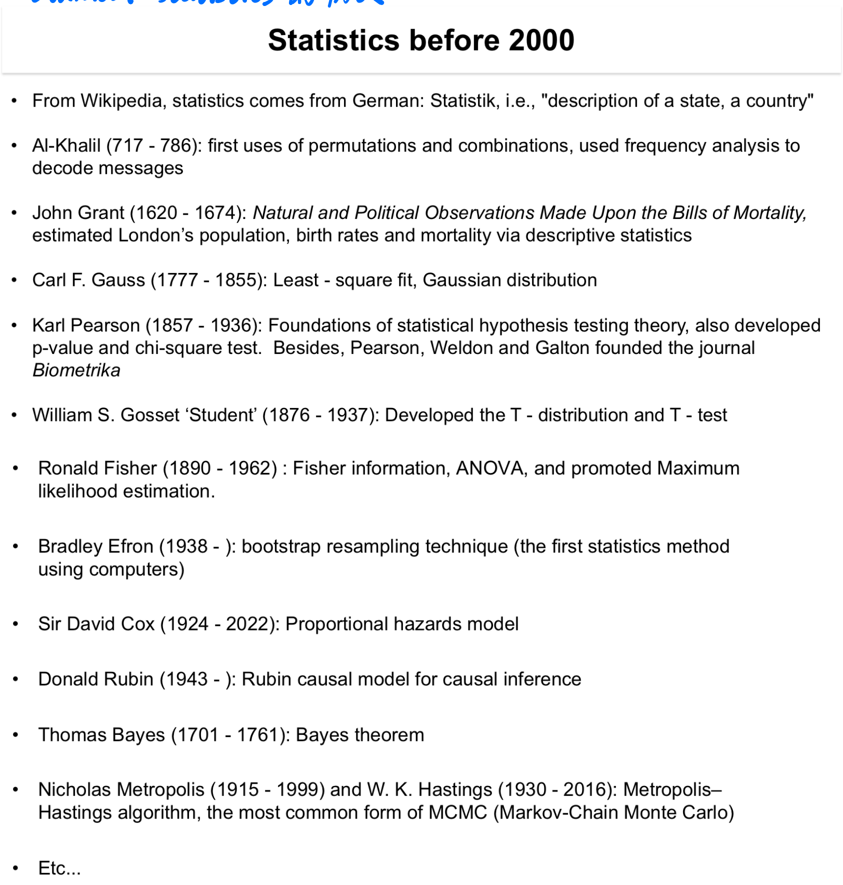
Lecture 1

§1 Introduction to statistics

1. Classical statistics 的历史



Statistics before 2000

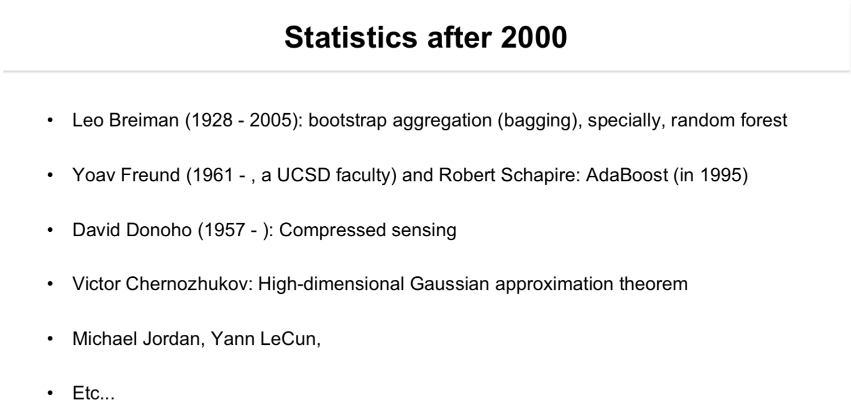
- From Wikipedia, statistics comes from German: Statistik, i.e., "description of a state, a country"
- Al-Khalil (717 - 786): first uses of permutations and combinations, used frequency analysis to decode messages
- John Grant (1620 - 1674): *Natural and Political Observations Made Upon the Bills of Mortality,* estimated London's population, birth rates and mortality via descriptive statistics
- Carl F. Gauss (1777 - 1855): Least - square fit, Gaussian distribution
- Karl Pearson (1857 - 1936): Foundations of statistical hypothesis testing theory, also developed p-value and chi-square test. Besides, Pearson, Weldon and Galton founded the journal *Biometrika*
- William S. Gosset 'Student' (1876 - 1937): Developed the T - distribution and T - test
- Ronald Fisher (1890 - 1962) : Fisher information, ANOVA, and promoted Maximum likelihood estimation.
- Bradley Efron (1938 -): bootstrap resampling technique (the first statistics method using computers)
- Sir David Cox (1924 - 2022): Proportional hazards model
- Donald Rubin (1943 -): Rubin causal model for causal inference
- Thomas Bayes (1701 - 1761): Bayes theorem
- Nicholas Metropolis (1915 - 1999) and W. K. Hastings (1930 - 2016): Metropolis–Hastings algorithm, the most common form of MCMC (Markov-Chain Monte Carlo)
- Etc...

2. Classical statistics 的特征

- 数据较少
- 统计模型/算法易于分析，结果优雅
- 数据集清晰 (missing data 较少，data structure 简单，e.g. 实数或实向量)
- i.i.d 假设永远成立
- 注重 inference

注：本课程将重点研究 classical statistics

3. Modern statistics 的历史



Statistics after 2000

- Leo Breiman (1928 - 2005): bootstrap aggregation (bagging), specially, random forest
- Yoav Freund (1961 - , a UCSD faculty) and Robert Schapire: AdaBoost (in 1995)
- David Donoho (1957 -): Compressed sensing
- Victor Chernozhukov: High-dimensional Gaussian approximation theorem
- Michael Jordan, Yann LeCun,
- Etc...

4. Modern statistics 的特征

- 数据较多

- 数据源多，数据形式多样
- 除了 inference，还注重 prediction 与 model simplification
- 电脑的应用

§1 Basic ideas in statistics

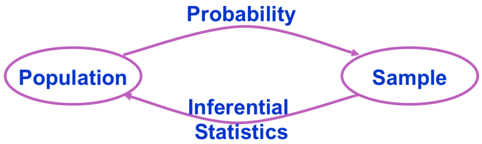
1. Probability 与 statistics

1° Probability （概率）

- 对于 samples 的产生有一个明确的 machanism
- no modelling

2° Statistics （统计）

- 已知 samples，需要猜测并证实产生这些样本的 model
- require modelling



2. Population，sample，与 sampling bias

1° Population （全体）

一个有限的，明确定义的，包括 all objects 的 group，尽管可能很大，但理论上可被 enumerated.

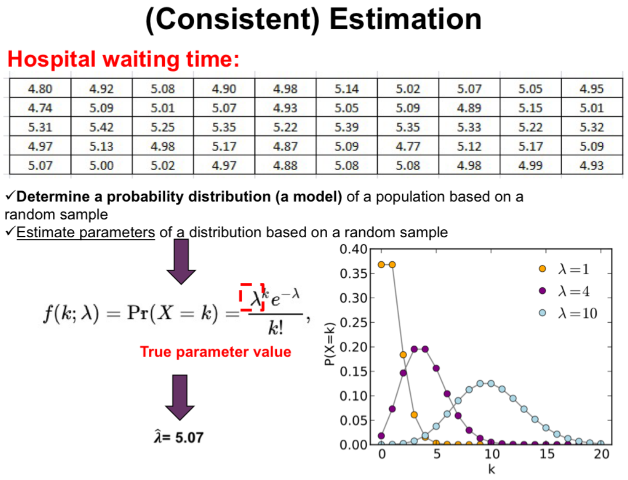
2° Sample （样本）

Population 的一个子集.

3° Sampling bias （抽样偏差）

样本不能完全反映全体

3. 一些例子



(Consistent) Estimation

Hospital waiting time:

4.80	4.92	5.08	4.90	4.98	5.14	5.02	5.07	5.05	4.95
4.74	5.09	5.01	5.07	4.93	5.05	5.09	4.89	5.15	5.01
5.31	5.42	5.25	5.35	5.22	5.39	5.35	5.33	5.22	5.32
4.97	5.13	4.98	5.17	4.87	5.09	4.77	5.12	5.17	5.09
5.07	5.00	5.02	4.97	4.88	5.08	5.08	4.98	4.99	4.93

✓**Determine a probability distribution (a model)** of a population based on a random sample

✓Estimate parameters of a distribution based on a random sample

$$f(k;\lambda) = \Pr(X = k) = \frac{\boxed{\lambda}^k e^{-\lambda}}{k!},$$

True parameter value

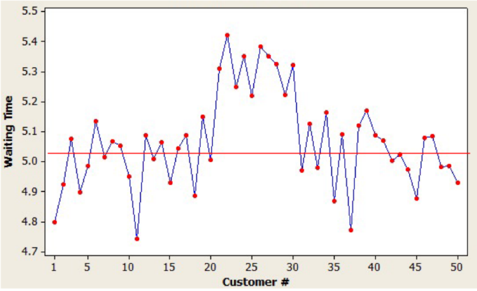
$\hat{\lambda}$= 5.07

Confidence interval

How confident we are given the variability of data?

✓Construct confidence intervals for parameters of a distribution

5.07



$$\boldsymbol{\lambda} \in [\mathbf{5.07 - 0.16. 5.07 + 0.16}]$$

Test a hypothesis for the population

Given the average wait time 5.07

Null hypothesis $\boldsymbol{\lambda \leq 5}$

Alternative hypothesis $\boldsymbol{\lambda > 5}$

Which one is true?

Data

↓

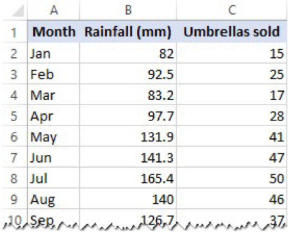
Statistics

↓

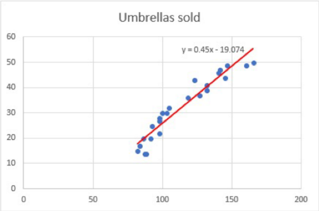
Decision

Regression

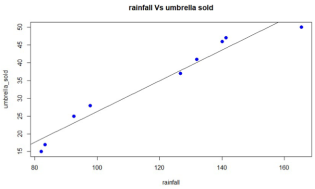
✓ Predict a response variable based on one or more predictor variables

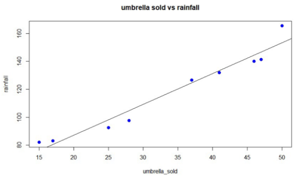


	A	B	C
1	Month	Rainfall (mm)	Umbrellas sold
2	Jan	82	15
3	Feb	92.5	25
4	Mar	83.2	17
5	Apr	97.7	28
6	May	131.9	41
7	Jun	141.3	47
8	Jul	165.4	50
9	Aug	140	46
10	Sep	126.7	37



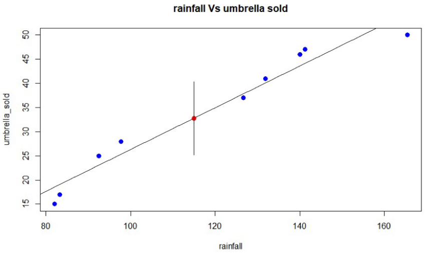
Causal inference





- Correlation does not imply causality
- Find which implies which

Prediction & Predictive inference



- For a new x - value (rainfall here), estimate a y - value
- Provide an interval that the new y is ‘likely’ to be in this interval with