Introduction to Statistics

Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

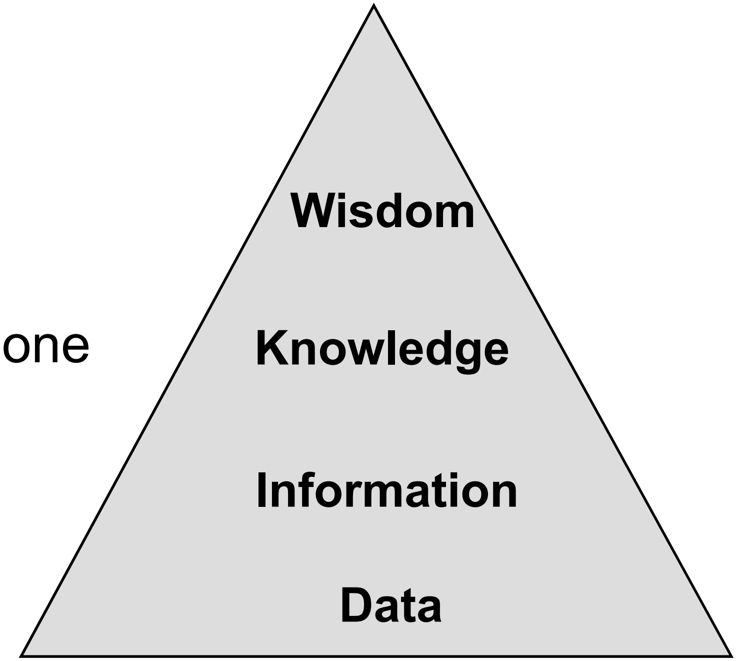
What is Statistics?

- Statistics is the field of study concerned with the collection, analysis, and interpretation (making decisions on) of uncertain data
  - E.g., the explanation of social or economic trends through the analysis of data
- Or, in more common usage, statistics refers to numerical facts of the data
  - E.g., the age of a student, the allowance of a student, the height of a student, etc.
- Another definition: Statistics is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data

統計學：“以偏概全” + “有所本” ??

Information Hierarchy

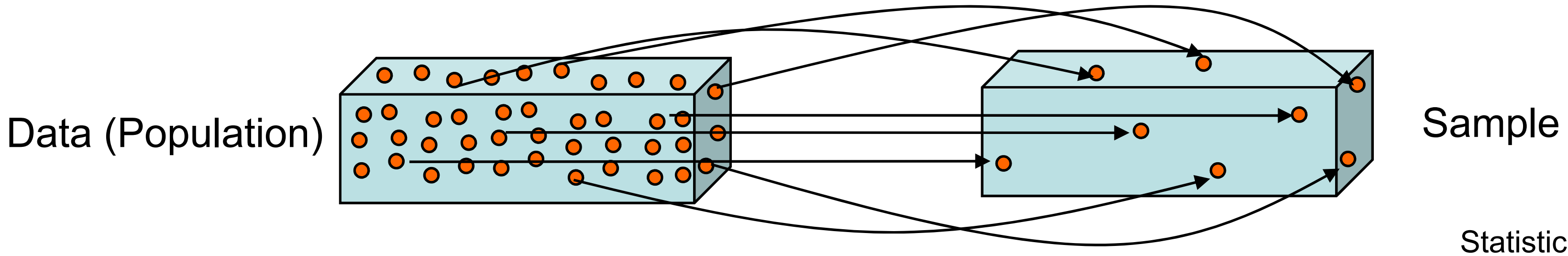
- Data
  - The raw material of information
- Information
  - Data organized and presented by someone
- Knowledge
  - Information read, heard or seen and understood
- Wisdom
  - Distilled and integrated knowledge and understanding



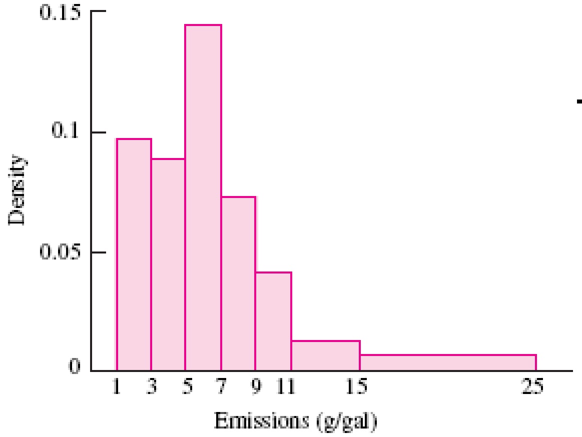
Types of Statistics (1/4)

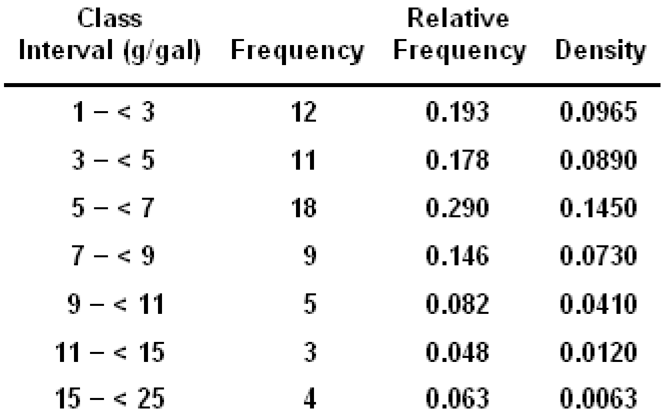
- Broadly speaking, statistics can be divided into two areas
  - Descriptive statistics (敘述統計學)
  - Inferential statistics (推論統計學)

- Descriptive Statistics
  - To be concerned with the methods of collecting data and of summarizing clearly the basic information they contain
    - Collecting data refers to sampling, i.e., choosing a subset of data (a sample)
    - Summarizing data refers to organizing, displaying, and describing data by tables, graphs, and summary measures



Types of Statistics (2/4)





Class Interval (g/gal)	Frequency	Relative Frequency	Density
1 – < 3	12	0.193	0.0965
3 – < 5	11	0.178	0.0890
5 – < 7	18	0.290	0.1450
7 – < 9	9	0.146	0.0730
9 – < 11	5	0.082	0.0410
11 – < 15	3	0.048	0.0120
15 – < 25	4	0.063	0.0063

- Histogram and Frequency table for PM emissions of 62 vehicles driven at high altitude

- Inferential statistics
  - Concerned with the methods that use sample results to help make decisions or predictions about the data (population)
  - Or, the methods that draw conclusions from the data

Types of Statistics (3/4)

- Example 1
  - A machine makes 1000 steel rods per hour, with a specification of 0.45 ± 0.02 cm
  - An engineer would like determine the quality/quantity of the production process by randomly draw a sample of rods (say, 50 rods)
  - Given that 92% of the sample meet the specification
    - How likely is the size of difference between the sample proportion and the population proportion?

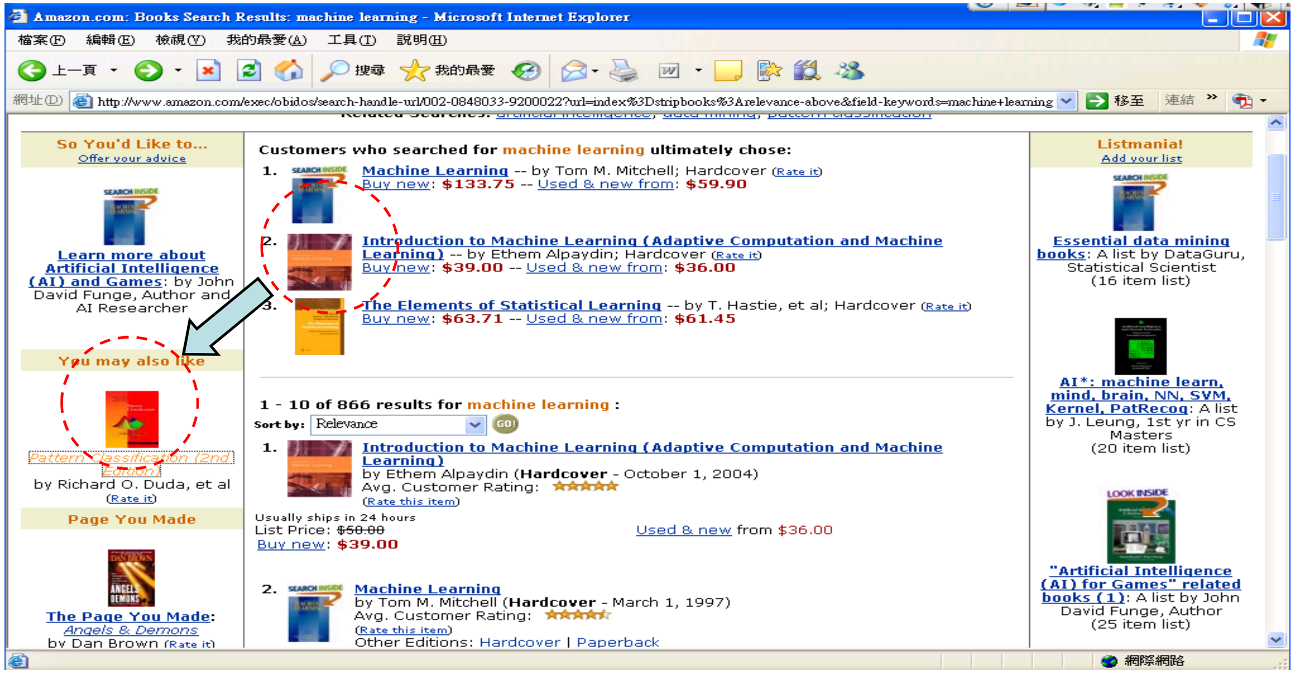
      Standard derivation (Chapters 2 and 4)
    - How is he confident that the true population proportion will be in 92% ± *x*%

      Confidence interval (Chapter 5)
    - Can he draw a conclusion that the percentage of good rods is at least 90%

      Hypothesis testing (Chapter 6)
    - ….

Types of Statistics (4/4)

- Example 2: relationship between two factors/populations


Amazon.com: Books Search Results: machine learning - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁
搜尋
我的最愛
網址(D)
http://www.amazon.com/exec/obidos/search-handle-url/002-0848033-9200022?url=index%3Dstripbooks%3Arelevance-above&field-keywords=machine+learning
移至
連結
網際網路

So You'd Like to...
Offer your advice

Learn more about Artificial Intelligence (AI) and Games: by John David Funge, Author and AI Researcher

You may also like

Pattern Classification (2nd Edition)
by Richard O. Duda, et al
(Rate it)

Page You Made

The Page You Made: *Angels & Demons*
by Dan Brown (Rate it)

Customers who searched for **machine learning** ultimately chose:

1. **Machine Learning** -- by Tom M. Mitchell; Hardcover (Rate it)
   Buy new: **$133.75** -- Used & new from: **$59.90**
2. **Introduction to Machine Learning (Adaptive Computation and Machine Learning)** -- by Ethem Alpaydin; Hardcover (Rate it)
   Buy new: **$39.00** -- Used & new from: **$36.00**
3. **The Elements of Statistical Learning** -- by T. Hastie, et al; Hardcover (Rate it)
   Buy new: **$63.71** -- Used & new from: **$61.45**

1 - 10 of 866 results for machine learning :

Sort by: Relevance GO!

1. **Introduction to Machine Learning (Adaptive Computation and Machine Learning)**
   by Ethem Alpaydin (**Hardcover** - October 1, 2004)
   Avg. Customer Rating: ★★★★★
   (Rate this item)

   Usually ships in 24 hours
   List Price: ~~$50.00~~
   Buy new: **$39.00**
   Used & new from $36.00

2. **Machine Learning**
   by Tom M. Mitchell (**Hardcover** - March 1, 1997)
   Avg. Customer Rating: ★★★★☆
   (Rate this item)
   Other Editions: Hardcover | Paperback

Listmania!
Add your list

Essential data mining books: A list by DataGuru, Statistical Scientist
(16 item list)

AI*: machine learn, mind, brain, NN, SVM, Kernel, PatRecog: A list by J. Leung, 1st yr in CS Masters
(20 item list)

"Artificial Intelligence (AI) for Games" related books (1): A list by John David Funge, Author
(25 item list)

- Association Rule:

  P(buying “Pattern Classification”| buying “Machine Learning”) = ?

Popular Software Packages for Statistics

- SPSS
- SAS
- MINITAB
- Microsoft Excel
- …

Textbook and Reference

- Textbook
  - William C. Navidi, “Statistics for Engineers and Scientists,” McGraw-Hill (2 edition, 2007)

- References
  - Prem S. Mann, "Introductory Statistics," Wesley, (6 edition, 2007)
  - D. P. Bertsekas, J. N. Tsitsiklis, “Introduction to Probability,” Athena Scientific (2002)

Topics to be Covered

- Descriptive Statistics (Chapter 1)
- Probability and Common Used Distributions (Chapters 2 & 4, quick review)
- Propagation of Error (Chapter 3)
- Confidence Intervals (Chapter 5)
- Hypothesis Testing (Chapter 6)
- Correlation and Simple Linear Regression (Chapter 7)
- More Topics:
  - Data Analysis and Dimension Reduction
  - Data Cleansing and Presentation
  - Bayesian Decision Theory
  - Parametric Methods - Bias and Variance of the Estimator
  - …

Grading (Tentatively)

- Midterm and Final: 50%
- Homework: 35%
- Attendance/Other: 15%

- TA: 劉家妏 同學 (碩一)
  - E-mail: acat103@yahoo.com.tw
  - Tel: 29322411ext 208 (資工系208室)