

Medical Statistics and Data Analysis

Lecture 1: Introduction and Overview

Zhiyu Wan, PhD (wanzhy@shanghaitech.edu.cn)

Assistant Professor of Biomedical Engineering

ShanghaiTech University

September 14, 2026

Welcome to Medical Statistics and Data Analysis

- You're sitting in BME1323
- Where & When:
 - SIST 1A106, 15:00-16:40 on Mondays
 - SIST 1D108, 10:15-11:55 on Wednesdays (Even Week)
- Office Hours: Upon Request
- Teaching Assistant:
 - Feiyang Zhou (zhoufy2025@shanghaitech.edu.cn)
- Instructor: Zhiyu Wan (wanzhy@shanghaitech.edu.cn)

Goals of This Lecture

- Know more about this course
 - Instructor
 - Objectives
 - Grading Policies
 - Schedule
- Overview of the concepts
 - Statistics
 - Medical Statistics
 - Medical Data Analysis
 - Variables
 - Errors



<https://www.magnific.com/vectors/ai-healthcare-analytics/2>

- Hi! I'm Zhiyu Wan.
 - BEng, **Automation**. XJTU (Gifted Young)
 - MS & PhD, **Computer Science**, Vanderbilt
 - Postdoc, **Biomedical Informatics**, VUMC

- Faculty in **Biomedical Engineering**

- Health Information Safety and Intelligence Research Lab

(<https://zhiyuwan.com/hisir-lab/>)

- Research Areas

- Biomedical Data Privacy
- AI in Medicine
- Trustworthy AI
- AI Ethics
- Computational Game Theory
- ...



BME1323: Medical Statistics and Data Analysis

Schedule

Projects

Public Datasets

Who / When / Where
Instructor: [Zhiyu Wan](#)
Teaching Assistant: Feiyang Zhou
Semester: Fall 2026
Location & Time:
– SIST 1A106, 15:00-16:40 on Mondays
– SIST 1D108, 10:15-11:55 on Wednesdays (Even Week)
Office Hours: Upon request, Location: BME Building, Room 228

Course Syllabus ([PDF](#))

First Day of Class: September 14, 2026

Description

Medical Statistics and Data Analysis is a professional elective course for undergraduate students majoring in Biomedical Engineering. The course focuses on applying the principles and methods of mathematical statistics, together with statistical tools, to the analysis of biomedical big data in the field of biomedical engineering, with an emphasis on practical applications.

Through 48 hours of instruction, the course aims to enable students to: (1) understand and master fundamental statistical concepts, theories, and mathematical foundations; (2) understand the origins, development, research scope, and roles of medical statistics; (3) accurately collect, organize, analyze, and extract information and data according to specific research objectives; (4) master important methods in medical statistics and conduct appropriate statistical design and research; and (5) use R and Python programming languages to perform basic statistical analyses.

By mastering Python-based methods for processing and analyzing medical big data and gaining hands-on experience through project-based practice, students will establish a solid foundation for future research and engineering applications in the field of medical artificial intelligence.

BME1323: Medical Statistics and Data Analysis (Schedule)

Syllabus

Who / When / Where
Instructor: [Zhiyu Wan](#)
Teaching Assistant: Feiyang Zhou
Semester: Fall 2026
Location & Time:
– SIST 1A106, 15:00-16:40 on Mondays
– SIST 1D108, 10:15-11:55 on Wednesdays (Even Week)
Office Hours: Upon request, Location: BME Building, Room 228

Schedule

(Note: This schedule is tentative and subject to change.)

Week	Chapter	Teaching Contents	Readings	Assignments
Week 1, Monday Sept 14, 15:00-16:40 Lecture 1 (PDF)	I. Course Introduction (a)	Course introduction. Introduction to medical statistics. Concepts of statistics.	(Optional: 1. 《Fundamentals of Biostatistics》 Ch.1. 2. 《医学统计学（高教版）》 Ch.1. 3. 《医学统计学（人卫版）》 Ch.1.)	/
Week 2, Monday Sept 21, 15:00-16:40 Lecture 2 (PDF)	I. Course Introduction (b)	Statistical experimental design approaches. Descriptive statistics.	(Optional: 1. 《Fundamentals of Biostatistics》 Ch.2. 2. 《医学统计学（高教版）》 Ch.2. 3. 《医学统计学（人卫版）》 Ch.2.)	/

More to come (projects, homeworks, etc.) – links will be available from this schedule page

Overview of the Course

- **Medical Statistics and Data Analysis** is a professional elective course for undergraduate students in **Biomedical Engineering**, focusing on the application of statistical principles, methods, and computational tools to **biomedical data analysis** and **medical AI-related problems**.
- **What You Will Learn**
 - Fundamental concepts, theories, and mathematical foundations of statistics
 - Development, scope, and applications of medical statistics
 - Data collection, organization, analysis, interpretation, and information extraction
 - Statistical study design and commonly used medical statistical methods
 - Statistical programming and data analysis using **R and Python**



Overview of the Course (Cont.)

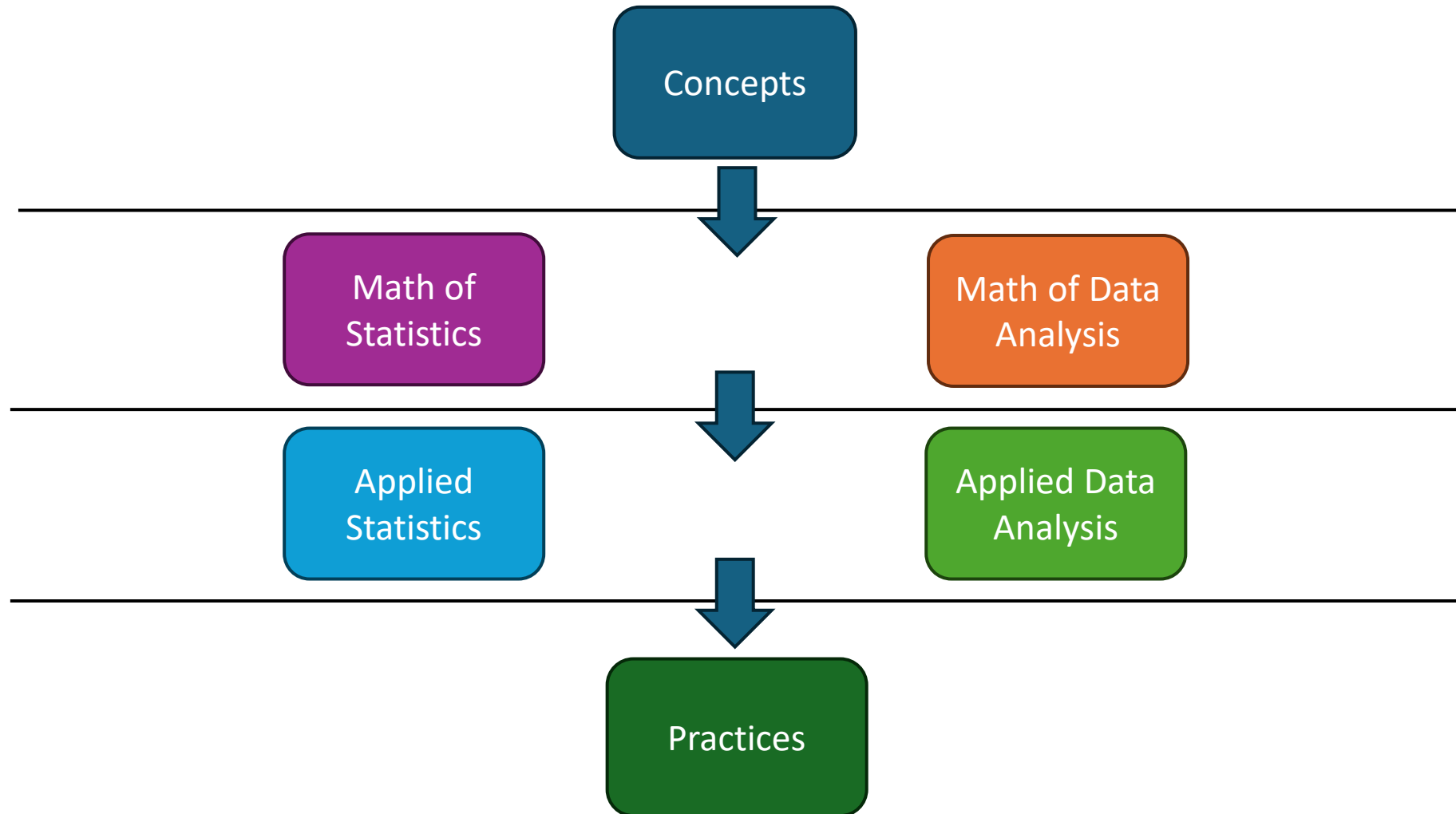


- **How You Will Learn**
 - **48 hours of instruction** combining lectures, case-based learning, programming practice, and project-based exercises with real-world biomedical data.
- **What You Will Be Able to Do**
 - **Design studies → Analyze biomedical data → Interpret results → Communicate findings → Solve practical biomedical problems**
- **Course Perspective**
 - Build a solid statistical and computational foundation for future research and engineering applications in **biomedical big data, data science, and medical artificial intelligence**.

Course Objectives

- After this course, you should be able to apply medical statistics to biomedical research from three complementary perspectives:
 - Statistical Thinkers: Understand fundamental statistical concepts and methods, including statistical inference, ANOVA, regression, classification, clustering, dimensionality reduction, time-series analysis, and data visualization.
 - Data Analysts: Design appropriate studies; collect, organize, analyze, interpret, and visualize biomedical data using **R and Python**.
 - Biomedical Researchers: Apply statistical methods to real-world biomedical problems, integrate knowledge across **mathematics, biology, medicine, and computer science**, and communicate findings effectively in collaborative research.

Course Modules



Expectations

- You are expected to have a **working knowledge** of the Internet, word processing (WPS, Word, ...), and data processing tools (Excel, ...)
- You are preferred to be **competent** in a programming language (Python, R, ...)
- Reading & Writing in English.

Instructional Pedagogy

- **Inputs:** Lectures, group discussions, case-based learning, expert/guest lectures.
- **Learning Activities & Assessments:** In-class quizzes, homework assignments, case discussions, and course projects, and a final examination.
- **Pedagogy:** Understanding-based, heuristic, and practice-oriented teaching aligned with international standards.
- **Learning Philosophy:** Promote active learning, independent thinking, hands-on practice, and interest in scientific research.
- **Theory-to-Practice:** Integrate fundamental statistical theories and methods with real-world biomedical cases and data-analysis problems.
- **Language:** Spoken instruction primarily in **Chinese**; written materials primarily in **English**.

Overview of Grading Policies

- A substantial portion of your grade will be based on your “final” project.

Criteria	% of Total Grade
Course Project	30%
Homework Assignments	30%
Final Exam	30%
In-class quiz	5%
Attendances and Class Participation	5%
Total	100%

More details: <https://zhiyuwan.com/bme1323/#grading>

Homework Policy

- Please do your own homework.
- Do not plagiarize without proper attribution.
- You can use AI assistant. However, you need to disclose and document the version number.

Final Projects

- This course assesses students' mastery of the course content through course projects.
- Each team will consist of **1 to 2 students**.
- Teams will select a topic of interest from a list prepared by the course coordinator, conduct a literature review, identify existing solutions, summarize unresolved problems, propose their own approaches, and locate relevant datasets for experimental validation.
- Each team is required to produce data-driven experimental results, deliver an **oral presentation** and an **poster presentation**, and submit a **final project report** in the form of an English-language academic paper.
- With prior notification and approval from the course coordinator, students may also propose and investigate a project topic outside the provided list.
- Do not be afraid to discuss your project ideas with the instructor!

Course Project Schedule

Week	Course Project Planning
10	Announce course project topics, and students begin forming teams.
10	Students complete team formation and begin selecting course project topics.
12	Each team completes project topic selection and begins preparing a project proposal.
14	Each team submits a 4-page project proposal .
14	The instructor and teaching assistants conduct a mid-term evaluation of each team's project proposal and provide feedback and guidance to each team. Students then prepare their final course project based on the feedback.
16	Each team must record and submit an approximately 10-minute video of oral presentation for the course project in advance. All teams will present their posters, and the order of evaluation will not be announced in advance. Note: All teams must submit their presentation slides and recorded videos before the poster presentation begins. The course instructor and teaching assistants will evaluate and rank all team presentations and posters in a fair and impartial manner.
17	Students submit the final course project paper (an updated 10-page paper. Note: co-authors are listed without ranking, but each co-author's responsibilities and contributions must be specified in the report).
18	The instructor and teaching assistants conduct a final evaluation of the course project papers.

Grading Policies for the Course Project

- The grading breakdown for the course project is as follows: **project proposal (10%) + oral presentation (20%) + poster presentation (30%) + course project paper (40%)**.
- The grading for the **project proposal** and the **project paper** is assisted by the teaching assistant and is finally determined by the instructor.
- The scores for the **oral presentation** performance and the **poster presentation** performance are obtained through a weighted average from the instructor (**60%**) and TAs/experts (**40%**).

Grading Policies for Other Issues

- **Assignments & Quiz**

- Tests and submitted work are primarily in **English**.
- **5 homework assignments**: Q&A, calculation, programming, etc.; each carries **equal weight**.
- **No collaboration** is allowed.
- Use of **generative AI tools** (e.g., ChatGPT, DeepSeek) must be **disclosed**, including the **model/version used**.
- **1 closed-book in-class quiz** in **Week 2**, covering Week 1 material.
- Grades and feedback are normally provided within **4 weeks**.

- **Attendance & Participation**

- Attendance may be checked by **QR code** and/or random in-class questioning.
- Unexcused absence: **–1% per occurrence**, up to **–5%** of the overall grade.
- Excused absences may receive **no deduction**, as appropriate.
- Lack of active participation: **–1%** of the overall grade.
- Serious classroom disruption: up to **–5%** of the final course grade.

Grading Policies for Other Issues

- **Late Submission Policy**

- Each student receives a total **1-week grace period** across all submissions, including homework, project proposal, and final project paper.
- After the grace period: approximately **–1 point (–1% of overall grade) per additional day**.
- Any delay of **less than one day counts as one full day**.

- **Other Notes**

- Homework may include **bonus questions / bonus points**.

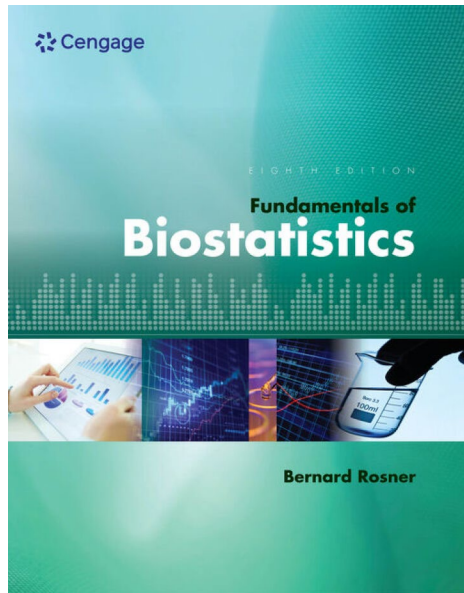
- **Final Grading Conversion**

- It is recommended that if the class size exceeds **30 students**, no more than 40% should receive an A- or above, and the failure rate should not exceed 10%.

Grade	A+	A	A-	B+	B	B-	C+	C	C-	F	EXC	P	NP	N
Grade Point	4.0	4.0	3.7	3.3	3.0	2.7	2.3	2	1.7	0	N/A	N/A	N/A	N/A
Reference Percentage-to-Grade Conversion Scale	95-100	90-94	85-89	80-84	75-79	70-74	67-69	63-66	60-62	0-59	As per the course syllabus	>=60	<60	No grade record

Textbook & Recommended Reading

Book Name: Fundamentals of Biostatistics	Authors: Bernard Rosner	ISBN: 9781305268920	Book Name: 医学统计学	Authors: 李晓松	ISBN: 9787040540468
Publisher: Cengage Learning	Publication Month: 2015.08	Version: 8	Publisher: 高等教育出版社	Publication Month: 2020.05	Version: 4



- Professor, Medicine, Harvard Medical School
- Professor, Biostatistics, Harvard T.H. Chan School Of Public Health
- Professor, Department of Medicine, Brigham And Women's Hospital



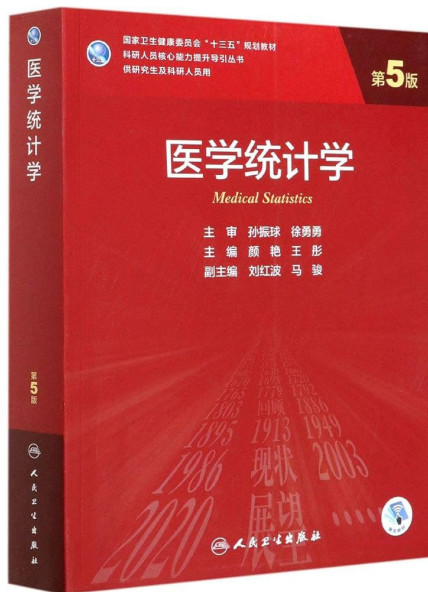
- 卫生统计学家、公共卫生教育学家，国务院政府特殊津贴获得者，四川省学术和技术带头人，四川大学二级教授、博士生导师。
- 生前任四川大学华西公共卫生学院（华西第四医院）院长。兼任国际生物统计学会中国分会副理事长、英国皇家统计学会会士，中国卫生信息与健康大数据学会卫生统计学教育专业委员会主任委员。



Textbook & Recommended Reading

Book Name: 医学统计学	Authors: 颜艳、王彤	ISBN: 9787117303859
Publisher: 人民卫生出版社	Publication Month: 2021.01	Version: 5

Book Name: 医学统计学 习题解答	Author: 颜艳、王彤	ISBN: 9787117337465
Publisher: 人民卫生出版社	Publication Month: 2022.10	Version: 5



- 颜艳，医学博士，中南大学湘雅公共卫生学院二级教授、博导，现任流行病与卫生统计学系教师。她长期从事出生队列研究、健康数据管理与评价、统计预测模型及医学统计方法应用等研究。曾赴英国伦敦大访学。
- 她同时担任中国统计教育学会常务理事、中国卫生信息与健康医疗大数据相关学术组织常务理事等职务。



- 王彤，山西医科大学教授、博士生导师，现任公共卫生学院院长、卫生统计教研室主任，公共卫生与预防医学一级学科带头人。
- 现任教育部公共卫生与预防医学类专业教学指导委员会委员、中国卫生信息与健康医疗大数据学会统计理论与方法专业委员会主任委员、国家食品与药品监督管理局新药审评专家。



Textbook & Recommended Reading

Book Name: 医学统计学实习指导	Author: 赵星、张菊英	ISBN: 9787040549386
Publisher: 高等教育出版社	Publication Month: 2020.8	Version: 4

Book Name: Python医学实用统计分析	Authors: 潘兴强、陈根浪	ISBN: 9787117341417
Publisher: 人民卫生出版社	Publication Month: 2023.01	Version: 1



- 赵星，中国卫生统计学专家，四川大学华西公共卫生学院教授、博士生导师，流行病与卫生统计学系副主任。



- 现任中国卫生信息学会统计理论与方法专业委员会副主任委员、国际生物统计学会中国分会副秘书长，担任《Biostatistics & Epidemiology》《BMC Public Health》期刊副编辑。

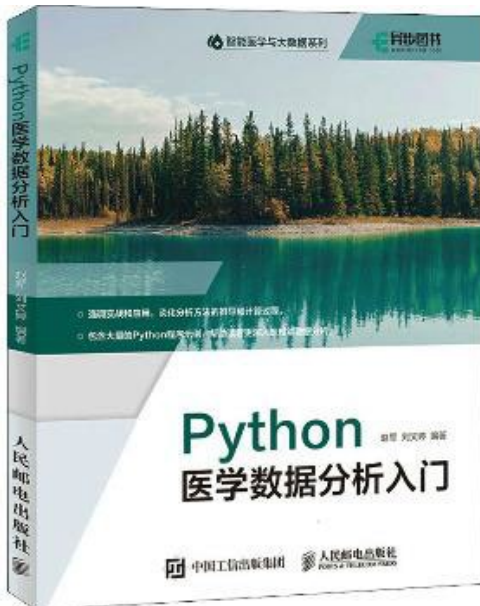


- 潘兴强，宁波市疾病预防控制中心公共卫生专家，主要从事免疫规划、疫苗效果与安全性评价、传染病流行病学及公共卫生数据分析研究；长期参与宁波市疫苗接种与传染病监测相关研究，并开展Python数据分析与机器学习等公共卫生统计方法培训。

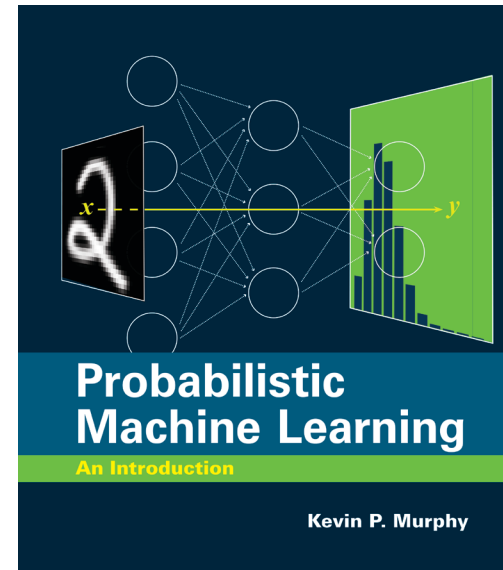
Textbook & Recommended Reading

Book Name: Python医学数据分析入门	Authors: 赵军、刘文婷	ISBN: 9787115575432
Publisher: 人民邮电出版社	Publication Month: 2022.01	Version: 1

Book Name: Probabilistic Machine Learning: An Introduction	Authors: Kevin Patrick Murphy	ISBN: 9780262046824/9780262369305
Publisher: The MIT Press	Publication Month: 2022.03	Version: 1



- 赵军，现为湖北医药学院公共卫生与健康学院副教授，研究领域主要包括流行病学、应用统计学和数据科学。他本科毕业于西安交通大学数学系，后获上海理工大学硕士学位，并于2018年获泰国宋卡王子大学医学院流行病学博士学位。
- 赵军长期从事医学统计与数据分析教学和研究，研究内容涉及临床与公共卫生数据分析、疾病风险预测、流行病学建模及机器学习应用。



<https://probml.github.io/pml-book/book1.html>



Kevin P. Murphy is a **Research Scientist at Google DeepMind** and a leading researcher in **machine learning, probabilistic modeling, and Bayesian methods**. He received his Ph.D. from **UC Berkeley** and previously served as a **professor at the University of British Columbia**.

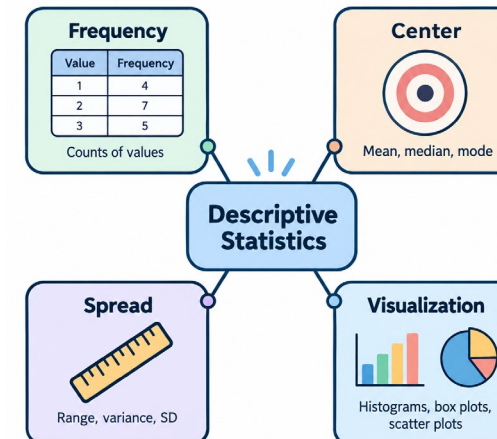
Schedule

- Let's look at the syllabus.

More details: <https://zhiyuwan.com/bme1323-schedule/>

Introduction to Medical Statistics (Sept 14, 21)

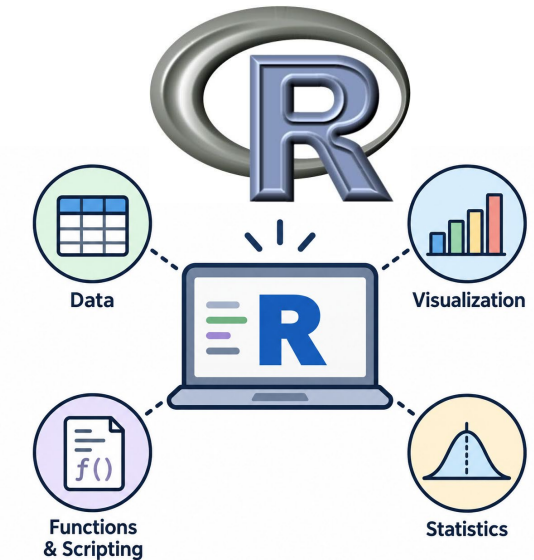
- Concepts of medical statistics:
 - Statistics
 - Medical Statistics
 - Medical Data Analysis
 - Variables
 - Errors
- Statistical experimental design approaches.
- Descriptive statistics:
 - Measures of Location
 - Measures of Spread
 - Grouped Data
 - Graphic Methods



Quiz 1

Data Visualization and Programming (Sept 23, 28)

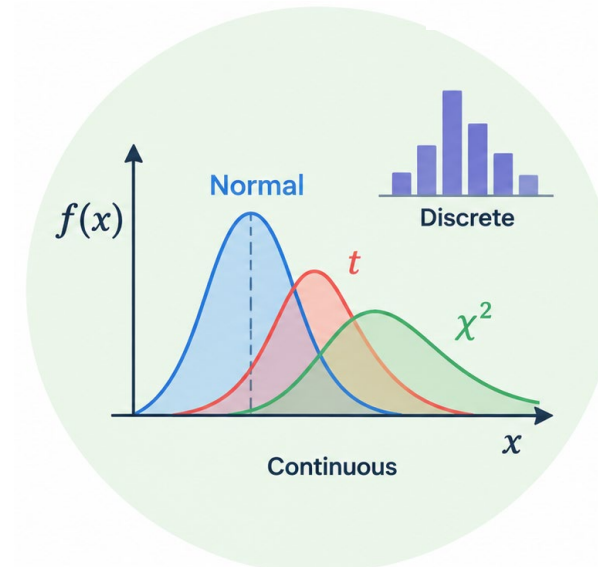
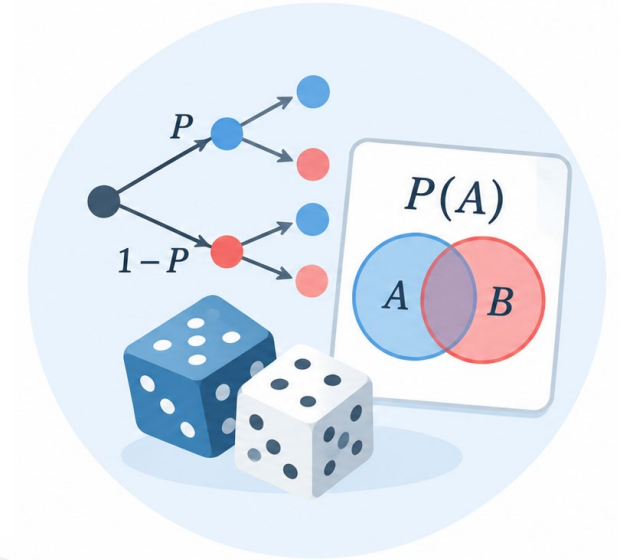
- Introduction to R.
- Introduction to Python.



HW 1

Probability and Probability Distributions (Oct 10, 12)

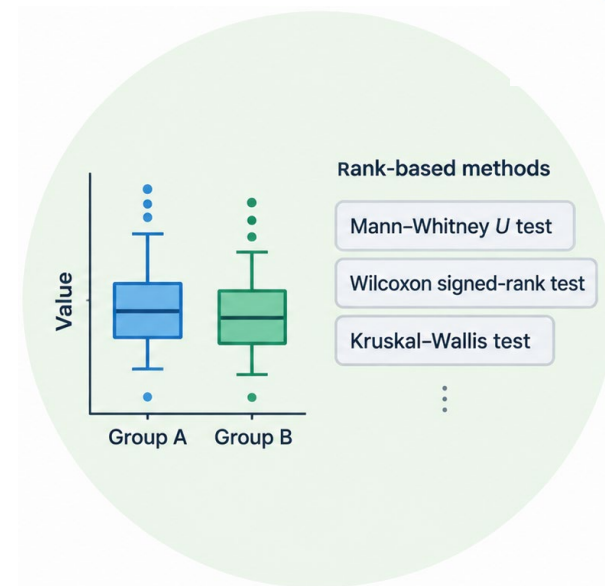
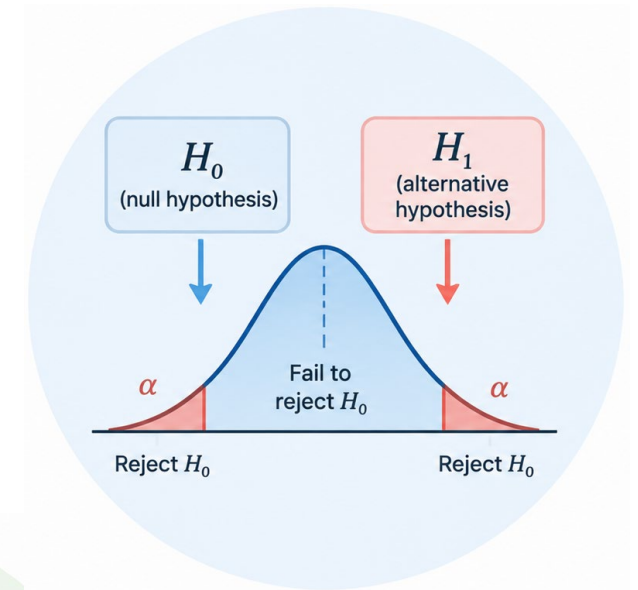
- Randomized experimental design and sampling.
- Bayesian inference.
- Receiver operating characteristic (ROC) curves.
- Discrete probability distributions.
- Continuous probability distributions.
- Point estimations.
- Confidence intervals.



HW 2

Hypothesis Testing and Nonparametric Tests (Oct 19, 21, 26)

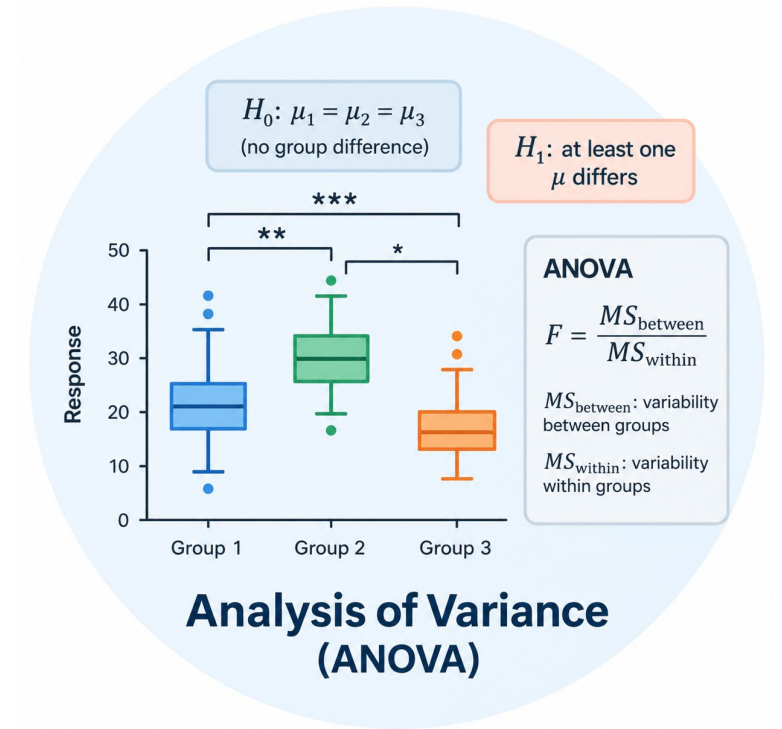
- One-Sample Inference.
- Two-sample inference.
- Statistical power and sample size estimation.
- Rank-sum tests.
- Chi-square tests.



HW 3

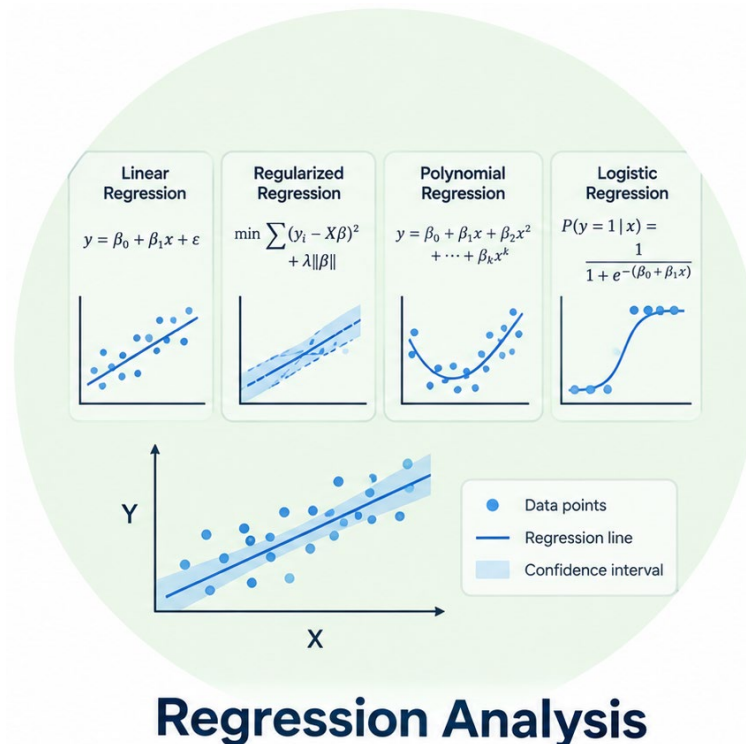
Analysis of Variance (Nov 2, 4)

- One-way analysis of variance.
- Two-way analysis of variance.
- Multiple comparisons in analysis of variance.
- 2×2 factorial design.
- Tests for homogeneity of variance.



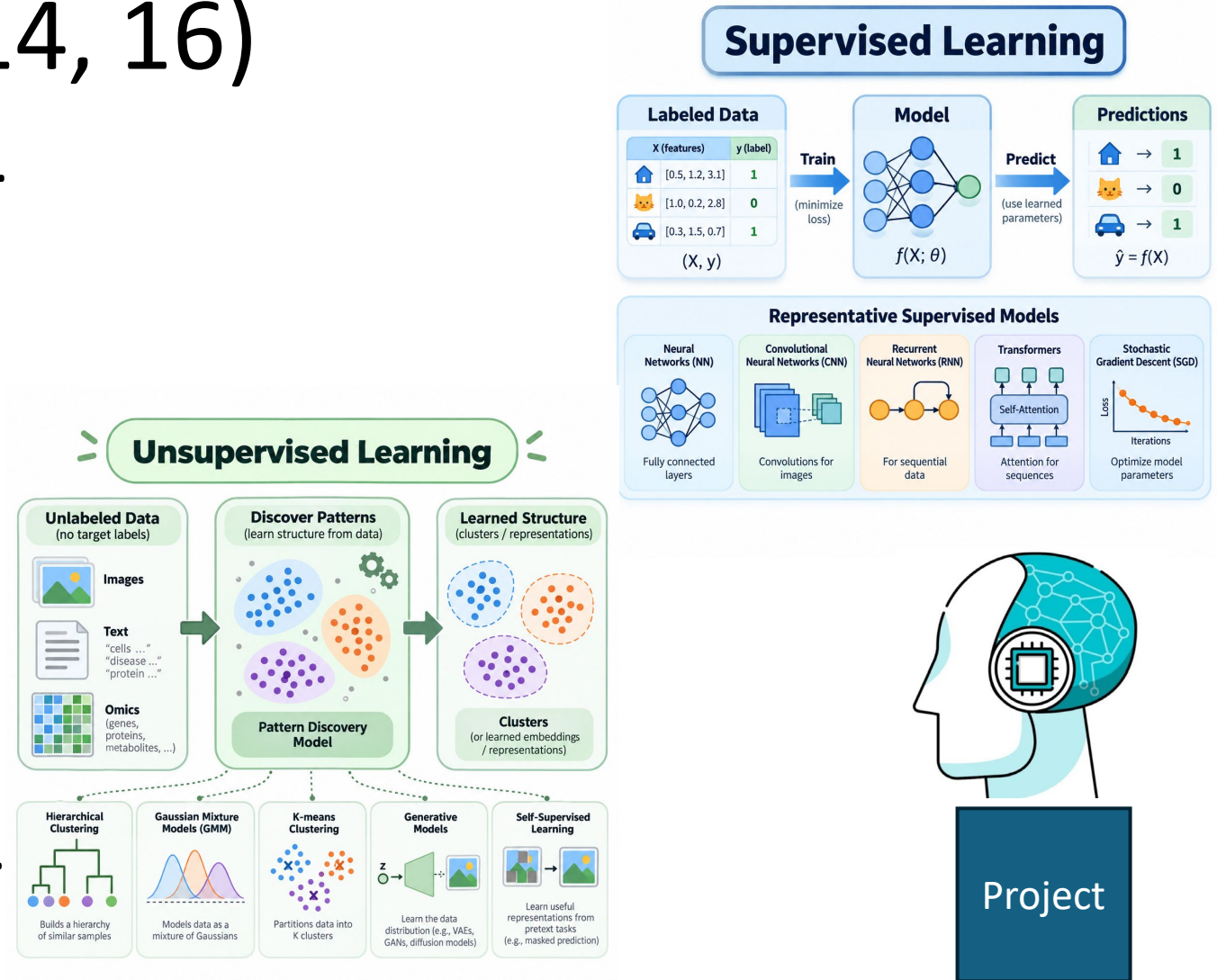
Data Classification and Regression Analysis (Nov 16, 18, 23, 30)

- K-nearest neighbors (KNN) algorithm.
- Support vector machines.
- Naive bayes.
- Decision trees.
- Random forests.
- Linear regression.
- Regularized regression.
- Polynomial regression.
- Logistic regression.



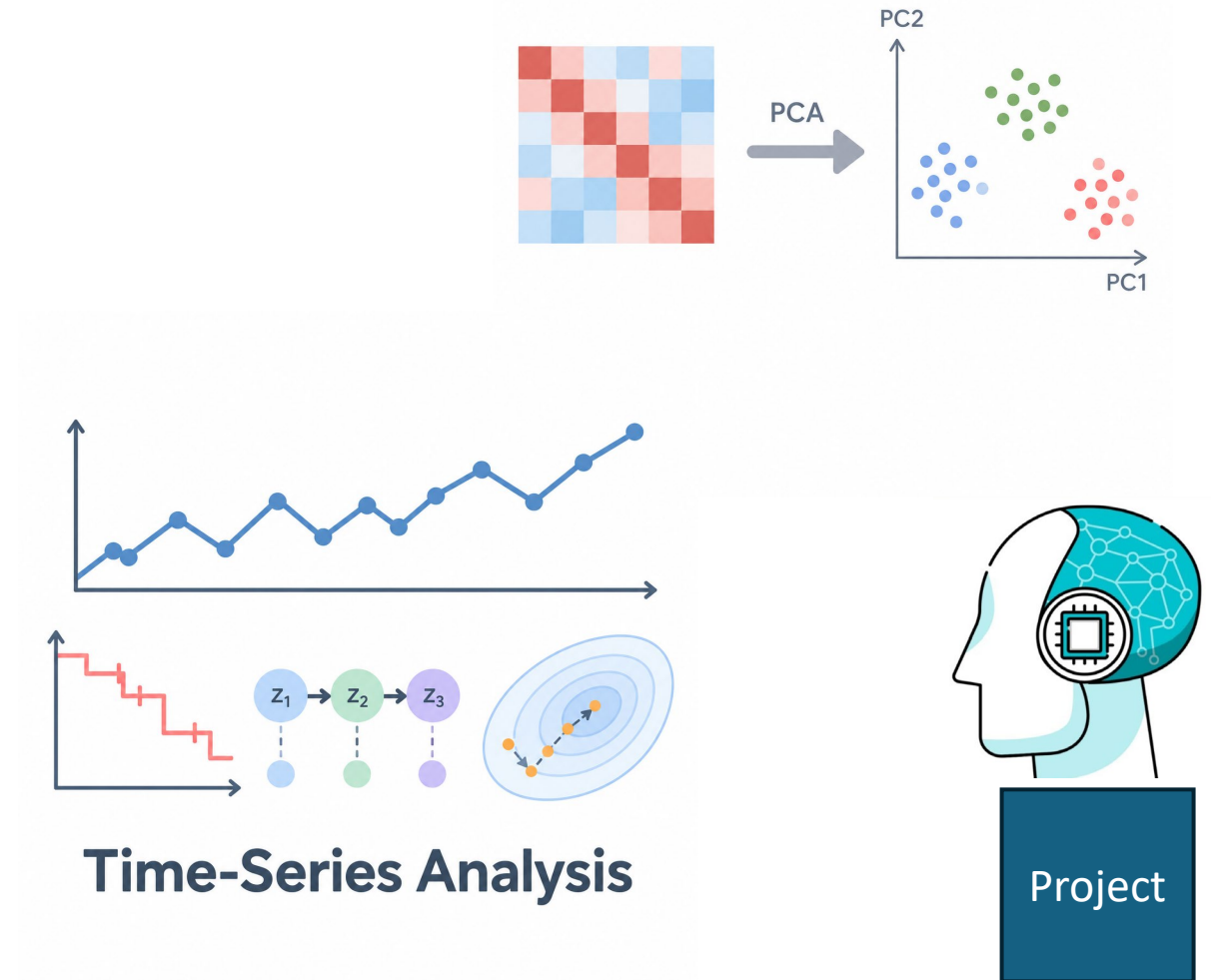
Supervised Learning and Unsupervised Learning (Dec 2, 7, 14, 16)

- Introduction to neural networks.
- Convolutional neural networks.
- Stochastic gradient descent.
- Recurrent neural networks.
- Transformers.
- Generative models.
- Hierarchical Clustering.
- Gaussian Mixture Models.
- K-means clustering.
- Self-supervised learning models.



Dimensionality Reduction and Time-Series Analysis (Dec 21, 28)

- Correlation analysis.
- Principal component analysis.
- Survival analysis.
- Hidden Markov models.
- Markov chain Monte Carlo.



Course Project Presentation (Dec 30)

- The students are in control
- You'll be graded by a committee of special reviewers



Introduction to Medical Statistics

- **Statistics:** a science whereby inferences are made about specific random phenomena on the basis of relatively limited sample material.
- **Mathematical statistics:**
 - Concerns the development of new methods of statistical inference and requires detailed knowledge of abstract mathematics for its implementation.
- **Applied statistics:**
 - Involves application of mathematical statistical methods to specific subject areas such as economics, psychology, and public health.

- **Biostatistics:** a branch of applied statistics that applies statistical methods to medical and biological problems.
- **Medical statistics:** a branch of biostatistics that applies statistical methods to medical problems.

Example of Study Question

- Are the methods of automatic vs. manual determination of blood pressure comparable? (Planning step)
- Questions:
 1. No. of machines
 2. No. of participants for each machines to be tested
 3. Order of taking measurements: Manual -> automated or vice versa
 4. Critical data to be captured in addition to the measurements
 5. Format of data
 6. Quality check of data

Example of Study Question

- Are the methods of automatic vs. manual determination of blood pressure comparable? (Planning step)
- Questions:
 1. No. of machines: 4
 2. No. of participants for each machines to be tested: 100
 3. Order of taking measurements: Manual -> automated or vice versa: random
 4. Critical data to be captured in addition to the measurements: Age, sex, size
 5. Format of data: ID, coding
 6. Quality check of data: outliers checking

Data Analysis

- Steps: Data collection → entry → editing → analysis
- Data obtained from the study can be summarized using descriptive statistics
- Descriptive material can be numeric or graphic
 - If numeric, data can be tabulated or presented as frequency distribution
 - If graphic, data can be summarized pictorially

- Choice of numeric or graphic descriptive statistics is dependent on type of distribution of data.

1. Continuous data:

- Where there are infinite number of possible values
- Means and standard deviations may be used

2. Discrete data

- Where there are only a few possible values
- Percentages of people for each value may be considered

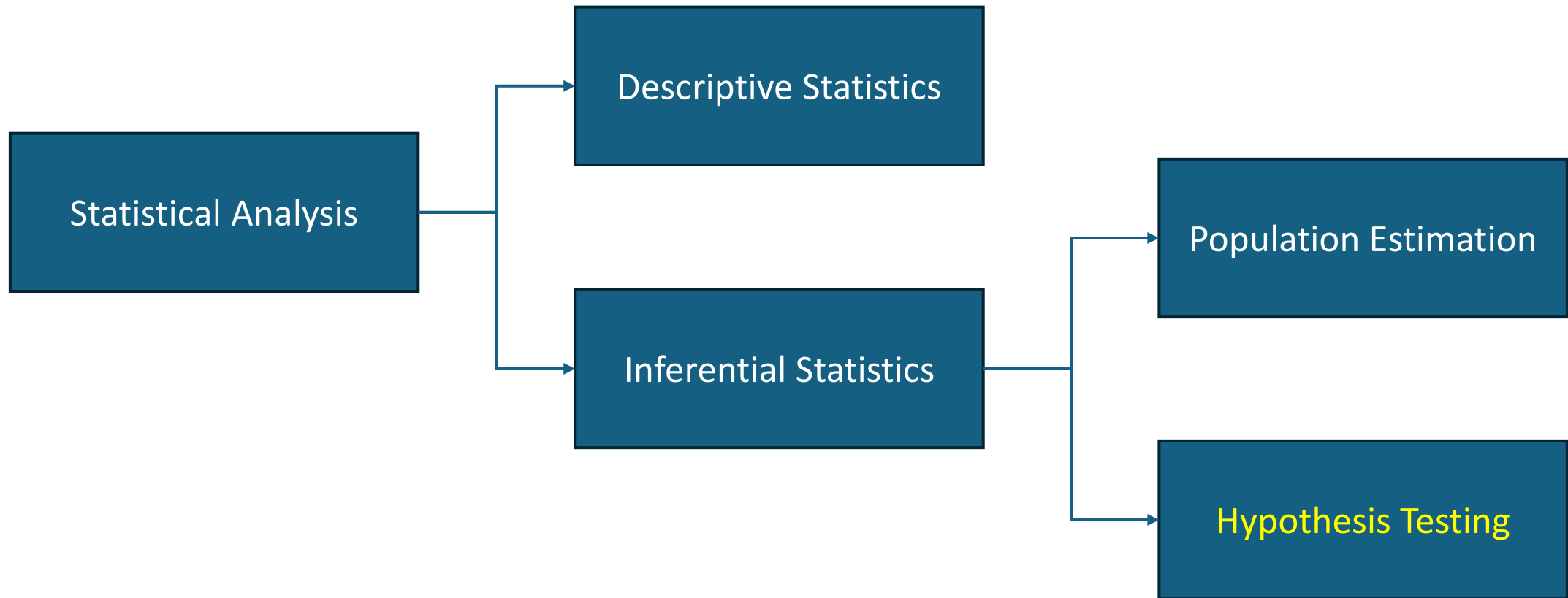
Tabulated results of study under consideration

- Missing data is common in medical statistics
- Notice the apparent difference in blood pressure readings between machines vs. manual measurements in locations C and D.

Table 1.1 Mean blood pressures and differences between machine and human readings at four locations

Location	Number of people	Systolic blood pressure (mm Hg)					
		Machine		Human		Difference	
		Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
A	98	142.5	21.0	142.0	18.1	0.5	11.2
B	84	134.1	22.5	133.6	23.2	0.5	12.1
C	98	147.9	20.3	133.9	18.3	14.0	11.7
D	62	135.4	16.7	128.5	19.0	6.9	13.6

Source: By permission of the American Heart Association, Inc.



Inferential Statistics

- Determining whether the difference in blood pressure readings is “real” or “by chance”
- Inferring the characteristics of a population from a sample is the central concern of statistical inference.
- If we can calculate the probability (*likelihood*) that the observed 14-mm Hg difference happens in a sample of 98 people if there were no real difference between the two methods over the entire population of users of the machine. (Student's *t*-test)
- A small enough probability would indicate that the difference between the two methods is real.

Final Step

- Publication of manuscript
- Essential data obtained in the data analysis step is included in the final manuscript for publishing

- Population: A population is the **entire set of homogeneous** observational units defined according to the **research objective**.
- E.g.:
 - The heights of healthy adult men in a certain region in 2025. (Finite)
 - The therapeutic effect of a certain drug in patients with hypertension.
- Note: medical research studies, unlike census, usually can not do population-level data collection.
- Sample: A sample is a set of observed values obtained from a subset of observational units **randomly selected** from the population.

- Probability: Probability describes the likelihood that a random event will occur.
- E.g.:
 - Among 300 patients treated with a new drug for a certain disease, 240 were cured, resulting in a cure rate of 80% (frequency).
 - When the probability of an event $P \leq 0.05$, it is referred to as a small-probability event. Characteristic: If an event has a very low probability of occurring, it is **generally** considered unlikely to occur in a single trial.

Types of variables

- Numerical Variable
 - Blood pressure
 - Height
 - Weight
- Categorical Variable
 - Sex

Types of variables

- Numerical Variable (Usually has unit)
 - Blood pressure
 - Height
 - Weight
- Categorical Variable
 - Nominal Variable
 - Sex
 - Blood type
 - Ordinal Variable
 - Treatment response grade (Cured, Markedly effective, Improved, Ineffective)

Types of variables

- Numerical Variable (Usually has unit)

Which of the following is categorical variable?

1. Weight

2. Blood Pressure

3. Heart rate

4. Race

- Ordinal Variable
 - Treatment response grade (Cured, Markedly effective, Improved, Ineffective)

Types of variables

- Numerical Variable (Usually has unit)

Which of the following is categorical variable?

1. Weight

2. Blood Pressure

3. Heart rate

4. Race✓

- Ordinal Variable
 - Treatment response grade (Cured, Markedly effective, Improved, Ineffective)

Types of variables

- **Numerical** Variable (Usually has unit)

- Blood pressure
- Height
- Weight

- t tests
- ANOVA
- Regression analysis
- Correlation analysis

- **Categorical** Variable

- **Nominal** Variable

- Sex
- Blood type

- Chi-square tests

- **Ordinal** Variable

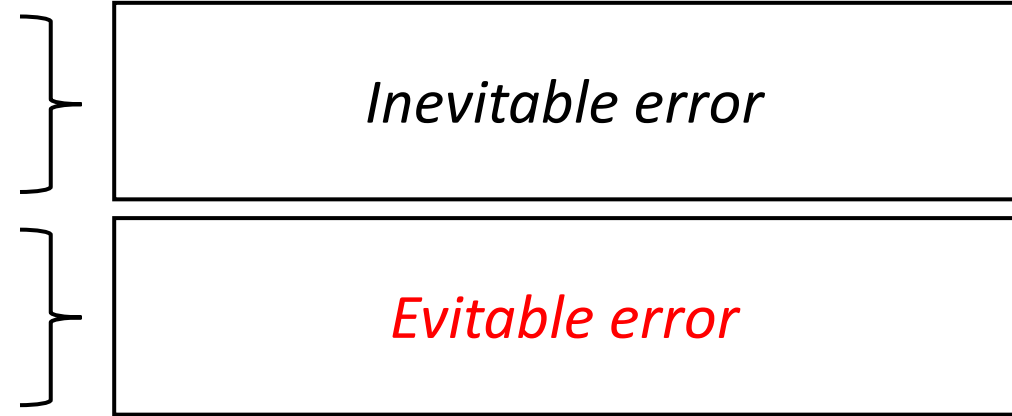
- Treatment response grade (Cured, Markedly effective, Improved, Ineffective)

- Rank-sum tests

Blood pressure -> High blood pressure or not -> Low, normal & high blood pressure

Types of Errors

- Sampling error
- Random error of measurement
- Systematic error
- Gross error



From data about people
to a healthier, smarter world

A Brief History of Statistics

From census records and probability theory to modern statistical inference, data science, and AI

"Statistics is not just a branch
of mathematics, it is a science about
uncertainty and a key to a deeper
understanding of the world."

— R. A. Fisher

Data
Drives
Healthier
Lives
and a Brighter
Future

1 Ancient Statistics and Early Records



Ancient Times – 16th Century

Used for population counts, land management, taxation, and administrative records.

Ancient civilizations (e.g., Egypt, China, Rome) laid the foundations of statistical records.

2 The Birth of Probability Theory



17th Century

Driven by problems in mortality analysis, insurance, and gambling, leading to the emergence of probability theory.

Key Figures



John Graunt
(1620–1674)
Father of vital statistics

John Locke
(1620–1674)
Early ideas on probability

Jacob Bernoulli
(1654–1705)
Law of large numbers

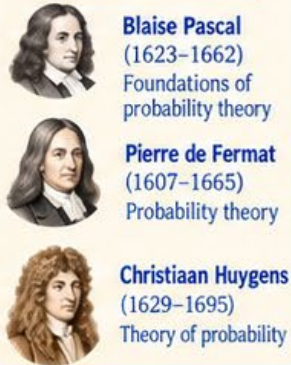
3 Foundation of Statistical Inference



17th Century

From gambling problems to general principles, establishing the basis of statistical inference.

Key Figures



Blaise Pascal
(1623–1662)
Foundations of probability theory

Pierre de Fermat
(1607–1665)
Probability theory

Christiaan Huygens
(1629–1695)
Theory of probability

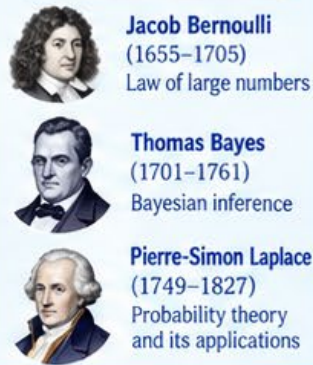
4 The Law of Large Numbers and Probability Distributions



17th – 18th Century

Established the law of large numbers, developed probability distributions, and formed a theoretical foundation for statistics.

Key Figures

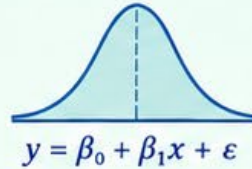


Jacob Bernoulli
(1655–1705)
Law of large numbers

Thomas Bayes
(1701–1761)
Bayesian inference

Pierre-Simon Laplace
(1749–1827)
Probability theory and its applications

5 Statistical Models and Inference



17th – 18th Century

Developed regression analysis, maximum likelihood, and statistical inference, forming a modern statistical framework.

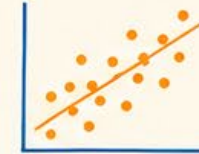
Key Figures



Carl F. Gauss
(1777–1855)
Least squares and the normal distribution

Adrien-Marie Legendre
(1752–1833)
Method of least squares

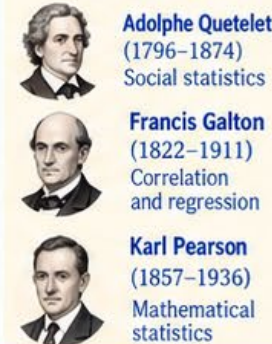
6 Modern Statistical Methods



19th Century

Statistical methods flourished, with wide applications in scientific research, social studies, economics, and medicine.

Key Figures



Adolphe Quetelet
(1796–1874)
Social statistics

Francis Galton
(1822–1911)
Correlation and regression

Karl Pearson
(1857–1936)
Mathematical statistics

7 Modern Statistical Inference

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

- ✓ Hypothesis testing
- ✓ Confidence intervals
- ✓ Experimental design

20th Century – First Half

Established hypothesis testing, confidence intervals, experimental design, and modern statistical inference.



William S. Gosset
(1876–1937)
"Student's t-distribution"

Ronald A. Fisher
(1890–1962)
Modern statistical inference

Jerzy Neyman
(1894–1981)
Neyman–Pearson theory

Egon Pearson
(1895–1980)
Inference and testing

8 Data Science, Statistics and the AI Era



20th Century – Present

Integration with computer science, big data, and AI. Statistics drives data science, machine learning, and AI, powering breakthroughs across disciplines.

Key Figures



John Tukey
(1915–2000)
Exploratory data analysis

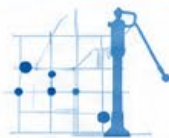
Bradley Efron
(1938–)
Bootstrap methods

Trevor Hastie
(1958–)
Statistical learning and high-dimensional data



John Snow
(1813–1858)

Cholera study and epidemiological analysis



Florence Nightingale
(1820–1910)

Pioneered the use of statistics in public health



Austin Bradford Hill
(1897–1991)

Clinical trials and causal inference in epidemiology



The development of statistics has advanced medicine, biology, public health, social science, and artificial intelligence, through evidence discovery and scientific decision-making.



Statistics for a Healthier World

