Chapter 1: Using Graphs to Describe Data

Department of Mathematics
Izmir University of Economics

Week 1
2014-2015

Introduction

In this chapter we will focus on

- the definitions of *population*, *sample*, *parameter*, and *statistic*,
- the classification of variables,
- the graphs to describe these variables.

Introduction

Statistics are used to predict or forecast sales of a new product, construction costs, customer satisfaction levels, the weather, election results, university enrollment figures, grade point averages, interest rates, currency-exchange rates, and many other variables that affect our daily lives. We need to absorb and interpret substantial amounts of data. Governments, businesses, and scientific researchers spend a lot of money and resources collecting data.

Introduction

For example

- Supermarkets want to know how and where their customers spend money,
- Investors want to know how to best investigate their money and capital,
- Politicians wants to know how they can obtain more power and more votes.

For each of these situations, we must carefully define the problem, determine what data are needed, collect the data, use statistics to summarize the data, and make inferences and decisions based on the data obtained.

Introduction

AMSTAT (The American Statistical Association) defines the term *statistics* as the scientific application of mathematical principles to the collection, analysis, and presentation of data.

In statistical applications we are usually interested in data that comes from a specific source. For example if we are interested in approval of a newly manufactured drug, we are especially interested in people/patients who will use this drug.

Random and Systematic Sampling

Definition:

A *population* is the **complete set of all items** that interest an investigator.

Definition:

A *sample* is an **observed subset** of population.

A sample is used to obtain useful information that have some validity about the population at large.

In general, we denote the population size by N (which is usually a very large number) and the sample size by n. While $n \leq N$ always holds, in general $n << N$ (n is much smaller than N).

Random and Systematic Sampling

There are two important types of sampling, namely (simple) random sampling and systematic sampling.

Simple random sampling is a procedure used to select a sample of n objects from a population in such a way that

- each member of the population is chosen strictly by chance,
- the selection of one member does not influence the selection of any other member,
- each member of the population is equally likely to be chosen, and
- every possible sample of a given size, n, has the same chance of selection.

The resulting sample is called a *random sample*.

Random and Systematic Sampling

Suppose that the population list is arranged in some fashion unconnected with the subject of interest. *Systematic sampling* involves the selection of every jth item in the population, where j is the ratio of the population size N to the desired sample size n, that is, $j = \frac{N}{n}$. We, then, randomly select a number from 1 to j to obtain the first item to be included in our systematic sample.

Random and Systematic Sampling

Example. Suppose that we wish to know the average age of university students in Turkey who are smokers. Describe the related population and suggest ways to construct a sample from this population.
Solution. Our population consists of university students in Turkey who are smokers ($N \approx 1000000$). Since N is very large to calculate the average age, we have to construct a sample of an appropriate size, say $n = 500$. We can construct this sample either randomly or systematically. To construct a systematic sample, we can fist sort these smoker students according to their names in an alphabetical order, then choose a number from 1 to $j = \frac{1000000}{500} = 2000$, say 1905, and finally select the 500 smoker students numbered 1905, 3905, 5905, 7905, and so on.

Random and Systematic Sampling

Definition:

A *parameter* is a numerical measure that describes a specific characteristic of a **population**.

Definition:

A *statistic* is a numerical measure that describes a specific characteristic of a **sample**.

Average age of all university students in Turkey that are smokers is an example to the term parameter, while average age of the sample (with size 500) constructed from university students in Turkey that are smokers is an example to the term statistic.

Random and Systematic Sampling

Example. One airline company claims that less than 1% of its scheduled flights out of Orlando International Airport (OIA) depart late. From a random sample of 200 flights, 1.5% were found to be departing later than the scheduled time.

a) What is the population?

b) What is the sample?

c) Is 1.5% a parameter or a statistic?

Descriptive and Inferential Statistics

There are two branches of statistics that are both used to change data into knowledge:

- *Descriptive statistics* focus on graphical and numerical procedures that are used to summarize and process data.
- *Inferential statistics* focus on using the data to make predictions, forecasts, and estimates to make better decisions.

Descriptive statistics transform data into information and inferential statistics transform this information into knowledge.

Descriptive and Inferential Statistics

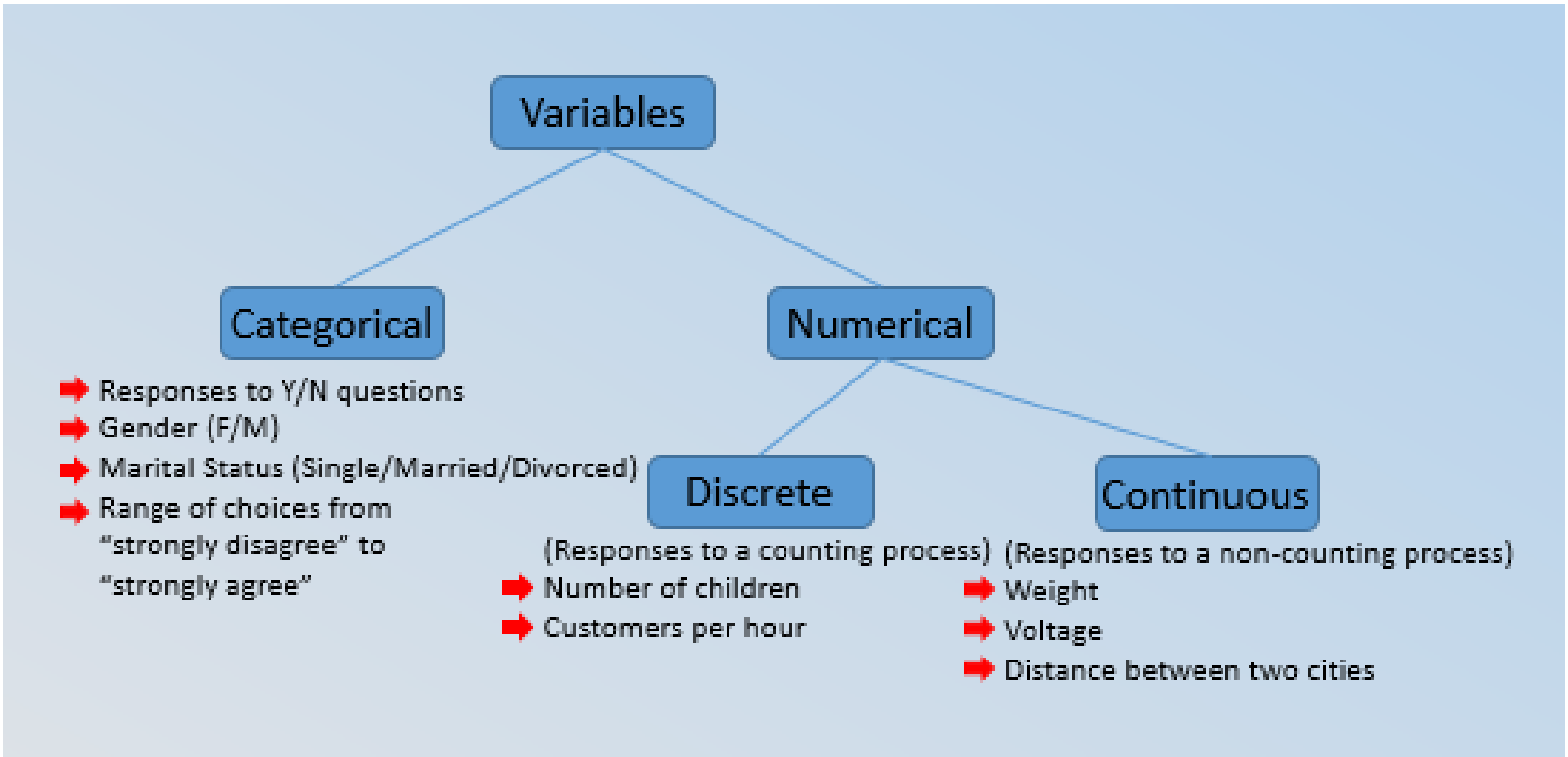
Example. Determine whether descriptive statistics or inferential statistics should be used to obtain the following information.

a) A graph that shows the number of defective bottles produced during the day-shift over one week's time.

b) An estimate of the percentage of the employees who arrive to work late.

Classification of Variables

A *variable* is a specific characteristic of an individual or an object. Variables can be classified in several ways. One method of classification refers to the type and amount of information contained in the data.



Classification of Variables

Example. A mortgage company randomly samples accounts of their time-share customers. State whether each of the following variables is categorical or numerical. If numerical, is it discrete or continuous?

a) The original purchase price of a customer's time-share unit.

b) The country of residence of a time-share customer.

c) A time-share owner's satisfaction level with the maintenance of the unit purchased (from 1:very dissatisfied to 5:very satisfied).

d) The number of times a customer's payment was late.

Classification of Variables

Example. Faculty at one university were asked a series of questions in a recent survey. State the type of data for each question.

a) Indicate your level of satisfaction with your teaching load (very satisfied, moderately satisfied, neutral, moderately dissatisfied, or very dissatisfied).

b) How many of your research articles were published in refereed journals during the last 5 years?

c) Did you attend the university faculty meeting?

d) Do you think that the teaching evaluation process needs to be revised?

Graphs to Describe Categorical Variables

Categorical variables can be described using frequency distribution tables and graphs like bar charts and pie charts.

Frequency distribution tables

A *frequency distribution* is a table used to organize data. We use classes to construct frequency distribution tables of a categorical variable and these are simply the possible responses on that variable. A frequency distribution table includes classes (all possible responses on the variable being studied) and frequencies (number of observations for each class).

A *relative frequency distribution* is obtained by dividing each frequency by the total number of observations and multiplying the resulting proportion by 100.

Frequency distribution tables

An example of a frequency distribution and a relative frequency distribution is

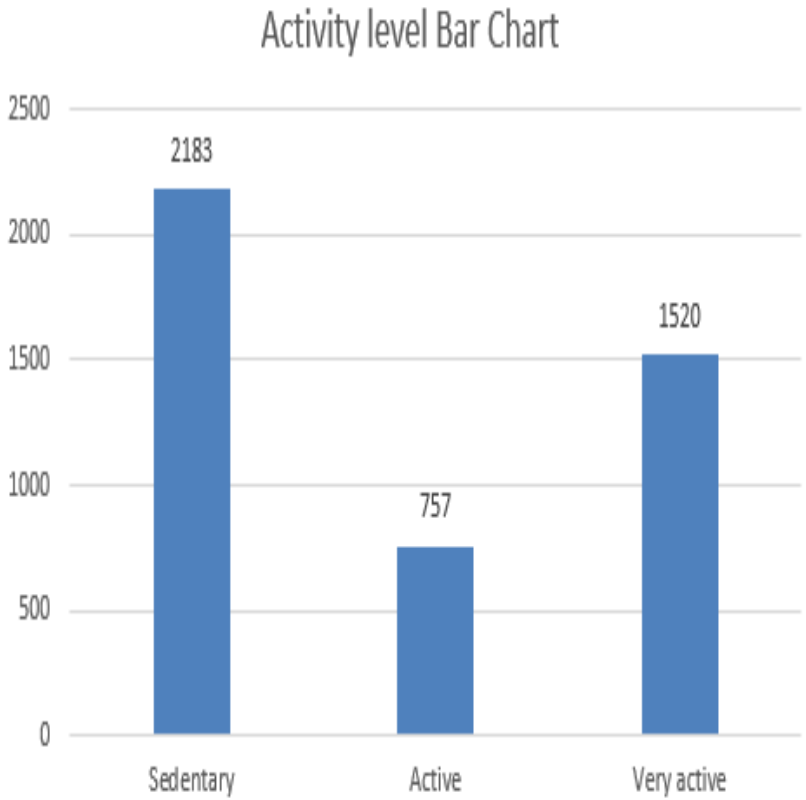
(Classes)	(Frequencies)	(Relative Frequencies)	
Activity level	**Number of participants**	**Percent**	
Sedentary	2183	48,9	(=(2183/4460)*100)
Active	757	17,0	(=(757/4460)*100)
Very active	1520	34,1	(=(1520/4460)*100)
Total	4460	100,0	

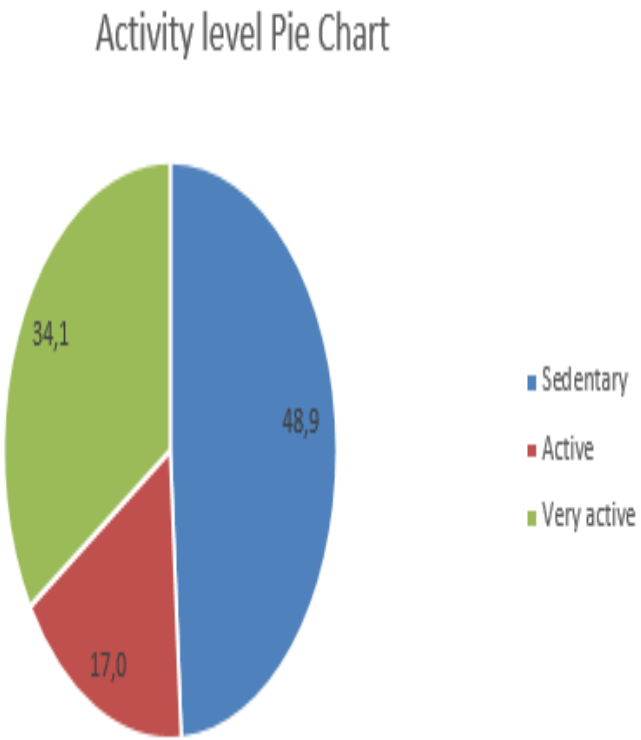
Bar Charts and Pie Charts

Bar charts and *pie charts* are commonly used to describe categorical data. If we want to draw attention to the frequency of each category, then we will probably use a bar chart. If we want to draw attention to the proportion of frequencies in each category (relative frequencies), then we will probably use a pie chart.

Bar Charts and Pie Charts

The respective bar and pie charts corresponding to the previous frequency distribution are





Bar Charts and Pie Charts

In a bar chart

- the height of a rectangle represent each frequency,
- there is no need for the bars to touch.

In a pie chart

- the circle (pie) represents the total and the segments (pieces of the pie) cut from its center indicates shares of the total,
- each segment is proportional to the corresponding frequency.

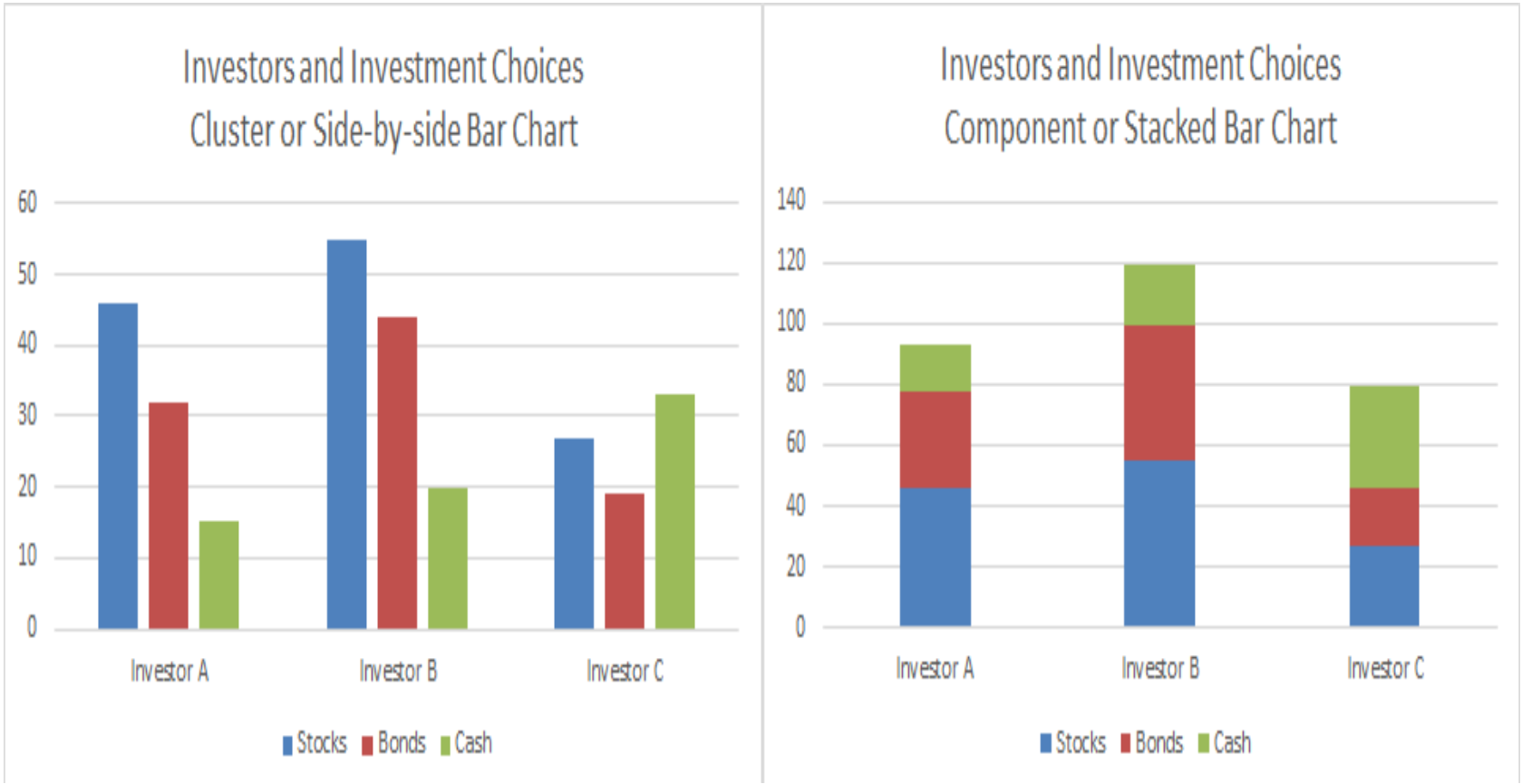
Cross Tables

A *cross table* lists the number of observations for every combination of values for **two categorical variables**. If there are r categories for the first variable and c categories for the second variable (r rows and c columns), the table is called an $r \times c$ cross table. The following is an example of 3×3 cross table

Investors / Investment Choices	Investor A	Investor B	Investor C	Total
Stocks	46	55	27	128
Bonds	32	44	19	95
Cash	15	20	33	68
Total	93	119	79	291

Cross Tables

We can graph this 3 $\times$ 3 cross table using either a cluster (side-by-side) or a component (stacked) bar chart.



Graphs to Describe Numerical Variables

Numerical variables can be described using frequency distributions, histograms, ogives, and stem-and-leaf displays.

Frequency distributions

Similar to a frequency distribution for categorical data, a frequency distribution for numerical data is a table that summarizes data by listing the classes in the left column and the frequencies in the right column. However, the classes for a frequency distribution of numerical data are not as easily identifiable.

Frequency distributions

Construction of a Frequency Distribution

Rule 1: Determine k (number of classes) using the following quick guide:

Sample size (n)	Number of classes (k)
< 50	5 - 7
50 - 100	7 - 8
101 - 500	8 - 10
501 - 1000	10 -11
1001 - 5000	11 - 14
> 5000	14 - 20

Rule 2: Choose the class width:

$$\text{class width} = \frac{\text{largest observation-smallest observation}}{\text{number of classes}} = \frac{\text{range}}{k}$$

Always round class width upward (depending on the decimal places in the data)!!!

Rule 3: Classes must be inclusive and nonoverlapping. Each observation must belong to one and only one class.

Frequency distributions

A *relative frequency distribution* is obtained by dividing each frequency by the total number of observations and multiplying the resulting proportion by 100.

A *cumulative frequency distribution* contains the total number of observations whose values are less than the upper limit for each class. We construct a cumulative frequency distribution by adding the frequencies of all frequency distribution classes up to and including the present class.

In a *relative cumulative frequency distribution*, cumulative frequencies can be expressed as cumulative proportions or percents.

Frequency distributions

Example. A manufacturer of insulation randomly selects 20 days and records the daily high temperature:

$$24, 35, 17, 21, 24, 37, 26, 46, 58, 30,$$

$$32, 13, 12, 38, 41, 43, 44, 27, 53, 27$$

Construct a frequency distribution and cumulative frequency distribution.

Histograms

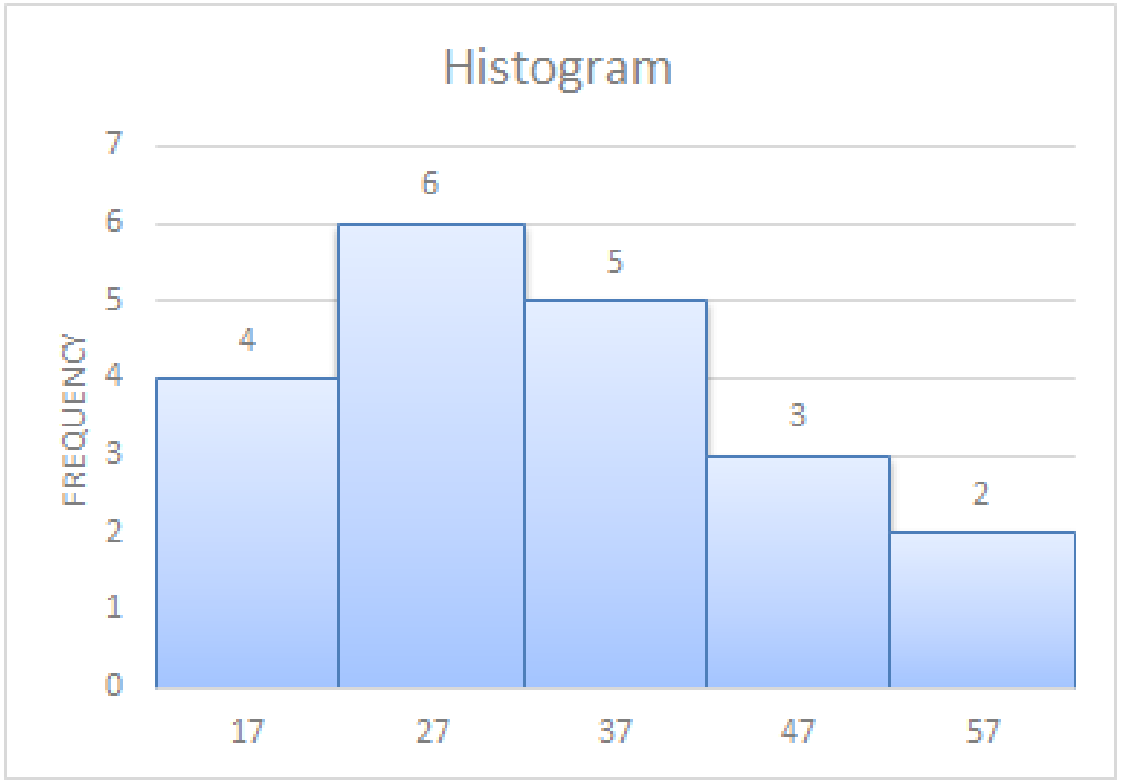
A *histogram* is a graph that consists of vertical bars constructed on a horizontal line that is marked off with intervals for the variable being displayed.

In a histogram

- the intervals correspond to the classes in a frequency distribution,
- the widths of the bars are equal,
- the height of each bar is proportional to the number of observations in that interval,
- the number of observations can be displayed above the bars,
- there are no gaps between the bars.

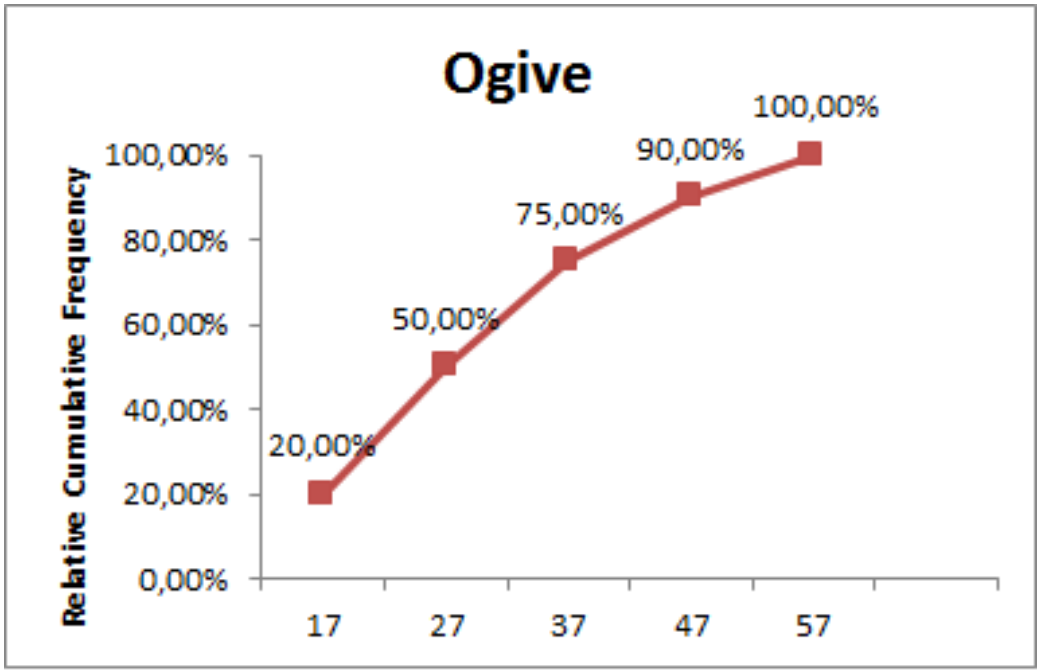
Histograms

The histogram related with the previous frequency distribution is



Ogives

An *ogive* is a line that connects points that are the cumulative percent of observations below the upper limit of each interval in a cumulative frequency distribution. The ogive related with the previous frequency distribution is



Stem-and-Leaf Displays

A *stem-and-leaf display* is an EDA (exploratory data analysis) graph that is an alternative to the histogram. Data are grouped according to their leading digits (called stems) and the final digits (called leaves) are listed separately for each member of a class. The leaves are displayed individually in ascending order after each of the stems. The stem-and-leaf display related with the previous data is

Stem	Leaf
1	2 3 7
2	1 4 4 6 7 7
3	0 2 5 7 8
4	1 3 4 6
5	3 8

Stem-and-Leaf Displays

Example. Describe the following sample with a stem-and-leaf display

$$88, 51, 63, 85, 79, 65, 79, 70, 73, 77$$

Scatter Plot

A *scatter plot* can be prepared by locating one point for each pair of **two variables** that represent an observation in the data set. The scatter plot provides a picture of the data including:

1. the range of each variable,
2. the pattern of values over the range,
3. a suggestion as to a possible relationship between the two variables, and
4. an indication of outliers (extreme points).

Scatter Plot

The scatter plot for the data

SAT Math Score	GPA
450	3.25
480	2.60
500	2.88
520	2.85
560	3.30
580	3.10
590	3.35
600	3.20
620	3.50
650	3.59
700	3.95

is

Scatter Plot