STATISTICS IS THE GRAMMAR OF SCIENCE

PROBABILITY AND STATISTICS

LECTURE – 3

PRESENTATION OF QUANTITATIVE DATA

PREPARED BY
HAZBER SAMSON
FAST NUCES ISLAMABAD

REPRESENTATION OF DATA

After collection of sample data, we must "get acquainted" with them. The best way to become acquainted with the data is to use an initial exploratory data-analysis. For this we need to organize the data first and then visualize and analyze it.

There are two major components of data Representation

- Data Organization
- Data Visualization

DATA ORGANIZATION

In order to visualize and analyze the data first step is to organize the data, we can organize data in two ways.

- Organizing Qualitative Data
- Organizing Quantitative Data

DATA VISUALIZATION

When we organize our data, we sometimes begin to discover patterns or relationships in our data. To better explore and discover patterns and relationships, we can visualize your data by creating various charts and special "displays." As is the case when organizing data, the techniques we use to visualize our data depend on the type of variable (categorical or numerical) of our data.

We can visualize data in two ways

- Qualitative Data Visualization
- Quantitative Data Visualization

Before going into details of data organization let's study the idea of frequency distribution

THE FREQUENCY DISTRIBUTION

FREQUENCY The number of times a particular distinct value occurs is called its frequency (or count).

DISTRIBUTION The pattern of variability displayed by the data of a variable. The distribution displays the frequency of each value of the variable.

FREQUENCY DISTRIBUTION A frequency distribution provides a table of the values of the observations and how often they occur.

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

TYPES OF FREQUENCY DISTRIBUTION There are two types of frequency distributions that are mostly used

- Categorical frequency distribution
- Grouped frequency distribution.

ORGANIZING QUANTITATIVE DATA

we organize numerical data by creating ordered arrays or distributions. The amount of data we have and what we seek to discover about your variables influences which methods we choose, as does the arrangement of data in our worksheet.

In case of quantitative data we can define frequency distribution as follows

FREQUENCY DISTRIBUTION A grouping of quantitative data into mutually exclusive and collectively exhaustive classes showing the number of observations in each class.

The procedures for constructing these distributions are shown now.

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, in what is called a grouped frequency distribution. For example, a distribution of the number of hours that boat batteries lasted is the following.

Class limits	Class boundaries	Tally	Frequency
24–30	23.5–30.5	///	3
31–37	30.5–37.5	/	1
38–44	37.5–44.5	𝍸	5
45–51	44.5–51.5	𝍸 ////	9
52–58	51.5–58.5	𝍸 /	6
59–65	58.5–65.5	/	1
			25

In this distribution, the values 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.

A **frequency distribution** is a table formed by classifying n data values into k **classes called bins.** The **class limits** or **bin limits** define the values to be included in each class. Usually, all the class widths are the same. The table shows the frequency of data values within each bin. Frequencies can also be expressed as relative frequencies or percentages of the total number of observations.

CONSTRUCTING FREQUENCY DISTRIBUTION FOR QUANTITATIVE DATA

The basic steps for constructing a frequency distribution are as follows
(1) sort the data in ascending order
(2) choose the number of classes
(3) choose the class width
(4) choose lower limit of the first class
(5) create the table.

Let's walk through these steps.

STEP-1 SORTING THE DATA IN ASCENDING ORDER

First of all sort the data in ascending order and identify minimum and maximum points.

STEP-2 CHOOSING THE NUMBER OF CLASSES

We use **Sturges' Rule** to find the number of classes "k", according to Sturges' Rule

$$Number\ of\ \ Classes = k = 1 + 3.3\log(n)$$

Sturges' formula suggests a reasonable number of classes, however, to get "nice" bin limits, you may choose more or fewer bins.

STEP-3 CHOOSING THE CLASS WIDTH

For guidance, find the approximate width of each class by dividing the data range by the number of classes.

$$Class\ Width = \frac{X_{\max} - X_{nin}}{k}$$

Round the bin width *up* to an appropriate value, then set the lower limit for the first bin as a multiple of the bin width. What does "appropriate" mean? If the data are discrete, then it makes sense to have a width that is an integer value. If the data are continuous, then setting a bin width equal to a fractional value may be appropriate. Experiment until you get aesthetically pleasing bins that cover the data range.

STEP-4 CHOOSE LOWER LIMIT OF THE FIRST CLASS OR STARTING POINT

Any convenient number that is equal to or less than the smallest value in the data set can be used as the lower limit of the first class.

In general, the lower limit is *included* in the class, while the upper limit is *excluded*. MINITAB follow this convention. However, Excel's histogram option *includes* the upper limit and *excludes* the lower limit.

STEP-5 CREATING THE TABLE

Create the table using all the values of data set by Put the data values in the appropriate class using Tally Method.

EXAMPLE-2 FREQUENCY DISTRIBUTION OF QUATITATIVE DATA

Consider a sample of 50 final exam scores taken from last semester's elementary statistics class. Table below lists the 50 scores. Construct a Frequency Distribution for this data set.

Statistics Exam Scores [TA02-06]

60	47	82	95	88	72	67	66	68	98	90	77	86
58	64	95	74	72	88	74	77	39	90	63	68	97
70	64	70	70	58	78	89	44	55	85	82	83	
72	77	72	86	50	94	92	80	91	75	76	78	

SOLUTION We apply Procedure discussed above

STEP-1 SORTING THE DATA IN ASCENDING ORDER

First of all we have sorted the data in ascending order and found that

$$Minimum\ Point = 39 \quad and \quad Maximim\ Point = 98$$

STEP-2 CHOOSING THE NUMBER OF CLASSES

We use **Sturges' Rule** to find the number of classes "k", according to Sturges' Rule

$$Number\ of\ Classes = k = 1 + 3.3 \times \log(50) = 6.61\ classes$$

So using Sturges' formula we have $Number\ of\ Classes = k = 7\ classes$

STEP-3 CHOOSING THE CLASS WIDTH

Find the approximate width of each class

$$Class\ width = \frac{98 - 39}{7} = 8.42$$

We shall select $class\ width = 10$ as it is most appropriate.

STEP-4 CHOOSING THE STARTING POINT

Pick a starting point. This starting point should be a little smaller than the lowest score, *L*. Suppose we start at 35; counting from there by tens (the class width), we get 35, 45, 55, 65, . . . , 95, 105. These are called the **class boundaries**.

STEP-5 CREATING THE TABLE

Standard Chart for Frequency Distribution

Class Number	Class Tallies	Boundaries	Frequency
1	||	$35 \le x < 45$	2
2	||	$45 \le x < 55$	2
3	||||| ||	$55 \le x < 65$	7
4	||||| ||||| |||	$65 \le x < 75$	13
5	||||| ||||| |	$75 \le x < 85$	11
6	||||| ||||| |	$85 \le x < 95$	11
7	||||	$95 \le x \le 105$	4
			50

VISUALIZING QUANTITATIVE DATA

In this section we shall discuss graphs that are used to summarize quantitative data. There are different types of graphs including

- Histogram
- Frequency Polygon
- Ogive

THE HISTOGRAM

Karl Pearson introduced the histogram in 1891. He used it to show time concepts of various reigns of Prime Ministers. A histogram for a frequency distribution based on quantitative data is similar to the bar chart showing the distribution of qualitative data. The classes are marked on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by the heights of the bars. However, there is one important difference based on the nature of the data. Quantitative data are usually measured using scales that are continuous, not discrete. Therefore, the horizontal axis represents all possible values, and the bars are drawn adjacent to each other to show the continuous nature of the data.

A histogram is also called a **frequency histogram**, a **relative frequency histogram**, or a **percentage histogram** depending on whether frequencies, relative frequencies, or percentages are marked on the vertical axis.

HISTOGRAM A graph in which the classes are marked on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by the heights of the bars, and the bars are drawn adjacent to each other.

PROCEDURE OF CONSTRUCTING A HISTOGRAM

To Construct a Histogram

Step 1 **Obtain a frequency (relative-frequency, percent) distribution of the data.**

Step 2 **Draw a horizontal axis on which to place the bars and a vertical axis on which to display the frequencies (relative frequencies, percents).**

Step 3 **For each class, construct a vertical bar whose height equals the frequency (relative frequency, percent) of that class.**

Step 4 **Label the bars with the classes, as explained in Definition 2.9, the horizontal axis with the name of the variable, and the vertical axis with "Frequency" ("Relative frequency," "Percent").**

EXAMPLE-8 RECORD HEIGHT TEMPERATURE

Construct a histogram to represent the data shown for the record high temperatures for each of the 50 states having class boundaries and frequencies as follows

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

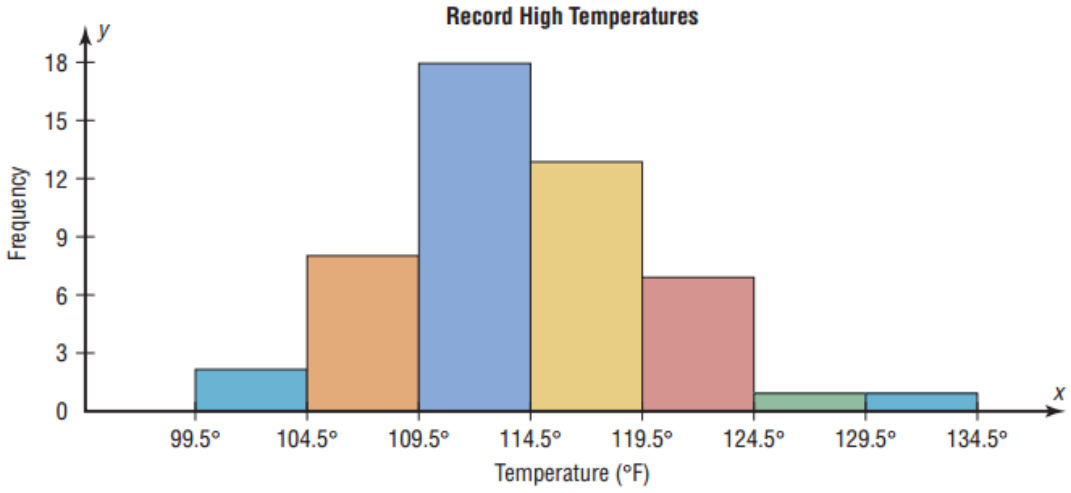
SOLUTION We apply Procedure discussed above

STEP-1 Draw and label the x and y axes. The x axis is always the horizontal axis, and the y axis is always the vertical axis.

STEP-2 Represent the frequency on the y axis and the class boundaries on the x axis.

STEP-3 Using the frequencies as the heights, draw vertical bars for each class. See Figure 2.11 below

FIGURE 2.11 HISTOGRAM



DIFFERENT SHAPES OF HISTOGRAMS

Histograms are valuable tools. For example, the histogram of a sample should have a distribution shape very similar to that of the population from which the sample was drawn. If the reader of a histogram is at all familiar with the variable involved, he or she will usually be able to interpret several important facts. Figure 2.12 presents histograms with specific shapes that suggest descriptive labels. Possible descriptive labels are listed under each histogram.

Briefly, the terms used to describe histograms are as follows:

Symmetrical: Both sides of this distribution are identical (halves are mirror images).

Normal: A symmetrical distribution is mounded up about the mean and becomes sparse at the extremes.

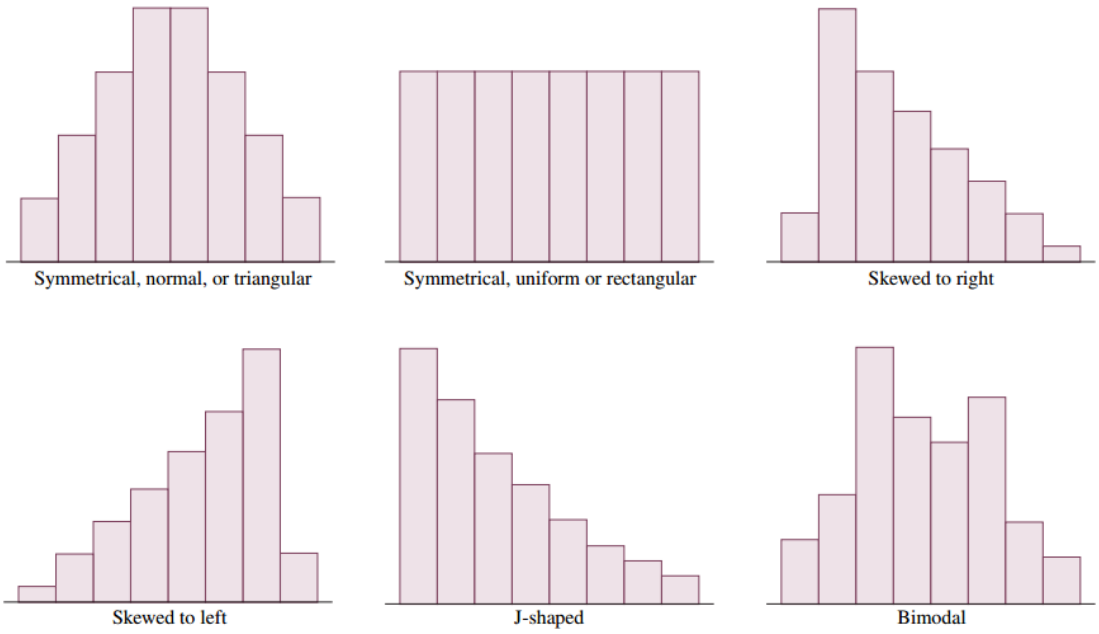
Uniform (rectangular): Every value appears with equal frequency.

Skewed: One tail is stretched out longer than the other. The direction of skewness is on the side of the longer tail.

J-shaped: There is no tail on the side of the class with the highest frequency.

Bimodal: The two most populous classes are separated by one or more classes. This situation often implies that two populations are being sampled.

FIGURE DIFFERENT SHAPES OF HISTOGRAM



THE FREQUENCY POLYGON

A frequency polygon also shows the shape of a distribution and is similar to a histogram. It consists of line segments connecting the points formed by the intersections of the class midpoints and the class frequencies. The midpoint of each class is scaled on the X-axis and the class frequencies on the Y-axis.

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

EXAMPLE-9 RECORD HEIGHT TEMPERATURE

Construct a frequency polygon to represent the data shown for the record high temperatures for each of the 50 states having class boundaries and frequencies as follows

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 We apply Procedure discussed above

STEP-1 Find the midpoints of each class. Recall that midpoints are found by adding the upper and lower boundaries and dividing by 2:

$$\frac{99.5 + 104.5}{2} = 102 \qquad \frac{104.5 + 109.5}{2} = 107$$

and so on. The midpoints are

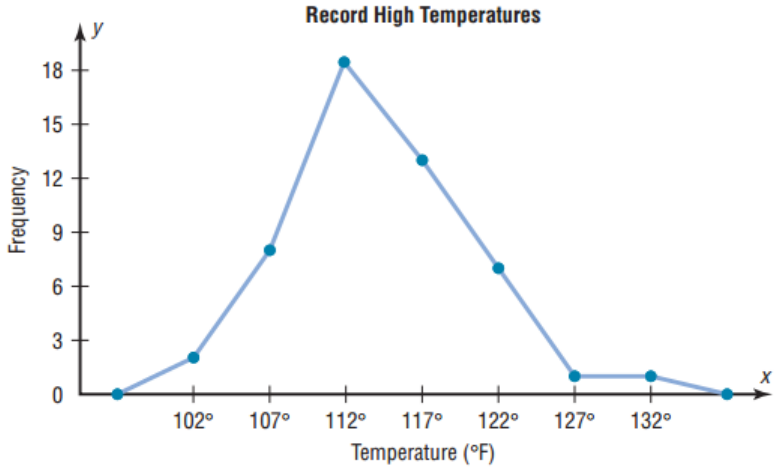
Class boundaries	Midpoints	Frequency
99.5–104.5	102	2
104.5–109.5	107	8
109.5–114.5	112	18
114.5–119.5	117	13
119.5–124.5	122	7
124.5–129.5	127	1
129.5–134.5	132	1

STEP-2 Draw the x and y axes. Label the x axis with the midpoint of each class, and then use a suitable scale on the y axis for the frequencies.

STEP-3 Using the midpoints for the x values and the frequencies as the y values, plot the points.

STEP-4 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, as shown in Figure below

FREQUENCY POLYGON



THE CUMULATIVE FREQUENCY POLYGON (OGIVE)

The cumulative frequency polygon, or ogive, uses the cumulative frequency distribution to display the variable of interest along the X-axis and the cumulative percentages along the Y-axis. 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-10 RECORD HEIGHT TEMPERATURE

Construct a cumulative frequency polygon to represent the data shown for the record high temperatures for each of the 50 states having class boundaries and frequencies as follows

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 We apply Procedure discussed above

STEP-1 Find the cumulative frequency for each class.

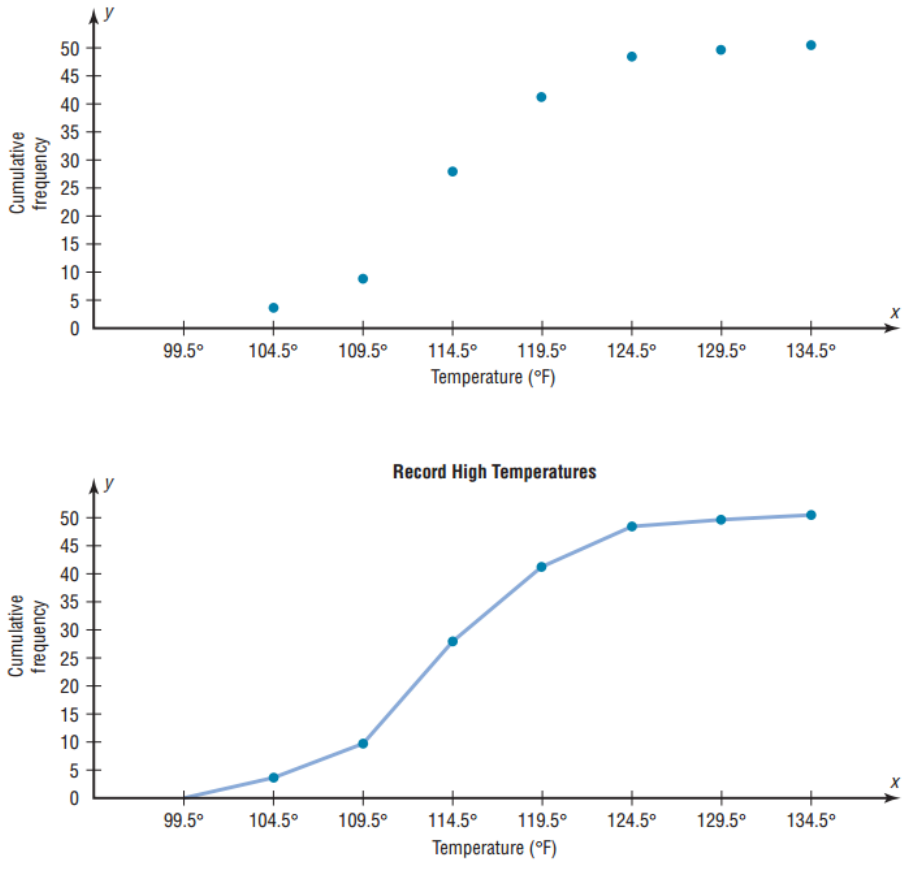
	Cumulative frequency
Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

STEP-2 Draw thex and y axes. Label the x axis with the class boundaries. Use an appropriate scale for the y axis to represent the cumulative frequencies.

STEP-3 Plot the cumulative frequency at each upper class boundary, as shown in Figure 2.13. Upper boundaries are used since the cumulative frequencies represent the number of data values accumulated up to the upper boundary of each class.

STEP-4 Starting with the first upper class boundary, 104.5, connect adjacent points with line segments, as shown in Figure 2.13. Then extend the graph to the first lower class boundary, 99.5, on the x axis.

FIGURE 2.13 CUMULATIVE FREQUENCY POLYGON



EXAMPLE-11 MILES RUN PER WEEK

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

Class boundaries	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2
	20

SOLUTION We apply Procedure discussed above

STEP-1 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 $\frac{2}{20} = 0.10$; for class 15.5–20.5, the relative frequency is $\frac{3}{20} = 0.15$; and so on.

Place these values in the column labeled Relative frequency.

Class boundaries	Midpoints	Relative frequency
5.5–10.5	8	0.05
10.5–15.5	13	0.10
15.5–20.5	18	0.15
20.5–25.5	23	0.25
25.5–30.5	28	0.20
30.5–35.5	33	0.15
35.5–40.5	38	0.10
		1.00

STEP-2 Find the cumulative relative frequencies. To do this first find the cumulative frequencies and then convert each one to a relative frequency.

	Cumulative frequency	Cumulative relative frequency
Less than 5.5	0	0.00
Less than 10.5	1	0.05
Less than 15.5	3	0.15
Less than 20.5	6	0.30
Less than 25.5	11	0.55
Less than 30.5	15	0.75
Less than 35.5	18	0.90
Less than 40.5	20	1.00

STEP-3 Draw each graph as shown in Figure 2.14. 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.

FIGURE 2.14 HISTOGRAM, FREQUENCY POLYGON AND OGIVE