Introduction

Applied Statistics

Fall 2025

目录

1 教材与考核 Textbook and Assessment

教材 Textbook

David S. Moore, George P. McCabe, and Bruce A. Craig, *Introduction to the Practice of Statistics*, 6th edition, 2006.

考核方式 Grading

- **作业 Homework**: 10%（按完成情况评分，不接受迟交）
  Grades are based on completion. Late submissions are not accepted.
- **实验考试 Lab Exam**: 45%
- **笔试 Written Exam**: 45%

2 Stata 学习资源 Stata Learning Resources

- **官方 Stata 网站 Official Stata Site**: 适合入门和查询命令，帮助文件比 Stata 内"help" 更详细。
  Good for introduction and looking up commands. The online help file is more detailed than the one from "help command" in Stata.
- **UCLA Stata Modules**: 提供最好的数据管理基础教程。
  Has the best basic introductions to data management.

3 统计学简史 A Brief History of Statistics

表 1: 统计学简史A Brief History of Statistics

中文	English
起源	**Origins**
统治者为了统计人口或应税土地而产生。	The earliest origins lie in the desire of rulers to count inhabitants or measure the value of taxable land.
约翰・格朗特 (1620-1674)	**John Graunt (1620-1674)**
首次系统整理伦敦的出生、死亡及死因数据，可视为现代人口统计学的开端。	First began systematically reviewing weekly bills of mortality (births, deaths, causes) in London, marking the beginning of modern population statistics.
中国历史上的统计	**History of Statistics in China**
自秦朝起，中央政府通过”户部”统计户籍与土地，用于税收管理。	Starting from the Qin dynasty, the central government used the ”Hu Bu” (Ministry of Revenue) to collect household and land statistics for taxation.

4 什么是统计学？ What is Statistics?

定义 Definition

从数据中获取信息的学科。

A way to get information from data.

描述性统计 Descriptive Statistics

对数据进行收集、整理和描述。

The collection and description of data.

- **工具 Methods**:
  - 图表 Graphs: 柱状图 bar chart, 直方图 histogram, 散点图 scatter plot
  - 数字摘要 Numerical Summaries: 均值 mean, 标准差 standard deviation

推断性统计 Inferential Statistics

基于样本对总体进行推断。
Making inferences from the data about a population.

- **理论基础 Theoretical Basis**: 抽样分布 sampling distribution, 重抽样方法 resampling method
- **主要方法 Main Methods**: 估计 estimation, 假设检验 hypothesis testing

5 总体与样本 Populations and Samples

表 2: 总体与样本基本概念Basic Concepts of Populations and Samples

中文术语	**English Term**	**描述 Description**
总体	Population	所有感兴趣对象的集合；通常规模很大，有时无限。 The group of all items of interest; frequently very large; sometimes infinite.
样本	Sample	从总体中抽取的一部分数据。 A set of data drawn from the population.
参数	Parameter	描述总体特征的量，如总体均值 μ。 A descriptive measure of population characteristics, e.g., population mean μ.
统计量	(Sample) Statistic	描述样本特征的量，如样本均值 $\bar{x}$。 A descriptive measure of sample characteristics, e.g., sample mean $\bar{x}$.

关键例子 Key Example

研究西安交通大学 2025 届毕业生的首份工作薪资。
To study the starting salary of the first job for Xi’an Jiaotong University graduates of the class of 2025.

- **总体 Population**: 所有 2025 届西交大毕业生。
  All XJTU graduates of the class of 2025.

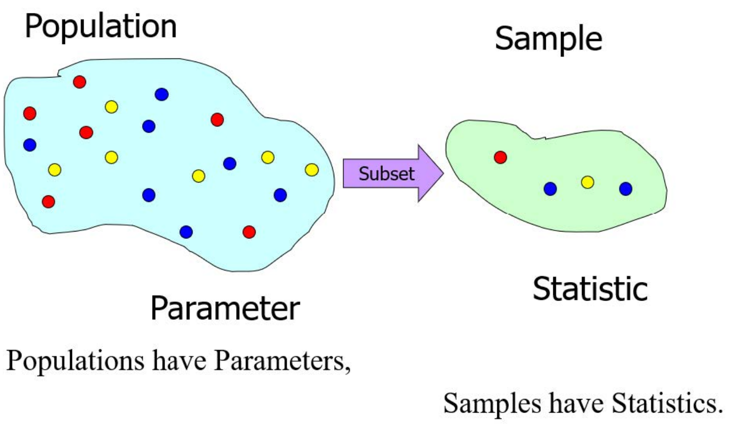

图 1: Populations & Samples

- **样本 Sample**: 随机抽取的 200 名毕业生。
  A randomly selected group of 200 graduates.

- **参数 Parameter**: 所有毕业生的平均起薪 μ。
  The average starting salary of all graduates (μ).

- **统计量 Statistic**: 200 名样本毕业生的平均起薪 $\bar{x}$。
  The average starting salary of the 200 sampled graduates ($\bar{x}$).

S&P 500 示例：参数 vs. 统计量

S&P 500 Example: Parameter vs. Statistic

总体设定 Population Setting: 所有 S&P 500 指数中的股票
The target population is all stocks in the S&P 500 index.

表 3: S&P 500 参数与统计量示例S&P 500 Parameter and Statistic Examples

描述 Description	**分类 Category**	**原因 Reason**
所有 500 支 S&P 指数股票的平均市盈率 The average price/earnings ratio for all 500 stocks in the S&P index.	参数 Parameter	描述了总体特征 (describes the population)
去年所有 S&P 500 指数股票中亏损股票的比例 The proportion of all stocks in the S&P 500 index that had negative earnings last year.	参数 Parameter	描述了总体特征 (describes the population)
随机抽取的 50 支股票中能源相关股票的比例 The proportion of energy-related stocks in a random sample of 50 stocks.	统计量 Statistic	描述了样本特征 (describes a sample)
经纪人推荐的 20 支股票的平均回报率 The average rate of return for 20 stocks recommended by a broker.	统计量 Statistic	描述了样本特征 (describes a sample)

6 数据类型与来源 Types and Sources of Data

表 4: 数据类型与来源Types and Sources of Data

数据类型	**Data Type**	**中文例子**	**English Example**
调查数据	Survey Data	中国健康与营养调查 (CHNS)	China Health and Nutritional Survey (CHNS)
行政数据	Administrative Data	美国出生与死亡数据; 新农合报销数据	U.S. Birth and Mortality Data; NCMS claims data
大数据	”Big Data”	Yelp 开放数据集; 尼尔森零售扫描数据	Yelp Open Dataset; Nielsen Retail Scanner Data

7 大数据时代的统计学 Statistics in the Age of Big Data

- **现象 Phenomenon**: 电子记录使得”总体数据”越来越容易获取。

  With widespread electronic recording, ”population” data are increasingly available.

- **推断统计依然重要 Inferential Statistics Still Relevant**:

  - **例子 Example**: 虽然已有全部历史购买记录，但下一次购买行为仍需基于历史数据进行预测。

    Although your entire purchasing history might be online, your next purchase is not yet. We still need to use previous data to infer the future.

  - **新视角 New Perspective**: 可将”所有潜在购买行为”视为总体，”已实现的购买行为”视为样本。

    Think of all potential purchasing behaviors as the population and those realized as the sample.

8 常见误区 Common Pitfalls to Avoid

表 5: 常见统计学误区Common Statistical Pitfalls

误区	**中文描述**	**English Description**	**例子 Example**
小样本/非随机样本 Small/Non-random Samples	基于小样本或非随机样本得出普遍结论。	Drawing general conclusions from small or non-random samples.	”摇滚明星死得早” ”Rock stars die young.”
调查方法不当 Poor Survey Methods	数据收集方式存在偏差，如自我报告偏差。	Biases in data collection, e.g., self-reporting bias.	实验者诱导被试给出特定答案。 Experimenter-induced responses.
相关不等于因果 Correlation Causation	两个变量相关并不意味着一个导致另一个。	Two variables being correlated does not mean one causes the other.	冰淇淋销量与溺水率正相关，真实原因是天气炎热（混杂变量）。 Drowning rates are higher when ice cream sales are high.
统计显著 vs. 实际显著 Statistical vs. Practical Significance	结果在统计上显著，但实际影响微小，没有实际意义。	A result is statistically significant but the effect size is too small to be meaningful.	新教学法使平均分从 75.0 升至 75.2 ($p < 0.01$)。 New teaching method increases average score from 75.0 to 75.2 ($p < 0.01$).

9 统计学与数据科学
Statistics vs. Data Science

表 6: 统计学与数据科学比较Statistics vs. Data Science Comparison

领域 Field	核心目标 Primary Goal	方法论特点 Methodology	例子 Example
传统统计学 (包括计量经济学) Traditional Statistics (incl. Econometrics)	解释数据 Explanation	关注生成模型和因果推断。重视模型的可解释性。 Focuses on explaining data (generative models, causal inference). Values model interpretability.	研究教育年限如何影响收入（寻找因果关系）。 Studying how years of education affect income (seeking causation).
数据科学 Data Science	预测结果 Prediction	目标是预测准确性。模型可以是” 黑箱”，不关心变量是否反映现实。 Focuses on prediction. The model can be a ”black box”; it doesn’t care if variables reflect underlying reality.	TikTok 推荐系统：不关心你为什么喜欢视频,只预测你会点击哪个。 TikTok’s recommendation system: doesn’t care why you like a video, only predicts which one you will click.