1 Overview and Descriptive Statistics

1

1.1 Populations, Samples, and Processes

2

Populations, Samples, and Processes

- Engineers and scientists are constantly exposed to collections of facts, or **data**, both in their professional capacities and in everyday activities.

- Discipline of statistics provides methods for
  - organizing and
  - summarizing data and
  - drawing conclusions based on information contained in data.

3

3

Populations, Samples, and Processes

- Investigation will typically focus on a well-defined collection of objects constituting a **population** of interest.
  - population might consist of all gelatin capsules (แคปซูลเจลาติน) of a particular type produced during a specified period.

  - Another investigation might involve population consisting of all individuals who received a bachelor of science (B.S.) in engineering during the most recent academic year.

4

4

Populations, Samples, and Processes

- When desired information is available for all objects in population, we have what is called a **census**.
  - Constraints on time, money, and other scarce resources usually make census impractical or infeasible.
- Instead, subset of the population—**a sample**—is selected in some prescribed manner.

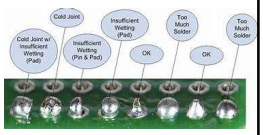
- Thus we might obtain sample of bearings from a particular production run as a basis for investigating
  - bearings are conforming to manufacturing specifications, or
  - selected sample of last year's engineering graduates to obtain feedback about quality of engineering curricula.

5

5

Populations, Samples, and Processes

- We are usually interested only in certain characteristics of objects in population:
  - number of flaws on surface of each casing,
  - thickness of each capsule wall,
  - gender of engineering graduate,
  - age at which individual graduated, and so on.



- Characteristics may be
  - Categorical, such as gender or type of malfunction, or
    - Value of characteristic is category such as female or insufficient solder
  - Numerical in nature.
    - Value of characteristic is number such as age = 23 or diameter = 0.502 cm

6

6

Populations, Samples, and Processes

- **Variable** is any characteristic whose **value** may change from one object to another in population.
- We denote variables by lowercase letters

Examples include

- x = brand of calculator owned by student
- y = number of visits to a particular Web site during a specified period
- z = braking distance of automobile under specified conditions

7

7

Populations, Samples, and Processes

- **Data** results from making observations either on a single variable or simultaneously on two or more variables.
  - **Univariate data**
  - **Bivariate data**
  - **Multivariate data**

8

8

Populations, Samples, and Processes

- **Univariate** data set consists of observations on a single variable.
- *Example I:* type of transmission, automatic (A) or manual (M), on each of ten automobiles

categorical data set ⇨ M A A A M A A M A A

Example II : Pulse rates (beats per minute) for patients recently admitted to adult intensive care unit is numerical univariate data set :

numerical univariate data set: ⇨ 88 80 71 103 154 132 67 110 60 105

9

9

Populations, Samples, and Processes

- **Bivariate** data : observations are made on each of two variables.
- Example :
  - Data set might consist of (height, weight) pair for each basketball player on a team
    - the first observation as (72, 168),
    - the second as (75, 212), and so on.
- If engineer determines the value of both
  x = component lifetime and
  y = reason for component failure

Data set is bivariate ⇨ (x, y)

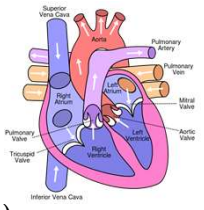
Numerical variable, Categorical variable

10

10

Populations, Samples, and Processes

- **Multivariate** data : observations are made on more than one variable (Note that : bivariate data is a special case of multivariate).



For example :

- Research physician might determine
  - systolic blood pressure,(ความดันโลหิตช่วงบน/ช่วงหัวใจบีบ)
  - diastolic blood pressure, and (ความดันโลหิตช่วงล่าง/ช่วงหัวใจคลาย)
  - serum cholesterol level (ระดับคอเลสเตอรอลในเลือด)
- for each patient participating in a study.
- Each observation would be a triple of numbers, such as

(120, 80, 146)

Numerical variable

ความดันโลหิตช่วงบน : ปกติ ต่ำกว่า 130 มม.ปรอท (ทอรร์)
ความดันโลหิตช่วงล่าง : ปกติ ต่ำกว่า 85 มม.ปรอท
ระดับคอเลสเตอรอลในเลือดรวม : ปกติ ต่ำกว่า 200 mg/dl 11

11

Populations, Samples, and Processes

- In many multivariate data sets, some variables are numerical and others are categorical. wheel), and so on.

Example :

- The annual automobile issue of *Consumer Reports* gives values of such variables as
- type of vehicle (small, sporty, compact, mid-size, large),
- city fuel efficiency (mpg),
- highway fuel efficiency (mpg),
- drivetrain type (rear wheel, front wheel, four wheel), and so on.

Note : mpg = Miles Per Gallon

(t, c, h, d)

Categorical variable Numerical variable Categorical variabl

12

12

Branches of Statistics

13

13

Branches of Statistics

- **Descriptive Statistics (สถิติเชิงพรรณนา)**

  consists of methods for
  - organizing,
  - displaying and
  - describing data by using tables, graphs, and
  - summary measures

- **Inferential Statistics (สถิติเชิงอนุมาน / สถิติอ้างอิง)**
  - process of describing population based on sample results

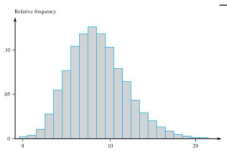
14

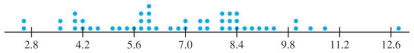
14

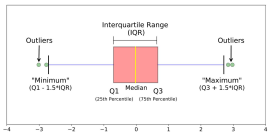
Branches of Statistics

Descriptive Statistics

- Investigator who has collected data may wish simply to summarize and describe important features of the data.
- Some of these methods are graphical in nature;
  - Histograms
  - Scatter Plot
  - Boxplots







- Other descriptive methods involve calculation of numerical summary measures, such as
  - Means, (ค่าเฉลี่ย)
  - Standard Deviations (ส่วนเบี่ยงเบนมาตรฐาน), and
  - Correlation Coefficients. (สัมประสิทธิ์สหสัมพันธ์)

15

15

Branches of Statistics

- The wide availability of statistical computer software packages has made these tasks much easier to carry out than they used to be.
- Computers are much more efficient than human beings at calculation and creation of pictures (once they have received appropriate instructions from the user!).
- This means that the investigator doesn't have to expend much effort on "grunt work" and will have more time to study data and extract important messages.
- Throughout this book, we will present output from various packages such as Minitab, SAS, S-Plus, and **R**.
  - R software can be downloaded without charge from the site http://www.r-project.org.

16

16

Example 1

- Charity is big business in the United States.
- The Web site charitynavigator.com gives information on roughly 6000 charitable organizations, and there are many smaller charities that fly below the navigator's radar screen.
- Some charities operate very efficiently, with fundraising and administrative expenses that are only a small percentage of total expenses, whereas others spend a high percentage of what they take in on such activities.



17

17

Example 1

cont'd

Here is data on fundraising expenses as a percentage of total expenditures for a random sample of 60 charities:

6.1	12.6	34.7	1.6	18.8	2.2	3.0	2.2	5.6	3.8
2.2	3.1	1.3	1.1	14.1	4.0	21.0	6.1	1.3	20.4
7.5	3.9	10.1	8.1	19.5	5.2	12.0	15.8	10.4	5.2
6.4	10.8	83.1	3.6	6.2	6.3	16.3	12.7	1.3	0.8
8.8	5.1	3.7	26.3	6.0	48.0	8.2	11.7	7.2	3.9
15.3	16.6	8.8	12.0	4.7	14.7	6.4	17.0	2.5	16.2

- Without any organization, it is difficult to get a sense of data's most prominent features—
  - what a typical (i.e. representative) value might be,
  - whether values are highly concentrated about typical value or quite dispersed,
  - whether there are any gaps in data,
  - what fraction of values are less than 20%, and so on.

18

18

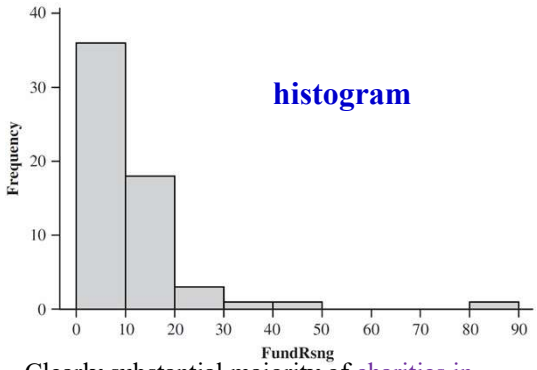
Example 1

cont'd

Figure 1.1 shows what is called a *stem-and-leaf display* as well as a *histogram*.

```
Stem-and-leaf of FundRsng N = 60
Leaf Unit = 1.0
     0 | 01111122223333333344
     0 | 55556666666778888
     1 | 0001222244
     1 | 55666789
     2 | 01
     2 | 6
     3 | 4
     3 |
     4 |
     4 | 8
     5 |
     5 |
     6 |
     6 |
     7 |
     7 |
     8 | 3
```

stem-and-leaf display



histogram

A Minitab stem-and-leaf display (tenths digit truncated) and histogram for the charity fundraising percentage data

- Clearly substantial majority of charities in sample spend less than 20% on fundraising, and only a few percentages might be viewed as beyond the bounds of sensible practice.

19

19

Branches of Statistics

- Having obtained **sample** from **population**, investigator would frequently like to use **sample information** to draw some type of **conclusion** (make an inference of some sort) about **population**.
- Techniques for generalizing from sample to population are gathered within the branch of our discipline called

Inferential statistics.

20

20

Branches of Statistics

- Main focus of this book is on presenting and illustrating methods of **Inferential Statistics** that are useful in engineering and scientific work
  - The most important types of inferential statistics procedures
    - Point estimation
    - Hypothesis testing
    - Estimation by confidence intervals
- Mastery of Probability leads to better understanding of
  - how inferential procedures are developed and used
  - how statistical conclusions can be translated into everyday language and interpreted
  - when and where pitfalls can occur in applying methods

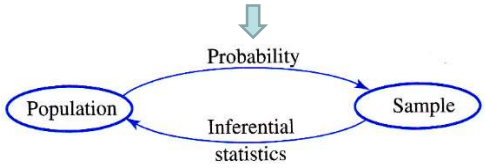
24

24

Branches of Statistics

- Probability and statistics both deal with questions involving **populations** and **samples**, but in an "inverse manner" to another
- The relation between **probability** and **inferential statistics**

In **probability problem**, properties of population under study are assumed **known** and questions regarding sample taken from population are posed and answered



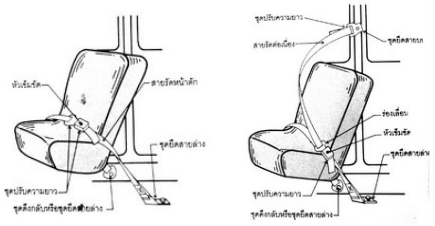
In **statistics problem**, characteristics of sample are available to experimenter, and this information enables experimenter to draw conclusions about population

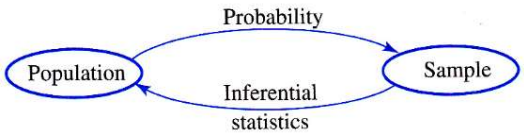
25

25

Branches of Statistics

Example : drivers' use of manual lap belts in cars equipped with automatic shoulder belt systems

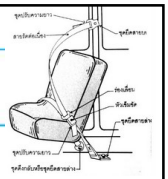




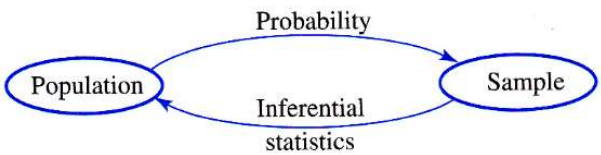
26

26

Branches of Statistics



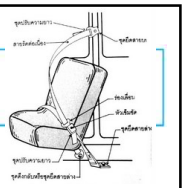
- **In probability**, we might assume that 50% of **all drivers** of cars in a certain metropolitan area regularly use their lap belt (**assumption** about **population**) – we might ask.
  - "How likely is it that sample of 100 drivers will include at least 70 who regularly use their lap belt?" or
  - "How many of drivers in sample of size 100 can we expect to regularly use their lap belt?"



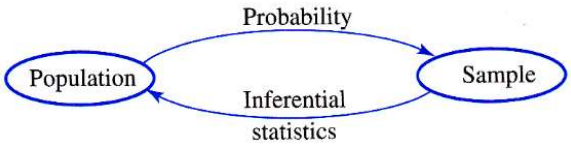
27

27

Branches of Statistics



- **In inferential statistics**, we have sample information available; for example, sample of 100 drivers of cars revealed that 65 regularly use their lap belt – we hight ask.
  - "Does this provide substantial evidence for concluding that more than 50% of **all drivers** in this area regularly use their lap belt?"



28

28