Topic 0: Class Logistics（课程信息） and Introduction of Statistics

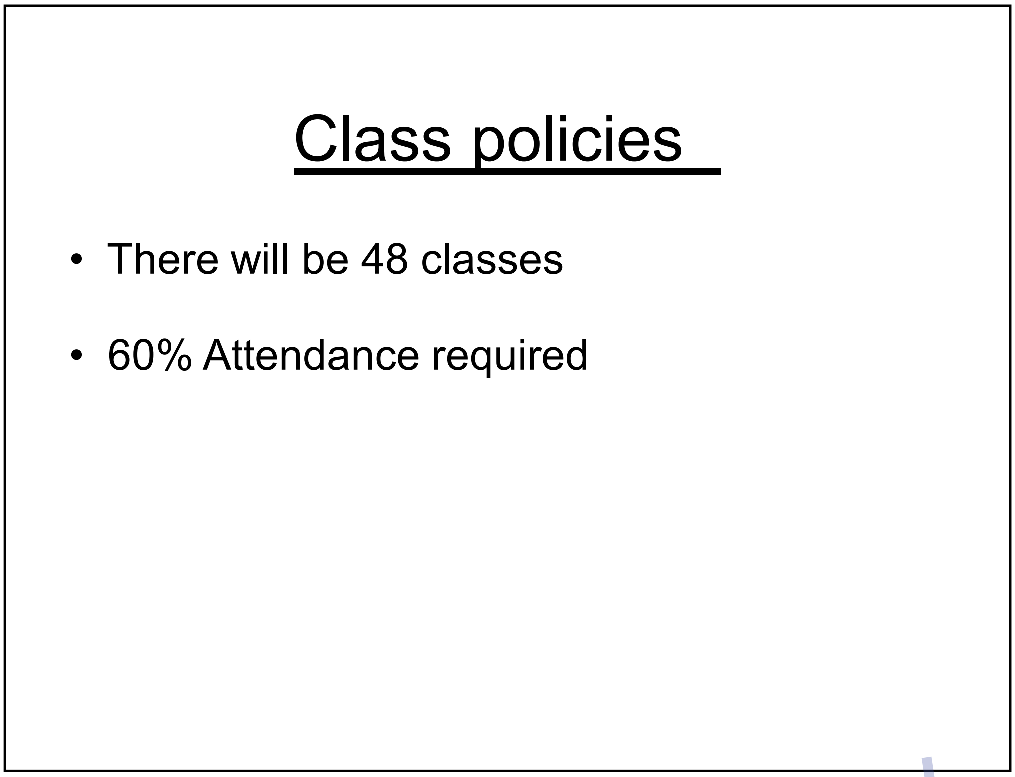
Instructor Information (教师信息）

- 姓名：王江典
- 地址：伟清楼203
- 邮箱：wangjiangdian@tsinghua.edu.cn
- 答疑： 3:00pm～4:00pm 周三

Class Logistics（课程信息）

Teaching Assistant Information

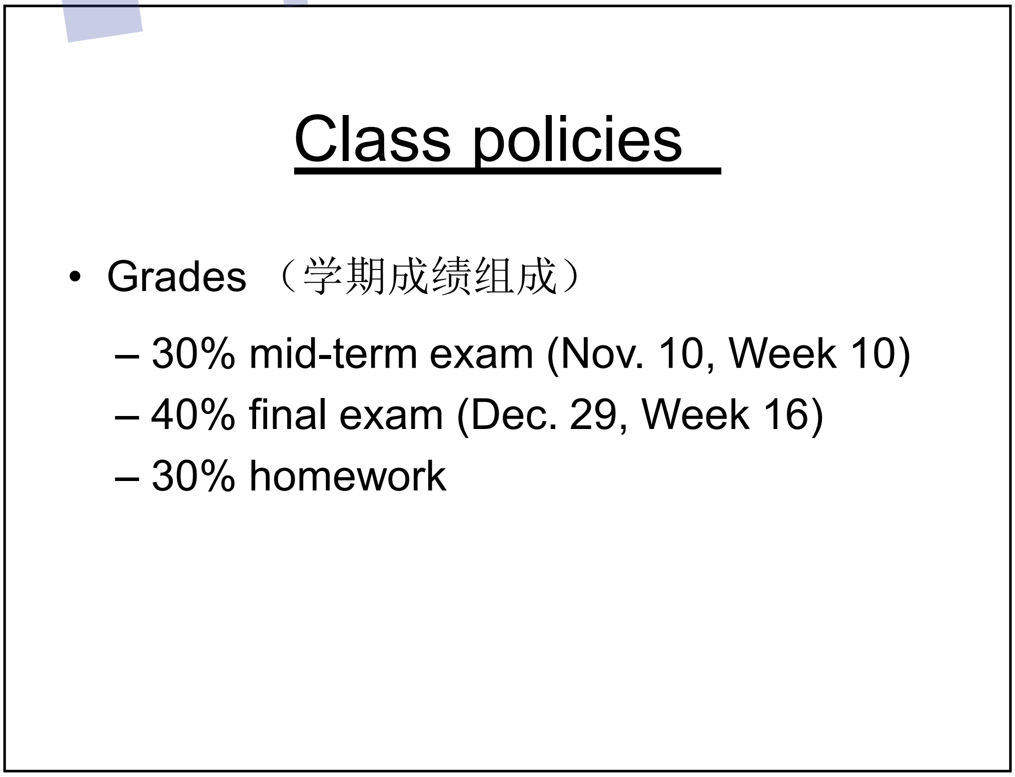
- 姓名：吴方维
- 地址：伟清楼209
- 邮箱：wfw19@mails.tsinghua.edu.cn
- 答疑：2:00pm – 3:00pm 周四



Class policies

- There will be 48 classes
- 60% Attendance required

5



Class policies

- Grades （学期成绩组成）
  - 30% mid-term exam (Nov. 10, Week 10)
  - 40% final exam (Dec. 29, Week 16)
  - 30% homework

6

Class policies

- Exams are
  - close book allowing a two-sided A4 or smaller handwriting cheat sheet.
  - will need a calculator with a square root function
  - 70% of grade

Class policies

- Homework assignments （作业）
  - Sent by TA
  - will cover material from prior week
  - expect 11 total but will drop lowest score
  - this should be mostly individual work
  - late HW within one week takes 40% penalty
  - latex, word (executable code, R is preferred)
  - do not allow to submit homework by taking a photo/picture of handwriting
  - 30% of grade

Homework Guidelines

- Each problem must be presented in order, including all relevant graphs and tables, readable and labeled
- Any graph or figure without comments will be ignored
- Providing distracting or irrelevant output can result in a loss of points

Overview

We will cover basic concepts of statistical inference, point estimation, confidence intervals and confidence regions, testing hypotheses, nonparametric inference and introduction of Bayesian inference and decision-making theory. The emphasis will be on

- Conceptual understanding
- Analysis methods and their intuitions
- Intuitive understanding of the theory
- Some methods for proofs
- Interpretation of results

Textbook

- Main textbook:

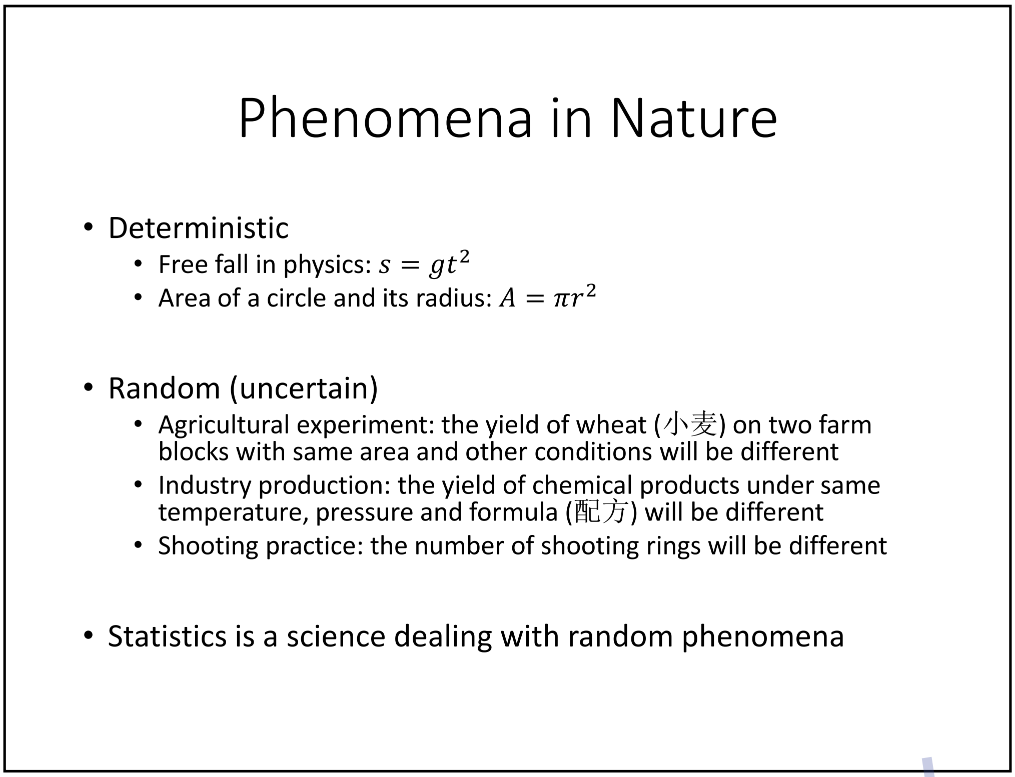
- An Introduction to Probability and Statistical Inference, 2th Ed. Academic Press, 2015
by George G. Roussas

- References:

- 数理统计（第二版），科学出版社，韦来生，2015

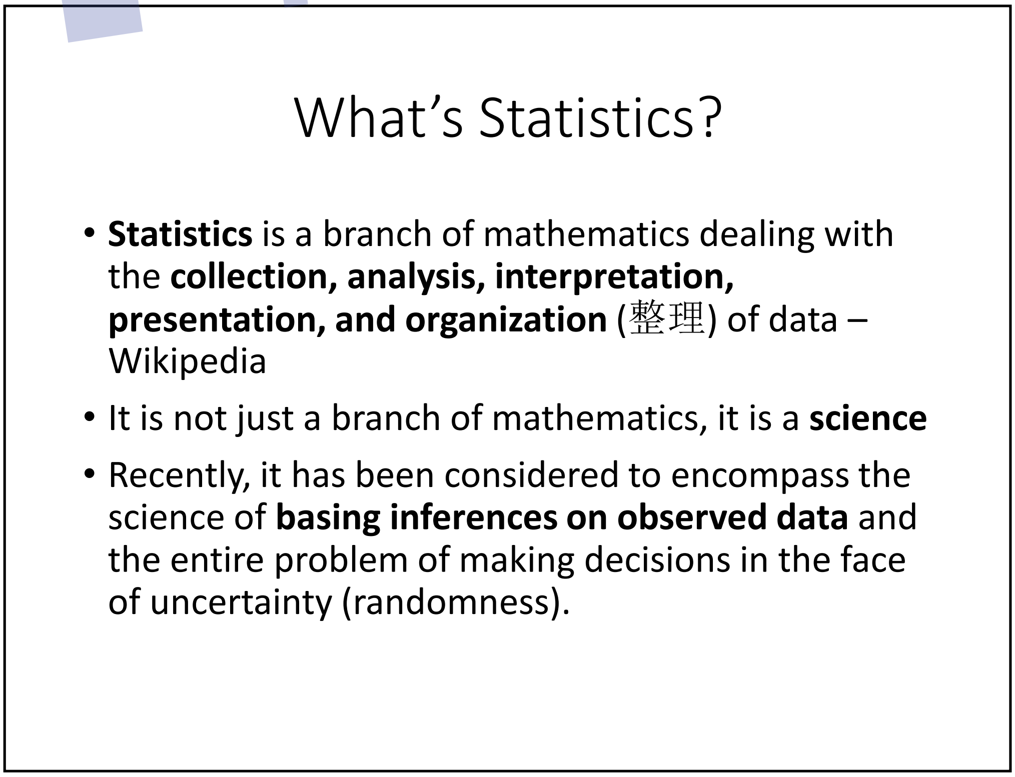
- Statistical inference. Duxbury Press, 2001, by G. Casella, R. L. Berger

Introduction of Statistics



Phenomena in Nature

- Deterministic
  - Free fall in physics: $s = gt^2$
  - Area of a circle and its radius: $A = \pi r^2$
- Random (uncertain)
  - Agricultural experiment: the yield of wheat (小麦) on two farm blocks with same area and other conditions will be different
  - Industry production: the yield of chemical products under same temperature, pressure and formula (配方) will be different
  - Shooting practice: the number of shooting rings will be different
- Statistics is a science dealing with random phenomena



What’s Statistics?

- **Statistics** is a branch of mathematics dealing with the **collection, analysis, interpretation, presentation, and organization** (整理) of data – Wikipedia
- It is not just a branch of mathematics, it is a **science**
- Recently, it has been considered to encompass the science of **basing inferences on observed data** and the entire problem of making decisions in the face of uncertainty (randomness).

What’s Statistics?

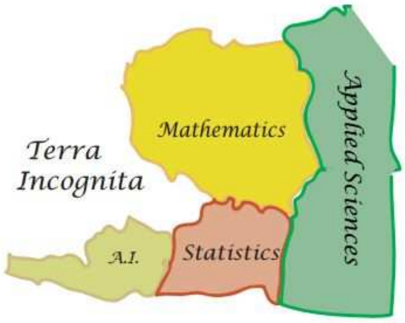
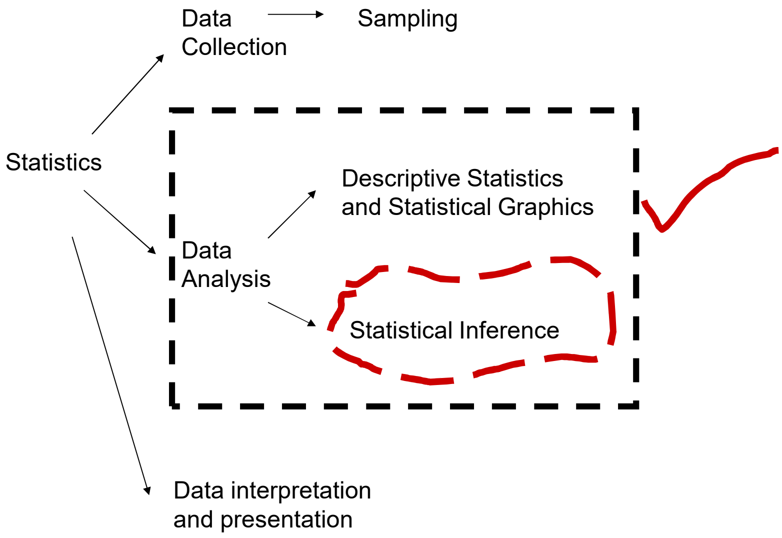


FIGURE 1. The greater world of mathematics and science.

Bradley Efron (2013), “A 250-YEAR ARGUMENT: BELIEF, BEHAVIOR, AND THE BOOTSTRAP”. Bulletin of the American Mathematical Society 50(1) .

What’s Statistics?



Data Collection

Sampling

Statistics

Descriptive Statistics and Statistical Graphics

Data Analysis

Statistical Inference

Data interpretation and presentation

Collect Data Efficiently

- Overall statistical survey (全面调查或普查)
  - Census (人口普查)
- Sampling survey (抽样调查)
  - Correct the results of census, require experts
  - **Example:** Study the income of 10000 farmers in a certain region. In this region, 70% of farmers live in plain areas (rich) and the remaining 30% live in mountain area (poor). How to sample 100 farmers from this region?
  - Sampling survey is an important branch of statistics
- Experiment (安排试验)

Collect Data Efficiently

- Experiment (安排试验)
  - **Example:** The yield of chemical product depends on temperature, pressure and formula. In order to increase the yield, we need experiment to find the best conditions for production.

	1	2	3	4
temperature	800°C	1000°C	1200°C	1400°C
pressure	10	20	30	40
Formula(配方)	A	B	C	D

  - 64 experiments in totally – cost too much
  - Experiment design and analysis is another important branch of statistics

Use Data Efficiently

- Extract information from data efficiently and make conclusion/inference on the investigated problem – statistical inference
- In order to make statistical inference, we must establish a statistical model for the data, propose inference methods, and use certain criterion to compare these methods
- Choose appropriate method based on the problem in hand

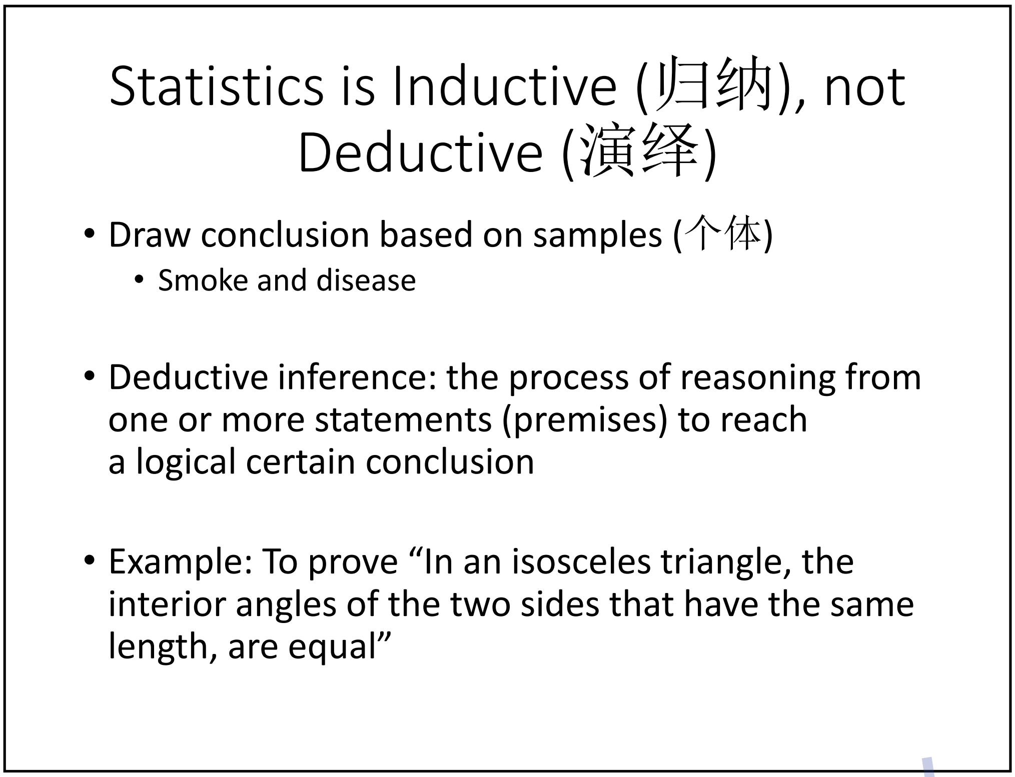
Use Data Efficiently

- **Example:** Investigate whether the 100 farmers in a village get rid of poverty. The standard for getting rid of poverty is Average Annual Income (AAI) is more than 10000 Yuan. After an overall survey, there are 90 farmers whose AAI is 5000 Yuan, and 10 farmers whose AAI is 100000 Yuan.
  - Mean

$$\bar{x} = \frac{90\times0.5+10\times10}{100} = 1.45(万元)$$

  - Median

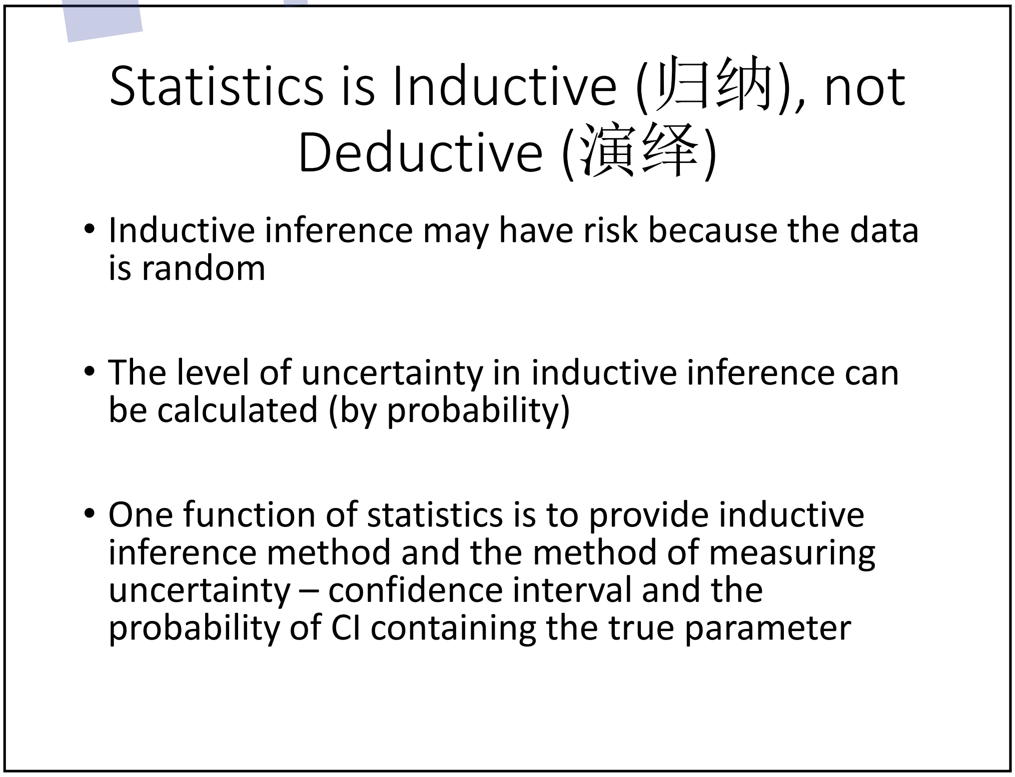
$$\frac{x_{(50)}+x_{(51)}}{2} = 0.5(万元)$$



Statistics is Inductive (归纳), not Deductive (演绎)

- Draw conclusion based on samples (个体)
  - Smoke and disease
- Deductive inference: the process of reasoning from one or more statements (premises) to reach a logical certain conclusion
- Example: To prove “In an isosceles triangle, the interior angles of the two sides that have the same length, are equal”

21



Statistics is Inductive (归纳), not Deductive (演绎)

- Inductive inference may have risk because the data is random
- The level of uncertainty in inductive inference can be calculated (by probability)
- One function of statistics is to provide inductive inference method and the method of measuring uncertainty – confidence interval and the probability of CI containing the true parameter

22

Applications of Statistics

- Government decision-making
- Agriculture and industry
  - Yield of wheat, industry product
  - Quality control, product reliability
- Economic and finance
  - Econometrics (计量经济学)
- Biology, medical science, genetics
- ……

Difference between Probability and Statistics

	Probability	
1	We have a **fair** coin.	

Difference between P and S

	Probability	
1	We have a **fair** coin.	
2	Flip the fair coin ten times.	

Difference between P and S

	Probability	
1	We have a **fair** coin.	
2	Flip the fair coin ten times.	
3	**P({all are heads}) = ?**	

Difference between P and S

	Probability	Statistics
1	We have a **fair** coin.	We have a coin.
2	Flip the fair coin ten times.	
3	**P({all are heads}) = ?**	

Difference between P and S

	Probability	Statistics
1	We have a **fair** coin.	We have a coin.
2	Flip the fair coin ten times.	Flip the coin ten times.
3	**P({all are heads}) = ?**	

Difference between P and S

	Probability	Statistics
1	We have a **fair** coin.	We have a coin.
2	Flip the fair coin ten times.	Flip the coin ten times.
3	**P({all are heads}) = ?**	**All heads are obtained**, then **is it a fair coin?**

Difference between P and S

- So, in the same random experiment
  - a probabilitist will only ask the probability of getting a certain event under some probabilistic model assumptions **before** doing the experiment, (kind of mathematics approach), while
  - a statistician will make some conclusion about the probability model **after** the experiment (kind of physics approach)

Difference between P and S

- Refer to the above example of tossing a coin.
- A probabilitist will tell you that if the coin is fair, then
  P({all are heads}) = $(0.5)^{10}$=0.0009765625.
- So, in some sense, probability is about **looking forward**.
- For a statistician, if all heads are obtained, then s(he) will make a conclusion that the coin is NOT fair; otherwise, it is very unlikely to get ten heads in a row.
- So, we can say that statistics is about **looking backward**.

A Brief History of Statistics

- Start at beginning of 20th century, two stage
  - **First stage:** to the end of the Second World War
  - 英国学派: R.A. Fisher (Chi-square distribution) and K. Pearson (F distribution)

  - W.S. Gosset (t distribution), J. Neyman, E.S. Pearson (sun of K. Pearson), A. Wald, P.L. Hsu, …(hypothesis testing)

A Brief History of Statistics

- Starting mark
  - Pearson, Karl (1900). "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling" (PDF). *Philosophical Magazine Series 5*. **50** (302): 157–175. doi:10.1080/14786440009463897
  - Fisher, R. A. (1922). "On the Mathematical Foundations of Theoretical Statistics". *Philos. Trans. R. Soc. London, Ser. A*. **222A**: 309–368.
- Mature (成熟): H. Cramer (瑞典). Mathematical Methods of Statistics, Princeton University Press, 1946.
- **Second stage:** Second World War to now
  - Branches of statistics
  - New branches: Decision-making theory, Bayes
  - Computer development – Markov Chain Monte Carlo, etc.

A Brief History of Statistics

- In the last 40 years, statistics has changed enormously under the impact of several forces:
  - The generation of what were once unusual types of data such as text, images, voice, video, html, and other types of combinatorial objects
  - The generation of enormous amounts of data – terabytes (10^12 characters), petabytes – **big data**
  - The possibility of implementing computations of a magnitude that would have once been unthinkable

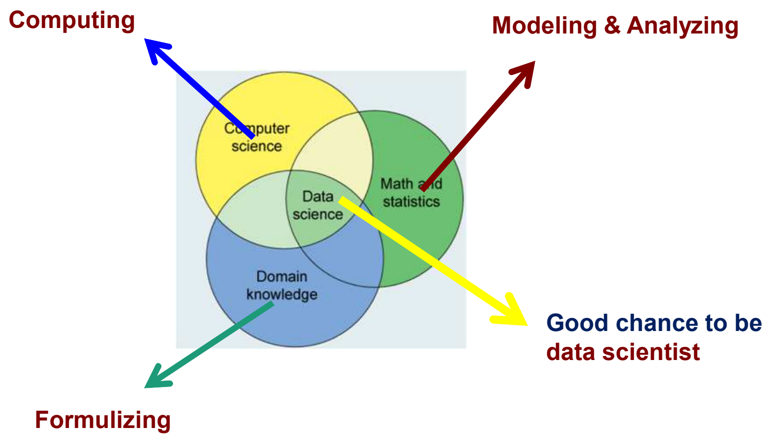
A Brief History of Statistics

- As a consequence the emphasis of statistical theory has shifted away from small sample optimality results in a number of directions:
  - Methods of inference based on larger number of observations and minimal assumptions – asymptotic methods in high-dimensional models, in non- and semiparametric models (models with infinite number of parameters)
  - The construction of models for temporal spatial series, and other complex data structures
  - The use of methods of inference involving simulations as a key element such as bootstrap and Markov Chain Monte Carlo

A Brief History of Statistics

- The development of techniques not describable in “close mathematical form” or “statistical model” but rather elaborate algorithms for which problems of existence of solutions are important and far from obvious – machine learning (SVM, random forest, bayes network, Deep Learning, Deep Neural Network)
- Interplay between numerical and statistical considerations – some methods are theoretically attractive but cannot be implemented in a human lifetime
- Interplay between the number of observations and the number of parameters of a model and the beginning of appropriate asymptotic theories

Statistics and Data Science



Computing

Modeling & Analyzing

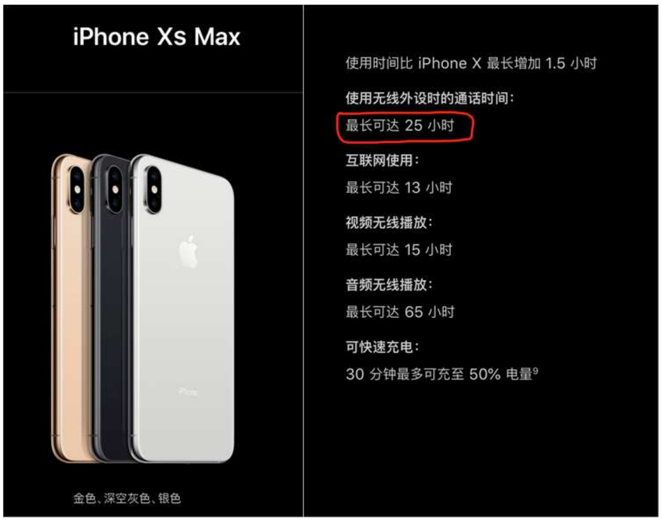
Good chance to be data scientist

Formulizing

Class Topic

- Basic concepts of statistical inference
- Point estimation
- Confidence intervals and confidence regions
- Testing hypotheses (parametric and nonparametric inference)

Example: iPhone Xs Max



iPhone Xs Max

使用时间比 iPhone X 最长增加 1.5 小时

使用无线外设时的通话时间：

最长可达 25 小时

互联网使用：

最长可达 13 小时

视频无线播放：

最长可达 15 小时

音频无线播放：

最长可达 65 小时

可快速充电：

30 分钟最多可充至 50% 电量[9]

金色、深空灰色、银色

iPhone Xs Max

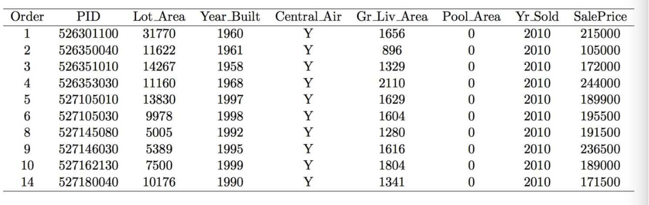
- **Question:** Could the Talk time (wireless) of iPhone Xs Max sold in China be up to 25 hours?
- Overall statistical survey?
  - Too expensive
  - may be unnecessary (e.g., we require an accuracy of 99%)
- Sampling survey?
  - how large the sample size?
- Suppose we took a sample of size 100
  - How to estimate the talk time? Mean or median?
  - How to evaluate the estimates? Which one is better?
  - Interval estimates?

iPhone Xs Max

- How to make a decision/inference on the statement/hypothesis: talk time is up to 25 hours?
  - If the estimate is 25.1 hours, could we say the talk time is up to 25 hours?
  - What about the estimate is 24.8 hours?
  - the estimate is 20 hours?
- Statistical inference will tell you
  - how small the estimated talk hours is enough to reject the hypothesis
  - how large the error we will make about our conclusion

Application of Statistical Inference

- Raleigh Housing Data
  - Information from the Raleigh Assessor's Office used in computing assessed values for individual residential properties sold in Raleigh, NC (北卡州那州) from 2006 to 2010
  - 2000 observations, 82 variables
  - 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables, 2 ID



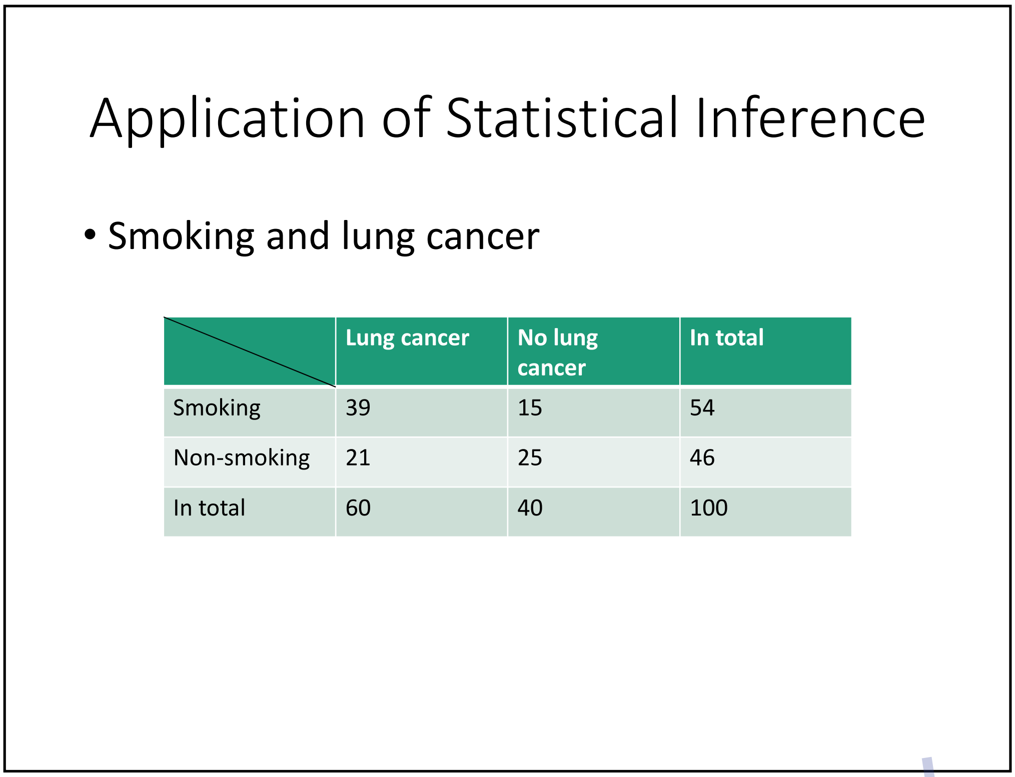
Order	PID	Lot_Area	Year_Built	Central_Air	Gr_Liv_Area	Pool_Area	Yr_Sold	SalePrice
1	526301100	31770	1960	Y	1656	0	2010	215000
2	526350040	11622	1961	Y	896	0	2010	105000
3	526351010	14267	1958	Y	1329	0	2010	172000
4	526353030	11160	1968	Y	2110	0	2010	244000
5	527105010	13830	1997	Y	1629	0	2010	189900
6	527105030	9978	1998	Y	1604	0	2010	195500
8	527145080	5005	1992	Y	1280	0	2010	191500
9	527146030	5389	1995	Y	1616	0	2010	236500
10	527162130	7500	1999	Y	1804	0	2010	189000
14	527180040	10176	1990	Y	1341	0	2010	171500

Application of Statistical Inference

- Raleigh Housing Data
  - Determine what drives the price of a house
    - Larger houses will fetch a higher price? But *how much more* does the price increase for each additional square foot?
    - Does the siding material (i.e., brick vs non-brick) have a significant impact on price?
    - Does it matter in which neighborhood the house is located?
  - Predict house price

Application of Statistical Inference

- Smoking and lung cancer



Application of Statistical Inference

- Smoking and lung cancer

	Lung cancer	No lung cancer	In total
Smoking	39	15	54
Non-smoking	21	25	46
In total	60	40	100