Statistics I – Chapter 1
What is Statistics?

Ling-Chieh Kung

Department of Information Management
National Taiwan University

September 12, 2012

What is Statistics?

- The **science** of gathering, analyzing, interpreting, and presenting **numerical** data.
- Using mathematics (particularly **probability**).
- To achieve better decision making.
- Scientific management.

What is Statistics?

- Some things are unknown...
  - Consumers' tastes.
  - Quality of a product.
  - Stock prices.
  - Employers' preferences.
- We want to understand these unknowns.
- We use statistical methods to gather, analyze, interpret, and present data to obtain **information**.
- Harder to apply on non-numerical data.

What is Statistics?

- The study of Statistics includes:
  - Descriptive Statistics.
  - Probability.
  - Inferential Statistics: Estimation.
  - Inferential Statistics: Hypothesis testing.
  - Inferential Statistics: Prediction.

Road map

- **Basic statistical concepts**.
  - Populations v.s. samples.
  - Descriptive v.s. inferential Statistics.
  - Parameters v.s. statistics.
- Variables and data.
- Data measurement.

Populations v.s. samples

- A **population** is a collection of persons, objects, or items.
  - A **census** is to investigate the whole population.
- A **sample** is a portion of the population.
  - A **sampling** is to investigate only a subset of the population.
  - We then use the information contained in the sample to infer ("guess") about the population.

Populations v.s. samples

- All students in NTU form a population.
  - All students in the business school form a sample.
  - 1000 students out of them form a sample.
- All students in the business school form a population.
  - All male students in the school form a sample.
- All chips made in one factory form a population.
  - Those made in a production lot form a sample.
- All packets passing a router form a population.
  - Those having the same destination form a sample.
- Are these samples **representative**?

Descriptive v.s. inferential Statistics

- **Descriptive Statistics**:
  - Graphical or numerical summaries of data.
  - Describing (visualizing or summarizing) a **sample**.
- **Inferential Statistics**:
  - Making a "**scientific guess**" on unknowns.
  - Trying to say something about the **population** .
- Most of our efforts in this year will be for inferential Statistics.

Examples of descriptive Statistics

- The average monthly income of 1000 people.
  - 1000 people form a sample.
  - The average monthly income summarizes the sample.
- The histogram of the monthly income of 1000 people.
  - Another way of describing the sample.
  - In particular, we visualize the sample.

Examples of inferential Statistics

- Pharmaceutical research.
  - All the potential patients form the population.
  - A group of randomly selected patients is a sample.
  - Use the result on the sample to infer the result on the population.
- A new product.
  - All the consumers in Taiwan form the population.
  - May try the new product in some of the stores before selling it in all stores.

Some remarks on descriptive Statistics

- Descriptive methods can also be applied on populations.
- Chapter 2: Describing data through graphs. We may draw graphs for a sample or a population.
- Chapter 3: Describing data through numbers. We may calculate those numbers for a sample or a population.

Parameters v.s. statistics

- A descriptive measure of a population is a **parameter**.
  - The average height of all NTU students.
  - The average willingness-to-pay of a new product of all potential consumers.
- A descriptive measure of a sample is a **statistic**.
  - The average height of all NTU male students.
- Understanding a population typically requires one to understand the parameter.
- Typically by investigating some statistics.

Parameters v.s. statistics: an example

- A laptop manufacturer wants to know the largest weight one can put on a laptop without destroying it.
  - Denote this number as θ.
  - θ can be various for different laptop!
- Suppose 10000 laptops have been produced.
- The parameter: $\min[\theta]$.
  - This will be the number announced to the public.
- Can the manufacturer conduct a census?

Parameters v.s. statistics: an example

- So probably 50 laptops will be randomly chosen as a sample for one to do inferential Statistics.
- For each laptop, we do an experiment (by destroying the laptop) and get a number x_i, $i = 1, 2, ..., 50$.
- These x_is form a sample.
- What is a statistic?
  - Any descriptive summary of the sample.
  - E.g., $\bar{x} = \sum_{i=1}^{50} x_i$, $\min_{i=1,...,50} \{x_i\}$, etc.
- Which statistic is "closer to" the parameter?

Some remarks for the example

- A parameter is a **fixed** number.
  - The parameter is $\min[\theta]$, a fixed number we want to estimate.
  - θ is NOT a parameter! θ is **random** and can never be found, even with a census.
  - While $\min[\theta]$ describes the population, θ describes only one single laptop.
- Statistics is a field. A statistic is a **number** or a **function**. Two statistics are two numbers or two functions.
- The selection of statistics matters. The **sampling process** also matters.

Another example

- (Suppose) there is a new proposal of increasing the tuition in NTU.
- We want to know the percentage of students supporting it.
- What is the population?
- What kind of statistics may we collect?
- Is it fine to sampling by standing at the "small small commissary"? How about the "normal teaching building"?

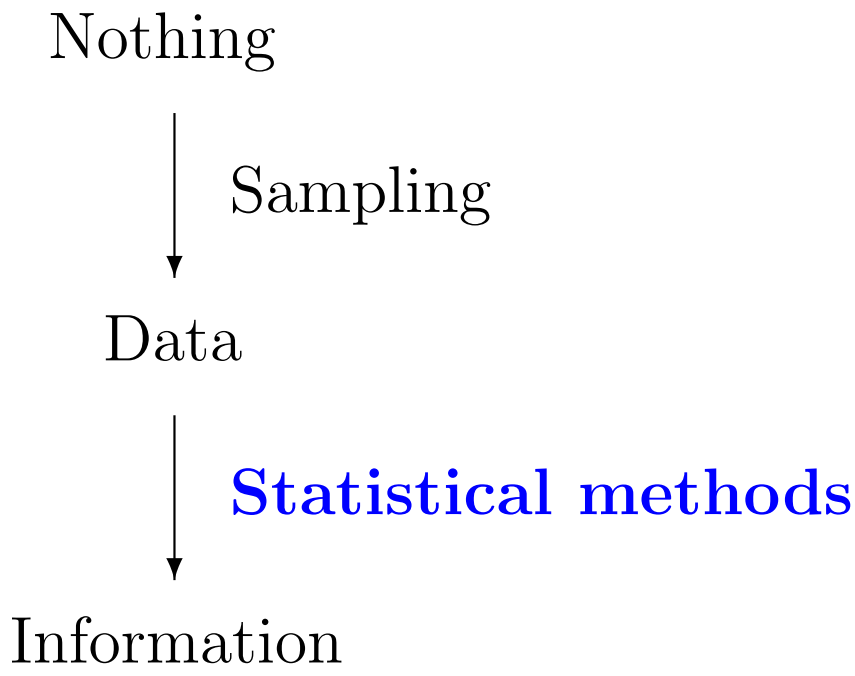
Road map

- Basic statistical concepts.
- **Variables and data**.
- Data measurement.

Variables and data

- A **variable** is an attribute of an entity that can take on **different values**, from entity to entity, from time to time.
  - The weight of a laptop.
  - The willingness-to-pay of a consumer for a product.
  - The result of flipping a coin.
- A **measurement** is a way of assigning values to variables.
- **Data** are those recorded values.

From data to information



Road map

- Basic statistical concepts.
- Variables and data.
- **Data measurement**.

Levels of data measurement

- In this year, most data we face will be numerical.
- Among all numerical data, there are some differences.
- Do identical numbers have an identical relation within different contexts?
  - In a post office, one package weights 60 kg while the other weights 80 kg.
  - In a baseball team, A's jersey number is 60 while B's is 80.
  - Is B heavier or bigger than A?

Levels of data measurement

- It is important to distinguish the following four levels of data measurement:
  - Nominal.
  - Ordinal.
  - Interval.
  - Ratio.

Nominal level

- A nominal scale classifies data into distinct categories in which **no ranking** is implied.
- Data are labels or names used to identify an attribute of the element.
- A non-numeric label or a numeric code may be used.
- Examples:

Categorical variables	Values (Categories)
Laptop ownership	Yes / No
Place of living	Taipei / Taoyuan / ...
Internet provider	AT&T / Comcast / Other

Coding for nominal data

- Let one's marital status be coded as:
  - Single = 1.
  - Married = 2.
  - Divorced = 3.
  - Widowed = 4.
- Because the numbering is arbitrary, arithmetic operations don't make any sense.
  - Does Widowed $\div$ 2 = Married?!

Ordinal level

- An **ordinal** scale classifies data into distinct categories in which **ranking** is implied.
- The order or rank of the data is meaningful.
- However, the differences between numerical labels DO NOT imply **distances**.
- Examples:

Categorical variables	Values (Categories)
Product satisfaction	Satisfied, neutral, unsatisfied
Professor rank	Full, associate, assistant
Ranking of scores	1, 2, 3, 4, ...

Coding for Ordinal data

- Ranking is meaningful for ordinal data.
  - A full professor is ranked higher than an associate professor.
  - A rank-10 student gets a higher grade than a rank-20 student.
- However, it is still not meaningful to do arithmetic on ordinal data.
  - Assistant + associate = full?!
  - The grade difference between no. 1 and no. 5 may not be equal to that between no. 11 and no. 15.

Interval and ratio levels

- An **interval** scale is an ordered scale in which the **difference** between measurements is a meaningful quantity but the measurements DO NOT have a true zero point.
- A **ratio** scale is an ordered scale in which the difference between measurements is a meaningful quantity and the measurements DO have a true **zero point**.
- For interval data:
  - Zero does not mean nothing; ratio is not meaningful.
  - E.g., Degrees in Celsius or Fahrenheit.
- For ratio data:
  - Zero means nothing; ratio is meaningful.
  - E.g., Degrees in Kelvin.

Interval and ratio levels

- Interval data are actually rare.
  - Another example: GRE or GMAT scores.
- Ratio data appear more often in the world.
  - Heights.
  - Weights.
  - Income.
  - Prices.

Comparisons of the four levels

- For each level, is it meaningful to calculate the ...

Level	Ranking	Distance	Ratio
Nominal	No	No	No
Ordinal	Yes	No	No
Interval	Yes	Yes	No
Ratio	Yes	Yes	Yes

- Nominal and ordinal data are called **qualitative data**.
- Interval and ratio data are called **quantitative data**.

Some remarks

- It is important to distinguish nominal from ordinal, from ordinal to interval, but NOT from interval to ratio.
- Most statistical methods are for quantitative data.
  - To apply these methods, typically one does not need to distinguish between interval and ratio data.
- Some method are for qualitative data.
  - To apply these methods, one need to distinguish between nominal and ordinal data.
  - Will be covered only in the Spring semester.