UECM1633 Probability and Statistics for Computing

Contents

Reference Books

1. Baron, M., 2013. *Probability and Statistics for Computer Scientists*. 2nd ed. Boca Raton: Chapman and Hall/CRC.

2. Hogg, R.V. & Tanis, E.A., 2008. *Probability and Statistical Inference*. 8th ed. Harlow: Pearson Education Limited.

3. Mann, P.S., 2010. *Introductory Statistics*. 7th ed. Hoboken, New Jersey: John Wiley & Sons.

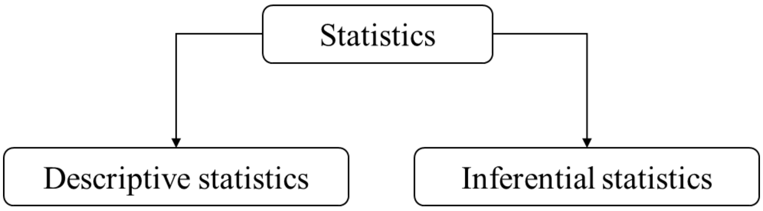
Method of Assessment

No.	Method of Assessment	Total
1.	Coursework a) Test 1 (20%) : Week 7 b) Test 2 (20%) : Week 11	40%
2.	Final Examination	60%

Chapter 1 Introduction to Statistics

1.1. Definition of Statistics

Statistics is a group of methods used to collect, analyze, present and interpret data and to make decisions.



Descriptive Statistics (or Deductive Statistics)

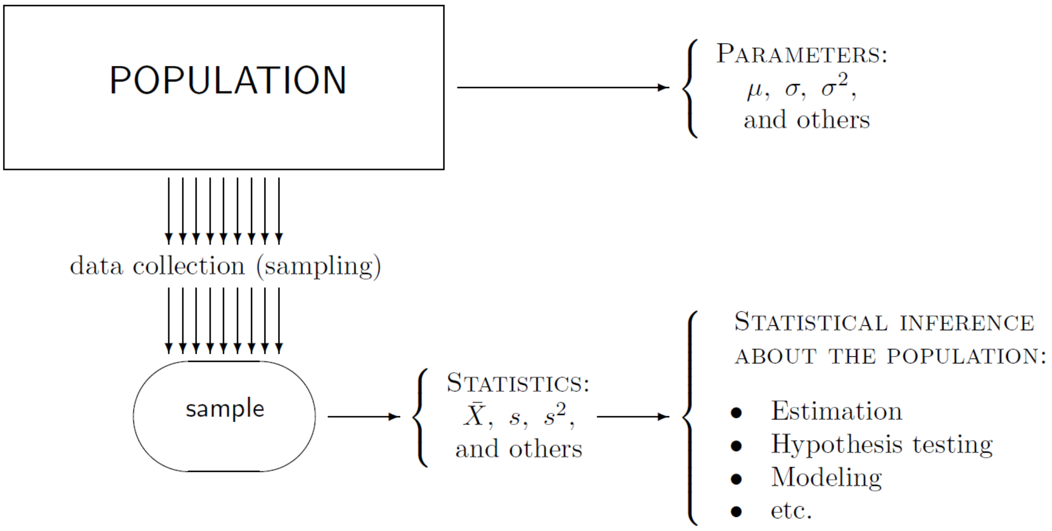
- Consists of methods for organizing, displaying, and describing data by using tables, graphs, and summary measures.
- Deals with the description and analysis of a given group of data.
- Present information in a convenient, usable and comprehensible form.

Inferential Statistics (or Inductive Statistics)

- Consists of methods that use sample results to make decisions or predictions about a population.
- Deals with the problems of making inferences or drawing conclusions about population based on information obtained from the samples taken from the population.

1.2. Population and Sample, Parameters and Statistics

Population parameters and sample statistics.



Population

- A *population* consists of all units of interest.
- A survey that includes every member of the population is called *census*.
- Any numerical characteristics of a population is a *parameter*.

Sample

- Consists of observed units collected from the population.
- *Sample survey* is the technique of collecting information from a portion of the population.
- Any function of a sample is called *statistic*.

<u>Some notations used for population parameters and sample statistics.</u>

Population Parameters	Sample Statistics
μ → population mean	$\bar{x}$ → sample mean
σ^2 → population variance	s^2 → sample variance
σ → population standard deviation	s → sample standard deviation
p → population proportion	$\hat{p}$ → sample proportion

Purpose of Statistics

Drawing conclusions about the *population* by studying the *sample*.

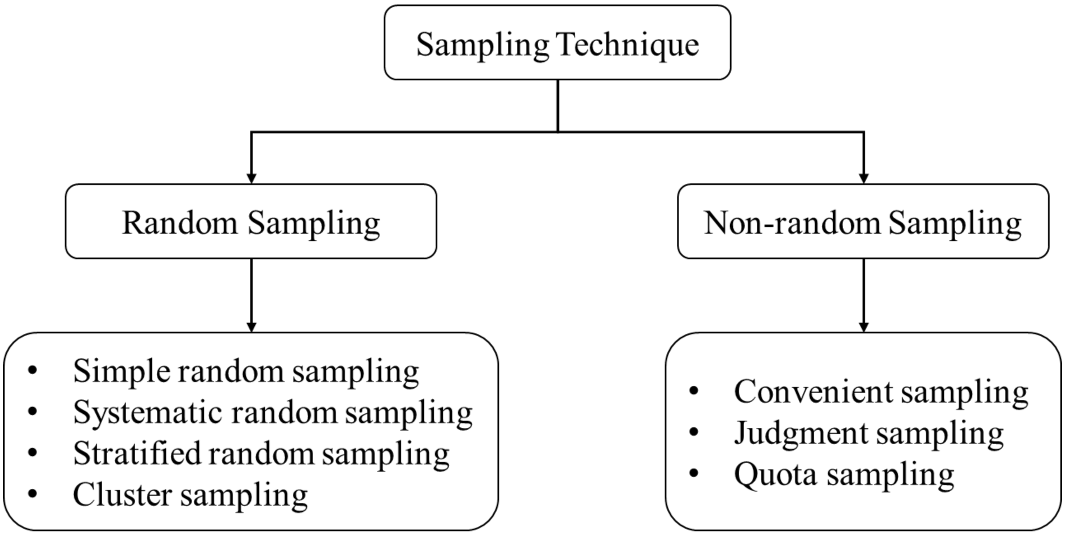
Collect information from a comparatively **SMALL** sample → Draw conclusions about a **LARGE** population

Example 1.1.

Suppose a television executive wants to know the percentage of television viewers who watch the program “ABC”. Noted that 100 million people may be watching television on a given evening, how are you going to find out the percentage of TV viewers who watch “ABC”?

To do so, we would interview a sample of 1,000 television viewers and calculate the sample percentage as an ***estimate*** of the percentage of all households that watch the program “ABC”.

1.3. Sampling Technique

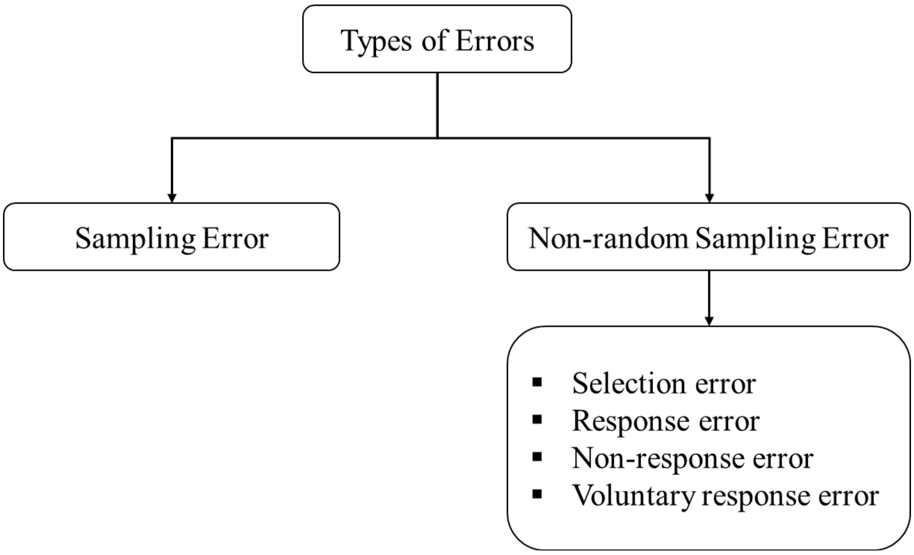


1.4. Sampling and Non-sampling Errors

Sampling and non-sampling errors refer to any discrepancy between a collected sample and a whole population.

Sampling Errors
Caused by the mere fact that only a sample, a portion of a population, is observed. For most of reasonable statistical procedures, sampling errors decrease (and converge to zero) as the sample size increases.

Non-sampling Errors
Caused by inappropriate sampling schemes or wrong statistical techniques. Often no wise statistical techniques can rescue a poorly collected sample of data.



Example 1.2.
Even if 80% of users are satisfied with their internet connection, it does not mean that exactly 8 out of 10 customers in the observed sample are satisfied. As we can see from the Binomial distribution table, with probability 0.0328, only a half of ten sampled customers are satisfied. In other words, there is a 3% chance for a random sample to suggest that contrary to the population parameter, no more than 50% of users are satisfied.

This example shows that a sample may give a rather misleading information about the population. *Sampling errors* are inevitable.

Example 1.3.
To evaluate the work of a Windows help desk, a survey of social science students of some university is conducted. This sample poorly represents the whole population of all Windows users. For example, computer science students and especially computer professionals may have a totally different opinion about the Windows help desk.

In this example, *selection error* occurs because the sampling frame is not representative of the population.

1.5. Basic Terms

Element or Member
An *element* or *member* of a sample or population is a specific subject or object about which the information is collected.

Variable
A *variable* is a characteristic under study that assumes different values for different elements.

Observation or Measurement
An *observation* or a *measurement* is the value of a variable for an element.

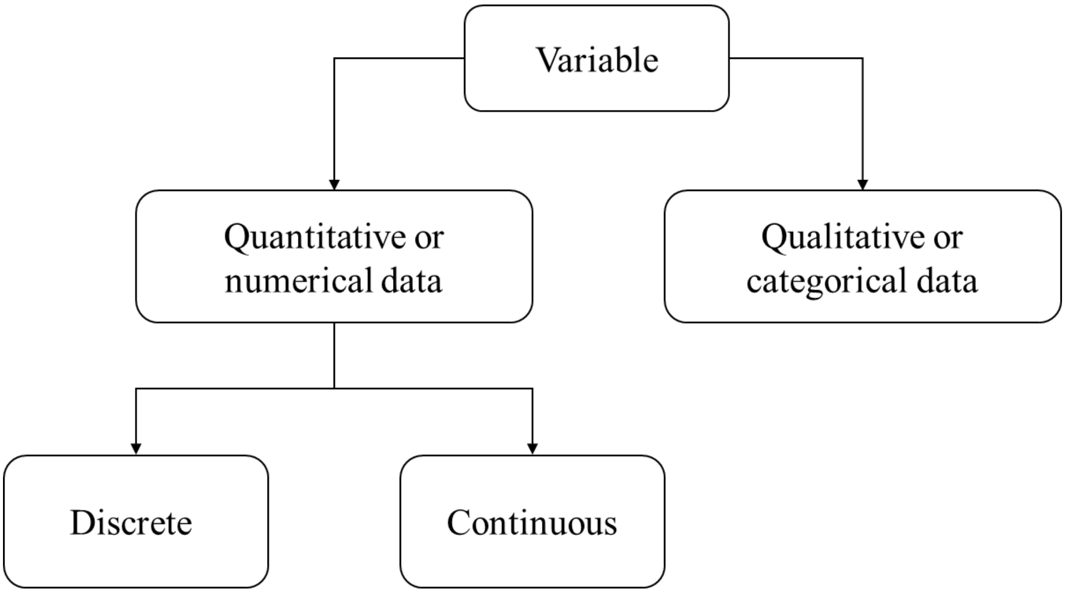
Data Set
Data set is a collection of observations on one or more variables.

Example 1.4.
The following table gives the number of dog bites reported to the police last year in five townships.

Townships	**Number of Bites**
Bandar Mahkota	32
Bandar Sungai Long	52
Bandar Tun Hussein Onn	43
Sungai Chua	44
Taman Saujana Impian	12

Briefly explain the meaning of a member, a variable, a measurement and a data set with reference to this table.

1.6. Types of Variables


Variable
Quantitative or numerical data
Qualitative or categorical data
Discrete
Continuous

1.7. Descriptive Statistics

Raw Data

Raw data is data recorded in the sequence in which they are collected and before they are processed or ranked.

Example:

Table 1: The weights of 20 students in *kg* (Quantitative raw data).

61	68	65	67	68	71	69	63	74	64
66	65	62	67	60	73	69	70	70	71

Table 2: The grades of UECM1633 of 20 students (Qualitative raw data).

A	B	C	A	C	B	B	A	B	C
B	A	B	B	B	A	C	D	D	B

Array

An arrangement of numerical raw data in *ascending order* or *descending order* of magnitude.

60	61	62	63	64	65	65	66	67	67
68	68	69	69	70	70	71	71	73	74

Ungrouped Data

Data set contains information on each member of a sample or population individually.
Example: Data presented in Table 1 and Table 2.

Grouped Data

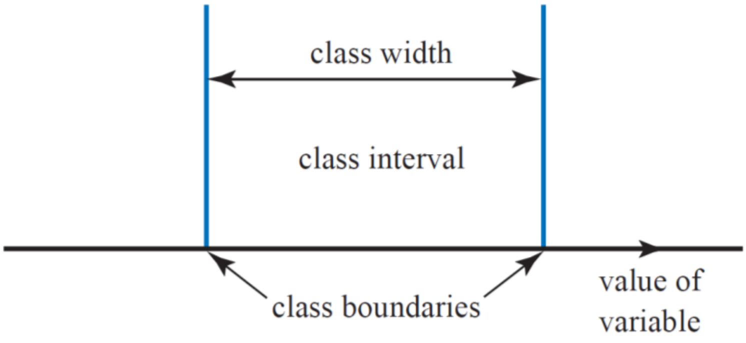
Data presented in classes or intervals.
Example:

Statistics Test Scores	10 – 12	13 – 15	16 – 18	19 – 21
Number of students	4	12	20	14

1.8. Grouped Data

Grouping means putting the data into a number of classes. The number of data items falling into any class is called the *frequency* for that class.

When numerical data are grouped, each item of data falls within a *class interval* lying between *class boundaries*.



Class
An interval that includes all the values that falls within two numbers, the lower and upper limits.

Class Limits
Endpoints of each interval.

Class Boundary
The *dividing line* between two classes. It is given by the midpoint of the upper limit of one class and the lower limit of the next higher class.

Class Width / Class Size
The difference between the upper and lower class boundary.

$$\text{class width} = \text{upper boundary} - \text{lower boundary}$$

Class Mark / Class Midpoint
The midpoint of the class interval.

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

Relative Frequency and Percentage Distributions
Tabular arrangement that lists the relative frequencies and percentages for all categories.

$$\text{Relative frequency of a category} = \frac{\text{frequency of that category}}{\text{sum of all frequencies}}$$

$$rf = \frac{f}{\sum f}$$

$$\text{Percentange} = \text{Relative frequency} \times 100\%$$

Example 1.5.

The following table gives the frequency distribution of ages for all 50 employees of a company.

Age	Number of Employees
18 to 30	8
31 to 43	22
44 to 56	14
57 to 69	6

(a) Find the class boundaries and class midpoints.
(b) Do all classes have the same width? If yes, what is that width?
(c) Prepare the relative frequency and percentage distribution columns.

Solution:

Age	Class boundaries	Class midpoint	*f*	*rf*	Percentage

Constructing frequency distribution tables.

1. Determine the number of classes, usually varies from 5 to 20, depending mainly on the number of observations in the data set.
   Find 2^k where *k* is the smallest number such that 2^k is greater than the number of observations *n*.

2. Determine the class size or width (*i*)
   Must cover at least the distance from the smallest value (*L*) in the raw data up to the largest value (*H*).

$$\text{approximate class width } = \frac{\text{largest value } (H) \text{ - smallest value } (L)}{\text{number of classes}}$$

- The class width is usually rounded up to some convenient number.
- The rounding of this number may slightly change the number of classes initially intended.

3. Determine the lower limit of the first class or the 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.

4. Some common practices for classes:

* Class(inclusive type)	**Class (exclusive type)			Class (open-ended)
0 - 9	0 - < 10	or	0 - 10	Below 20
10 - 19	10 - < 20		10 - 20	20 - < 30
20 - 29	20 - < 30		20 - 30	30 - < 40
30 - 39	30 - < 40		30 - 40	40 - < 50
40 - 49	40 - < 50		40 - 50	50 and above

Note.

*class (inclusive type) is mainly used for discrete data where there is a gap between classes.

**class (exclusive type) is mainly used for continuous data or discrete data which have been rounded to the nearest tens, hundreds, thousands, millions etc.

Example 1.6.

Sample of birth-weights (*oz*) from 50 consecutive deliveries is given below. Construct a frequency distribution table. Then, calculate the relative frequencies and percentages distributions for the data.

86	111	118	121	92	124	108	104	132	125
120	91	89	122	115	138	118	99	95	115
123	128	134	115	84	138	140	105	124	144
104	133	132	106	98	125	146	108	132	98
121	104	98	115	107	127	122	135	126	89

Solution:

Birth–weights (*oz*)	Frequency (*f*)	Relative frequency (*rf*)	Percentage

1.9. Single-Valued Classes

For some discrete data, it may not be necessary or desirable to group them.

Single-valued classes is used if the observations in a data set assume only a few distinct values (classes that are made of single values and not of intervals).

It is useful in cases of discrete data with only a few possible values.

Example 1.7.
A sample of 40 randomly selected households from a city produced the following data on the number of vehicles owned:

5	1	1	2	0	1	1	2	1	1
1	3	3	0	2	5	1	2	3	4
2	1	2	2	1	2	2	1	1	1
4	2	1	1	2	1	1	4	1	3

Construct a frequency distribution table for these data.
Solution:

Vehicles owned	Number of households (f)

1.10. Measures of Central Tendency

Simple *descriptive statistics* measuring the location, spread, variability, and other characteristics can be computed immediately.

Each statistic is a random variable because it is computed from random data. It has a so-called *sampling distribution*.

Each statistic estimates the corresponding population parameter and adds certain information about the distribution of X , the variable of interest.

A measure of central tendency gives the center of a histogram or a frequency distribution curve.

Three measures will be considered here:

1. Mean
2. Median
3. Mode

Mean

The *mean*, also called the *arithmetic mean*, is the most frequently used measure of central tendency. In general,

$$\text{mean} = \frac{\text{sum of all values}}{\text{number of all values}}.$$

For UNGROUPED data

The mean for population data $x_1, x_2, ..., x_N$ is denoted by μ and is defined as

$$\mu = \frac{x_1 + x_2 + ... + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i \ .$$

The mean for sample data $x_1, x_2, ..., x_n$ is denoted by $\overline{X}$ and is defined as

$$\overline{X} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

The mean for sample data in the form of frequency distribution of single-valued classes is denoted by $\overline{X}$ and is defined as

$$\overline{X} = \frac{x_1 f_1 + x_2 f_2 + ... + x_n f_n}{f_1 + f_2 + ... + f_n} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} .$$

Example 1.8.

The CPU time for $n=30$ randomly chosen jobs (in seconds) are recorded to evaluate effectiveness of a processor for a certain type of tasks.

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

Estimate the average (expected) CPU time μ. Comment on your result.

Solution:

Example 1.9.

Find the mean of the following frequency distribution.

x_i	f_i
2	1
5	3
6	4
8	2

Solution:

For GROUPED data

Let f_i be the frequency of class i, for $i=1,2,\ldots,k$

m_i be the midpoint of class i, for $i=1,2,\ldots,k$

$\sum_{i=1}^{k} f_i$ be the sum of all frequencies in k classes

The mean for population data is defined as $\mu = \dfrac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i}$.

The mean for sample data is defined as $\bar{X} = \dfrac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i}$.

Example 1.10.

A nurse keeps a record of the weights, measured to the nearest kilogram (*kg*), of a group of patients she treats. Her data are summarized in the following grouped frequency table.

Weight (*kg*)	Number of patients
40 – 44	4
45 – 49	6
50 – 54	10
55 – 59	16
60 – 64	20
65 – 69	14

Choose suitable mid-class values and calculate an estimate for the mean weight.

Solution:

Weight	midpoint, m	Frequency, f	$m \times f$
40 – 44		4	
45 – 49		6	
50 – 54		10	
55 – 59		16	
60 – 64		20	
65 – 69		14	
		70	$\sum mf =$

Effect of Outliers or Extreme Values

Outliers are the values that are very small or very large relative to the majority of the values in a data set.

One disadvantage of a sample mean is its sensitivity to extreme observations.

Example 1.11.

Reconsider *Example 1.8.*. If the first job in the sample is unusually heavy, and it takes 30 minutes to get processed instead of 70 seconds. Recalculate the average CPU time. Compare and comment on your result, for with or without this one extremely large observation.

Solution:

Median

Median is the value of the middle term in a data set that has been ranked in increasing order, hence measuring the central value.

<u>Note.</u>

Given *n* is the total number of observation in a data set:

1. If *n* is odd, then median is the value of the middle term in the ranked data.
2. If *n* is even, then median is the average value of the two middle terms.

<u>*Example 1.12.*</u>

Find the median of set A = { 10, 5, 19, 8, 3 } and set B = { 2, 7, 3, 6, 4, 5 }.

Solution:

<u>*Example 1.13.*</u>

Find the median of the following frequency distribution.

No. of children	0	1	2	3	4	5
Frequency	3	5	7	9	14	2

Solution:

<u>Note.</u>

Median is **not** influenced by the extreme value. It is less sensitive than the mean.

<u>*Example 1.14.*</u>

Reconsider *Example 1.8.*. Compute the median of CPU times.

Solution:

9	15	19	22	24	25	30	34	35	35
36	36	37	38	42	43	46	48	54	55
56	56	59	62	69	70	82	82	89	139

Mode

Mode is the value that occurs with the highest frequency in a data set.

Note.

1. Mode is not influenced by the extreme value.
2. Mode may not exist, exist one mode(unimode), two modes(bimodal) or more than two modes(multimodal).
3. Mode can be used for both quantitative and qualitative data.

Example 1.15.

Find the mode of each of the following data set.

i) 74, 9, 5, 8, 3, 8, 8

iii) 2, 6, 6, 6, 3, 8, 8, 8, 3

ii) 2, 2, 6, 6, 8, 8, 9, 9

iv) B, C, D, A, A, C, C, C, B, A

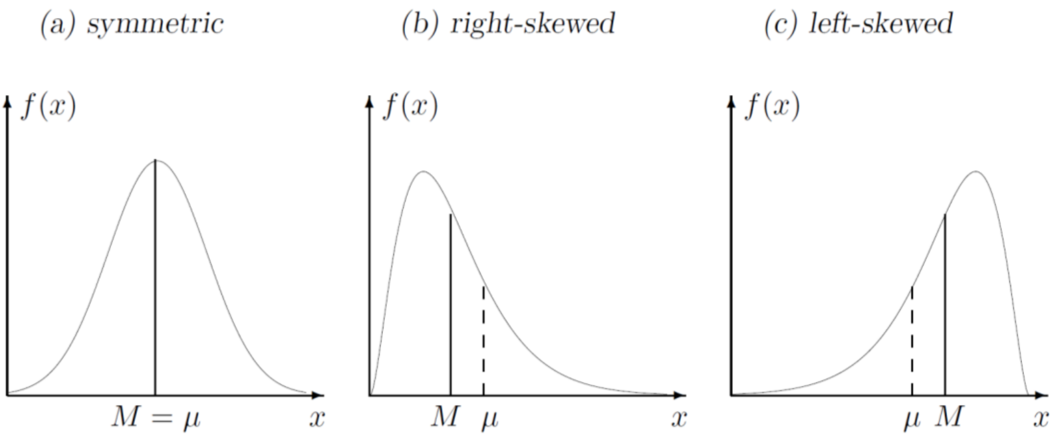
Example 1.16.

Find the mode of the following frequency distribution.

No. of children	0	1	2	3	4	5
Frequency	3	5	12	9	4	2

Understanding the shape of a distribution

Comparing the mean and the median, one can tell the distribution of X .



1.11. Measures of Dispersion

Statistics introduced in the previous sections showed where the central values of a population are located are not enough to reveal the whole picture of the distribution of a data set. This is because the measure of central tendency does not describe how the data is distributed.

Data set	Data	Mean	Median	Mode
A	1, 3, 6, 10, 10, 21, 26	11	10	10
B	7, 8, 10, 10, 10, 15, 17	11	10	10

Note. The mean, median and mode are the same for data set *A* and *B* but the distribution of the data are different.

Variance

Both population and sample variances are measured in squared units. Therefore, it is convenient to have standard deviations that are comparable with our variable of interest, X .

The formula for s^2 follow the same idea as that for σ^2. It is also the average squared deviation from the mean, this time computed for a sample. Like σ^2, sample variance measures how far the actual values of X are from their average.

For UNGROUPED data

The variance for population data $x_1, x_2, ..., x_N$ is denoted by σ^2 and is defined as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 .$$

Or, the shortcut formula $$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N}\left(x_i^2\right)\right] - \mu^2 .$$

The variance for sample data $x_1, x_2, ..., x_n$ is denoted by s^2 and is defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 .$$

Or, the shortcut formula $$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right].$$

Warning. $\sum_{i=1}^{n}\left(x_i^2\right) \neq \left(\sum_{i=1}^{n}x_i\right)^2$!

For GROUPED data

Let f_i be the frequency of class i, for $i = 1, 2, \ldots, k$

m_i be the midpoint of class i, for $i = 1, 2, \ldots, k$

$\sum_{i=1}^{k} f_i$ be the sum of all frequencies in k classes

The variance for population data is defined as

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i (m_i - \mu)^2}{\sum_{i=1}^{k} f_i} .$$

Or, the shortcut formula
$$\sigma^2 = \frac{\sum_{i=1}^{k} \left(m_i^2 f_i\right)}{\sum_{i=1}^{k} f_i} - \left[\frac{\sum_{i=1}^{k} \left(m_i f_i\right)}{\sum_{i=1}^{k} f_i} \right]^2 .$$

The variance for sample data is defined as

$$s^2 = \frac{\sum_{i=1}^{k} f_i (m_i - \overline{x})^2}{\left(\sum_{i=1}^{k} f_i\right) - 1} .$$

Or, the shortcut formula
$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} \left(m_i^2 f_i\right) - \frac{1}{n} \left\{ \sum_{i=1}^{k} \left(m_i f_i\right) \right\}^2 \right],$$

where $n = \sum_{i=1}^{k} f_i$.

Standard Deviation

The standard deviation is obtained by taking the positive square root of the variance.

The population standard deviation is defined as

$$\sigma = \sqrt{\sigma^2} .$$

The sample standard deviation is defined as

$$s = \sqrt{s^2} .$$

Note.

1. A small standard deviation means that the data are distributed closely to their mean.
2. A large standard deviation means that the data are widely scattered about their mean.
3. Standard deviation is influenced by extreme values.

Example 1.17.
Data shows the salary per day for *all* 6 employees of a small company.

29.50   16.50   35.40   21.30   49.70   24.60

Calculate the variance and standard deviation for these data.

Solution:

Method 1

$$\mu =$$

x_i	$x_i - \mu$	$(x_i - \mu)^2$	${x_i}^2$
29.50			
16.50			
35.40			
21.30			
49.70			
24.60			

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 =$$

Method 2

$$\sum\left({x_i}^2\right) =$$

$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N}\left({x_i}^2\right)\right] - \mu^{\,2} =$$

<u>*Example 1.18.*</u>

Find the variance from the following frequency distribution if it represent

a) population b) sample.

Height	Frequency, f
20 – 22	3
23 – 25	6
26 – 28	12
29 – 31	9
32 – 34	2

Solution:

Height	Midpoint, m	Frequency, f	mf	m^2f
20 – 22		3		
23 – 25		6		
26 – 28		12		
29 – 31		9		
32 – 34		2		

$$\sigma^2 = \frac{\sum_{i=1}^{k}\left(m_i^2 f_i\right)}{\sum_{i=1}^{k} f_i} - \left[\frac{\sum_{i=1}^{k}\left(m_i f_i\right)}{\sum_{i=1}^{k} f_i}\right]^2 =$$

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k}\left(m_i^2 f_i\right) - \frac{1}{n}\left\{\sum_{i=1}^{k}\left(m_i f_i\right)\right\}^2\right] =$$

Interquartile Range

An interquartile range is defined as the difference between the first and the third quartiles, namely,

$$IQR = Q_3 - Q_1$$

where Q_1 is the first quartile and Q_3 is the third quartile. It measures variability of data. Not much affected by outliers, it is often used to detect them.

Detection of outliers

A "rule of thumb" for identifying outliers is the rule of $\mathbf{1.5}\left(\boldsymbol{IQR}\right)$.

Remark: The rule of $1.5\left(IQR\right)$ comes from the assumption that the data are nearly normally distributed. If this is a valid assumption, then 99.3% of the population should appear within 1.5 interquartile ranges from quartiles.

Example 1.19.

Reconsider *Example 1.8.*. Is there any outlier(s) in this sample?

Solution:

9	15	19	22	24	25	30	34	35	35
36	36	37	38	*42*	*43*	46	48	54	55
56	56	59	62	69	70	82	82	89	139

Handling of outliers

Q: What should we do if the $1.5\left(IQR\right)$ rule suggests possible outliers in the sample?

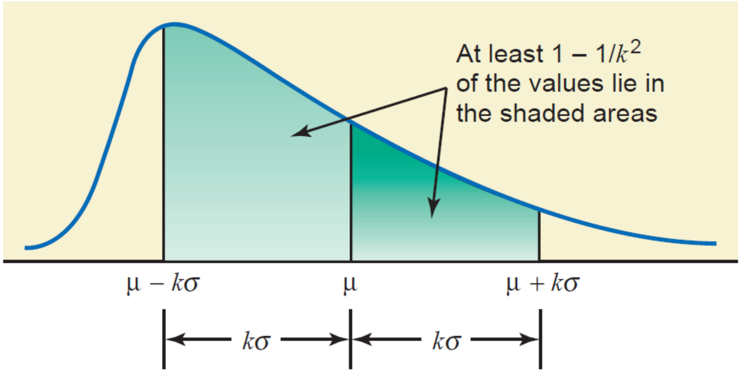
A: If it is confirmed that a suspected observation entered the data set by a mere mistake, it can be deleted.

1.12. Use of Standard Deviation

By using the mean and standard deviation, we can find the proportion or percentage of the total observations that fall within a given interval about the mean.

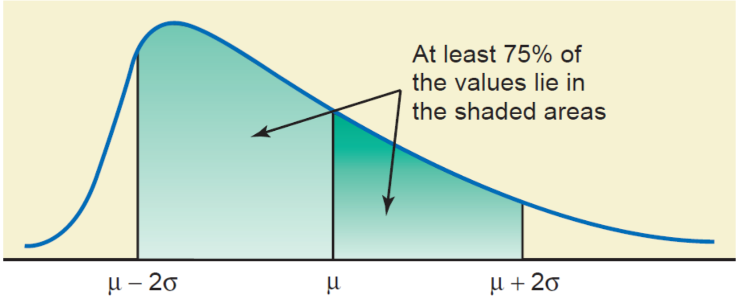
Chebyshev's Theorem

For any number k greater than 1, at least $\left(1-\frac{1}{k^2}\right)$ of the data values lie within k standard deviations of the mean.



Example 1.20.

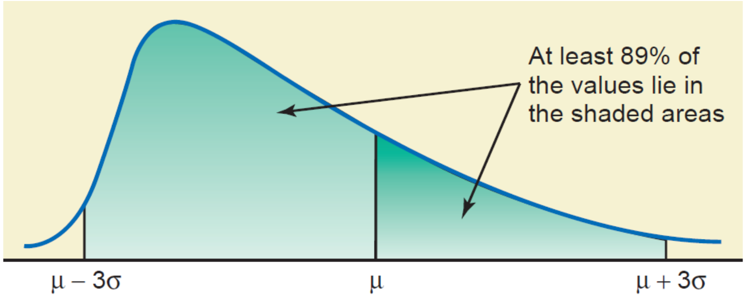
If $k=2$, then $\qquad 1-\frac{1}{k^2}=$



Therefore, according to Chebyshev's theorem, at least of the values of a data set lie within two standard deviations of the mean.

Example 1.21.

If $k=3$, then $\qquad 1-\frac{1}{k^2}=$



According to Chebyshev's theorem, at least of the values fall within three standard deviations of the mean.

Note.

1. The Chebyshev's theorem applies to both sample and population data.
2. The Chebyshev's theorem is applicable to a distribution of ANY shape.
3. However, Chebyshev's theorem can be used only for $k > 1$. This is so because when
   - a. $k = 1$, the value of $1 - \frac{1}{k^2}$ is zero,
   - b. $k < 1$, the value of $1 - \frac{1}{k^2}$ is negative.

Example 1.22.

A sample of 2000 observations has a mean of 74 and a standard deviation of 12. Using Chebyshev's theorem , find at least what percentage of the observations fall in the intervals

(a) $\bar{x} \pm 2s$

(b) $\bar{x} \pm 2.5s$

(c) $\bar{x} \pm 3s$

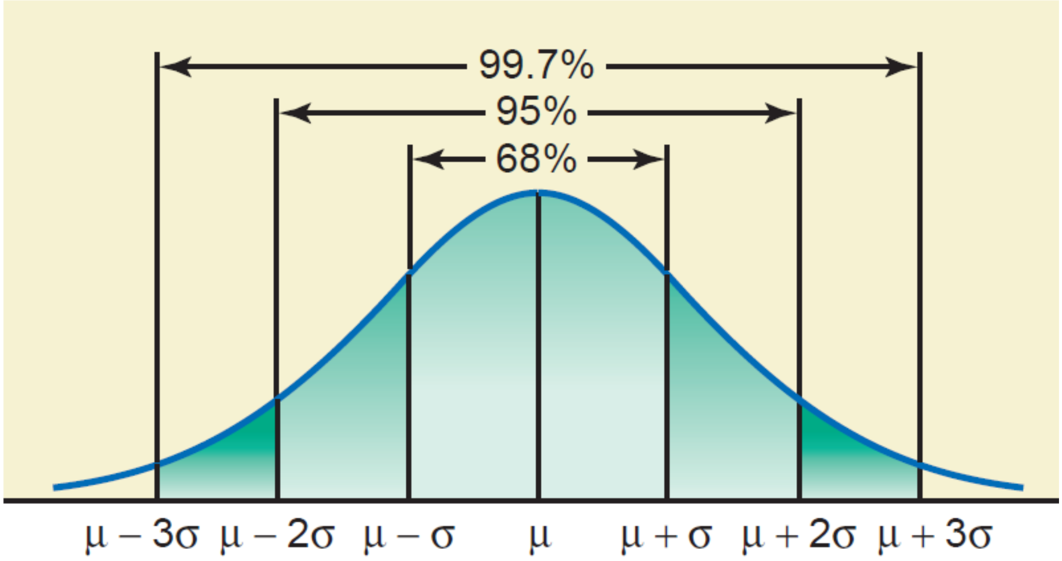
Solution:

Empirical Rule

Whereas Chebyshev's theorem is applicable to any kind of distribution, the *empirical rule* applies only to a specific type of distribution called a *bell-shaped* distribution.

For a bell-shaped distribution, approximately

1. 68% of the observations lie within one standard deviations of the mean.
2. 95% of the observations lie within two standard deviations of the mean.
3. 99.7% of the observations lie within three standard deviations of the mean.

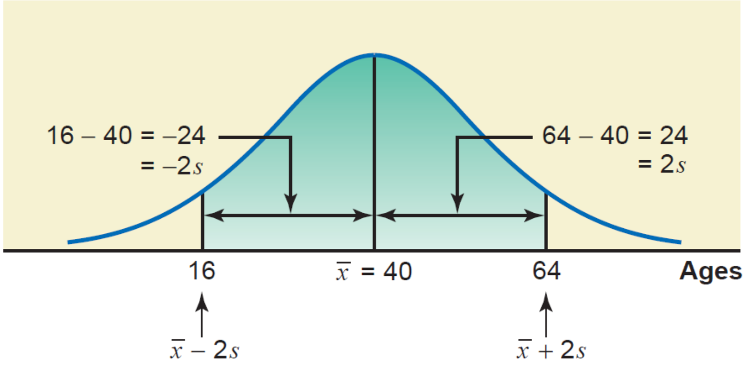


<u>*Example 1.23.*</u>

The age distribution of a sample of 5000 persons is bell-shaped with a mean of 40 years and a standard deviation of 12 years. Determine the approximate percentage of people who are 16 to 64 years old.

Solution:

$$\bar{x} = 40, \; s = 12$$



1.13. Graphical Statistics

Histogram

A *histogram* shows the shape of a *pmf* or a *pdf* of data, checks for homogeneity, and suggests possible outliers. To construct a histogram, we split the range of data into equal intervals, "bins", and count how many observations fall into each bin.

Three types of histogram:

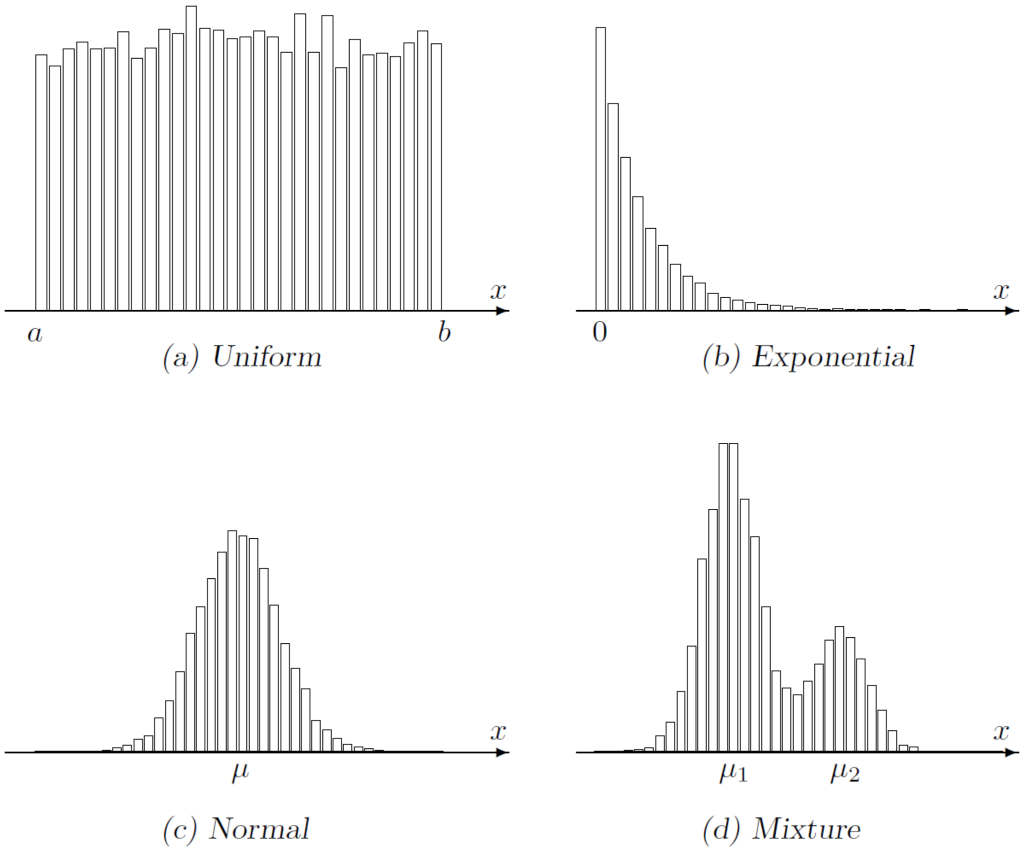
1. Frequency histogram
2. Relative frequency histogram
3. Percentage histogram

Shapes of Histograms

A histogram can assume any one of a large number of shapes. The most common shapes are

1. Symmetrical
2. Skewed
3. Uniform or rectangular

Example: Histograms of various sample.



(a) Uniform

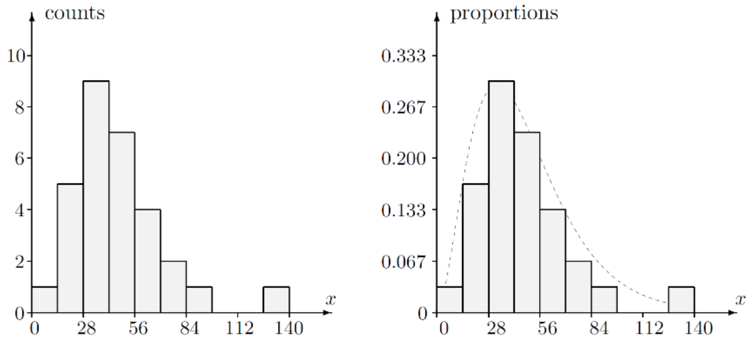
(b) Exponential

(c) Normal

(d) Mixture

Example 1.24.

Reconsider *Example 1.8.*. The sample of CPU times stretches from 9 to 139 seconds. Choosing intervals $0-14$, $14-28$, $28-42$, $\ldots$ as bins, the frequency histogram and the relative frequency histogram of CPU times are then constructed.



(a) Frequency histogram

(b) Relative frequency histogram

What information can we draw from these histograms?

Solution:

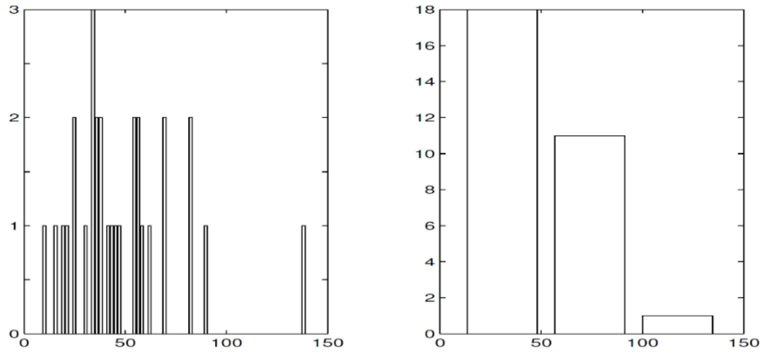
Procedures to draw a histogram.

1. Mark the class boundary of each interval on the horizontal axis.
2. For each class, mark the frequencies (or relative frequencies or percentages) on the vertical axis.
3. Draw a bar for each class so that its height represents the frequency of that class. (**No gap** between each bars)
4. Label the histogram.

The choice of bins.

Experimenting with histograms, you can notice that their shape depends on the choice of bins. One can hear various rules of thumb about a good choice of bins, but in general,

- There should not be too few or too many bins;
- Their number may increase with a sample size;
- They should be chosen to make the histogram informative, so that we can see shapes, outliers, etc.



(a) Too many columns

(b) Too few columns

Example 1.25.

The frequency distribution gives the weight of 35 objects, measured to the nearest *kg*. Draw a histogram to illustrate the data.

Weight (*kg*)	6 – 8	9 – 11	12 – 14	15 – 17	18 –20
Frequency	4	6	10	3	12

Stem-and-leaf Displays

In a *stem-and-leaf display*, each value is divided into two portions – a stem and a leaf. The leaves for each stem are shown separately in a display. It provides a basis for evaluating the "shape" of a data set with minimal loss of the original information.

To make a stem-and-leaf display:

1. Split each score/number into two parts, namely, **stem** and **leaf**.
2. Write the stems in a vertical column with the smallest at the top.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.
4. Provide a key/scale as a reference.

Example 1.26.

The following are the scores of 30 college students in a statistics test.

75	52	80	96	65	79	71	87	93	95
69	72	81	61	76	86	79	68	50	92
83	84	77	64	71	87	72	92	57	98

Construct a sorted stem-and-leaf display for these data.

Solution:

Example 1.27.

Show the following numbers on a sorted stem-and-leaf diagram with six branches.

0.212	0.242	0.223	0.248	0.226	0.253	0.230	0.253
0.233	0.250	0.259	0.237	0.262	0.241	0.266	0.256

Solution:

Example 1.28.

The following stem-and-leaf diagram showing the ages of a sample of 56 cyclists (sorted).

1	6 6 7 7 7 8 8 9
2	0 0 1 1 2 2 3 3 3 5 5 6 6 6 7 8 8 8 8 9 9
3	0 0 0 1 1 1 2 2 2 3 5 6 6 8 8 9
4	1 2 4 4 5 6 6 7 7 9
5	
6	6

Key: 1|6 = 16 years

Stretch the stem-and-leaf diagram by using steps of five years between the levels rather than ten.

Solution: