Elementary Statistics 15060101

Statistics:

is the **science** of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data.

Statistics

Descriptive

consists of the collection, organization, summarization, and presentation of data.

Inferential Statistics

consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions

Note: Inferential statistics uses probability

A variable: is a characteristic or attribute that can assume different values.

Data are the values (measurements or observations) that the variables can assume.

Variables whose values are determined by chance are called **random variables.**

A **population** consists of all subjects (human or otherwise) that are being studied.

A **sample** is a group of subjects selected from a population.

Population

Sample

Variables and Types of Data

Variables can be classified as qualitative or quantitative.

Qualitative variables: are variables that can be placed into distinct categories, according to some characteristic or attribute.
Example, if subjects are classified according to gender (male or female), then the variable gender is qualitative.
Example, Geographic locations.

Quantitative variables: are numerical and can be ordered or ranked.
Example, the variable age is numerical, and people can be ranked in order according to the value of their ages.
Example, heights, weights, and body temperatures.

Also, quantitative variables can be further classified into two groups: discrete and continuous.

Discrete variables can be assigned values such as 0, 1, 2, 3 and are said to be countable. Discrete variables assume values that can be counted.

Example, the number of children in a family, the number of students in a classroom, and the number of calls received by a switchboard operator each day for a month.

Continuous variables,
assume an infinite number of values between any two specific values [interval]. Not countable. They often include fractions and decimals.

Example, Temperature is a continuous variable, since the variable can assume an infinite number of values between any two given temperatures.

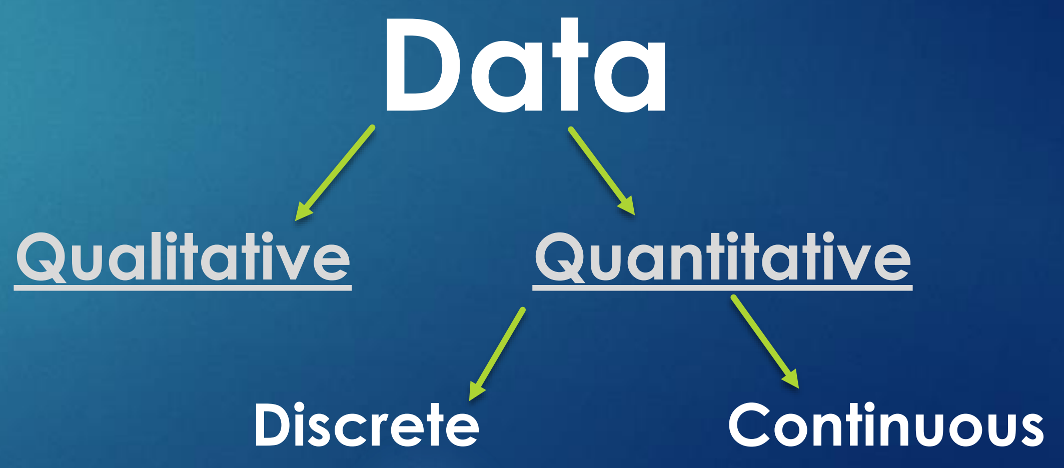
Qualitative or Quantitative, variables can be classified by how they are categorized, counted, or measured:
Nominal and Ordinal

The **nominal level of measurement** classifies data into mutually exclusive (**nonoverlapping**), exhausting categories in which **no order** or ranking can be imposed on the data.
Example: Gender, Location, Name, ….

The **ordinal level of measurement classifies** data into **categories** that can be ranked; however, precise differences between the ranks do not exist.
Example: Educational level, age level, job level, ….

Variables can also classified into interval or ratio level of measurements:
The **interval level** of measurement **ranks** data, and precise differences between units of measure do exist; however, there is **no meaningful zero**
Example: Temperature, Intelligence quotient (IQ) test

The **ratio level** of measurement possesses all the characteristics of interval measurement, and there **exists a true zero**. In addition, **true ratios** exist when the same variable is measured on two different members of the population.
Example, height, weight, area, and number of phone calls received.



Example:

Table 1–2 Examples of Measurement Scales

Nominal-level data	Ordinal-level data	Interval-level data	Ratio-level data
Zip code	Grade (A, B, C, D, F)	SAT score	Height
Gender (male, female)	Judging (first place, second place, etc.)	IQ	Weight
Eye color (blue, brown, green, hazel)	Rating scale (poor, good, excellent)	Temperature	Time
Political affiliation	Ranking of tennis players		Salary
Religious affiliation			Age
Major field (mathematics, computers, etc.)			
Nationality			

Section 1.3: Data Collection and Sampling Techniques

An example of the importance of collecting data and making a statistical analysis:
A manufacturer might want to know something about the consumers who will be purchasing his product so he can plan an effective marketing strategy.

Data can be collected in a variety of ways.
Surveys is the most common one.

Surveys can be done by using a variety of methods as:
the telephone survey ----- the mailed questionnaire ----- the personal interview

Researchers use samples to collect data and information about a particular variable from a large population.
Samples saves time and money and in some cases enables the researcher to get more detailed information about a particular subject.

Statisticians use four basic methods of sampling:
random, systematic, stratified, and cluster sampling.

Random Sampling:

are selected by using chance methods or **random** numbers.

Use: random number generator or pick a random card…

Systematic Sampling

Researchers obtain systematic samples by numbering each subject of the population and then selecting every kth subject (k=N/n). For example, suppose there were 2000 subjects in the population and a sample of 50 subjects were needed. Since 2000/50 = 40, then k = 40, and every 40th subject would be selected; however, the first subject (numbered between 1 and 40) would be selected at **random**. Suppose subject 12 were the first subject selected; then the sample would consist of the subjects whose numbers were 12, 52, 92, etc., until 50 subjects were obtained

Stratified Sampling

Researchers obtain stratified samples by dividing the population into groups (called **strata**) according to some characteristic that is important to the study, then sampling from each group. Samples within the strata should be **randomly** selected.

Cluster Sampling

Here the population is divided into groups called **clusters** by some means such as geographic area or schools in a large school district, etc. Then the researcher **randomly** selects some of these clusters and uses **all members** of the selected clusters as the subjects of the samples.

Table 1–4	Summary of Sampling Methods
Random	Subjects are selected by random numbers.
Systematic	Subjects are selected by using every *k*th number after the first subject is randomly selected from 1 through *k*.
Stratified	Subjects are selected by dividing up the population into groups (strata), and subjects are randomly selected within groups.
Cluster	Subjects are selected by using an intact group that is representative of the population.

Observational and Experimental Studies

There are several different ways to classify statistical studies.

Observational study, the researcher merely observes what is happening or what has happened in the past and tries to draw conclusions based on these observations.

Experimental study, the researcher manipulates one of the variables and tries to determine how the manipulation influences other variables.

The **independent variable** in an experimental study is the one that is being manipulated by the researcher. The independent variable is also called the **explanatory** variable.

The **resultant variable** is called the **dependent** variable or the **outcome** variable.

Chapter 2: Frequency distributions and Graphs (Textbook pages 35-62)

Categorical Frequency distribution

Use the following raw data to construct a frequency table:

DATA FROM A SAMPLE OF 50 SOFT DRINK PURCHASES

Coca-Cola	Coca-Cola	Coca-Cola	Sprite	Coca-Cola
Diet Coke	Dr. Pepper	Diet Coke	Dr. Pepper	Diet Coke
Pepsi	Sprite	Coca-Cola	Pepsi	Pepsi
Diet Coke	Coca-Cola	Sprite	Diet Coke	Pepsi
Coca-Cola	Diet Coke	Pepsi	Pepsi	Pepsi
Coca-Cola	Coca-Cola	Coca-Cola	Coca-Cola	Pepsi
Dr. Pepper	Coca-Cola	Coca-Cola	Coca-Cola	Coca-Cola
Diet Coke	Sprite	Coca-Cola	Coca-Cola	Dr. Pepper
Pepsi	Coca-Cola	Pepsi	Pepsi	Pepsi
Pepsi	Diet Coke	Coca-Cola	Dr. Pepper	Sprite

FREQUENCY DISTRIBUTION OF SOFT DRINK PURCHASES

Soft Drink	Frequency
Coca-Cola	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	50

Grouped Frequency distribution

Discover the frequency table contents: Class limit, class boundaries, lower and upper limit (or boundary), class width, class midpoint , frequencies and cumulative frequency

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

How you can convert raw data to frequency table:

- Determine Number of classes (assumed by researcher)
- Compute Class width (round up)= (max{x}-min{x})/Number of classes

$$\text{Approximate class width} = \frac{\text{Largest data value} - \text{Smallest data value}}{\text{Number of classes}}$$

The approximate class width given above can be rounded up. Example: 9.28 might be rounded to 10

Example

YEAR-END AUDIT TIMES (IN DAYS)

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

class width=(33 − 12)/5 = 4.2 → 5

The smallest value is 12. we can start The first class from 10

Note:
class width=15-10=5
class width=20-15=5
class width=25-20=5
class width=30-25=5
class width=35-30=5

FREQUENCY DISTRIBUTION FOR THE AUDIT TIME DATA

Audit Time (days)	Frequency
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

Class midpoint: is the value halfway between the lower and upper class limits.

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

or

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

Audit Time (days)	Frequency	Relative Frequency	Percent Frequency	Class midpoint
10–14	4	4/20=0.20	20%	(10+14)/2=12
15–19	8	8/20=0.40	40%	(15+19)/2=17
20–24	5	5/20=0.25	25%	(20+24)/2=22
25–29	2	2/20=0.10	10%	(25+29)/2=27
30–34	1	1/20=0.05	5%	(30+34)/2=32
Total	20	1.00	100%	

Cumulative Distributions

FREQUENCY DISTRIBUTION FOR THE AUDIT TIME DATA

Audit Time (days)	Frequency
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

CUMULATIVE FREQUENCY, CUMULATIVE RELATIVE FREQUENCY, AND CUMULATIVE PERCENT FREQUENCY DISTRIBUTIONS FOR THE AUDIT TIME DATA

Audit Time (days)	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency
Less than or equal to 14.5	4	.20	20
Less than or equal to 19.5	12	.60	60
Less than or equal to 24.5	17	.85	85
Less than or equal to 29.5	19	.95	95
Less than or equal to 34.5	20	1.00	100

Exercise:

Consider the following frequency distribution (Given by Black Color)

Construct a Relative Frequency, Percent Frequency, Cumulative Frequency, Cumulative Frequency, Relative Cumulative Frequency, Class midpoint and Class Width

Class Limit	Frequency	Relative Frequency	Percent Frequency	Cumulative Frequency	Relative Cumulative Frequency	Percent Cumulative Frequency	Class midpoint	Class Width
10–19	10	0.2	20%	10	0.2	20%	14.5	10
20–29	14	0.28	28%	24	0.48	48%	24.5	10
30–39	17	0.34	34%	41	0.82	82%	34.5	10
40–49	7	0.14	14%	48	0.96	96%	44.5	10
50–59	2	0.04	4%	50	1	100%	54.5	10
Total	50							

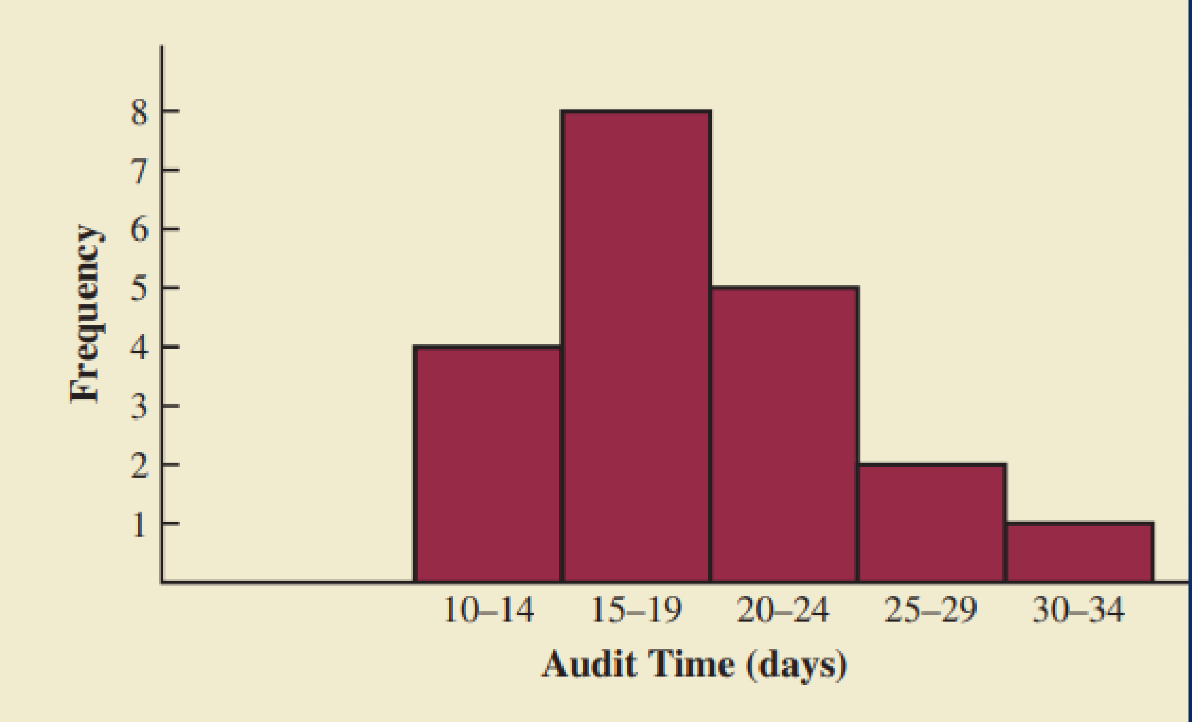
Histogram

A common graphical display of quantitative data is a histogram. This graphical display can be prepared for data previously summarized in either a frequency, relative frequency, or percent frequency distribution

Class boundaries are used to construct histogram

Audit Time (days)	Frequency	Class boundaries
10–14	4	9.5-14.5
15–19	8	14.5-19.5
20–24	5	19.5-24.5
25–29	2	24.5-29.5
30–34	1	29.5-34.5
Total	20	

FIGURE 2.5 HISTOGRAM FOR THE AUDIT TIME DATA



A histogram contains **no natural separation** between the rectangles of adjacent classes.
By making the class upper limit = next class lower limit (Class boundaries)

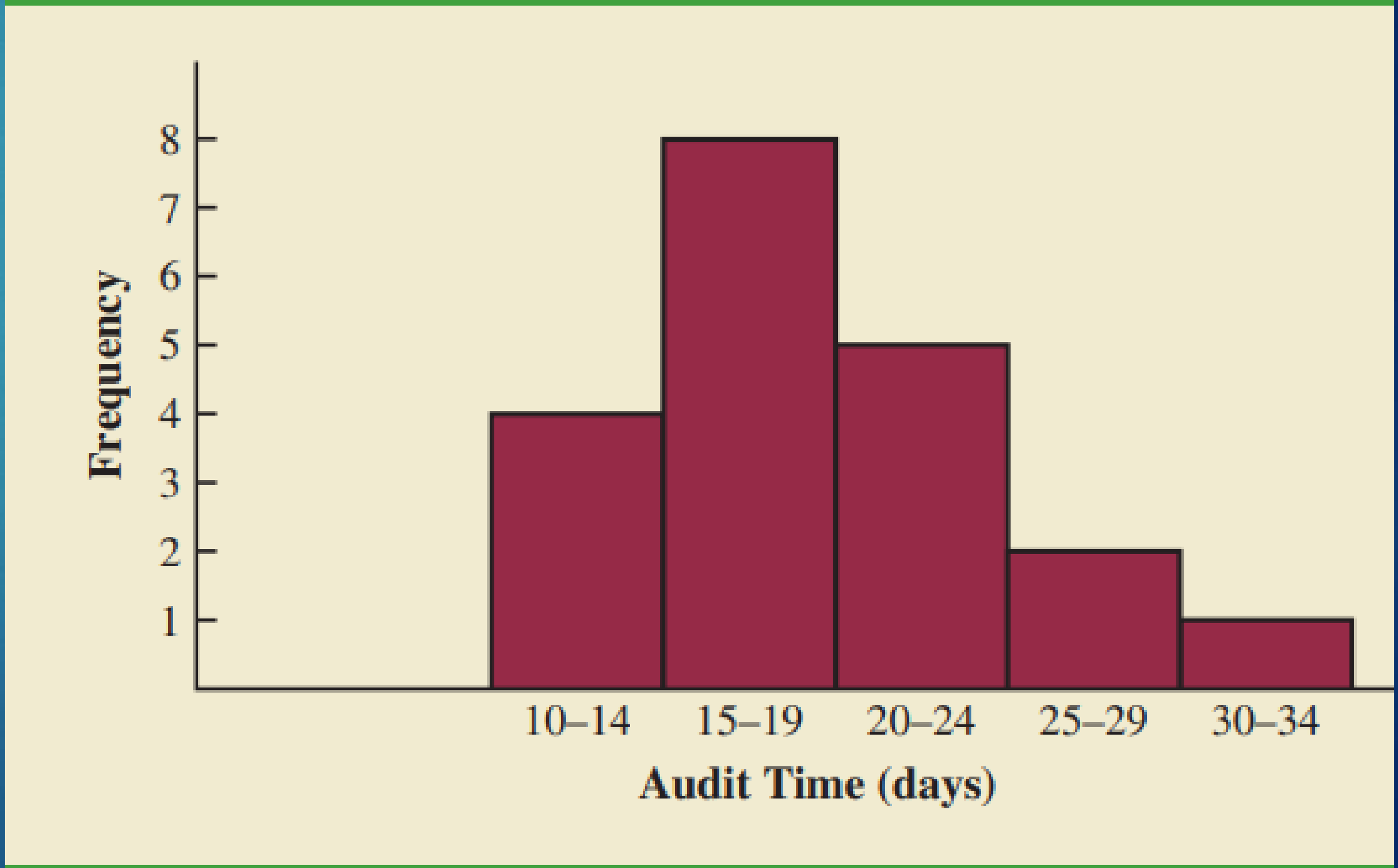
In our Example: the audit time data are stated as 10–14, 15–19, 20–24, 25–29, and 30–34, one-unit spaces of 14 to 15, 19 to 20, 24 to 25, and 29 to 30

TABLE 2.5

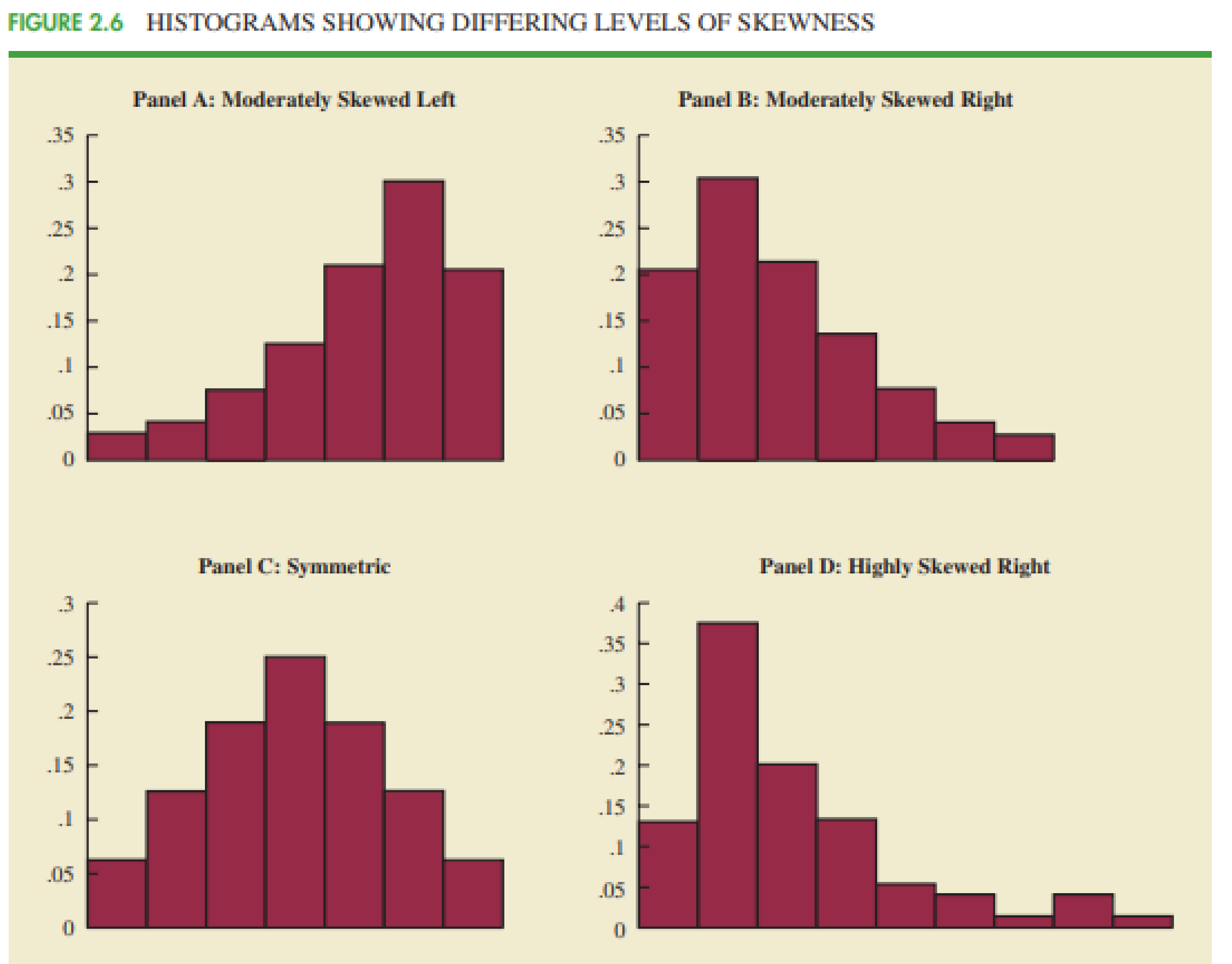
FREQUENCY DISTRIBUTION FOR THE AUDIT TIME DATA

Audit Time (days)	Frequency
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

FIGURE 2.5 HISTOGRAM FOR THE AUDIT TIME DATA

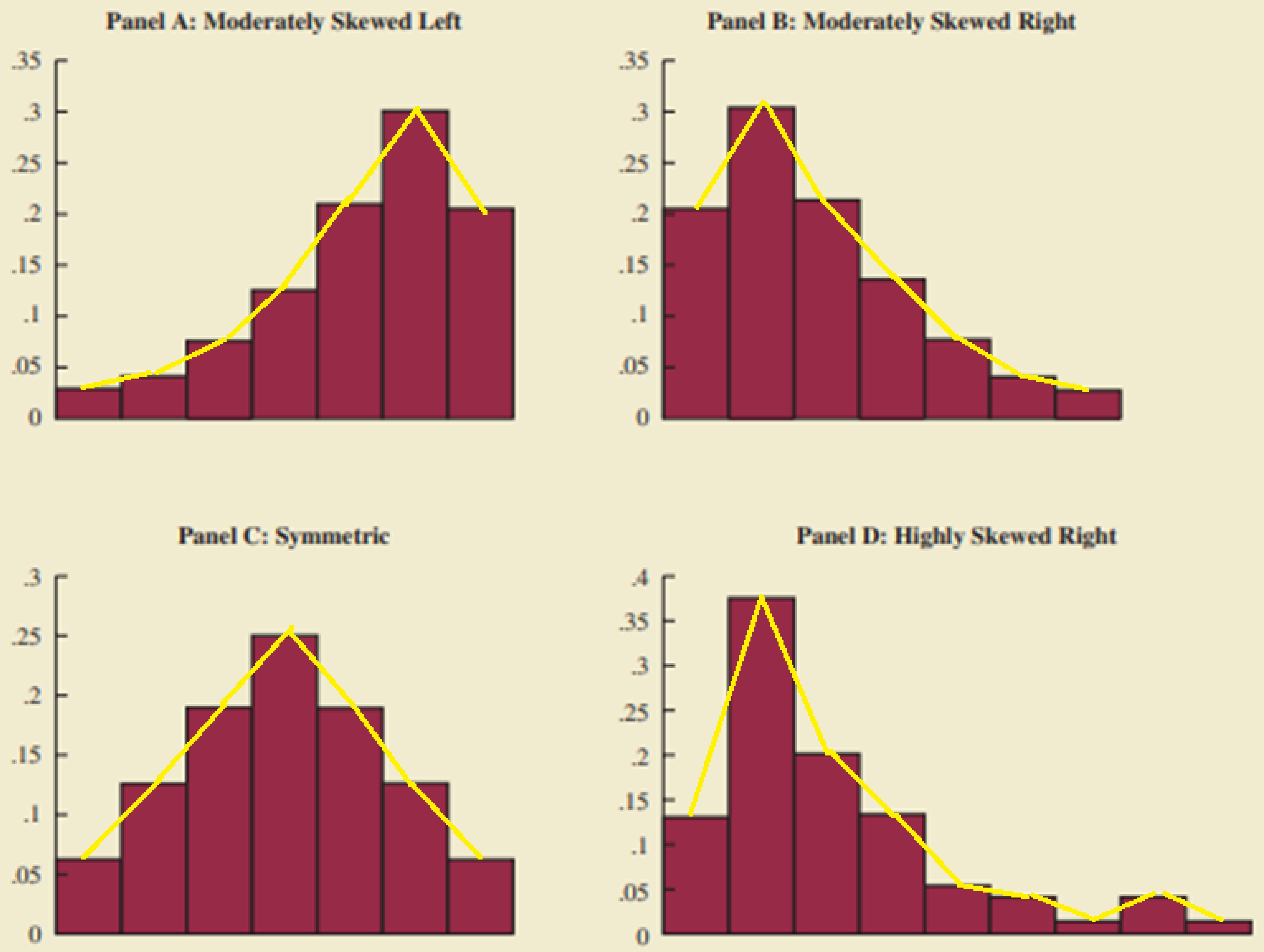


Histogram: Symmetric, Skewed to the left and Skewed to the right

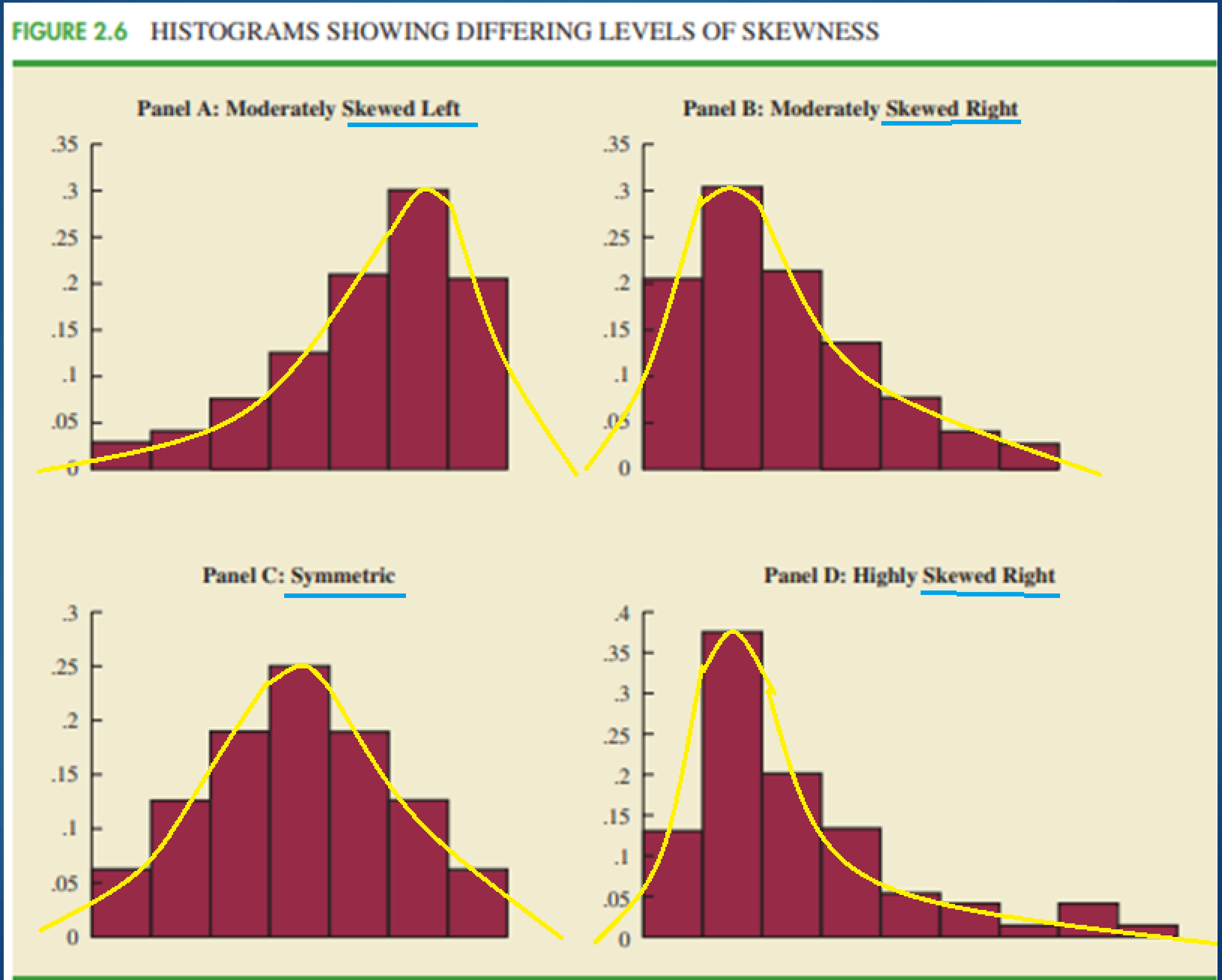


Frequency Polygon: Symmetric, Skewed to the left and Skewed to the right

FIGURE 2.6 HISTOGRAMS SHOWING DIFFERING LEVELS OF SKEWNESS



Frequency Ogives: Symmetric, Skewed to the left and Skewed to the right



Review:

Classify the following into: Descriptive / Inferential Statistics:

1) Plot a Histogram for student marks
2) Test the hypothesis: the student marks depends on the number of studying hours
3) Find the maximum and minimum mark for student

Choose the most correct answer:

The variable (number of student in a class) classifies to:

A. Quantitative variable
B. Qualitative variable
C. Discrete variable
D. Continuous variable
E. Ordinal variable
F. Nominal variable
G. Ratio level
H. Interval level
I. A, D, F, and H
J. B, C, E and H
K. A, C, E and G-- Answer
L. All (A to H)

Review:

Compute the systematic interval (k) if the population size is 600 and the sample size is 30?

K=N/n = 600/30=20

Determine : Dependent / independent variable :

1) Student marks - Y

   studying hours – X

2) Gender -X

   smoking - Y

Chapter 3:/ Section: 3.1 Measures of central tendency page (103-
(Mean, Median, Mode, Weighted mean and Midrange)

if the measures are computed for data from a sample, they are called **sample statistics**.

if the measures are computed for data from a population, they are called **population parameters**.

Statistic

SAMPLE MEAN

$$\bar{x} = \frac{\Sigma x_i}{n}$$

Parameter

POPULATION MEAN

$$\mu = \frac{\Sigma x_i}{N}$$

in statistical inference, a **sample statistic** is referred to as the point estimator of the corresponding population parameter.

i.e. Statistic is the point estimator of the corresponding population parameter

Note: The mean is a central tendency measure. Thus, the mean value must be between The lowest and highest values

The mean is sometimes referred to as the arithmetic mean

$$\bar{x} = \frac{\sum x_i}{n}$$

Compute the mean of: 46, 54, 42, 46, 32

$x_1 = 46 \qquad x_2 = 54 \qquad x_3 = 42 \qquad x_4 = 46 \qquad x_5 = 32$

Hence, to compute the sample mean, we can write

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{46 + 54 + 42 + 46 + 32}{5}$$

$$= 44$$

Compute the mean of: 46, 114, 42, 46, 32

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{46 + 114 + 42 + 46 + 32}{5}$$

$$= \frac{280}{5} = 56$$

The properties of arithmetic mean :

1. The sum of deviation between each value and the arithmetic mean equals zero.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

2. Sum of the squares of the deviations is minimum when deviations are taken from arithmetic mean

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 < \sum_{i=1}^{n}(x_i - A)^2$$, where A is any value such that $\mathbf{A} \neq \bar{x}$

Example: Suppose $\sum_{i=1}^{10}(x_i - 49) = 20$, then $\bar{X} = ???$

3. Mean is the measure of central tendency do affect the outliers. (disadvantage)

The median (Med)

is another measure of central tendency .
The median is the value in the middle when the data are arranged in ascending order (smallest value to largest value).

Also; **The median is: the midpoint of the data array. The symbol for the median is MD.**

With an odd number of observations, the median is the middle value.

An even number of observations has no single middle value. in this case, we follow convention and define the median as the average of the values for the middle two observations

Example
Compute the median of values: 46, 54, 42, 32, 46
Start by arranged values in ascending order:

value	32	42	46	46	54
order	1st	2nd	3rd	4th	5th

Median=46

Example
Compute the median of values: 13, 8, 44, 32, 34, 10
Start by arranged values in ascending order:

value	8	10	13	32	34	44
order	1st	2nd	3rd	4th	5th	6th

Median$= \frac{13+32}{2} = \frac{45}{2} = 22.5$

Mode (M)

Another measure of **central tendency** is the mode.

The mode is the value that occurs with <u>greatest frequency</u>

Example

Compute the mode of values: <u>46</u>, 54, 42, 32, <u>46</u>

Mode=46

Unimodal

Example

Compute the mode of values: <u>46</u>, <u>54</u>, 42, 32, <u>46</u>, <u>54</u>

Mode=46 and 54

Bimodal

Note: Median and **mode** are the two measure of central tendency **<u>do not affect</u>** the **outliers**. **(<u>advantage</u>)**

Midrange

$$MR = \frac{\text{lowest value} + \text{highest value}}{2}$$

Example 3–15

Find the midrange. 2, 3, 6, 8, 4, 1

Solution

$$MR = \frac{1 + 8}{2} = \frac{9}{2} = 4.5$$

Hence, the midrange is 4.5.

Weighted Mean

in the formulas for the sample mean and population mean, each x is given equal importance or weight. As follows:

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{1}{n}\left(\sum x_i\right) = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}(x_1) + \frac{1}{n}(x_2) + \cdots + \frac{1}{n}(x_n)$$

This shows that each observation in the sample is given a weight of $\frac{1}{n}$

The weighted mean is computed as follows

$$\bar{X}_w = \frac{w_1X_1 + w_2X_2 + \cdots + w_nX_n}{w_1 + w_2 + \cdots + w_n} = \frac{\Sigma wX}{\Sigma w}$$

where $w_1, w_2, \ldots, w_n$ are the weights and $X_1, X_2, \ldots, X_n$ are the values.

Example:

Compute the mean for a student marks:

Mark (x)	Weighted (w)	w*x
70	0.3	21
60	0.1	6
70	0.1	7
80	0.1	8
90	0.4	36
Total	1	78

$$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i}$$

$$\bar{x} = \frac{78}{1} = 78$$

Example:

Suppose that a manager wanted to know the mean cost per pound of the raw material

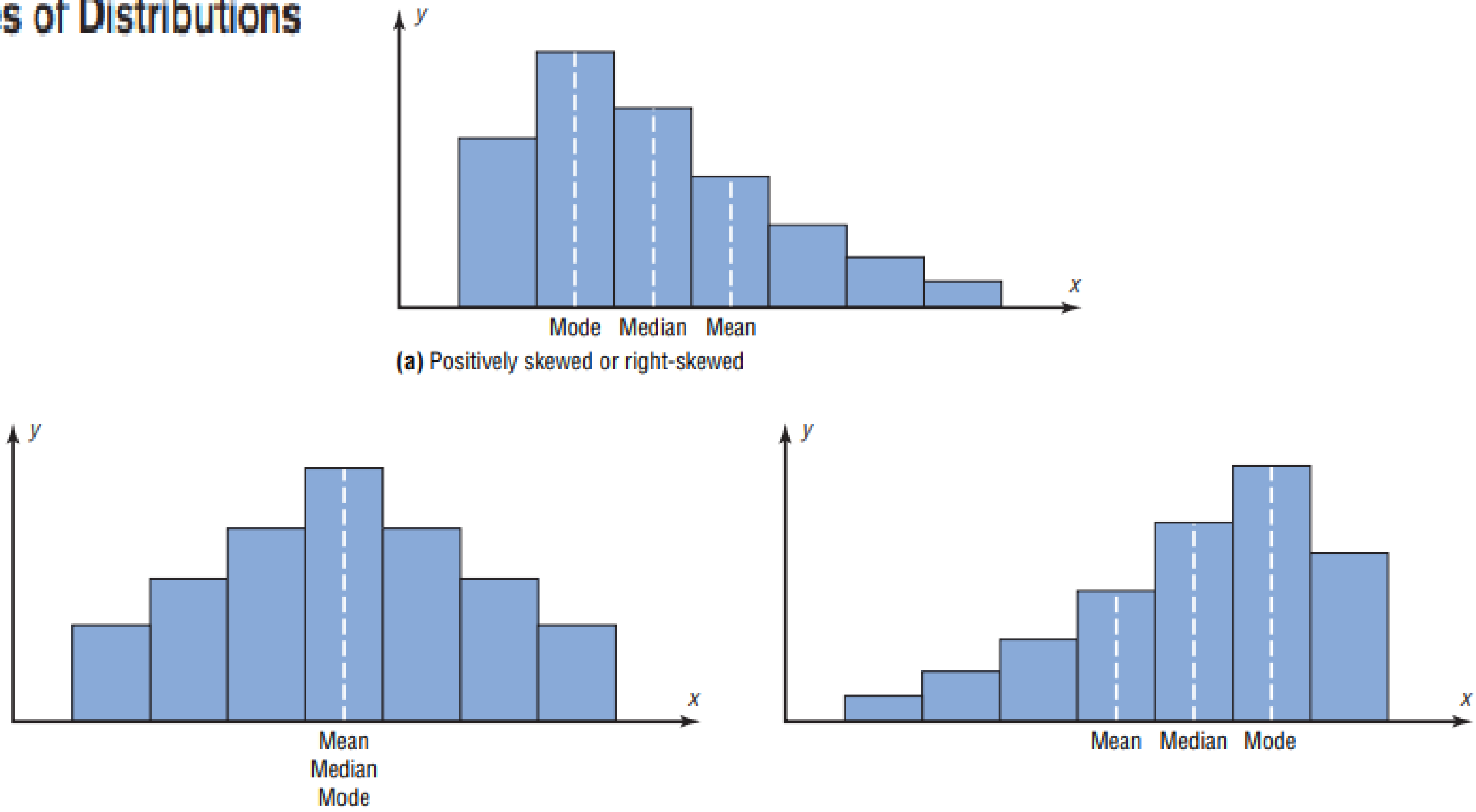
As an example of the need for a weighted mean, consider the following sample of five purchases of a raw material over the past three months.

Purchase	Cost per Pound ($)	Number of Pounds
1	3.00	1200
2	3.40	500
3	2.80	2750
4	2.90	1000
5	3.25	800

$$\bar{x} = \frac{1200(3.00) + 500(3.40) + 2750(2.80) + 1000(2.90) + 800(3.25)}{1200 + 500 + 2750 + 1000 + 800}$$

$$= \frac{18{,}500}{6250} = 2.96$$

Types of Distributions



Percentiles: (position measures) **Percentiles divide the data set into 100 equal groups.**

* the pth percentile is the value that approximately p% of the observations are less than the pth percentile

* and approximately (100 – p)% of the observations are greater than the pth percentile.

Note that the 50th percentile is also the median.

To find the pth percentile begin by arranging the sample values in ascending order then locate it using the corresponding value

LOCATION OF THE *P*TH PERCENTILE

$$L_p = \frac{p}{100}(n + 1)$$

Rule:

P% value= lower value+ (Lp-lower Location)*(higher-lower)

Example: Compute the 80th percentile of values:
3920, 3880, 3940, 3710, 3850, 3755, 3880, 4325, 4050, 3950, 4130, 3890

$$L_{80} = \frac{p}{100}(n + 1) = \left(\frac{80}{100}\right)(12 + 1) = 10.4$$

	3710	3755	3850	3880	3880	3890	3920	3940	3950	4050	4130	4325
Position	1	2	3	4	5	6	7	8	9	10	11	12

$$\begin{aligned}\text{80th percentile} &= 4050 + .4(4130 - 4050)\\ &= 4050 + .4(80) = 4082\end{aligned}$$

Quartiles (position measures)

Q1 = first quartile, or 25th percentile
Q2 = second quartile, or 50th percentile (also the median)
Q3 = third quartile, or 75th percentile

Example: Compute Q1, Q2, and Q3 of values: 3920, 3880, 3940, 3710, 3850, 3755, 3880, 4325, 4050, 3950, 4130, 3890

	25% of the data			25% of the data			25% of the data			25% of the data		
	3710	3755	3850	3880	3880	3890	3920	3940	3950	4050	4130	4325
Position	1	2	3	4	5	6	7	8	9	10	11	12

$Q_1 = 3857.5$ $\qquad$ $Q_2 = 3905$ (Median) $\qquad$ $Q_3 = 4025$

$$L_{25} = \frac{p}{100}(n+1) = \left(\frac{25}{100}\right)(12+1)$$
$$= 3.25$$
$$Q_1 = 3850 + .25(3880 - 3850)$$
$$= 3850 + .25(30) = 3857.5$$

$$L_{75} = \frac{p}{100}(n+1) = \left(\frac{75}{100}\right)(12+1)$$
$$= 9.75$$
$$Q_3 = 3950 + .75(4050 - 3950)$$
$$= 3950 + .75(100) = 4025$$

Case of frequency table

The mean

$$\bar{x} = \frac{\sum x_i \ f_i}{\sum f_i}$$

x_i: *calass ith midpoint*

f_i: *calass ith frequency*

Example:
Compute the mean of student marks

	fi	xi	xi fi
mark	Frequency	Class Midpoint	
1-8	4	4.5	18
9-16	6	12.5	75
17-24	2	20.5	41
25-32	7	28.5	199.5
33-40	1	36.5	36.5
Total	20		370

$$\bar{x} = \frac{\sum x_i \ f_i}{\sum f_i} = \frac{370}{20} = 18.5$$

Case of frequency table

Example:
Compute the mean of student marks

	fi	xi	xi fi
mark	Frequency	Class Midpoint	
0-10	3	5	15
10-20	8	15	120
20-30	6	25	150
30-40	5	35	175
40-50	3	45	135
	25		595

$$\bar{x} = \frac{\sum x_i \ f_i}{\sum f_i} = \frac{595}{25} = 23.8$$

The Mode (M)

The mode is the midpoint of a class have greatest frequency

Example:
Compute the mode of student marks

	fi	xi
mark	Frequency	Class Midpoint
1-8	4	4.5
9-16	6	12.5
17-24	2	20.5
25-32	7	28.5
33-40	1	36.5
Total	20	

Mode=28.5

Example:
Compute the mode of student marks

	fi	xi
mark	Frequency	Class Midpoint
1-8	4	4.5
9-16	7	12.5
17-24	2	20.5
25-32	7	28.5
33-40	1	36.5
Total	21	

Mode=12.5 and 28.5

The median (Med)

is the value in the middle when the data are arranged in ascending order (smallest value to largest value).

1) Rank of Median$= \frac{\sum f_i}{2}$

2) Construct cumulative frequency table (using all class boundaries)

3) Apply the <u>proportion</u> or the <u>rule</u>:

$$\textit{Median} = \begin{matrix}\textit{Lower boundary}\\ \textit{of Median class}\end{matrix} + \left(\begin{matrix}\textit{Upper boundary}\\ \textit{of Median class}\end{matrix} - \begin{matrix}\textit{Lower boundary}\\ \textit{of Median class}\end{matrix}\right) \frac{\left(\begin{matrix}\textit{Median}\\ \textit{Rank}\end{matrix} - \begin{matrix}\textit{Lower Rank}\\ \textit{of Median class}\end{matrix}\right)}{\left(\begin{matrix}\textit{Upper Rank}\\ \textit{of Median class}\end{matrix} - \begin{matrix}\textit{Lower Rank}\\ \textit{of Median class}\end{matrix}\right)}$$

Example1:
Compute the median of student marks

Rank of Median$= \frac{\sum f_i}{2} = \frac{20}{2} = 10$

Marks	Frequency	Class boundaries
1-8	4	0.5-8.5
9-16	6	8.5-16.5
17-24	2	16.5-24.5
25-32	7	24.5-32.5
33-40	1	32.5-40.5
	20	

Cumulative Class	Cumulative Frequency (Rank)
less than or equals to 0.5	0
less than or equals to 8.5	4
less than or equals to **16.5**	**10**
less than or equals to 24.5	12
less than or equals to 32.5	19
less than or equals to 40.5	20

Median=16.5

Example 2:
Compute the median of student marks

Rank of Median$= \frac{\sum f_i}{2} = \frac{18}{2} = 9$

Marks	Frequency	Class boundaries
1-8	2	0.5-8.5
9-16	3	8.5-16.5
17-24	8	16.5-24.5
25-32	1	24.5-32.5
33-40	4	32.5-40.5
	18	

Cumulative Class	Cumulative Frequency (Rank)
less than or equals to 0.5	0
less than or equals to 8.5	2
less than or equals to 16.5	5
less than or equals to 24.5	13
less than or equals to 32.5	14
less than or equals to 40.5	18

Cumulative Class	Cumulative Frequency (Rank)
less than or equals to 0.5	0
less than or equals to 8.5	2
less than or equals to 16.5	5
Med=?	9
less than or equals to 24.5	13
less than or equals to 32.5	14
less than or equals to 40.5	18

$$\frac{\text{Median} - 16.5}{24.5-16.5} = \frac{9-5}{13-5}$$

$$\frac{\text{Median} - 16.5}{8} = \frac{4}{8}$$

$$\text{Median} = 16.5 + 8 \times \frac{4}{8} = 16.5+4 = 20.5$$

Another solution way for Example2:
Compute the median of student marks

Rank of Median= $\frac{\sum f_i}{2} = \frac{18}{2} = 9$

Marks	Frequency	Class boundaries
1-8	2	0.5-8.5
9-16	3	8.5-16.5
17-24	8	16.5-24.5
25-32	1	24.5-32.5
33-40	4	32.5-40.5
	18	

Cumulative Class	Cumulative Frequency (Rank)
less than or equals to 0.5	0
less than or equals to 8.5	2
less than or equals to 16.5	5
less than or equals to 24.5	13
less than or equals to 32.5	14
less than or equals to 40.5	18

Cumulative Class	Cumulative Frequency (Rank)
less than or equals to 0.5	0
less than or equals to 8.5	2
less than or equals to 16.5	5
Med=?	9
less than or equals to 24.5	13
less than or equals to 32.5	14
less than or equals to 40.5	18

$$Median = \begin{matrix} Lower\ boundary \\ of\ Median\ class \end{matrix} + \left(\begin{matrix} Upper\ boundary \\ of\ Median\ class \end{matrix} - \begin{matrix} Lower\ boundary \\ of\ Median\ class \end{matrix} \right) \frac{\left(\begin{matrix} Median \\ Rank \end{matrix} - \begin{matrix} Lower\ Rank \\ of\ Median\ class \end{matrix} \right)}{\left(\begin{matrix} Upper\ Rank \\ of\ Median\ class \end{matrix} - \begin{matrix} Lower\ Rank \\ of\ Median\ class \end{matrix} \right)}$$

$$Median = 16.5 + (24.5 - 16.5)\ \frac{(9-5)}{(13-5)} = 16.5 + (8)\ \frac{4}{8} = 16.5 + 4 = 20.5$$

Percentiles

* the p^{th} percentile is the value that approximately p% of the observations are less than the p^{th} percentile

* **and** approximately (100 – p)% of the observations are greater than the p^{th} percentile.

Note: The 50th percentile is also the median.

To find the p^{th} percentile:

1) **Rank of percentile P** $= \frac{P}{100} \sum f_i$
2) **Construct cumulative frequency table (using all class boundaries)**
3) **Apply the proportion or the rule**

Percentile p^{th} $value = \begin{matrix} Lower\ boundary \\ of\ p^{th}\ class \end{matrix} + \left(\begin{matrix} Upper\ boundary \\ of\ p^{th}\ class \end{matrix} - \begin{matrix} Lower\ boundary \\ of\ p^{th}\ class \end{matrix} \right) \frac{\left(\begin{matrix} p^{th} \\ Rank \end{matrix} - \begin{matrix} Lower\ Rank \\ of\ p^{th}\ class \end{matrix} \right)}{\left(\begin{matrix} Upper\ Rank \\ of\ p^{th}\ class \end{matrix} - \begin{matrix} Lower\ Rank \\ of\ p^{th}\ class \end{matrix} \right)}$

Example 3:
Compute the percentile 20 of student marks (i.e. p=20)

Rank of $20^{th} = \frac{20}{100}\sum f_i = \frac{20}{100}(18) = 3.6$

Marks	Frequency	Class boundaries
1-8	2	0.5-8.5
9-16	3	8.5-16.5
17-24	8	16.5-24.5
25-32	1	24.5-32.5
33-40	4	32.5-40.5
	18	

Cumulative Class	Cumulative Frequency (Rank)
less than or equals to 0.5	0
less than or equals to 8.5	2
less than or equals to 16.5	5
less than or equals to 24.5	13
less than or equals to 32.5	14
less than or equals to 40.5	18

Cumulative Class	Cumulative Frequency (Rank)
less than or equals to 0.5	0
less than or equals to 8.5	2
20th Percentile = ?	3.6
less than or equals to 16.5	5
less than or equals to 24.5	13
less than or equals to 32.5	14
less than or equals to 40.5	18

$$\frac{\text{Percentile } 20^{th} - 8.5}{16.5-8.5} = \frac{3.6-2}{5-2}$$

$$\frac{\text{Percentile } 20^{th} - 8.5}{16.5-8.5} = \frac{3.6-2}{5-2}$$

$$\frac{\text{Percentile } 20^{th} - 8.5}{8} = \frac{1.6}{3}$$

$$\text{Percentile } 20^{th} - 8.5 = 8 \times \frac{1.6}{3} = 8.5 + 4.24 = 12.76$$

Another solution way for Example 3:
Compute the percentile 20 of student marks (i.e. p=20)

Rank of $20^{th} = \frac{20}{100}\sum f_i = \frac{20}{100}(18) = 3.6$

Marks	Frequency	Class boundaries
1-8	2	0.5-8.5
9-16	3	8.5-16.5
17-24	8	16.5-24.5
25-32	1	24.5-32.5
33-40	4	32.5-40.5
	18	

Cumulative Class	Cumulative Frequency (Rank)
less than or equals to 0.5	0
less than or equals to 8.5	2
less than or equals to 16.5	5
less than or equals to 24.5	13
less than or equals to 32.5	14
less than or equals to 40.5	18

Cumulative Class	Cumulative Frequency (Rank)
less than or equals to 0.5	0
less than or equals to 8.5	2
20th Percentile = ?	3.6
less than or equals to 16.5	5
less than or equals to 24.5	13
less than or equals to 32.5	14
less than or equals to 40.5	18

Percentile p^{th} $value = \begin{matrix} Lower\ boundary \\ of\ p^{th}\ class \end{matrix} + \left(\begin{matrix} Upper\ boundary \\ of\ p^{th}\ class \end{matrix} - \begin{matrix} Lower\ boundary \\ of\ p^{th}\ class \end{matrix} \right) \frac{\left(\begin{matrix} p^{th} \\ Rank \end{matrix} - \begin{matrix} Lower\ Rank \\ of\ p^{th}\ class \end{matrix} \right)}{\left(\begin{matrix} Upper\ Rank \\ of\ p^{th}\ class \end{matrix} - \begin{matrix} Lower\ Rank \\ of\ p^{th}\ class \end{matrix} \right)}$

Percentile 20^{th} $value = 8.5 + (16.5 - 8.5)\ \frac{(3.6-2)}{(5-2)}$=8.5+ (8) $\frac{1.6}{3}$ =8.5+4.24=12.76

Example 4:
Compute the percentile 90 of student marks (i.e. p=90)

Rank of $90^{th} = \frac{90}{100}\sum f_i = \frac{90}{100}(18) = 16.2$

Marks	Frequency	Class boundaries
1-8	2	0.5-8.5
9-16	3	8.5-16.5
17-24	8	16.5-24.5
25-32	1	24.5-32.5
33-40	4	32.5-40.5
	18	

Cumulative Class	Cumulative Frequency (Rank)
less than or equals to 0.5	0
less than or equals to 8.5	2
less than or equals to 16.5	5
less than or equals to 24.5	13
less than or equals to 32.5	14
less than or equals to 40.5	18

Cumulative Class	Cumulative Frequency (Rank)
less than or equals to 0.5	0
less than or equals to 8.5	2
less than or equals to 16.5	5
less than or equals to 24.5	13
less than or equals to 32.5	14
90th Percentile = ?	16.2
less than or equals to 40.5	18

Percentile p^{th} $value = \begin{matrix} Lower\ boundary \\ of\ p^{th}\ class \end{matrix} + \left(\begin{matrix} Upper\ boundary \\ of\ p^{th}\ class \end{matrix} - \begin{matrix} Lower\ boundary \\ of\ p^{th}\ class \end{matrix} \right) \frac{\left(\begin{matrix} p^{th} \\ Rank \end{matrix} - \begin{matrix} Lower\ Rank \\ of\ p^{th}\ class \end{matrix} \right)}{\left(\begin{matrix} Upper\ Rank \\ of\ p^{th}\ class \end{matrix} - \begin{matrix} Lower\ Rank \\ of\ p^{th}\ class \end{matrix} \right)}$

Percentile 90^{th} $value = 32.5 + (40.5 - 32.5)\ \frac{(16.2-14)}{(18-14)} = 32.5 + (8)\ \frac{2..2}{4} = 32.5 + 4.4 = 36.9$

Exercise:
Compute the percentile 25, 50, and 75 for the student marks

Marks	Frequency	Class boundaries
1-8	2	0.5-8.5
9-16	3	8.5-16.5
17-24	8	16.5-24.5
25-32	1	24.5-32.5
33-40	4	32.5-40.5
	18	

Percentile 25^{th} = 15.17 =P25 = Q1

Percentile 50^{th} = 20.5 =P50 = Q2 = Median

Percentile 75^{th} = 28.5 =P75 = Q3

Quartiles

Q1 = first quartile, or 25th percentile

Q2 = second quartile, or 50th percentile (also the median)

Q3 = third quartile, or 75th percentile

Percentile $25^{th} = 1^{st}\ quartile$

Percentile $50^{th} = 2^{nd}\ quartile =$ median

Percentile $75^{th} = 3^{rd}\ quartile$

Recall

Case of Frequency table

The mean

The mean of values is the average $\bar{x} = \frac{\sum x_i \ f_i}{n} = \frac{\sum x_i \ f_i}{\sum f_i}$

The mode

The mode is the *midpoint* of a class having greatest frequency

The median

The median is the value in the middle of data (50%) of data less than it or greater than it when the data are arranged in ascending order (smallest value to largest value).

1) **Rank of Median=** $\frac{\sum f_i}{2}$
2) **Construct cumulative frequency table (using all class boundaries)**
3) **Apply the proportion or the rule**

p^{th} Percentile

* is the value that approximately p% of the observations are less than the p^{th}percentile.

- **and** approximately (100 – p)% of the observations are greater than the p^{th} percentile.

1) **Rank of percentile P** $= \frac{P}{100} \sum f_i$
2) **Construct cumulative frequency table (using all class boundaries)**
3) **Apply the proportion or the rule**

Quartiles

Q1 = first quartile, or 25th percentile = Q1 = P25

Q2 = second quartile, or 50th percentile (also the median) = Q2 = P50 = Median

Q3 = third quartile, or 75th percentile = Q3 = P75

Chapter 3 / Section 3.2: measures of variations

Range, Interquartile range, Variance and Standard Deviation, coefficient of Variation

Range

Range= Largest value - Smallest value

Interquartile Range (IQR):

IQR = Q3 – Q1

Dispersion measure

Variance:

Sample variance (S^2)

Population variance (σ^2)

Standard Deviation:

Sample standard deviation (S)

Population standard deviation (σ)

The mean absolute error (MAE)

Coefficient of variation

$$\text{COEFFICIENT OF VARIATION} = \left(\frac{\text{Standard deviation}}{\text{Mean}} \times 100\right)\%$$

Case of values:

Example:

For the sample values: 46, 52, 42, 48, 32, Compute

1) Range =52-32=20

2) IQR=Q3-Q1=50-37=13

3) Variance $S^2 = \frac{\sum(X-\bar{X})^2}{n-1} = \frac{(46-44)^2+(52-44)^2+(42-44)^2+(48-44)^2+(32-44)^2}{5-1} = \frac{232}{4} = 58$

4) Standard Deviation $S = \sqrt{58} = 7.62$

5) Coefficient of variation $CV = \frac{S}{\bar{x}} \times 100\% = \frac{7.62}{44} \times 100\% = 17.3\%$

6) Mean Absolute Error $MAE = \frac{\sum|\mathbf{x} - \bar{x}|}{\mathbf{n}} = \frac{|46-44|+|52-44|+|42-44|+|48-44|+|32-44|}{5} = \frac{28}{5} = 5.6$

Case of values:

Example:

For the **Population** values: 46, 52, 42, 48, 32, Compute

1) Range =52-32=20

2) IQR=Q3-Q1=50-37=13

3) Variance $\sigma^2 = \frac{\sum(X-\mu)^2}{N} = \frac{(46-44)^2+(52-44)^2+(42-44)^2+(48-44)^2+(32-44)^2}{5} = \frac{232}{5} = 46.4$

4) Standard Deviation $\sigma = \sqrt{46.4} = 6.81$

5) Coefficient of variation $CV = \frac{\sigma}{\mu} \times 100\% = \frac{6.81}{44} \times 100\% = 15.5\%$

6) Mean Absolute Error $MAE = \frac{\sum|x-\mu|}{N} = \frac{|46-44|+|52-44|+|42-44|+|48-44|+|32-44|}{5} = \frac{28}{5} = 5.6$

	Sample Variance Formula
For Ungrouped Data	$S^2 = \frac{\sum(x - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$
For Grouped Data	$S^2 = \frac{\sum(x_i - \bar{x})^2 f_i}{(\sum f_i) - 1} = \frac{\sum x_i^2 f_i - (\sum f_i)\bar{x}^2}{(\sum f_i) - 1}$

	Population Variance Formula
For Ungrouped Data	$\sigma^2 = \frac{\sum(x - \mu)^2}{N} = \frac{\sum x_i^2 - N\mu^2}{N}$
For Grouped Data	$\sigma^2 = \frac{\sum(x_i - \mu)^2 f_i}{(\sum f_i)} = \frac{\sum x_i^2 f_i - (\sum f_i)\mu^2}{(\sum f_i)}$

	MAE: mean absolute error
For Ungrouped Data	$MAE = \frac{\sum\lvert \mathbf{x} - \bar{x}\rvert}{\mathbf{n}}$ or $MAE = \frac{\sum\lvert \mathbf{x} - \mu\rvert}{N}$
For Grouped Data	$MAE = \frac{\sum\lvert x_i - \bar{x}\rvert f_i}{\sum f_i}$

Case of frequency table

Range (R):
R = Upper boundary of last class value – Lower boundary of first class value

Interquartile Range (IQR):
IQR = Q3 – Q1

Variance

Standard deviation

MAE

Finding the Sample Variance and Standard Deviation for Grouped Data

Step 1 Make a table as shown, and find the midpoint of each class.

A	B	C	D	E
Class	Frequency	Midpoint	$f \cdot X_m$	$f \cdot X_m^2$

Step 2 Multiply the frequency by the midpoint for each class, and place the products in column D.

Step 3 Multiply the frequency by the square of the midpoint, and place the products in column E.

Step 4 Find the sums of columns B, D, and E. (The sum of column B is n. The sum of column D is $\Sigma f \cdot X_m$. The sum of column E is $\Sigma f \cdot X_m^2$.)

Step 5 Substitute in the formula and solve to get the variance.

$$s^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\left(\sum f_i\right) - 1} \quad \textbf{or} \quad s^2 = \frac{\sum x_i^2 f_i - (\sum f_i)\bar{x}^2}{\left(\sum f_i\right) - 1}$$

Step 6 Take the square root to get the standard deviation.

Example: Assuming the Sample of student marks, Compute

$$\bar{x} = \frac{\sum x_i \; f_i}{\sum f_i} = 18.5$$

1) Range = 40.5 - 0.5 = 40

2) IQR = Q3 - Q1 = 27.93 - 9.83 = 18.1

3) $Variance\ S^2 = \frac{2032}{19}$=106.95

$$S^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{(\sum f_i) - 1} = \text{OR} = \frac{\sum x_i^2 f_i - (\sum f_i)\bar{x}^2}{(\sum f_i) - 1}$$

4) Standard Deviation $S = \sqrt{106.95} = 10.34$

5) MAE=$\frac{184}{20}$=9.2

6) Coefficient of variation $CV = \frac{S}{\bar{x}} \times 100\% = \frac{10.34}{18.5} \times 100\% = 55.9\%$

	f_i	x_i	$x_i \; f_i$	$x_i^2 \; f_i$	$(x_i - \bar{x})^2 \; f_i$	$\lvert x_i - \bar{x}\rvert \; f_i$
Student Marks	# of Students	Class Midpoint				
1-8	4	4.5	18	81	784	56
9-16	6	12.5	75	937.5	216	36
17-24	2	20.5	41	840.5	8	4
25-32	7	28.5	199.5	5685.75	700	70
33-40	1	36.5	36.5	1332.25	324	18
Total	20		370	8877	2032	184