Probability and Statistics

Peng HUA 华鹏

Professor, School of Science

huapeng@hit.edu.cn

办公室 Office: G620

Prof. Zhenghui FENG contributed to the lecture notes

Course Chat Group

群聊：Probability & Statistics2025-Hua

All course materials and information will be distributed in the chat grou

The two teaching assistant are already in the chat gro

该二维码7天内(9月14日前)有效，重新进入将更新

Introduction of the instructor

Peng Hua 华鹏 (call me Teacher Hua, 华老师, Prof. Hua, Dr. Hua, or Peng)

B.Eng., Mechanical Engineering, Wuhan University, China (2009-2013)

M.Sc., Mechanical Engineering, Imperial College London, UK (2014-2015)

Ph.D., Mechanical Engineering, The Hong Kong University of Science and Technology (HKUST), Hong Kong (2016-2019)

Postdoc, HKUST, 2019-2022

Research Assistant Professor, HKUST, 2022-2023

Professor, Harbin Institute of Technology, Shenzhen, 2023-present

Homepage: https://faculty.hitsz.edu.cn/huapeng

IELTS: 7; **TOEFL**: 104;

Publication of the instructor

Publication list

1. Peng Hua, Minglu Xia, Yusuke Onuki, Qingping Sun*, “Nanocomposite NiTi shape memory alloy with high strength and fatigue resistance”, Nature Nanotechnology, 16, (2021) 409–413.
2. Peng Hua, Kangjie Chu, Fuzeng Ren, Qingping Sun*, “Cyclic phase transformation behavior of nanocrystalline NiTi at microscale”, Acta Materialia, 185 (2020) 507–517.
3. Peng Hua, Kangjie Chu, Qingping Sun*, “Grain refinement and amorphization in nanocrystalline NiTi micropillars under uniaxial compression”, Scripta Materialia, 154 (2018) 123–126.
4. Peng Hua, Hongyang Lin, Qingping Sun*, “Ultrahigh cycle fatigue deformation of polycrystalline NiTi micropillars”, Scripta Materialia, 203 (2021) 114108.
5. Peng Hua, Bing Wang, Chao Yu, Yilong Han, Qingping Sun*, “Shear-induced amorphization in nanocrystalline NiTi micropillars under large plastic deformation”, Acta Materialia, 241 (2022) 118358.
6. Xueshi Li, Peng Hua*, Qingping Sun*, “Continuous and efficient elastocaloric air cooling by coil-bending”, Nature Communications, 14 (2023) 7982.
7. Hongyang Lin, Peng Hua*, Qingping Sun*, “Effects of grain size and partial amorphization on elastocaloric cooling performance of nanostructured NiTi”, Scripta Materialia, 209 (2022) 114371.
8. Dingshan Liang, Qiuhong Wang, Kangjie Chu, Junyu Chen, Peng Hua*, Fuzeng Ren*, Qingping Sun*, “Ultrahigh cycle fatigue of nanocrystalline NiTi tubes for elastocaloric cooling”, Applied Materials Today, 26 (2022) 101377.
9. Hongyang Lin, Peng Hua*, Kai Huang, Qiao Li, Qingping Sun*, “Grain boundary and dislocation strengthening of nanocrystalline NiTi for stable elastocaloric cooling”, Scripta Materialia. 226 (2023) 115227.
10. Pengbo Wei[1], Peng Hua[1], Minglu Xia, Kai Yan, Hongyang Lin, Shenghui Yi, Jian Lu*, Fuzeng Ren*, Qingping Sun*, “Bending fatigue life enhancement of NiTi alloy by pre-strain warm surface mechanical attrition treatment”, Acta Materialia, 240 (2022) 118269.
11. Cheng, S.; Xiao, Y.; Li, X.; Lin, H.; Hua, P.; Sheng, L., Buckling prevention of a single long NiTi tube compressive elastocaloric regenerator, International Journal of Solids and Structures, (2023) 112263.
12. Mostafa KARAMI; Zeyuan Zhu; Ka Hung Chan; Hua, Peng; Nobumichi Tamura; Chen, Xian, Nondissipative Martensitic Phase Transformation after Multimillion Superelastic Cycles, Phys. Rev. Lett. , 132（2024） 066101.
13. Guoan Zhou; Zexi, Li; Qiuhong Wang; Yuxiang Zhu; Hua, Peng; Yao, Shuhuai; Sun, Qingping, A multi-material cascade elastocaloric cooling device for large temperature lift, Nature Energy, (2024)
14. Guoan Zhou, Lingyun Zhang, Zexi Li, Peng Hua, Qingping Sun, Shuhuai Yao. Achieving kilowatt-scale elastocaloric cooling by a multi-cell architecture. Nature 639, 87–92 (2025).

What you need to complete this course

1. **Ears and attention: all important information about how to complete this course will be elaborated in classes or in the chat group. Don't miss them.**
2. **Hands and notes: Print the lecture slides in advance. Bring some papers or notebooks. Take notes when you hear important information and you have questions.**
3. **Time and assignment: Finish all assignment in time. All contents in the final exam come from the assignment.**

Course Objectives

- Provide you with a comprehensive understanding of statistical concepts and their applications
- Develop your quantitative reasoning skills and equip you with the tools necessary to analyze and interpret data for decision-making

What you can learn from this course?

- Basic probability and statistical concepts and methods; 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.
- Emphases on what, how, when and why certain statistical methods can and cannot be applied.

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. (Optional. All optional contents are excluded from the final exam)
6. **Correlation and Regression Analysis**: Relationship between variables, linear regression, and interpretation of regression models.

Assessment

Quizzes & Attendance

2 quizzes

2 attendances (will be checked randomly)

(10%)

Assignments

5 homework

(30%)

1 Group project

Will be assigned after mid-term

(20%)

Final examination

Contents will be revealed in the last class (don't miss)

(40%)

Assignment (30%)

- You are encouraged to discuss homework with others, but must **write your homework independently (AI is forbidden)**.
- **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.**
- **Each homework** should be **completed within 1 week**.

Group Project (20%)

- 4-5 students in one group.
- Please **give the group name list to TA**.
- Submit a report **before the final exam.**
- **Conduct the research in group, but write the report independently.**

Rubric

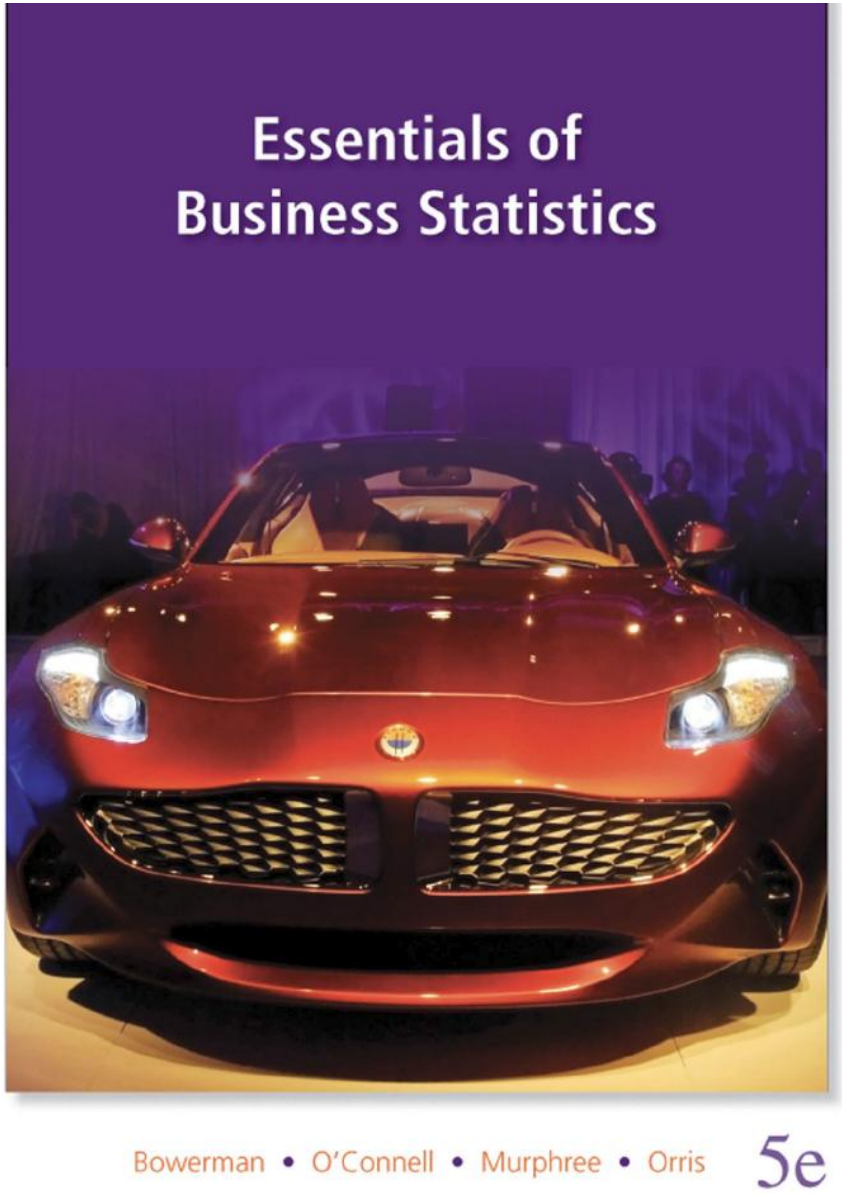
Generic Rubrics for Assessment of Project Assignment (presented in formal report format)

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

Essentials of Business Statistics

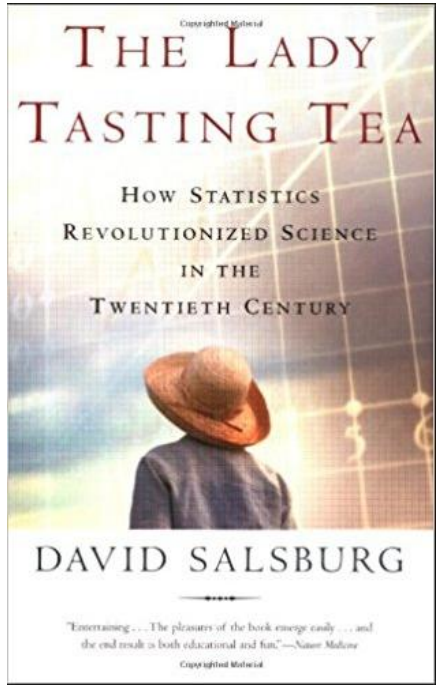
- Authors: Bruce L. Bowerman, Richard T. O'Connell, Emily S. Murphree, J.B. Orris
- Fifth Edition
- ISBN: 978-0-07-802053-7
- Publisher: McGraw-Hill Education

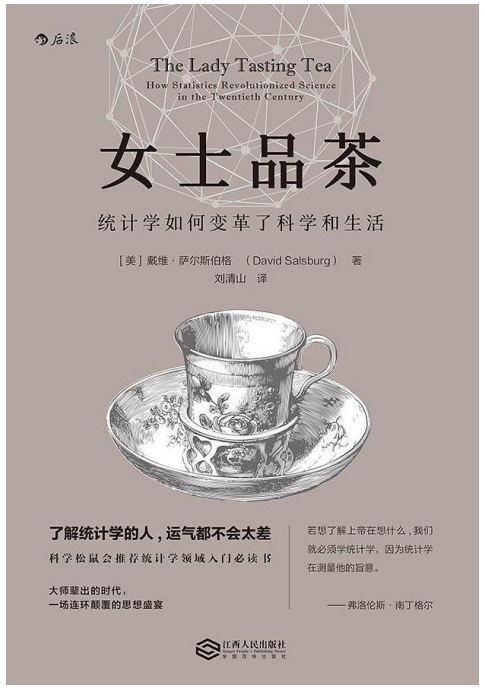


References

- 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.

Recommended book: The Lady Tasting Tea






Ronald Fisher in 1913

In the 1920s at a tea party in Cambridge, England. A lady claimed that she could taste the difference between a cup of tea into which milk had been poured and one into which milk had been added after the tea.

The scientists present were skeptical, but statistician Ronald Fisher designed an experiment to test her claim. This experiment not only marked the beginning of modern statistical methods but also serves as a gateway to the book's exploration of key statistical concepts.

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 are **data**?

- **Data** are numerical facts and figures from which conclusions can be drawn. Such conclusions are important in the decision-making processes of many professions and organizations.
- Examples:
  - How many sisters/brothers do you have? 0, 1, 2, 3 ...
  - 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) ? ”

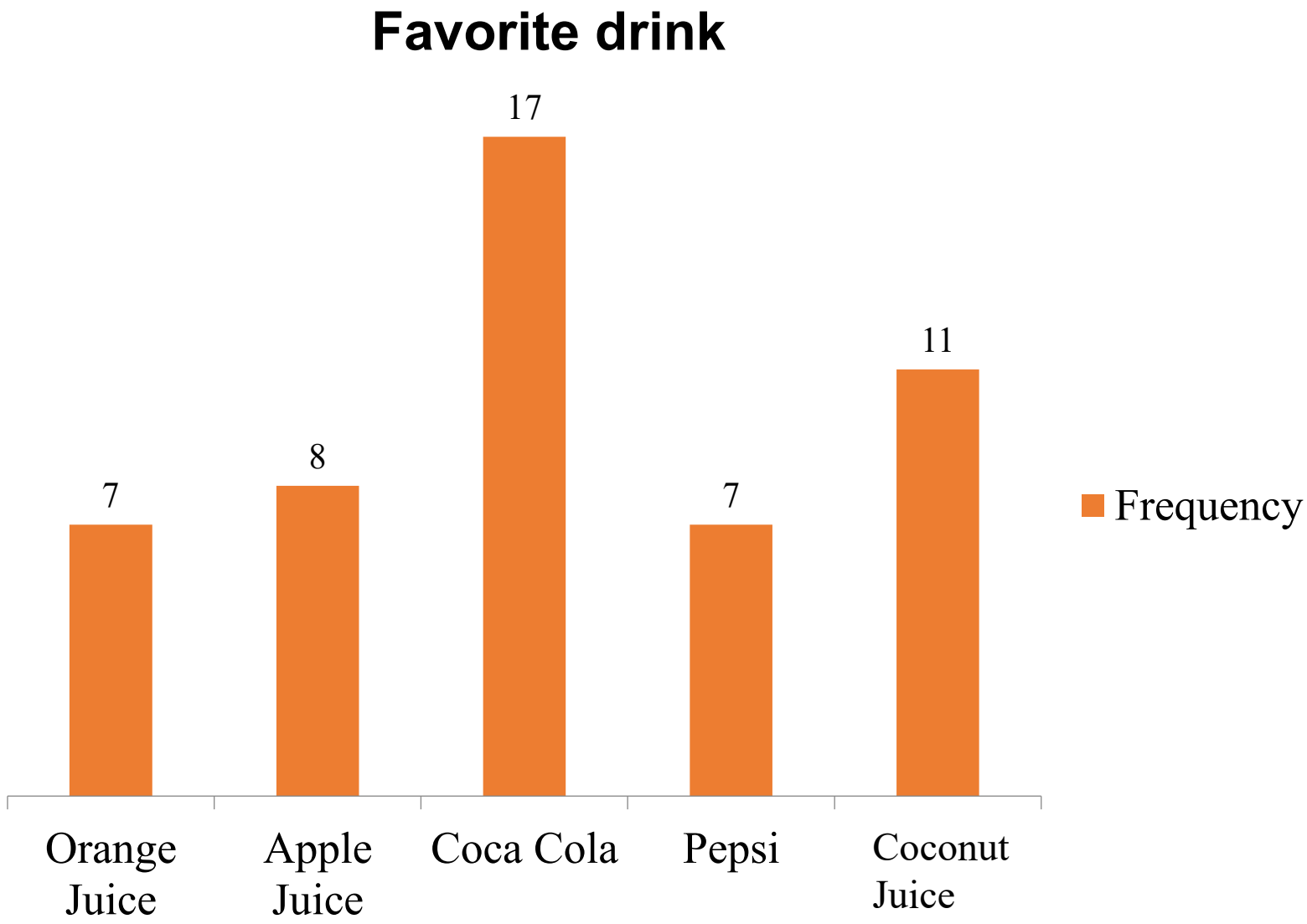
Data collected are

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

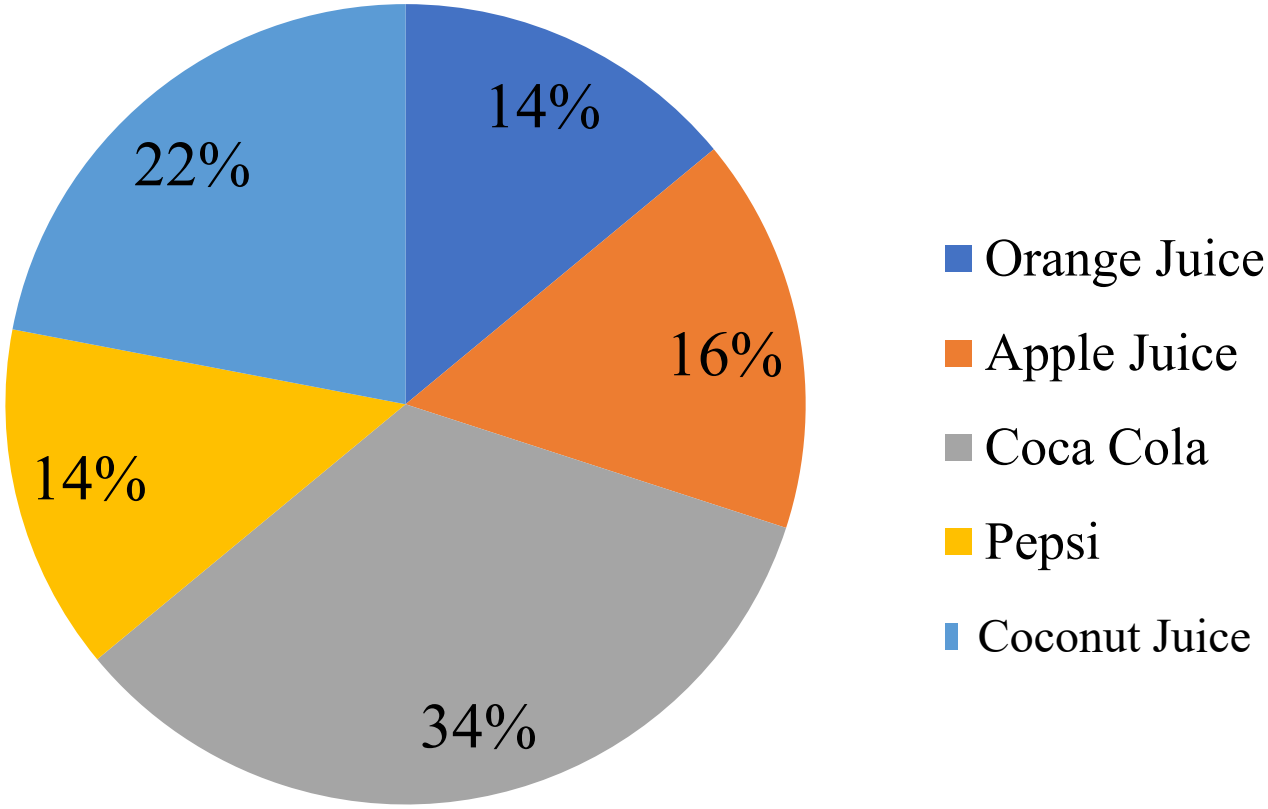
Question: Which one is the most popular?

Example

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 drinking tea, a colleague of Fisher's at Rothamsted Experiment Station near London claimed *she could distinguish whether milk or tea was added to the cup first*. To test her claim, Fisher designed an experiment in which she tasted eight cups of tea.

- Four cups had milk added first, and the other four had tea added first. She was told there were four cups of each type and she should try to select the four that had milk added first. The cups were presented to her in random order.

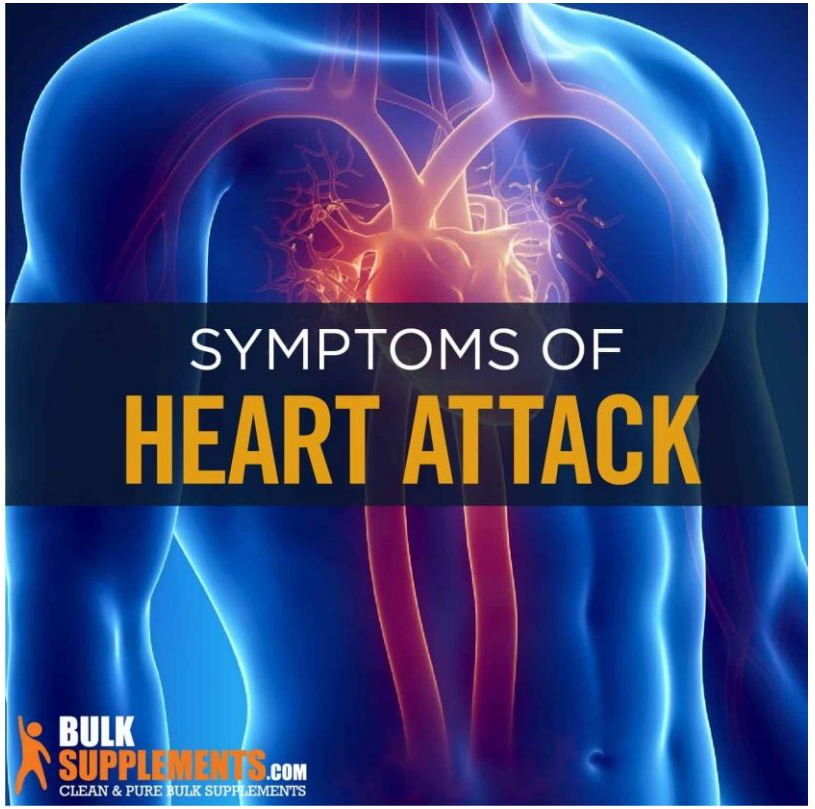
Example: Fisher's Tea Taster

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

CAN ASPIRIN PREVENT HEART ATTACKS?



Example: Aspirin and Heart Attacks

- Table 2.3 is from a report on the relationship between aspirin use and myocardial infarction (heart attacks) by the Physicians' Health Study Research Group at Harvard Medical School.
- The Physicians' Health Study was a five-year randomized study *testing whether regular intake of aspirin reduces mortality from cardiovascular disease.*
- Every other day, the male physicians participating in the study took either one aspirin tablet or a placebo. The study was "blind"- the physicians in the study did not know which type of pill they were taking.

Example: Aspirin and Heart Attacks

- Table 2.3 A report on the relationship between aspirin use and myocardial infarction (heart attacks) by the Physicians' Health Study Research Group at Harvard Medical School.

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.

1.71%

0.94%

Example: Female Horseshoe Crabs and their Satellites

- A study of nesting horseshoe crabs (J. Brockmann, *Ethology*,**102**: 1–21, 1996). Each female horseshoe crab in the study had a male crab attached to her in her nest.
- The study investigated factors that affect whether the female crab had any other males, called *satellites*, residing nearby her. The response outcome for each female crab is her number of satellites.

Example: Female Horseshoe Crabs and their Satellites

Example: Female Horseshoe Crabs and their Satellites

Fig. 1: Two pairs nest side by side on the beach. The female is in front, the attached male behind. The pair on the right has four satellite males around them whereas the pair on the left has none. (Picture drawn by Daryl HARRISON from a video)

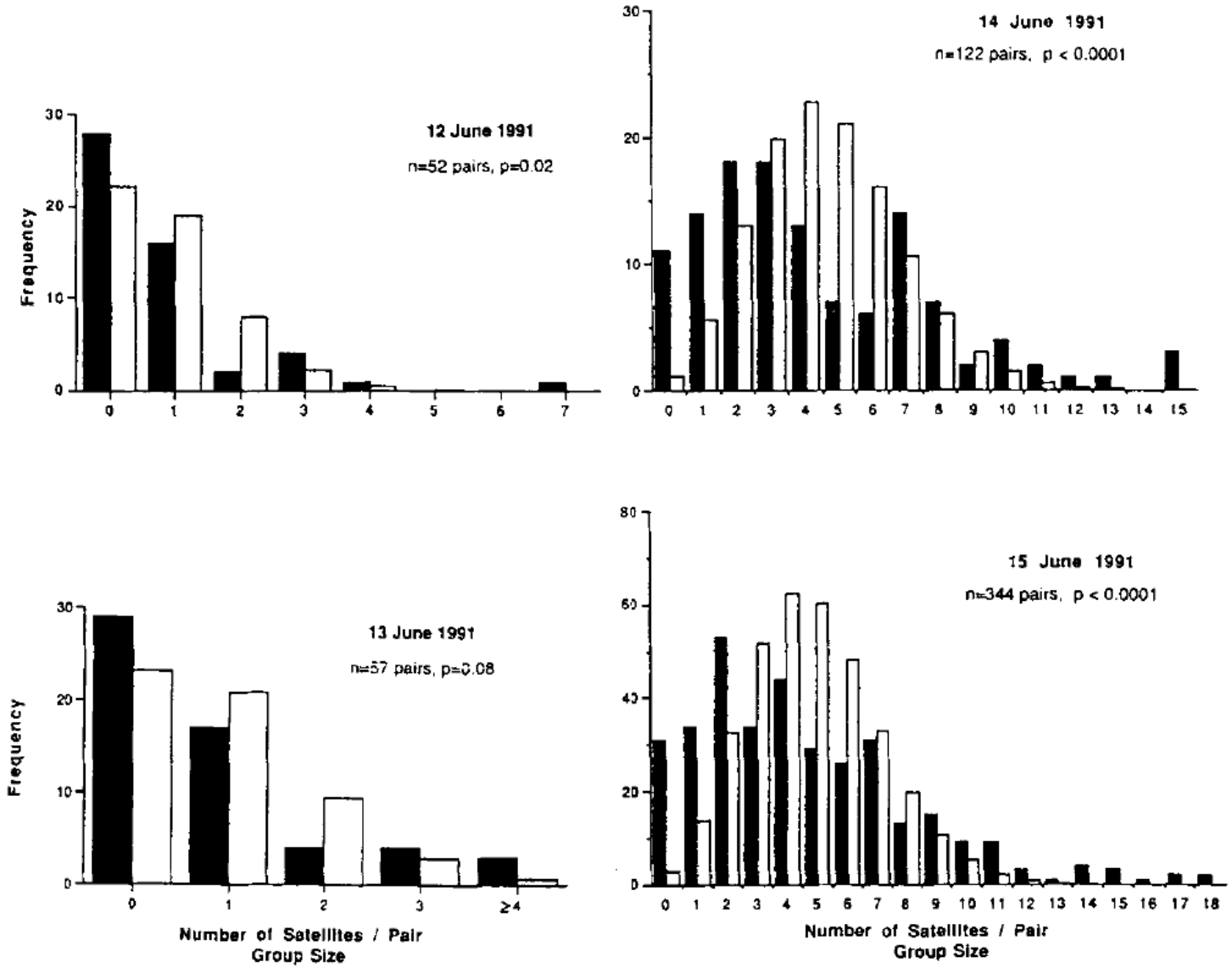
Example: Female Horseshoe Crabs and their Satellites

Example: Female Horseshoe Crabs and their Satellites

Table 1: Distribution of satellite males with nesting couples. Unless indicated, the census covers 1 km of beach. One census was taken per night from the h of the maximum high tide to 1 h after. Using the Bonferroni procedure (McCLAVE & DIETRICH 1988), a p value of 0.002 is needed for significance

	No. of satellites per pair (group size)[1]																				Total no.				
Date	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	≥19	Pairs	Sats	$\bar{X}$	χ^2 (df)	p
June 1991																									
7	16	25	25	11	13	10	10	8	5	2	1	1	3	0	1	0	0	1	0	1	133	491	3.7	99 (8)	**
8a	52	29	23	8	4	4	4	1													125	166	1.3	28 (4)	**
8b	12	18	21	11	6	8	8	3	4	1	2	4	1	0	1	0	0	2	1	1	104	428	4.1	132 (8)	**
9	47	25	17	14	6	3	3	7	2	2	0	0	0	0	0	1	0	0	1	1[2]	129	295	2.3	156 (6)	**
10	27	2	0	2	0	0															31	8	0.3	4.8 (2)	0.09
11	95	20	5	12	5	3	2	1	0	0	1										144	130	0.9	114 (4)	**
12	28	16	2	4	1	0	0	1													52	43	0.8	9.8 (3)	0.02
13	29	17	4	4	2	0	0	1													57	52	0.9	6.9 (3)	0.08
14	11	14	18	18	13	7	6	14	7	2	4	2	1	1	0	1	1	0	1	1[2]	122	556	4.6	128 (9)	**
15	31	34	53	34	44	29	26	31	13	15	9	9	3	1	4	3	1	2	0	2[3]	344	1612	4.7	405 (11)	**
May and June 1992																									
31	46	20	8	10	5	4	3	5	2	2	2										197	213	2.0	120 (5)	**
1	35	12	7	0	1																55	30	0.6	2.8 (2)	0.2
2[4]	86	51	29	27	21	15	6	3	1	2											241	432	1.9	131 (6)	**
3	54	61	38	40	18	10	6	2	0	0	1										230	439	1.9	34 (6)	**
7	45	38	22	16	4	5	3	0	1												134	197	1.5	30 (5)	**
11	7	5	4	2	0	0	1														19	25	1.3	1.1 (3)	0.8
12	28	18	10	13	7	4	1	1	1	0	0	0	0	1							84	159	1.9	32 (5)	**
14	59	19	12	8	10	3	3	1	3	3	3										124	228	1.8	131 (5)	**

Example: Female Horseshoe Crabs and their Satellites



Extensions

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

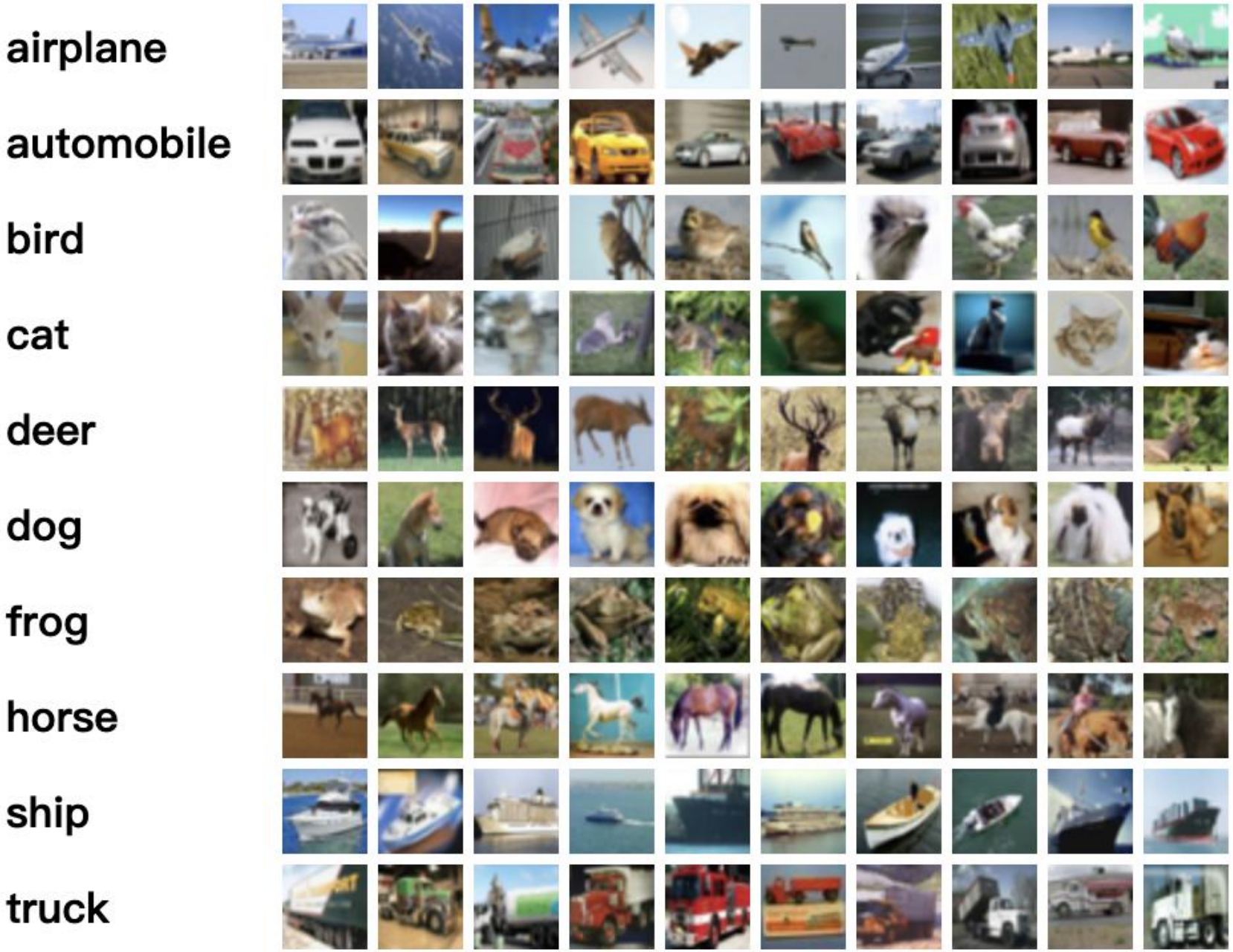
- **MNIST**

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

MNIST images of numbers

- **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 is the future of statistics? Will AI replace data analysts?

  What can we do to make our work more valuable?

Thank you!