GMBA 7098: Statistics and Data Analysis

Introduction

Ling-Chieh Kung

Department of Information Management
National Taiwan University

Road map

- **What is statistics**?
- Syllabus.
- In-class brainstorming.
- Basic statistical concepts.

Coffee pricing

- How to **set the price** p of a cup of coffee?
- We look for a balance between profit margin and demand volume.
- Suppose that you know:
  - The unit cost of making one cup of coffee is 10 NTD.
  - The demand as a function of the price:

Price (NTD)	40	50	60	70	80
Demand (cup)	300	280	230	200	220

  Then simple calculation helps you find a profit-maximizing price.
- But **how** would you find the demand function?

Coffee pricing

- The first thing: Collect **data**!
- Given this set of data, how would you estimate the demand function?
- Do you need **more data**?
  - More records?
  - More attributes?

Day	Price	Demand	Day	Price	Demand
1	40	312	16	60	198
2	40	307	17	60	239
3	40	267	18	60	271
4	40	287	19	70	165
5	40	343	20	70	178
6	40	331	21	70	194
7	50	276	22	70	202
8	50	290	23	70	188
9	50	275	24	70	210
10	50	300	25	80	240
11	50	243	26	80	233
12	50	266	27	80	167
13	60	212	28	80	198
14	60	234	29	80	179
15	60	256	30	80	225

Measuring unknowns in the world

- It is always challenging to **measure unknowns** in the world.
- To help us measure unknowns, people develop the field of **Statistics**.
- Statistics is the **science** of collecting, analyzing, interpreting, and presenting (**numerical**) data.
  - For texts: text mining, natural language processing, etc.
  - For images: image recognition, digital image processing, etc.
- Mathematics (particularly probability) is helpful.
  - E.g., to help us model and measure the uncertainty when estimating consumer demands.
- Ultimate goal (of Business Statistics): to achieve better decision making.

What is Statistics?

- Many things are unknown...
  - Consumers' tastes.
  - Quality of a product.
  - Stock prices.
  - The effectiveness of a new way of teaching/training.
- The study of Statistics includes:
  - Descriptive Statistics.
  - Probability.
  - Inferential Statistics: Estimation.
  - Inferential Statistics: Hypothesis testing.
  - Inferential Statistics: Prediction.
- In summary: To **estimate**, **test**, and **predict** those unknowns.

Road map

- What is statistics?
- **Syllabus**.
- In-class brainstorming.
- Basic statistical concepts.

The instructing team

- Instructor:
  - Ling-Chieh Kung.
  - Assistant professor.
  - Department of Information Management.
  - Office: Room 413, Management Building II.
  - Office hour: by appointment.
  - E-mail: lckung@ntu.edu.tw.
- Teaching assistant:
  - Jeff Lee (r04725023@ntu.edu.tw).
  - Share Lin (r04725037@ntu.edu.tw).
  - First-year master students.
  - Department of Information Management.
  - Office: Room 320C, Management Teaching and Research Building.

Language and references

- Language: **"All" English**.
  - All materials are in English (except some original names in Chinese).
  - The instructor and TAs speak Chinese and English.
- Main references:
  - *Business Statistics: For Contemporary Decision Making* by Ken Black.
  - Any introductory Business Statistics textbook at the same level is fine.
  - Students are not required to buy one.
- Other references:
  - *Freakonomics* by Steven Levitt and Stephen Dubner.
  - *Big Data* by Viktor Mayer-Schönberger and Kenneth Cukier.
  - *Learn R in a Day* by Steven Murray (Amazon Kindle e-books only).
  - Not related to homework or exam.

"Flipped classroom"

- Lectures in **videos**, then practices and discussions in classes.
- Before each Monday, the instructor uploads a few videos of lectures.
  - The videos in total will be around 60 minutes.
  - Students are expected to watch the videos before that Monday.
- During the lecture, we do three things:
  - Discussing the lecture materials.
  - Doing lecture problems (to earn points).
  - Further discussions.
- Teams:
  - Students form teams by themselves to work on lecture problems.
  - Each team should have **two or three** students.
  - For different weeks, one may have different teammates.

Homework, project, and exam

- About five homework assignments will be assigned.
  - Discussions are encouraged.
  - One should submit her own work.
  - Assignments are uneasy and time-consuming.
- Exams:
  - In-class and open whatever you have (including all electronic devices).
  - No information is allowed to be transferred among students.
  - The final exam is comprehensive.
- Final project:
  - Students form teams to apply the techniques learned in this course to a self-selected problem.
  - Each team does an oral presentation in one of the last two weeks.
  - All team members must be in class for the team to present.

Grading

- Class participation: 10%.
- Lecture problems: 10%.
- Homework: 20%.
- Exams (one of the following two ways):
  - Midterm 15% and final 15%.
  - Midterm 10% and final 20%.
- Final project: 30%.
- The final letter grades will be given according to the following conversion rule:

Letter	Range	Letter	Range	Letter	Range
A+	$[90, 100]$	B+	$[77, 80)$	C+	$[67, 70)$
A	$[85, 90)$	B	$[73, 77)$	C	$[63, 67)$
A−	$[80, 85)$	B−	$[70, 73)$	C−	$[60, 63)$

Important dates and websites

- Important dates:
  - Week 3 (2015/9/28): No class: Mid-autumn Festival.
  - Week 9 (2015/11/9): Midterm exam.
  - Week 16 (2015/12/28): Final exam.
  - Weeks 17 and 18 (2016/1/4 and 2016/1/11): Project presentations.
- CEIBA.
  - Viewing your grades.
  - Receiving announcements.
- `http://www.im.ntu.edu.tw/~lckung/courses/SDA15/`.
  - Downloading course materials.
  - Linking to lecture videos.
- `https://piazza.com/ntu.edu.tw/fall2015/mba7098/`.
  - On-line discussions.

Road map

- What is statistics?
- Syllabus.
- **In-class brainstorming**.
- Basic concepts.

Road map

- What is statistics?
- Syllabus.
- In-class brainstorming.
- **Basic concepts**.

Populations vs. 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.
  - **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.
- What are samples for the following populations?
  - All students in NTU.
  - All students in the business school.
  - All chips made in one factory.
  - All consumers who have bought iPhone 6.
- Two important questions:
  - **Why** sampling?
  - Is a sample **representative**?

Descriptive vs. inferential statistics

- **Descriptive statistics**:
  - Graphical or numerical summaries of data.
  - Describing (visualizing or summarizing) a set of data.
- **Inferential statistics**:
  - Making a "scientific guess" on unknowns.
  - Trying to say something about the population.
- Which is descriptive and which is inferential?
  - Calculating the average height of 1000 randomly selected NTU students.
  - Using this number to estimate the average height of all NTU students.
- Another example (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.

Parameters vs. statistics

- A numerical summary of a population is a **parameter**.
  - The average height of all NTU students.
  - The expected coffee demand when the price is 50 NTD.
- A numerical summary of a sample is a **statistic**.
  - The average height of all NTU male students.
  - The average coffee demand when the price is 50 NTD in the past 6 days.
- Almost always people use a statistic to infer a parameter.
  - Some statistics are "good" while some are "bad."

Parameters vs. statistics: an example

- What is the average height of all NTU students?
- While a census is possible, it is still quite costly.
- It is natural to:
  - Sample some NTU students.
  - Calculate a statistic.
  - Use that statistic to estimate the average height (the parameter).
- Some (good or bad) samples and statistics:
  - The average height of all students in this classroom.
  - The average height of 100 students randomly drawn from all students.
  - The maximum height of 100 students randomly drawn from all students.
  - The sum of heights of 100 students randomly drawn from all students.
  - The average height of 60 male and 40 female students randomly drawn from the population.

Parameters vs. statistics

- A parameter is a **fixed number**.
  - E.g., the average height of all NTU students.
- A statistic is a **random number** depending on the sample.
  - Two different random samples typically generate two values of a statistic.
  - The sampling process matters.

Levels of data measurement

- Most data we will play with are numerical.
- Numerical data may be categorized to three levels:
  - Nominal.
  - Ordinal.
  - Quantitative: interval or ratio.

Nominal level

- A **nominal** scale classifies data into categories with **no ranking**.
- Data are labels or names used to identify an attribute of the element.
- The label may be numeric or non-numeric label.
- Examples:

Categorical variables	Values (Categories)
Laptop ownership	Yes / No
Citizenship	Taiwan / Japan / ...
Country code	886 / 86 / 1 / ...

- Arithmetic operations **cannot** be applied on nominal data.

Ordinal level

- An **ordinal** scale classifies data into categories with **ranking**.
- The order or rank of the data is meaningful.
- However, **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, ...

- 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.

Quantitative (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 have a true **zero point**.
- Ratio data appear more often in the world.
  - Heights, weights, income, prices.
- Interval data are actually rare.
  - Degrees in Celsius or Fahrenheit.
  - GRE or GMAT scores.
- How about degrees in Kelvin?

Some remarks

- Nominal and ordinal data are called **qualitative data**.
- Interval and ratio data are called **quantitative data**.
- Most statistical methods are for quantitative data; some are for qualitative data.
  - Distinguishing nominal and ordinal scales is important.
  - Distinguishing interval and ratio scales is not.
- Sometimes quantitative data are called **numeric** data.

A short summary

- Understand these terms:
  - Populations vs. samples.
  - Parameters vs. statistics.
  - Inferential statistics vs. descriptive statistics.
- For each scale of measurement, is it meaningful to calculate the following numbers?

Level	Ranking	Distance
Nominal	No	No
Ordinal	Yes	No
Quantitative	Yes	Yes