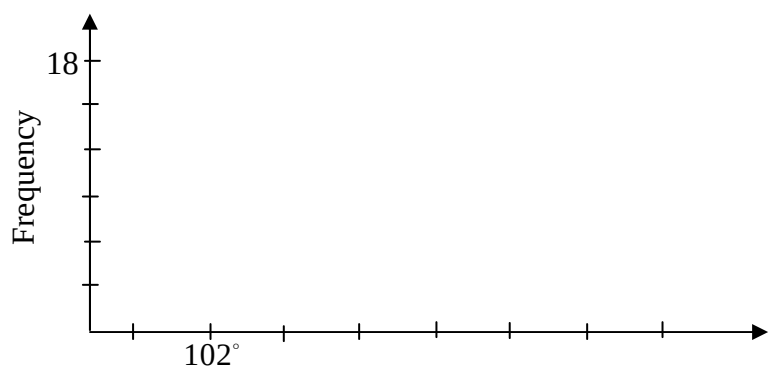
(i). Find the midpoints (class marks) of each class to get the results as shown below

Class boundaries	**Midpoints**	**Frequency**
99.5-104.5	102	2
104.5-109.5	107	8
109.5-114.5	112	18
114.5-119.5	117	13
119.5-124.5	122	7
124.5-129.5	127	1
129.5-134.5	132	1

(ii). Draw and label the *x* and *y* axis. Label *x* axis with the midpoints of each class, and then use a suitable scale on the *y* axis for the frequencies.

(iii). Using the midpoints for the *x* values and the frequencies as the *y* values, plot the points.

(iv). Connect adjacent points with line segments. Draw a line back to the *x* axis at the beginning and end of the graph, at the same distance that the previous and next midpoints would be located.



3.3 The Ogive

The third type of graph that can be used represents the cumulative frequencies for the classes. This type of graph is called the cumulative frequency graph or ogive. The cumulative frequency is the sum of the frequencies accumulated up to the upper boundary of a class in the distribution.

The **Ogive** is a graph that represents the cumulative frequencies for the classes in a frequency distribution.

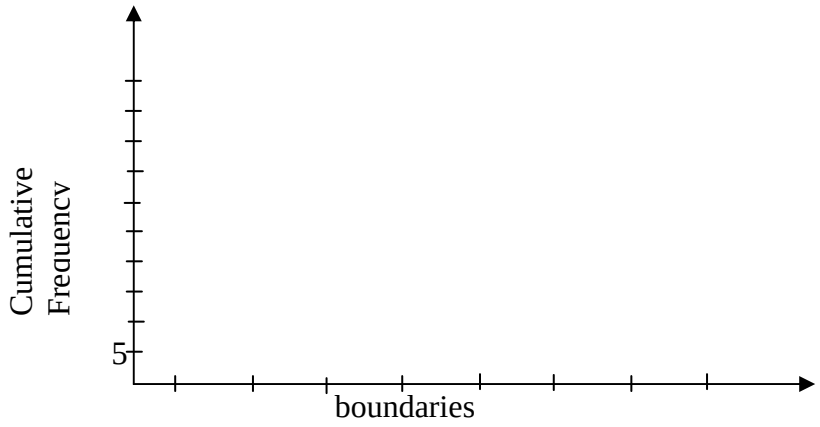
Example

Construct an Ogive for the frequency distribution in example above.
Solution

(i). Find the cumulative frequency for each class.

Class boundaries	**Frequency**	**Cumulative Frequency**
99.5-104.5	2	2
104.5-109.5	8	10
109.5-114.5	18	28
114.5-119.5	13	41
119.5-124.5	7	48
124.5-129.5	1	49
129.5-134.5	1	50

(ii). Draw and label the *x* and *y* axis. Label *x* axis with the class boundaries, Use an appropriate scale for the *y* axis to represent the cumulative frequencies.
(iii). Plot the cumulative frequency at each upper class boundary. Upper boundaries are used since the cumulative frequencies represent the number of data values accumulated up to the upper boundary of each class.



The steps for drawing these three types of graphs are shown as follows
<u>Step1:</u> Draw and label the *x* and *y* axis
<u>Step2:</u> Choose a suitable scale for the frequencies or cumulative frequencies, and label it on the *y* axis.
<u>Step3:</u> Represent the class boundaries for the histogram or Ogive, or the midpoint for the frequency polygon, on the *x* axis.
<u>Step4:</u> Plot the points then draw the bars or lines.

3.4 Relative Frequency Graphs

The histogram, the frequency polygon, and the Ogive shown previously were constructed by using frequencies in terms of the raw data. These distributions can be converted to distributions using *proportions* instead of raw data as frequencies. These types of graphs are called **relative frequency graphs.**

To convert a frequency into a proportion or relative frequency, divide the frequency for each class by the total of frequencies. These graphs are similar to the ones that use raw data as frequencies, but the value on *y* axis are in terms of proportions.

Example

Construct a histogram, frequency polygon, and Ogive using relative frequencies for the distribution (shown below) of the miles that 20 randomly selected runners ran during a given week

Class boundaries	Frequency	Cumulative frequency
5.5-10.5	1	1
10.5-15.5	2	3
15.5-20.5	3	6
20.5-25.5	5	11
25.5-30.5	4	15
30.5-35.5	3	18
35.5-40.5	2	20

Solution

Step1: Convert each frequency to a proportion or relative frequency by dividing the frequency for each class by the total number of observations. For class 5.5-10.5, the relative frequency is $\frac{1}{20} = 0.05$, for class 10.5-15.5, the relative frequency is $\frac{2}{10} = 0.10$ and so on.

Step2: Find the cumulative relative frequencies.

Class boundaries	Midpoints	Relative frequency	Cumulative relative frequency
5.5-10.5	8	0.05	0.05
10.5-15.5	13	0.10	0.15
15.5-20.5	18	0.15	0.30
20.5-25.5	23	0.25	0.55
25.5-30.5	28	0.20	0.75
30.5-35.5	33	0.15	0.90
35.5-40.5	38	0.10	1.00

Step3: Draw each graph. For the histogram and ogive, use the class boundaries along the *x* axis. For the frequency polygon, use the midpoints on the *x* axis. The scale on the *y* axis uses proportions.

4. Other Types of Graphs

4.1 Pareto Charts

When the variable displayed on the horizontal axis is qualitative or categorical, a Pareto Chart can be used.

A **Pareto chart** is use to represent a frequency distribution for a categorical variable, and the frequencies are displayed by the heights of vertical bars, which are arranged in order from highest to lowest.

Example

The table shown here displays the number of crimes investigated by law enforcement officers in U.S. national parks during 1995. Construct a Pareto chart for the data.

Type	**Number**
Homicide	13
Rape	34
Robbery	29
Assault	164

Solution

Step1: Arrange the data from the largest to the smallest according to frequency

Type	**Number**
Assault	164
Rape	34
Robbery	29
Homicide	13

Step2: Draw and label *x* and *y* axes.

Step3: Draw the bar corresponding the frequencies.

4. 3 The Pie Graph

Pie graphs are used extensively in statistics. The purpose of the pie graph is to show the relationship of the parts to the whole by visually comparing the sizes of the sectors. Percentages or proportions can be used. The variable is nominal or categorical.

A **pie graph** is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution.

Example

This frequency distribution shows the number of pounds of each snack food eaten during the 1998 Super Bowl. Construct a pie graph for the data.

Snack	Million pounds
Potato Chips	11.2
Tortilla Chips	8.2
Pretzels	4.3
Popcorn	3.8
Snack nuts	2.5

Solution:

Step1: Since there are 360° in a circle, the frequency for each class must be converted into a proportional part of the circle. This conversion is done by using the formula

$$\text{Degrees} = \frac{f}{n}360^\circ$$

where *f* is the frequency of each class and *n* is the sum of the frequencies. Hence, the following conversions are obtained.

$$\text{Potato chips}\frac{11.2}{30}\times 360^\circ = 134^\circ$$

.....

$$\text{Snack nut } \frac{2.5}{30}\times 360^\circ = 30^\circ$$

Step2: Each frequency must also be converted to a percentage. Use the formula

$$\% = \frac{f}{n}\times 100\%$$

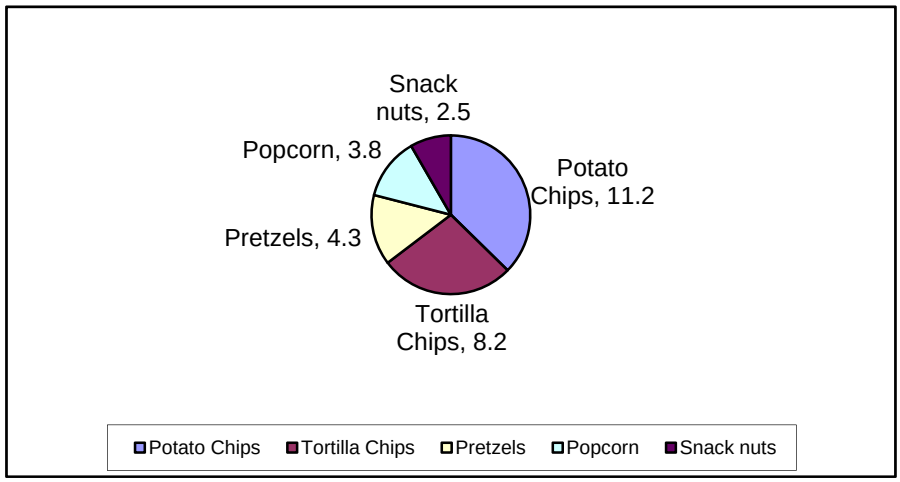
Hence, the following percentages are obtained.

$$\text{Potato chips}\frac{11.2}{30}\times 100\% = 37.3\%$$

.....

$$\text{Snack nut } \frac{2.5}{30}\times 100\% = 8.3\%$$

Step3: Using the protractor and a compass, draw the graph and label each section with the name and percentages.



4.4 Stem and Leaf Plots

The stem and leaf plot is a method of organizing data and is a combination of sorting and graphing. It has the advantage over a grouped frequency distribution of retaining the actual data while showing them in graphic form.

> A **stem and leaf plot** is a data plot that uses part of the data value as the stem and part of the data value as the leaf to form groups or classes.

Example
At an outpatient testing center, the number of cardiograms performed each day for 20 days is shown. Construct a tem and leaf plot for the data.

25	31	20	32	13
14	43	02	57	23
36	32	33	32	44
32	52	44	51	45

Solution

Setp1: Arrange the data in order:
02, 13, 14, 20, 23, 25, 31, 32, 32, 32, 32, 33, 36, 43, 44, 44, 45, 51, 52, 57
Note: Arranging the data in order is not essential and can be cumbersome when the data set is large; however, it is helpful in construction a stem and leaf plot.

Step2: Separate the data according to the first digit, as shown
02 13, 14 20, 23, 25 31, 32, 32, 32, 32, 33, 36
43, 44, 44, 45 51, 52, 57

Step3: A display can be made by using the leading digit as the *stem* and the trailing digit as the *leaf*. For example, for the value 32, the leading digit, 3, is the stem and the trailing digit, 2, is the leaf.

Leading digit (stem)	**Trailing digit (stem)**
0	2
1	3 4
2	0 3 5
3	1 2 2 2 2 3 6
4	3 4 4 5
5	1 2 7

Example
An insurance company researcher conducted a survey on the number of car thefts in a large city for a period of 30 days last summer. The raw data are shown below. Construct a stem and leaf plot by using classes 50 – 54, 55 – 59, 60 – 64, 65 – 69, 70 – 74, and 75 – 79.

52	62	51	50	69
58	77	66	53	57
75	56	55	67	73
79	59	68	65	72
57	51	63	69	75
65	53	78	66	55

Solution
Separate the data according to the classes 50, 51, 51, 52, 53, 53, 55, 55, 56, 57, 57, 58, 59, 62, 63, 65, 65, 66, 66, 67, 68, 69, 69, 72, 73, 75, 75, 77, 78, 79

Leading digit (stem)	**Trailing digit (leaf)**
5	0 1 1 2 3 3
5	5 5 6 7 7 8 9
6	2 3
6	5 5 6 6 7 8 9 9
7	2 3
7	5 5 7 8 9

When the data values are in the hundreds, such as 325, the stem is 32 and the leaf is 5. For example, the stem and leaf plot for the data values 325, 327, 330, 332, 335, 341, 345, and 347 looks like this.

32	5 7
33	0 2 5
34	1 5 7