Chapter 1

Basic Concepts of Statistics

Tony CHAN, Ph.D.

1.1 Statistical Data

There are 2 sources of statistical data:

- **Primary data**
  This source of data has been purposely collected for particular enquiries, e.g. sample surveys, observations, interviews.

- **Secondary data**
  Secondary data are of two kinds: internal and external. This source of data has been collected for some other enquiries than the one of immediate interest (i.e. primary data). Secondary data are generated by a company or an organization such as a university or a chamber of commerce. People can refer to these published secondary data (historical data) for their own purposes.

1.1 Statistical Data (Cont'd)

- Official Statistics
  Published by Census & Statistics Department, Hong Kong

  The official statistics include:
  Hong Kong Monthly Digest of Statistics,
  Hong Kong Annual Digest of Statistics,
  Demographic trends in Hong Kong,
  2011 Population Census, etc.

- Statistics Published by Universities and Semi-official Organizations

  Examples
  Social Indicators
  (The former Public Opinion Program, The University of Hong Kong)

  Visitor Arrivals in Hong Kong in July 2012
  Hong Kong Tourism Board

3

1.1 Statistical Data (Cont'd)

- **Merits of Primary Data**
  - The investigator can know where the data came from, the circumstances under which they were collected.
  - The investigator can know any limitations or inadequacies in the data.
  - They relate directly to that person's research or study.

- **Demerits of Primary Data**
  - Primary data are usually more difficult, costly and time consuming to collect.
  - It often takes a long time to process the primary data.
  - A research based on only secondary data is least reliable and may have biases in, say, sampling method, method of data collection, etc.

4

1.1 Statistical Data (Cont'd)

- **Merits of Secondary Data**
  - Secondary data can usually be obtained more quickly and at lower cost than primary data.
  - Sometimes primary data do not exist in such situation one has to confine the research on secondary data.

    e.g. Social and economic researchers consider secondary data essential because it is impossible to conduct a survey to capture past changes and/or developments.
  - Secondary data can help plan the collection of primary data.

5

1.1 Statistical Data (Cont'd)

- **Demerits of Secondary Data**
  - The results from secondary data may not exactly fit one's research questions.
  - The researcher cannot check the data, and thus, their reliability may be questionable.
  - Extra caution is required when using secondary data.

6

1.1 Statistical Data (Cont'd)

When interpreting secondary data, the following must be noted:

- Is the target population the same as yours?
- Is the time period covered the same?
- Is the sampling method appropriate to your need?
- Is the method of data collection the same?

etc.

1.2 Population and Sample

A population includes all members of a defined group that we are studying or collecting information for data-driven decisions.

- The process of obtaining information from the whole population is called a ***census***.
- It is used to get a full picture of demographic characteristics (e.g. age, gender, marital status, etc.) of the people of a nation or region and their geographical distribution.

1.2 Population and Sample (Cont'd)

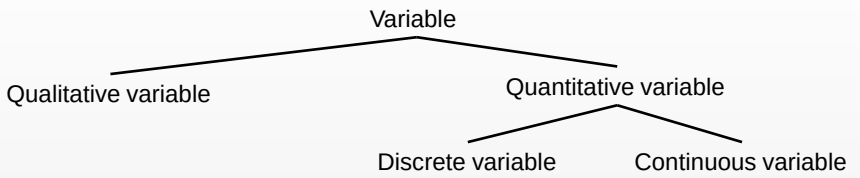
Conditions for a Census

- A census is feasible when the population is small.
- It is necessary when the elements are quite different from each other.
- Conducting a census every 10 years is an international practice to update the data of a region or country.

1.2 Population and Sample (Cont'd)

- A ***sample*** is a group of items drawn from a population.
- A ***part*** of the population is called a "sample". It is a proportion of the population, ***and all its characteristics***.
- A sample is a scientifically drawn group that actually ***possesses the same characteristics*** as the population – if it is drawn randomly.
- A ***sampling frame*** is a list of ***all*** members (elements) of the population. It can be used for selecting the sample.

1.3 Types of Variables



Variable

A characteristic that ***varies from one person or thing to another*** is called a variable.

e.g. height, weight, gender, marital status, age, race, educational attainment

Data

Information obtained by ***observing values*** of a variable.

e.g.: measure the weights (variable under study) of a class of students

11

1.4 Types of Data

- **Qualitative Data / Categorical Data**
  Data obtained from a variable that yields non-numerical data
  e.g.: gender, race, educational attainment

- **Quantitative Data**
  Data obtained from a variable that yields numerical data
  e.g.: height, weight, age

- **Discrete Data**
  Discrete data whose possible values can be listed, even the list may continue indefinitely.

  e.g.: number of iphones, number of brothers, number of cars, etc.

12

1.4 Types of Data (Cont'd)

- **Continuous Data**

Continuous data whose possible values form some interval of numbers.

e.g. Peter's weight is 50kg. Actually, his weight may lie in the range:

$[49.5\text{kg}, 50.5\text{kg}) = 49.5\text{kg} \leq \text{weight} < 50.5\text{kg}$

1.5 Levels of Measurement Scales

- **Nominal Scale**

  If the observed data are merely classified into various distinct categories in which no ordering is implied, a nominal level of measurement is achieved.

  Example:
  gender (male, female), answers to a question (true or false), country of origin (China, U.S., Canada, etc.), etc.

- **Ordinal Scale**

  If the observed data are classified into distinct categories in which ordering is implied, an ordinal level of measurement is attained.

  Example:
  educational attainment (primary, secondary, post-secondary, university), size (small, medium, large), etc.

1.5 Levels of Measurement Scales (Cont'd)

Interval Scale

- An interval scale is an ordered scale in which the difference between the measurements is a meaningful quantity.
- There is no true zero point in the interval scale. The "0" demarcation is arbitrary, not real.

Example:

A temperature reading of 67℃ is 2℃ warmer than 65℃. This 2℃ difference is the same as the difference between 35 ℃ and 33 ℃, in terms of warmth.

Hence, the difference has the same meaning anywhere on the scale.

15

1.5 Levels of Measurement Scales (Cont'd)

Examples of interval data

Temperature (℃, F)

Year:

The distances between each year are the same, but there is no absolute zero. We can only guess about the origins of the earth.

Most personality measures

IQ score:

Every point more on an intelligence test is the same. There is no absolute zero in an intelligence test.

16

1.5 Levels of Measurement Scales (Cont'd)

Ratio Scale

- A ratio scale is an ***ordered scale*** in which the ***difference*** between the measurements is a meaningful quantity.
- There is a true zero point.

Examples:

- A person who is 1.8m tall is twice as tall as someone who is 0.9m tall.
- 20kg mass of an object is twice as heavy as 10kg mass of another object.

17

1.6 Parameters and Statistics

Parameter

*A **parameter** is a **numerical measure** that describes a characteristic of a **population**. It is often **unknown but fixed**. It is usually represented by a **Greek letter.***

Examples of parameters:

population mean (μ) income of all Hong Kong residents;
population proportion (*P*) of all OUHK students who like the canteen;
population total (τ) number of residents in Hong Kong;
population standard deviation (σ) of incomes of all Hong Kong residents;
population variance (σ^2) of incomes of all Hong Kong residents.

If we measure the weights of ***all*** students in a class of 40, and obtain the average weight as 46.7kg, then
population mean weight (μ) = 46.7kg is a parameter.

18

1.6 Parameters and Statistics

Statistics

*A **statistic** is a **numerical measure** that describes a characteristic of a **sample**. It is usually represented by a **English letter**.*

Examples of sample statistics are:

sample mean ($\bar{x}$);

sample proportion (p);

sample standard deviation (s);

sample variance (s^2).

If we randomly select a sample of 10 students from a class of 40, and calculate the average weight as 45.3kg, then

sample mean ($\bar{x}$) = 45.3kg is a statistic.

Note: "statistics" is the plural form of "statistic".

1.7 Overview of a Survey

What is a Survey?

The word "survey" is most often used to describe a method of gathering information from a sample in order to learn something about the population from which the sample is to be drawn.

Objectives of a Survey:

A survey is widely used as a means of data collection to gather information required for policy formulation on public issues, business decision-making purposes and social studies.

1.7 Overview of a Survey (Cont'd)

Uses of Survey Results

Government: planning and policy formulation

Commerce: deciding their market strategy and making other important business decisions

Academics: testing their theories and discover new theories

21

1.7 Overview of a Survey (Cont'd)

Reasons for Sampling

- Since the size of a sample is much smaller than that of a population, the cost of collecting data and analyzing them is much lower.
- A sample can provide information about the population faster than by taking a complete census.
- The nature of the sampling may result in the sampling unit's destruction.
- The population may be infinite or countably infinite in size.
- The administrative problems involved in the collection of data, in processing, and in the supervision of fieldwork are more manageable.

22

1.7 Overview of a Survey (Cont'd)

- Each observation taken from a population contains a certain amount of information about the population parameter(s).
- Since information costs money, a researcher must determine how much information he/she should buy. Too little information prevents the experimenter from making good estimates, whereas too much information results in a waste of money.
- The quality of information obtained in the sample depends on
  - the number of observations sampled, and
  - the amount of variation in the data.

1.8 Jargons in Survey Sampling

An Example

In Hong Kong, an opinion poll will be conducted to determine public sentiment toward a certain candidate in upcoming legislative councilors' election.

Objective of Survey

To estimate the proportion (percentage) of registered voters in Hong Kong who favor the candidate.

Definition 1.1

An element is an object on which a measurement is taken.

An element here is a registered voter in Hong Kong.

The measurement taken on an element is the voter's preference on the candidate.

1.8 Jargons in Survey Sampling (Cont'd)

Definition 1.2

A population is a collection of elements about which we wish to make an inference.

The population here includes all registered voters in Hong Kong.

Definition 1.3

Sampling units are non-overlapping collections of elements from the population that cover the entire population.

A sampling unit may be a registered voter in Hong Kong. A more efficient process may be to sample households.

Remark

If each sampling unit contains one and only one element of the population, then a sampling unit and an element from the population are identical.

25

1.8 Jargons in Survey Sampling (Cont'd)

Definition 1.4

A sampling frame is a list of sampling units. It should be comprehensive, complete, and up-to-date to keep bias to a minimum.

Examples of Sampling Frames:

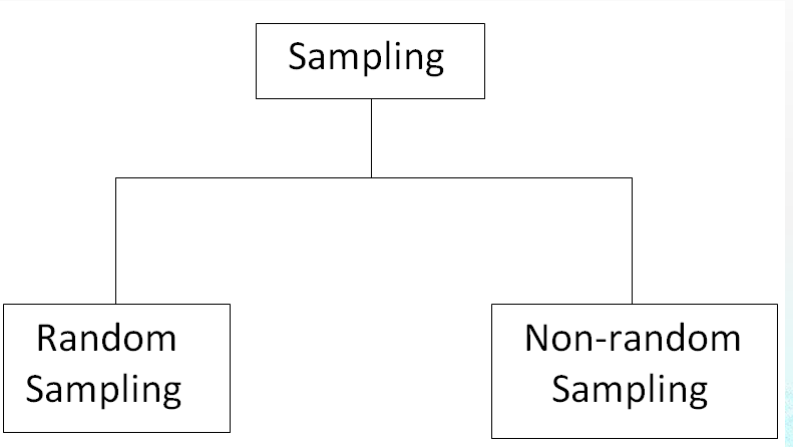
- A list of housing units in a city
- A list of retail establishments in a country
- A list of students in a university
- A list of all registered voters

Remarks

- If we take the household as a sampling unit, then a telephone directory, a city directory, or a list of household heads obtained from censuses data can serve as a sampling frame.
- All frames have some inadequacies:
  - The lists will not be up-to-date.
  - They will contain many names of unregistered household heads.

26

1.9 Random & Non-random Sampling



27

1.9.1 Random Sampling

With proper randomness in the sampling design, properties of the estimates can be assessed probabilistically. A sample obtained by a sample design that is based on planned randomness are called a ***random sample***.

Reasons for Using Random Sampling:

1. Subjective selection biases can be avoided.

2. The level of precision of the sample estimates (i.e., sample statistics) can be assessed.

3. Statistical inference can be drawn from the sample results.

Remark

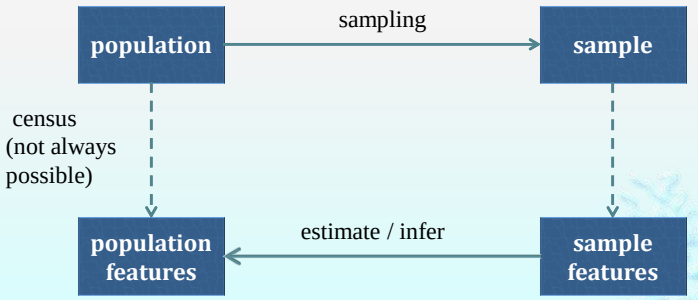
Sampling error is measurable in random sampling only.

28

1.9.2 Random Sampling & Inference

Purpose of Random Sampling:

To infer the population features (population parameters) based on a random sample.



29

1.9.3 Non-random Sampling

Properties

- Every element has an ***unknown*** probability of being selected.
- Sampling error cannot be estimated, and therefore, valid statistical inference cannot be made. Consequently, inferential results based on a sample cannot be generalized to the whole population.
- The sample is restricted to a part of the population that is readily accessible.
  e.g. A sample of coal from an open wagon may be taken from the top 6 to 9 inches.
- The sample is selected haphazardly.

  e.g. In picking 10 rabbits from a large cage in a laboratory, the researcher may take those that close to his hands, without concise planning.

30

1.9.3 Non-random Sampling (Cont'd)

- With a small but heterogeneous population, the sampler inspects the whole of it and selects a small sample of "typical" units that are close to his impression of the average of the population.
- The sample consists essentially of volunteers, in studies in which the measuring process is unpleasant or troublesome to the person being measured

31

1.9.4 Random Sampling vs. Non-random Sampling

- *Sample should be chosen at random and be large enough so as to be representative.* However, this will introduce high cost.
- Non-random sampling methods are alternatives to reduce costs. In addition, survey results can be achieved quickly by non-random sampling methods.
- Results obtained on the basis of a non-random sampling method cannot be generalized to the whole population.
- Statistical inference cannot be made. Instead, descriptive statistics can be employed to present the results of a non-random sample.

32

1.10 Tabulation of Data

The process of placing grouped data into a table format is known as tabulation.

Example 1.1

The table below shows the weights of 40 students:

43	62	45	50	57	53	54	46
52	53	31	46	44	43	54	48
53	52	42	57	38	35	55	52
50	41	64	56	40	45	30	56
47	58	38	45	47	58	47	51

33

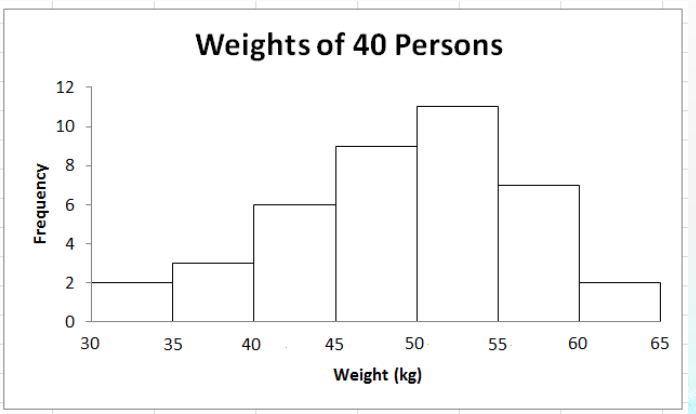
1.10 Tabulation of Data (Cont'd)

Without further tabulation or calculation, it is difficult to interpret the data quickly and easily. We may arrange them into a table, as shown below:

Weight (kg)	Frequency
30 - < 35	2
35 - < 40	3
40 - < 45	6
45 - < 50	9
50 - < 55	11
55 - < 60	7
60 - < 65	2
Total	40

34

1.10 Tabulation of Data (Cont'd)



35

1.11 Measures of Central Tendency

The purpose of a measure of central tendency is to determine the center of the data values or possibly the "most typical" data value.

36

1.11.1 Types of Measures of Central Tendency

There are several methods to measure the central tendency using different averages. The main types of averages that we commonly used are as follows:

- Arithmetic mean/mean;
- Weighted mean;
- Median;
- Mode.

1.11.2 Summation Notation

In order to save time and space, we will use the following symbol for summing n terms:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

Σ = summation symbol (the sum of)

x_i = i-th datum

i = the running index (i. e. counter)

"$i = 1$" = the starting value of the sum

n = the last value to be added in the sum

Example 1.2

Let $x_1 = 3, x_2 = 7, x_3 = 10$. Then

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3 = 3 + 7 + 10 = 20$$

39

1.11.3 Operations on Σ Notation

- $\sum_{i=1}^{n} 1 = 1 + 1 + \cdots + 1 = n$
- $\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$

$\sum_{j=1}^{n} x_j = x_1 + x_2 + \cdots + x_n$

$\therefore \sum_{i=1}^{n} x_i = \sum_{j=1}^{n} x_j$

40

1.11.3 Operations on Σ Notation (Cont'd)

$$\sum_{i=1}^{n} 1 = 1 + 1 + \cdots + 1 = n$$

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

$$\sum_{j=1}^{n} x_j = x_1 + x_2 + \cdots + x_n$$

$$\sum_{i=1}^{n} x_i = \sum_{j=1}^{n} x_j$$

$$\sum_{i=1}^{n} kx_i = kx_1 + kx_2 + \cdots + kx_n = k(x_1 + x_2 + \cdots + x_n) = k\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + \cdots + (x_n + y_n)$$
$$= (x_1 + x_2 + \cdots + x_n) + (y_1 + y_2 + \cdots + y_n)$$
$$= \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

41

Exercise 1

(1) Let $x_1 = 1$, $x_2 = 7$, $x_3 = 4$, $x_4 = 5$, $x_5 = 10$.

(a) Find the value of n.

(b) Compute $\sum_{i=1}^{5} x_i$.

(c) Find $\overline{x}$.

42

Solution

(a) $n = 5$.

(b) $\sum_{i=1}^{5} x_i = 1+7+4+5+10 = 27$

(c) $\bar{x} = \frac{1}{n}\sum_{i=1}^{5} x_i = \frac{1}{5} \times 27 = 5.4$

43

1.11.4 Arithmetic Mean

The arithmetic mean is the sum of values divided by the number of values. Its formula is given by

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

The above formula can be rewritten as

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

This means that each x_i carries an equal weight $1/n$. Thus, the arithmetic-mean formula can only be used for ***equal-weighting situation.***

44

Remarks

Business Managers often use a mean to represent a set of data values, such as

- mean sales
- mean price
- mean salary
- mean production per hour

45

In economics, the term "per capita" is a measure of central tendency. The following

- income per capita of a certain district
- number of mobile phones per capita

are examples of a mean.

46

Example 1.3

A sample of 10 was taken to determine the typical completion time (in months) for the construction of automobiles:

4.1, 3.2, 2.8, 2.6, 3.7, 3.1, 9.4, 2.5, 3.5, 3.8

Find the mean completion time.

Solution

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(4.1 + \cdots + 3.8) = 3.87 \text{months}$$

47

Use Scientific Calculator to Compute Mean

We compute mean by CASIO FX50FII:

i) Press **mode mode 4** to enter SD Mode

ii) Data input:

4.1, M+

.

.

.

3.8, M+

iii) "Press **shift 2 1 =**" : $\bar{x} = 3.87$

Note: Clear memory: "**shift 9 1 =**"

48

1.11.5 Weighted Mean

The entry test of a clerical position of a bank consists of three parts: typing, computer skills, and commercial English. The ***relative importance*** (i.e., ***weighting/weight***) of each part is different.

Item	Mark	Percentage (weight)
Typing	x_1	w_1
Computer Skills	x_2	w_2
Commercial English	x_3	w_3

Since the weights of the components are different, we can't use the formula for arithmetic mean. Instead, we may resort to the weighted mean below:

$$\bar{x} = \frac{\sum_{i=1}^{3} w_i x_i}{\sum_{i=1}^{3} w_i} = \frac{w_1 x_1 + w_1 x_2 + w_3 x_3}{w_1 + w_2 + w_3}$$

49

Example 1.4

Peter's scores in the assignment , test, and final examination of a subject are displayed in the table below. The corresponding weightings are already shown in the rightmost column:

Assessment	**Mark**	**Weighting**
Assignment	60	20%
Test	70	30%
Final Examination	80	50%

Find the mean mark of Peter in the subject.

50

Solution

$w_1 = 20\%, w_2 = 30\%, w_3 = 50\%$

$$\bar{x} = \frac{\sum_{i=1}^{3} w_i x_i}{\sum_{i=1}^{3} w_i x_i} = \frac{20\% \times 60 + 30\% \times 70 + 50\% \times 80}{20\% + 30\% + 50\%} = 73$$

His mean mark is 73 in the subject.

51

1.11.6 Median

The median is the value of a variable that divides the distribution in such a way that the number of items below it is (approximately) equal to the number of items above it. Hence, median is a ***positional average***. It represents about 50% of the data.

52

1.11.6 Median (Cont'd)

Since median is a positional average, we have to arrange the data in ascending order of magnitudes before finding its value.

- If the number of observations (n) is odd, then
  median = middle value of the data;

- If n is even, then
  median = mean of the middle 2 values

Example 1.5

A sample of 10 automobiles was taken to determine the typical completion time (in months) for the construction of automobiles:

4.1, 3.2, 2.8, 2.6, 3.7, 3.1, 9.4, 2.5, 3.5, 3.8

Find the median completion time.

Solution

The ranked data are as follows:

2.5, 2.6, 2.8, 3.1, 3.2, 3.5, 3.7, 3.8, 4.1, 9.4

$$\tilde{x} = \frac{1}{2}(3.2 + 3.5) = 3.35 \text{ months}$$

If we delete the last value, then *n* becomes odd:

2.5, 2.6, 2.8, 3.1, 3.2, 3.5, 3.7, 3.8, 4.1

middle one

Median = 3.2

55

1.11.7 Mode

The mode of a distribution is the value at the point around which the items tend to most heavily concentrate.

Example 1.6

Dataset 1: 1,1, 2,3, 6, 8, 9, 9, 9 Mode = 9

Dataset 2: 2, 2, 2, 7,7,7, 9, 11, 23 Mode = 2, 7

Dataset 3: 1, 3, 5, 7, 9, 11, 13, 15

Mode does not exist

56

1.12 Measures of Dispersion

As useful as the measure of central location is in providing some understanding about the data in a distribution, total reliance on the information conveyed by the mean, the median, and the mode can be misleading, as we now illustrate.

Suppose that a test is given to 2 sections of a class, A and B. The table gives the scores recorded in the 2 sections.

Section A	**Section B**
Score	Score
56	30
58	35
60	60
60	60
60	60
62	85
64	90

By observation and simple calculations, we have

Measure of Central Tendency	Section A	Section B
Mean	60	60
Median	60	60
Mode	60	60

Do the 3 measures imply that the performance of the Sections A and B are the same?

59

The answer to the question above is negative. After scrutiny of the data in both sections, we can see that:

- In Section A, the group of students is homogeneous, all of them scoring in the vicinity of 60 points.
- In Section B, on the other hand, the performance is very erratic, from a low of 30 points to a high of 90 points.

Definition

Dispersion is defined as the extent of the scatteredness of observations on each side of a measure of central tendency, i.e., the mean, the median, the mode, etc.

60

Standard Deviation

The commonly used measure of dispersion/spread/variability is the standard deviation. Its definition is given by

- Population Standard Deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}, \quad \text{where N = population size}$$

- Sample Standard Deviation

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \text{where n = sample size}$$

Remarks:

- If the dispersion is small, it indicates a high uniformity of observations in the distribution.
- A large dispersion shows little uniformity.

61

Example 1.7

A sample of 10 automobiles was taken to determine the typical completion time (in months) for the construction of automobiles:

4.1, 3.2, 2.8, 2.6, 3.7, 3.1, 9.4, 2.5, 3.5, 3.8

Find the standard deviation of the completion times.

62

Solution

We use CASIO FX50FHII to compute the SD:

Data input:
4.1 DATA
3.2 DATA
.
.
.
3.8 DATA

SD = 2.01 (press: "shift 2 3 =")

63

Estimator and Estimate

- An ***estimator*** refers to a statistic that is used to generate an estimate ***once data are collected***. Thus, the estimator is the tool that can be used.
- An ***estimate*** is the product of an application of the tool (estimator).

An Example

The sample variance, $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ is an estimator of the population variance. s = 7.12 might be an estimate of the variance of a target population derived from a sample.

64

Estimator and Estimate (Cont'd)

- A ***capital*** letter is used to refer to an estimator.

  Examples:

  $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (sample mean estimator)

  $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ (sample SD estimator)

- A ***small*** letter is employed to refer to an estimate.

  Examples:

  $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (sample mean estimator)

  $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (sample SD estimator)

65