Probability and Statistics

Lectured by Dr. FENG Zhenghui

fengzhenghui@hit.edu.cn

Autumn, 2024

Course Information

- Probability and Statistics (MATH1004E)
  Venue: Mon. (10:30-12:15, T2-811), Wen. (10:30-12:15, T2-811)
- E-mail: fengzhenghui@hit.edu.cn
  Office hours: TBD

- TA: Dr. HUANG Zijian (Oscar)
  E-mail: huangzijian@hit.edu.cn
  Office: G-7th floor

Course Description

- This course is designed to provide undergraduate students with a comprehensive understanding of statistical concepts and their applications. This course aims to develop students' quantitative reasoning skills and equip them with the tools necessary to analyze and interpret data for decision-making.

Course Objectives

- Understand the fundamental concepts of descriptive and inferential statistics.
- Learn various techniques for data collection, organization, and presentation.
- Develop the ability to analyze and interpret data using appropriate statistical methods.
- Apply statistical tools to solve real-world problems.
- Enhance critical thinking skills by evaluating the validity and reliability of statistical results.

- Basic probability and statistical concepts and methods.
- Emphases on what, how, when and why certain statistical methods can and cannot be applied.
- Solving simple real-life problems by statistics.

Course Topics

1. Introduction to Statistics: Importance and role in decision-making, ethical considerations.
2. Descriptive Statistics: Measures of central tendency, variability, and graphical representation of data.
3. Probability: Basic concepts, probability distributions, and their applications.
4. Sampling and Estimation: Sampling techniques, confidence intervals, and sample size determination.
5. Hypothesis Testing: Null and alternative hypotheses, type I and type II errors, significance levels, and p-values.
6. Correlation and Regression Analysis: Relationship between variables, linear regression, and interpretation of regression models.

Assessment

Quizzes & Attendence

(10%)

Assignments

(30%)

Group projects

(20%)

Final examination

(40%)

Assignment (30%)

- You are encouraged to discuss homework problems with others, but must write your homework **independently**.
- **Duplicated** homework or solution with no supporting work receives **no credit**.
- Please complete and hand in your assignments **before** the deadline. If not**, no credit.**

Group Project (20%)

- 4-5 students in one group.
- After mid-term test, please **give the group name list to TA**.
- The deadline will be **the 3rd day after final exam.**
- **Each group should submit a report.**

Rubric

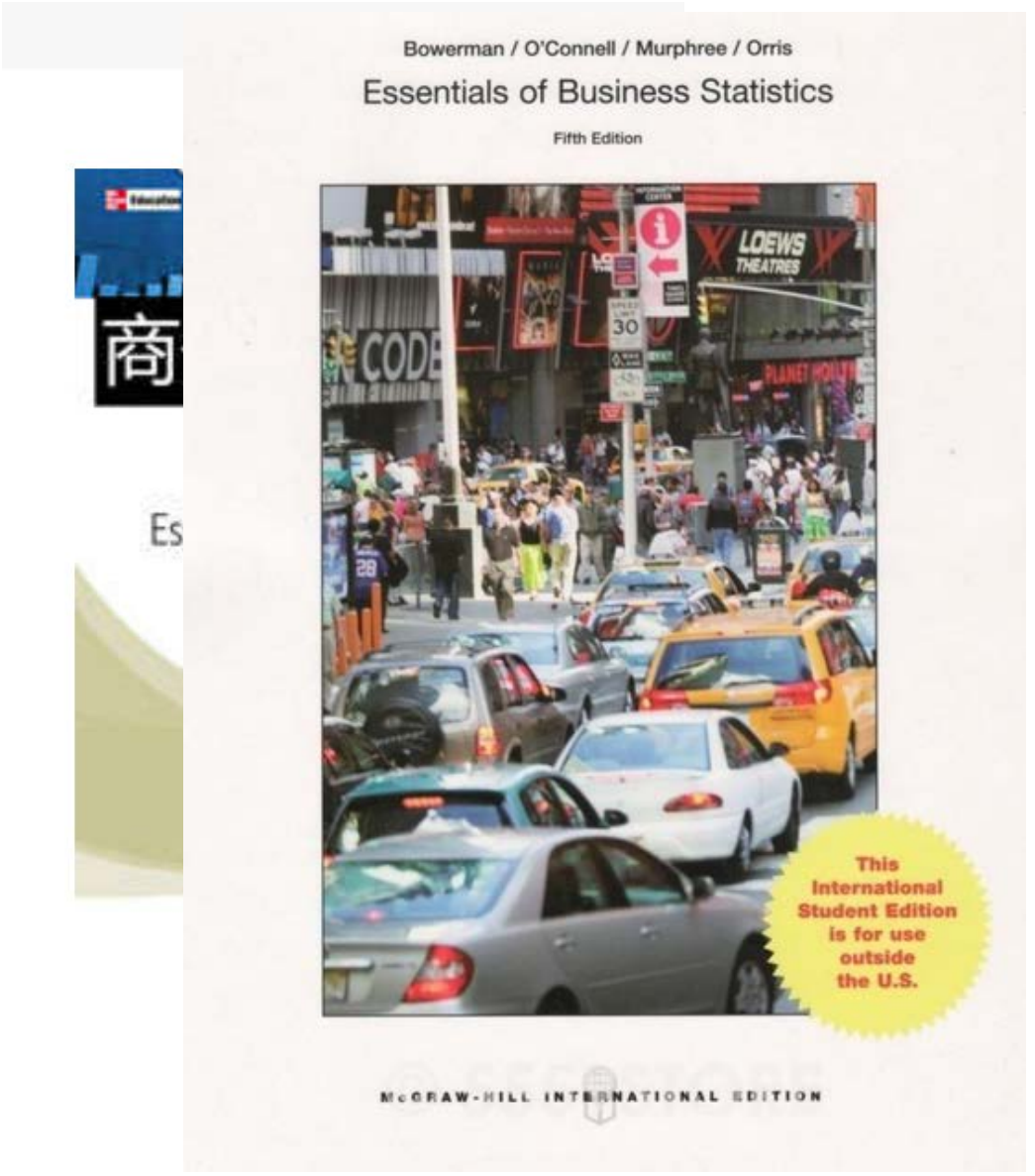
Generic Rubrics for Assessment of Project Assignment (presented in formal report format)

Criteria for assessment	***Performance levels***				
	Excellent 4	**Good 3**	**Satisfactory 2**	**Marginal Pass 1**	**Fail 0**
Focus and Contents (_50_ % weighting)	Clearly identifies the essence of the topic, a good and strong logical progression from the problem's introduction, analysis and to its conclusion / solution; information is relevant; main ideas are well supported by detailed, accurate and updated information.	Main ideas are clear; demonstrate a basic logical progression from the problem's introduction, analysis and to its conclusion/solution; information is relevant; main ideas are supported by information but are not in sufficient details or information is not updated.	Identifies the essence of the topic with some degree of confusion; a weak logical progression from the problem's introduction, analysis and to its conclusion/solution; main ideas are not well supported by information.	Main ideas are unclear; poorly identifies the essence of the topic, a poor logical progression from the problem's introduction, analysis and to its conclusion/solution; main ideas are barely supported by information.	Does not demonstrate the minimum understanding of the topic.
Organization and Presentation (_30_ % weighting)	Organization is well structured; showing good transitions between ideas; the length and depth of writing is appropriate.	Organization is clear; less transitions shown between ideas; the length and depth of writing is appropriate.	Basic organization is apparent; transitions connect ideas are somewhat mechanical; length and depth of work is either too little or too much.	Organization is weak; transitions connect ideas are weak; length and depth of work is either too little or too much.	There is no clear organization; no apparent transitions connecting ideas; length and depth of work is either too little or too much.
Sentence, Structure, Grammar, Mechanics, Spelling, neatness (_20_ % weighting)	All sentences are well constructed and have varied structure and length; no errors made in grammar, mechanics, and/or spelling; neatly bound in a report cover; illustrations provided.	Most sentences are well constructed and have varied structure and length; a few errors made in grammar, mechanics, and/or spelling, but they do not interfere with understanding; well-formed characters;	Most sentences are well constructed, but they have a similar structure and/or length; several errors in grammar, mechanics, and/or spelling that interfere with understanding; legible writing,	Most sentences are not well constructed and they have a similar structure and/or length; many errors in grammar, mechanics, and/or spelling that interfere with understanding; some ill-formed letters, print too small or too large; papers stapled together	Sentences sound Awkward and they are distractingly repetitive or are difficult to understand; numerous errors made in grammar, mechanics, and/or spelling that interfere with understanding; Illegible writing; loose pages.

Textbook

- Bowerman, O'Connell, Murphree and Orris. *Essentials of Business Statistcs*

- Bruce L. Bowerman, Richard T. O’ Connell , J.B. Orris, Emily S. Murphree
- Essentials of Business Statistics (Third Edition)
- ISBN-10：0-07-337368-0 Copyright © 2010 by McGraw-Hill Education (Asia) .



Reference

- D. Freedman, R. Pisani and R. Purves, Statistics, 3rd Ed., Norton, 1998.
- Chapter 1-18, 20-23, 26-28.
- W. Feller, Introduction to Probability Theory and Its Applications, Vol. 1, 3rd Ed., Wiley, 1968. page 1-258 J.E. Freund, Mathematical Statistics, 5th Ed., Prentice Hall, 1992, Chapter 1-9.
- Salsburg, The Lady Tasting Tea. How Statistics Revolutionized Science in the Twentieth Century, Freeman, 2001, Chapter 11-12
- Viktor Mayer-Schönberger and Kenneth Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Eamon Dolan/Mariner Books, 2014.
- Charles Wheelan, Naked Statistics: Stripping the Dread from the Data, W. W. Norton & Company, 2014.
- Joel Best and Patrick Lawlor, Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists, University of California Press; First Edition, 2012.
- Dana K. Keller, The Tao of Statistics: A Path to Understanding (With No Math), SAGE Publications, Inc, 2005.
- Gary Smith, Standard Deviations: Flawed Assumptions, Tortured Data, and Other Ways to Lie with Statistics, Overlook Hardcover, 2014.
- Matthew B. Robinson, Lies, Damned Lies, and Drug War Statistics, State University of New York Press, 2nd edition, 2014.
- Introduction to Mathematical Statistics. ROBERT V.HOGG， ALLEN T. CRAIG
- Essentials of Business Statistics. Bowerman, O'Connell, Murphree and Orris. 2015. McGraw-Hill/Irwin. ISBN10: 0078020530.
- Business Statistics in Practice. Bowerman, O'Connell, Murphree and Orris. McGraw-Hill/Irwin. ISBN 978-0-07-352149-7.

Zhenghui Feng (冯峥晖)

Associate Professor

- Research Interest

Mixture Model, Big Data, Functional Data Analysis,
Dimension Reduction, Variable Selection, Applied Statistics

- Education

PHD Statistics, Department of Mathematics,
Hong Kong Baptist University

MSc Statistics, School of Mathematics,
Beijing Normal University, China

Email:

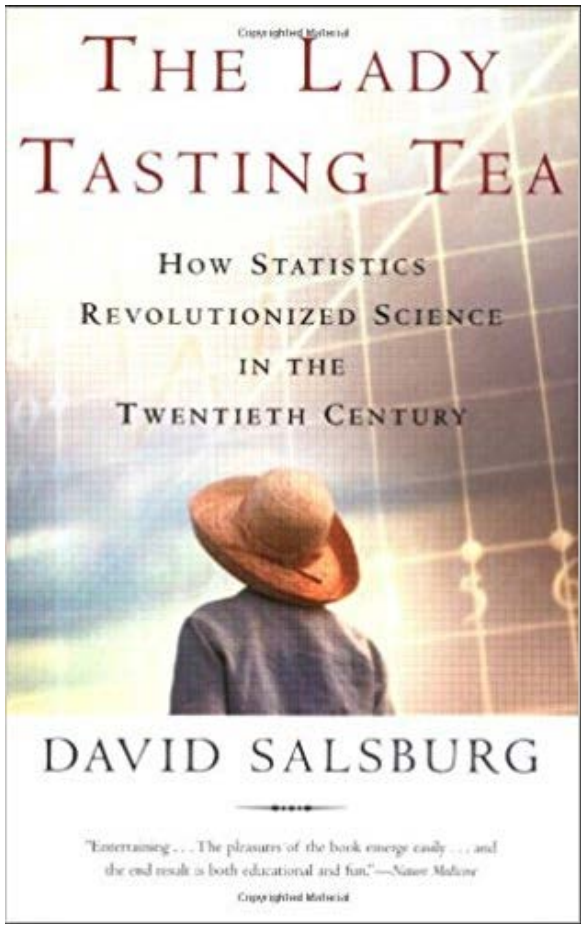
fengzhenghui@hit.edu.cn

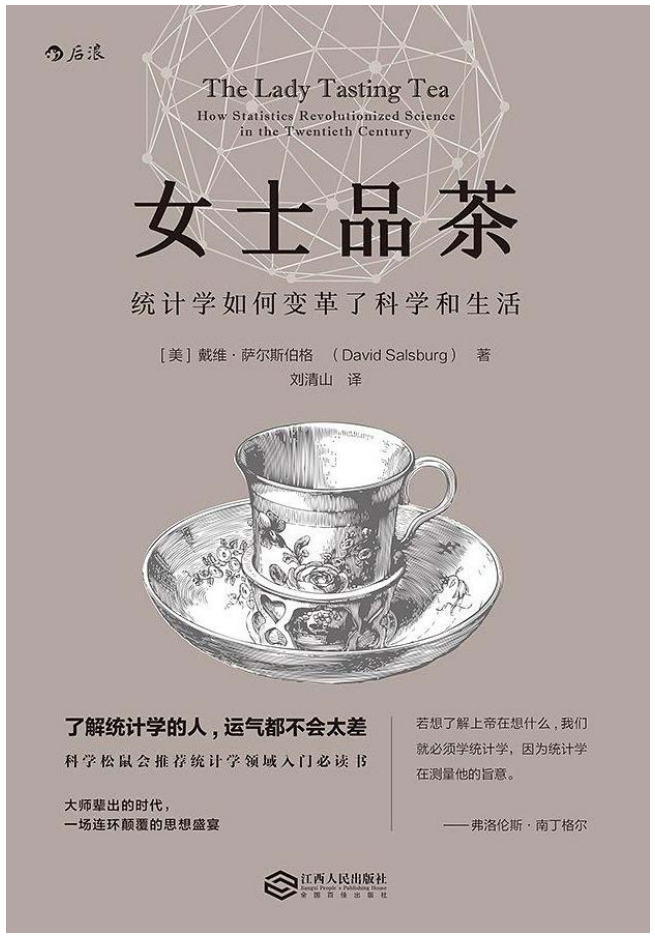
Office Hour: TBD
TA: HUANG Zijian, Oscar

The Lady Tasting the tea

Ronald Fisher in 1913

Books


THE LADY TASTING TEA
HOW STATISTICS REVOLUTIONIZED SCIENCE IN THE TWENTIETH CENTURY
DAVID SALSBURG


后浪
The Lady Tasting Tea
How Statistics Revolutionized Science in the Twentieth Century
女士品茶
统计学如何变革了科学和生活
[美] 戴维·萨尔斯伯格 (David Salsburg) 著
刘清山 译
了解统计学的人，运气都不会太差
科学松鼠会推荐统计学领域入门必读书
大师辈出的时代，
一场连环颠覆的思想盛宴
若想了解上帝在想什么，我们就必须学统计学，因为统计学在测量他的旨意。
——弗洛伦斯·南丁格尔
江西人民出版社

Reading List

- Darrell Huff (1991) *How to Lie with Statistics* New Ed edition, ISBN 0-14-013629-0
- Rao C R. (1997) *Statistics and truth: putting chance to work* World Scientific.
- Salsburg, D. (2002) *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, W.H. Freeman / Owl Book. ISBN 0-8050-7134-2
- Reinhart A. (2015) *Statistics done wrong: The woefully complete guide* No Starch Press.
- Moore S. and Notz W. (2017) Statistics: Concepts and Controversies 9th Edition.

What is Statistics all about?

- The subject of **statistics** involves the study of how to collect, summarize, analyze and interpret data.

- **Statistics** is described as a mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data or as a branch of mathematics concerned with collecting and interpreting data.

What is Statistics all about?

- **Data** are numerical facts and figures from which conclusions can be drawn. Such conclusions are important to the decision-making processes of many professions and organizations.
- Examples:
  - How many sisters/brothers do you have?
  - How often do you exercise? ----Never(1); rarely(2); often (3); always (4)
  - How happy are you with your life? ----Very low (1); Very high (10)

Why Study Statistics?

- Generally, make “decision”
- Statistical techniques are used to make decisions that affect our daily lives (without knowing it)
  - Choose the university
  - To buy a car or a house
  - Choose common stock from the stock markets
- Numerical information is everywhere
- No matter what your career, you will make professional decisions that involve data
- Joy of Statistics https://www.bilibili.com/video/BV1TJ411y7Mp?from=search&seid=19044661072730491 31

Statistics

- **Statistics is the art of learning from data**

- Statistics is concerned with
  - the collection of data
  - their description or sum
  - their analysis, which often leads to the drawing of conclusions (interpretation)

Example

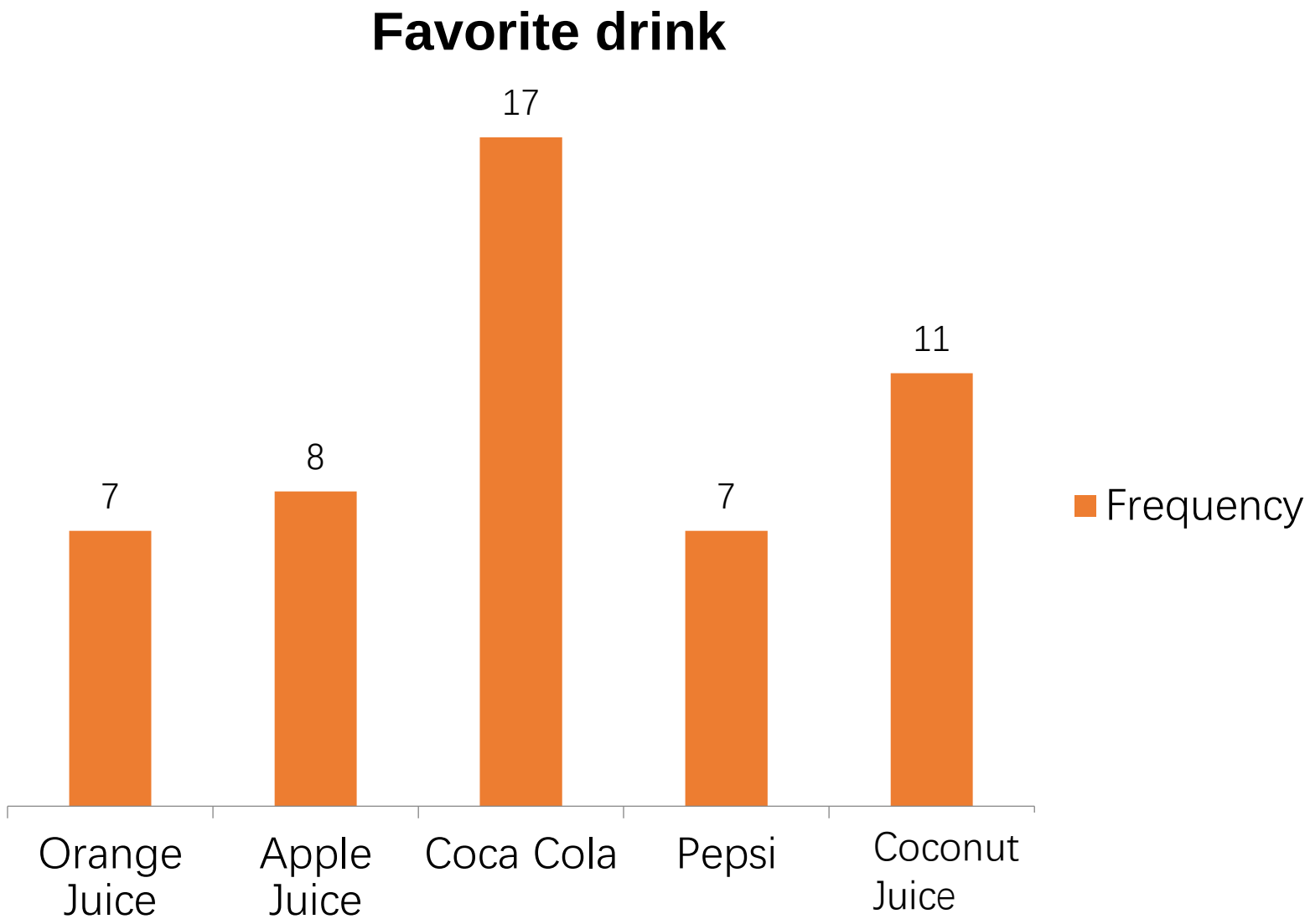
A survey is conducted among 50 people, asking them: “Which one of the following is your favorite, Orange juice(O), Apple juice(A), Coca Cola(C), Pepsi(P) or Coconut Juice (L) ? ”

Data collected are

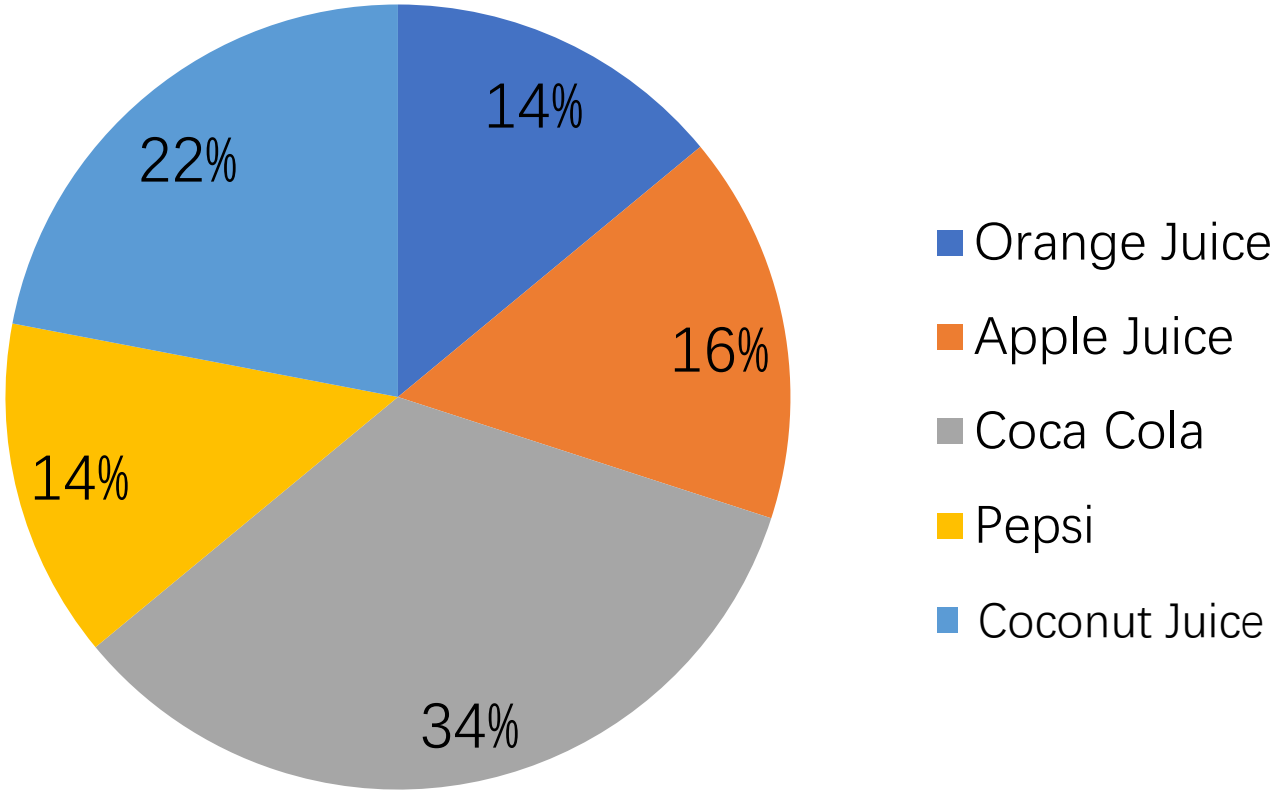
O, A, C, A, C, P, C, A, C, C, C, A, L, O, C, P, A, O, L, P, L, O, C, L, P, O, O, C, A, L, C, L, L, C, C, L, C, L, P, O, C, A, P, A, C, C, L, L, C, P

Question: Which one is the most popular?

Table: Frequency and percentage of favorite drink		
Drink	Frequency	Percentage
Orange Juice	7	14
Apple Juice	8	16
Coca Cola	17	34
Pepsi	7	14
Coconut Juice	11	22
Total	50	100


Favorite drink
17
11
8
7
7
Frequency
Orange Juice
Apple Juice
Coca Cola
Pepsi
Coconut Juice

Favorite drink



Example: Fisher's Tea Taster

- When
  Experin
  whethe
  Fisher
  tea.

- Four cu
  first. Sl
  should
  were p

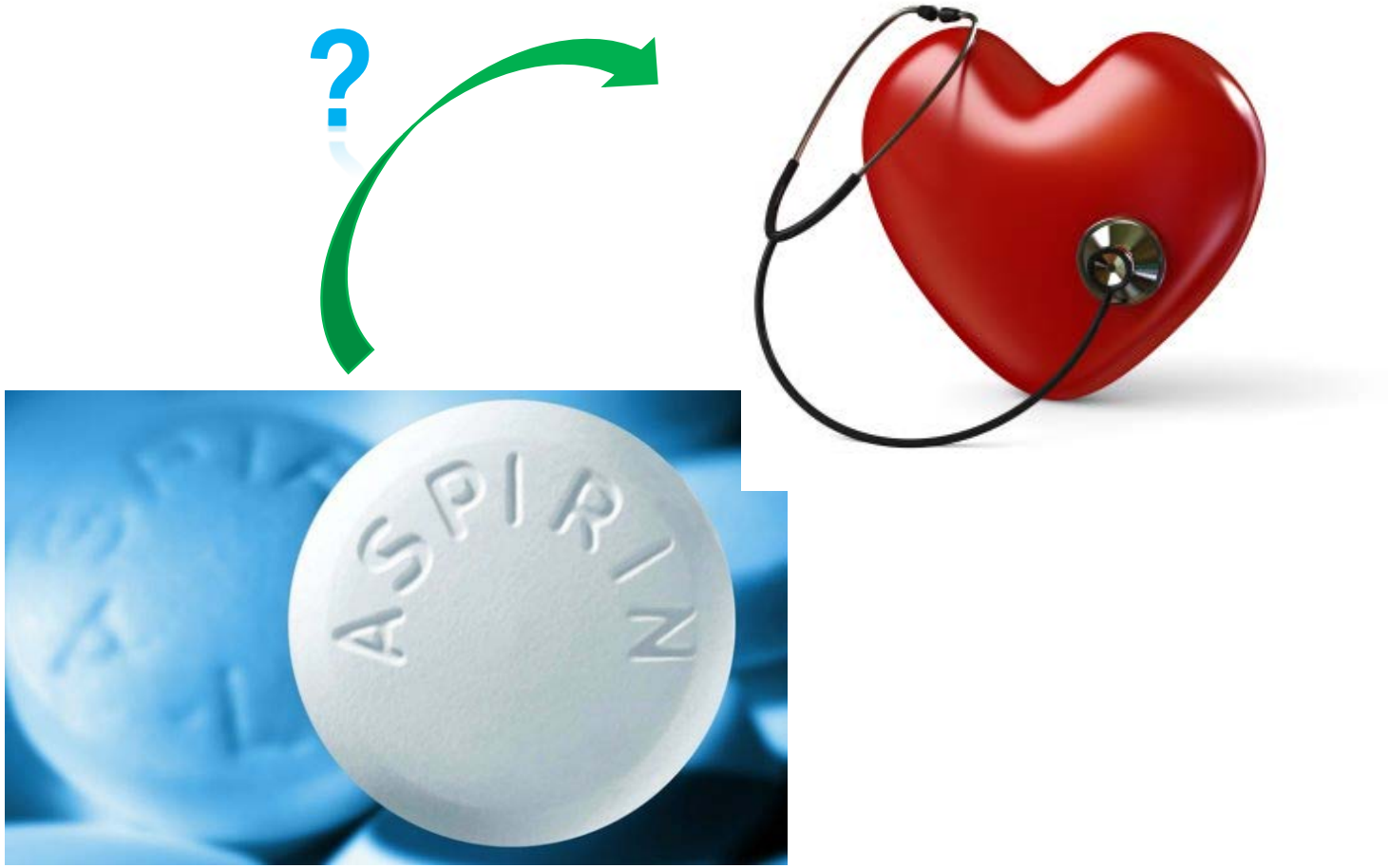
hamsted
stinguish
er claim,
cups of

a added
and she
The cups

Table 2.8. Fisher's Tea Tasting Experiment

	Guess Poured First		
Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	

Example: Aspirin and Heart Attacks



Example: Aspirin and Heart Attacks

- Table 2.3 is from a report on the relationship between aspirin use and myoc…lth Study Resea…
- The …dy *testing whet…liovascular disea…*
- Every …tudy took eithe…ind"- the physicians in the study did not know which type of pill they were taking.

Table 2.3. Cross Classification of Aspirin Use and Myocardial Infarction

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Source: Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. *New Engl. J. Med.*, **318**: 262–264, 1988.

Example: Female Horseshoe Crabs and their Satellites

- A study of r *thology*,**102**: 1–21, 1996). had a male crab attache

- The study in female crab had any oth by her. The response ou of satellites.


因为手创 与你相遇
手绘的乐趣

C	S	W	Wt	Sa	C	S	W	Wt	Sa
2	3	28.3	3.05	8	3	3	22.5	1.55	0
3	3	26.0	2.60	4	2	3	23.8	2.10	0
3	3	25.6	2.15	0	3	3	24.3	2.15	0
4	2	21.0	1.85	0	2	1	26.0	2.30	14
2	3	29.0	3.00	1	4	3	24.7	2.20	0
1	2	25.0	2.30	3	2	1	22.5	1.60	1
4	3	26.2	1.30	0	2	3	28.7	3.15	3
2	3	24.9	2.10	0	1	1	29.3	3.20	4
2	1	25.7	2.00	8	2	1	26.7	2.70	5
2	3	27.5	3.15	6	4	3	23.4	1.90	0
1	1	26.1	2.80	5	1	1	27.7	2.50	6
3	3	28.9	2.80	4	2	3	28.2	2.60	6
2	1	30.3	3.60	3	4	3	24.7	2.10	5
2	3	22.9	1.60	4	2	1	25.7	2.00	5
3	3	26.2	2.30	3	2	1	27.8	2.75	0
3	3	24.5	2.05	5	3	1	27.0	2.45	3
2	3	30.0	3.05	8	2	3	29.0	3.20	10
2	3	26.2	2.40	3	3	3	25.6	2.80	7

C: color S: spine condition W: width Wt: weight Sa: number of satellites

Extensions

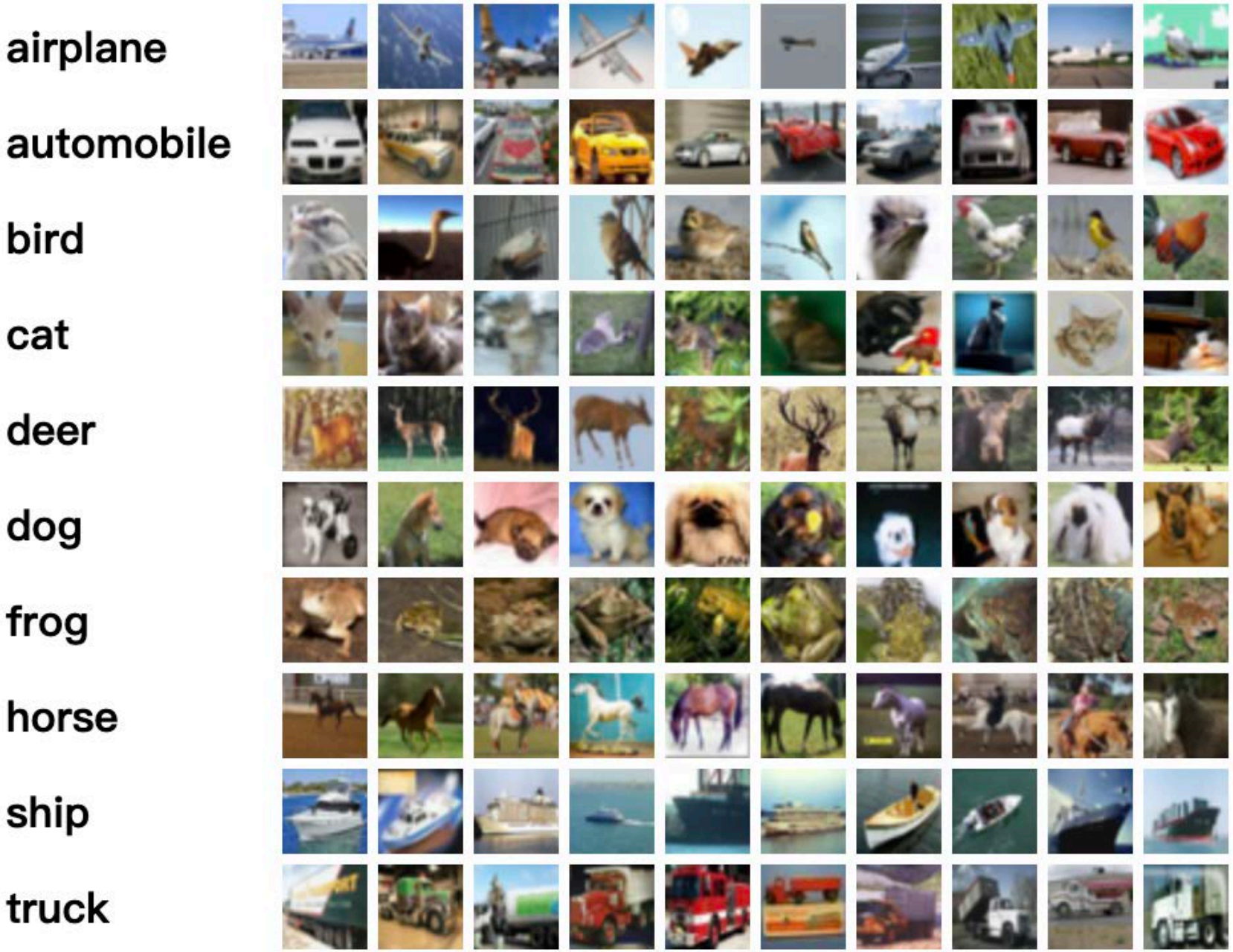
- **Big Data**
- **Machine Learning**
- **Deep learning**
- **Supervised**
- **Unsupervised**
- **Anomaly (Outlier) Detection**

- **MNIST**

The MNIST database (Modified National Institute of Standards and Technology database) is a large database of handwritten digits that is commonly used for training various image processing systems. It was created by "re-mixing" the samples from NIST's original datasets. The MNIST database contains 60,000 training images and 10,000 testing images. ----Wikipedia

- **CIFA-10**

The CIFAR-10 dataset (Canadian Institute For Advanced Research) is a collection of images that are commonly used to train machine learning and computer vision algorithms. It is one of the most widely used datasets for machine learning research. The CIFAR-10 dataset contains 60,000 32x32 color images in 10 different classes. The 10 different classes represent airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. There are 6,000 images of each class. ----Wikipedia


airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

Think

- What will we learn in this course?
- How to use them in real data analysis?
- What else can we do? (innovation)

Thank you!