



中國人民大學  
RENMIN UNIVERSITY OF CHINA

# 计算传播理论与实务

2019-2020秋季学期

## 第二讲

## 统计思维与实用机器学习

授课教师：范举副教授、塔娜讲师

时间：2019年9月23日

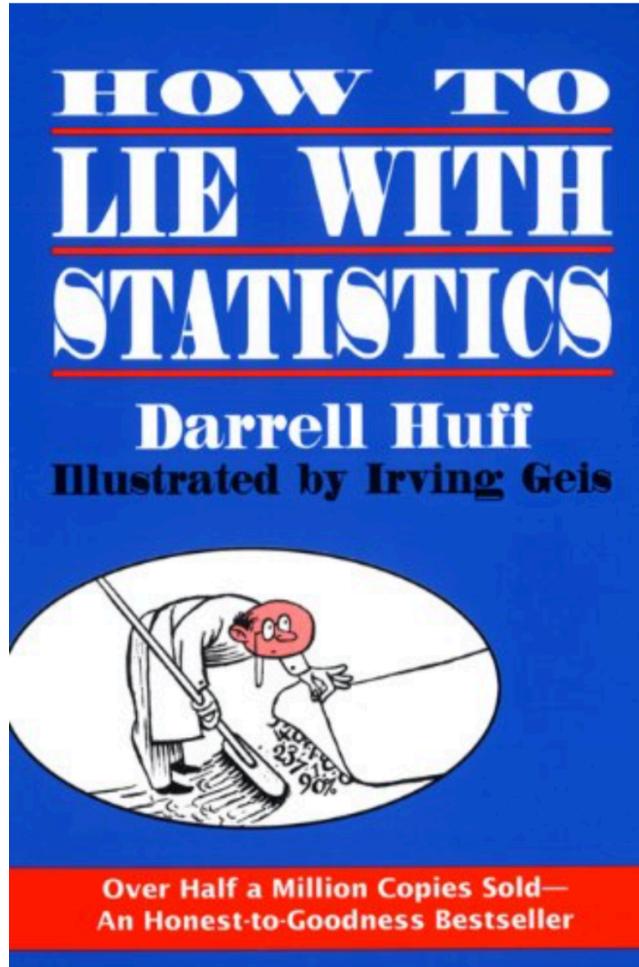
# 主要内容

- 统计思维
  - 描述统计 (Descriptive Statistics)
  - 推论统计 (Inferential Statistics)
- 实用机器学习
  - 研究设计 (Study Design)
  - 核心概念 (Conceptual Issues)
  - 代码实现 (Practical Implementation)

## 第2.1节

# 统计思维

# 为什么要有统计思维？



There are three kinds of lies:  
lies,  
damned lies,  
and statistics

你要做一个  
说谎  
的计算传播学家吗？

# 案例分析 - I

- 加州大学伯克利分校（UC Berkeley）的入学申请是否存在性别歧视？

	申请人数	录取率
男生	2691	45%
女生	1835	30%

~~YES!~~

# 案例分析 - I

- 加州大学伯克利分校 (UC Berkeley) 的入学申请是否存在性别歧视?

院系	男生		女生	
	申请人数	录取率	申请人数	录取率
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
合计	2691	45%	1835	30%

# 案例分析 - I

- 加州大学伯克利分校 (UC Berkeley) 的入学申请是否存在性别歧视?

院系	男生		女生	
	申请人数	录取率	申请人数	录取率
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
合计	2691	45%	1835	30%

NO!

- 女生录取率低是因为大部分女生去申请录取率低的院系!

# 案例分析 - 2

- 谁会赢得美国大选，希拉里还是特朗普？
  - 选前民调：希拉里支持率52%、特朗普支持率48%
  - 假设：所有选民在最终投票时不改变主意



~~WIN~~

# 案例分析 - 2

- 假设选前随机找了1000人进行民调
- 如果没有统计思维
  - 这1000人就是我要考虑的总体 (Population)
  - 统计支持希拉里的人数：520人
  - 统计支持特朗普的人数：480人
  - 结论：希拉里必胜！

# 案例分析 - 2

- 假设选前随机找了1000人进行民调
- 如果具备统计思维
  - 要考虑的**总体**是所有在当天投票的人
  - 这1000人是我抽取出的**样本** (Sample)
  - 通过样本估计出的支持率存在误差：
    - 希拉里： $52\% \pm 3\%$
    - 特朗普： $48\% \pm 2\%$
    -
  - 结论：特朗普也有胜算！

# 描述统计与推论统计

- 描述统计 (Descriptive Statistics)
  - 举例：选计算传播课同学的平均绩点
  - 目标：更好的**理解**数据
  - 方法：数据汇总、可视化、探索数据分析
- 推论统计 (Inferential Statistics)
  - 举例：选取100名同学估计学院的平均绩点
  - 目标：使用数据（样本）**学习**总体的特性
  - 方法：点估计、区间估计、假设检验

# 探索数据分析

- 探索数据分析 - Exploratory Data Analysis
  - 不断探索数据，进行描述统计分析的过程，目标是发现数据中的模式，从而更好地理解数据



Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those that we believe to be there

**From John Turkey**

# EDA需要警惕的“坑”

- 加州大学伯克利分校 (UC Berkeley) 的入学申请是否存在性别歧视？

	申请人数	录取率
男生	2691	45%
女生	1835	30%



院系	男生		女生	
	申请人数	录取率	申请人数	录取率
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
合计	2691	45%	1835	30%

- 不要迷信“客观”的数字
  - 基于描述统计，所有人给出的数字都是客观的
  - 只不过有的人可能更客观一些……

# EDA需要警惕的“坑”

	申请人数	录取率
男生	2691	45%
女生	1835	30%



院系	男生		女生	
	申请人数	录取率	申请人数	录取率
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
合计	2691	45%	1835	30%

- 辛普森悖论
  - 在分组比较中都占优势的一方，在总评中有时反而是失势的一方
  - 1951年，E.H.辛普森在他发表的论文中阐述此一现象



关我啥事……

# 课上练习 - I

- 假设你在某家研发移动端APP的公司实习
- APP用户中有10000人使用Android设备、5000人使用IOS设备，整体的付费转化率是5%。
- 细分发现其中：
  - IOS设备的转化率为4%
  - Android设备则是5.5%。
- “英明”的老板得出结论：IOS平台的用户付费转化率低下，建议放弃IOS平台的研发
- 可聪明的你通过生活经验知道IOS平台的用户更愿意付费，请问你如何用数据说服你的老板？



# 课上练习 - 2

- 步骤1：提出问题
  - iOS和Android平台的用户，谁付费转化率更高？
- 步骤2：收集数据
  - 请思考你要收集什么数据？

	Android手机	iOS手机	Android平板	iOS平板
转化	50	100	500	100
未转化	1950	3400	7500	1400

- 步骤3：分析数据

转化率	2.5%	2.9%	6.3%	6.7%
-----	------	------	------	------

# 课上练习 - 3

- 步骤4：原因分析
  - 汇总手机和平板的数据时，忽略了二者在“量”的差异——将“值与量”两个维度的数据，归纳成了“值”一个维度的数据

	Android手机	iOS手机	Android平板	iOS平板
转化	50	100	500	100
未转化	1950	3400	7500	1400
转化率	2.5%	2.9%	6.3%	6.7%

	Android设备	iOS设备
转化	550	200
未转化	9450	4800
转化率	5.5%	4%

普适的数据（如对比iOS和Android总体情况）没有多大参考意义，要细分到具体设备、获取渠道等再进行比对才有价值

# 课上练习 - 3

- 步骤5：形成报告
  - 假设你的老板看不懂这么复杂的图表.....
  - 对于占总体少数比例的样本加以更高的权重，也就是“逆概加权”（Inverse probability weighting）
  - 依旧是上面的例子，对于汇总的每个子群体加权，权重为该子群体在总群体里出现的概率的倒数

	iOS手机	iOS平板
转化	100	100
未转化	3400	1400

$$\frac{100 * \frac{5000}{3500} + 100 * \frac{5000}{1500}}{3500 * \frac{5000}{3500} + 1500 * \frac{5000}{1500}} = 4.8\%$$

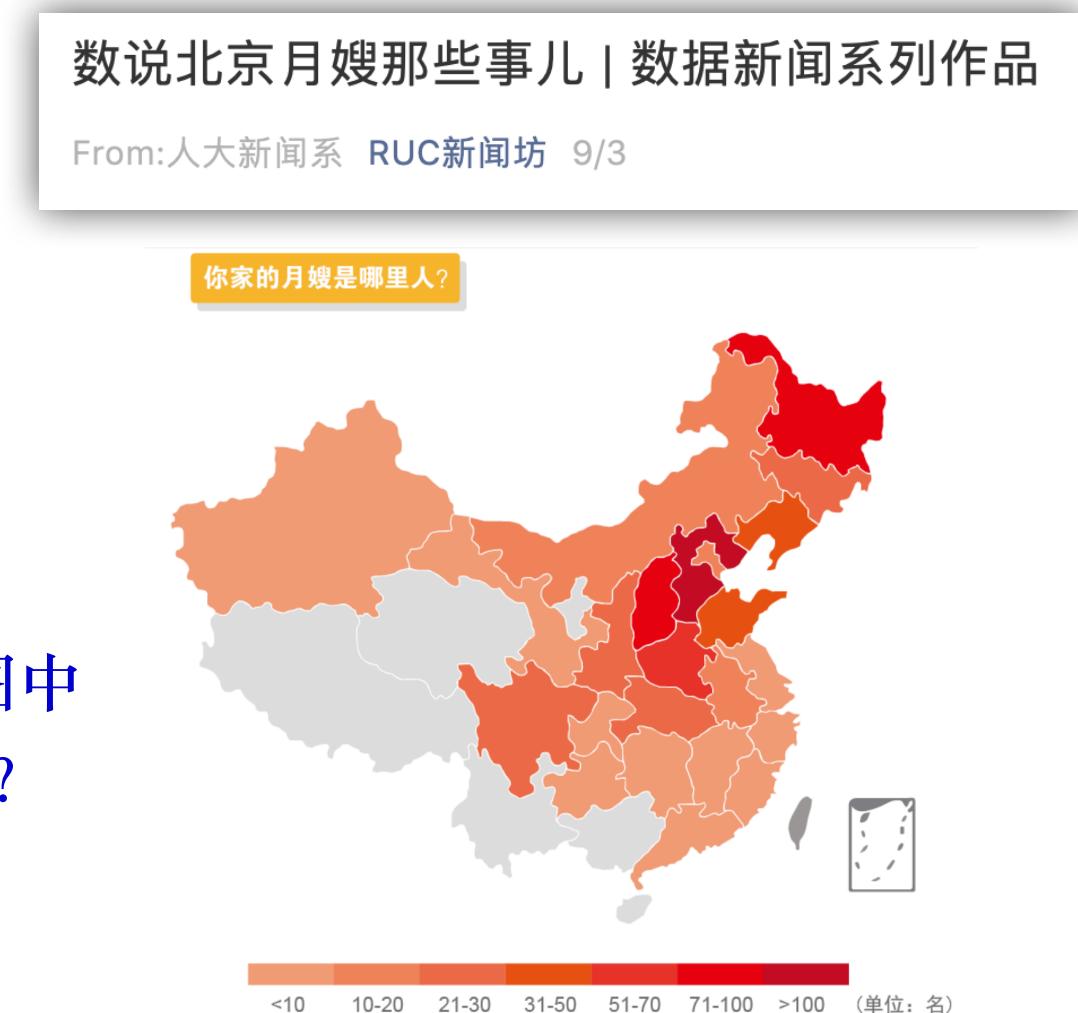
	Android手机	Android平板
转化	50	500
未转化	1950	7500

$$\frac{50 * \frac{10000}{2000} + 500 * \frac{10000}{8000}}{2000 * \frac{10000}{2000} + 8000 * \frac{10000}{8000}} = 4.4\%$$

# 探索数据的应用

- 数据新闻
  - 数据获取
  - 描述统计
  - 结果分析
  - 数据可视化

你能从右侧的图中  
得到什么结论？



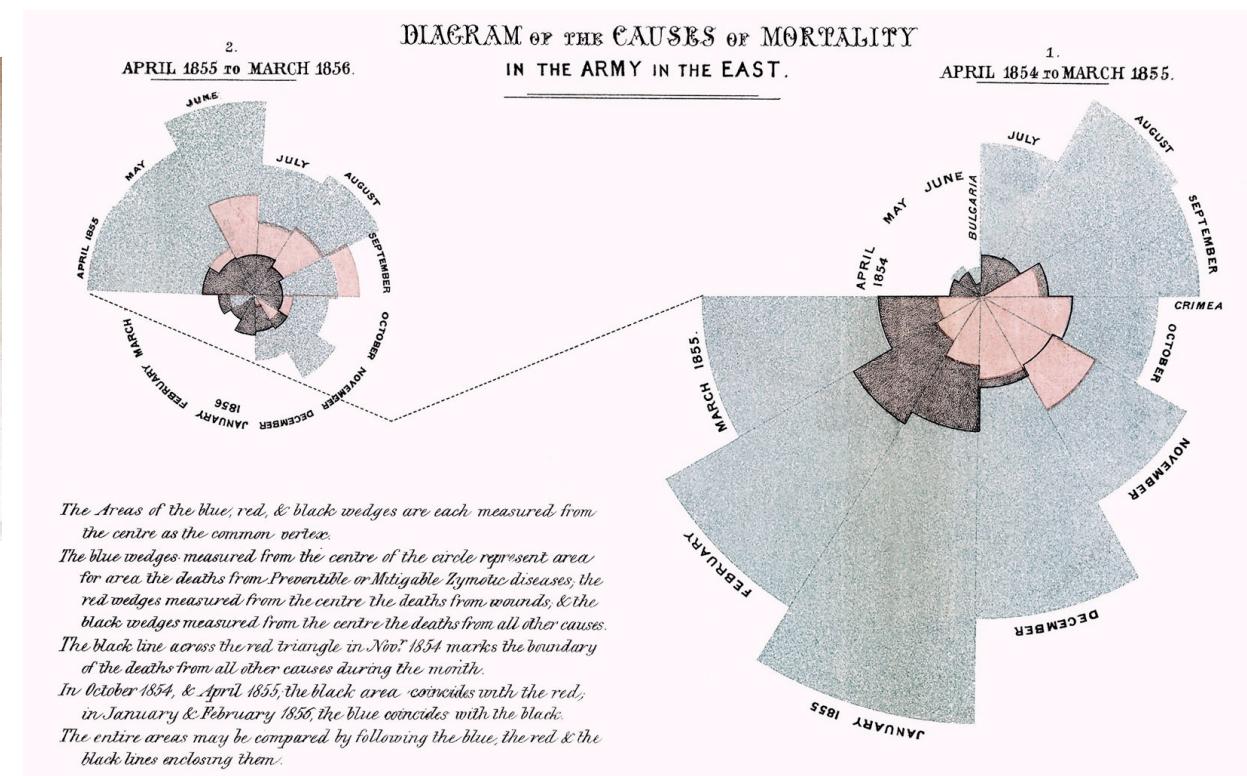
# 数据可视化

- 探索数据分析的重要手段



南丁格尔

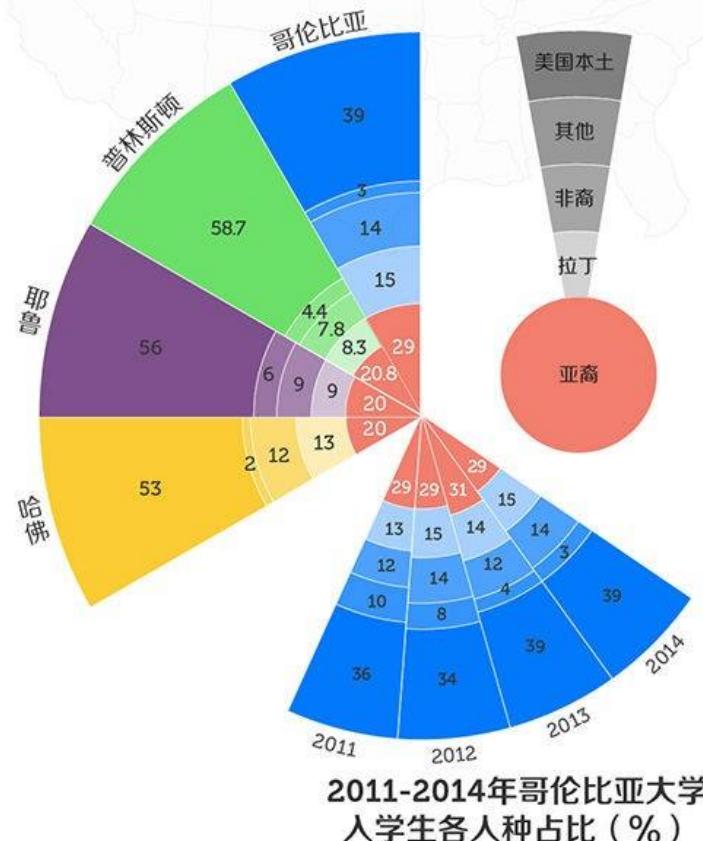
1820-1910



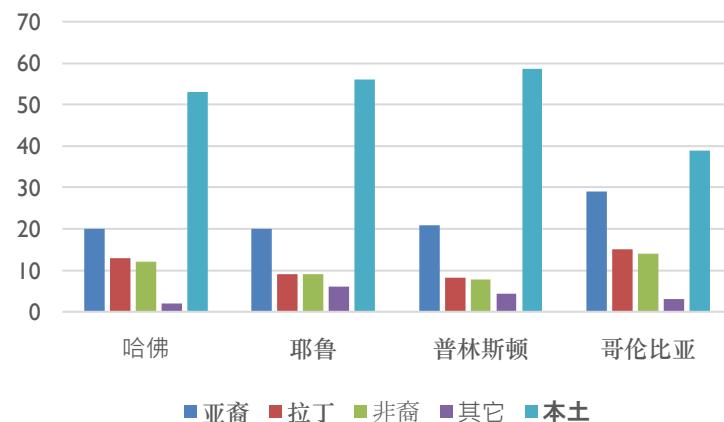
# 南丁格尔玫瑰图

美国名校已被中国学霸占领?  
没这回事

2014年北美名校入学生各人种占比 (%)



- 一种圆形的直方图



- 对比普通的直方图，你能发现什么？
  - 南丁格尔玫瑰图用长度表示数据，在面积上给人一定的误导！

# 思考题

- 观看数据可视化经典演讲
  - Hans Rosling's *200 Countries, 200 Years, 4 Minutes*
  - 视频链接: [http://www.iqiyi.com/w\\_19s0nozzyd.html](http://www.iqiyi.com/w_19s0nozzyd.html)
- 回答以下问题:
  - 图表中的横纵坐标、圆点、原点的大小分别表示什么含义
  - 图表的左下角和右上角分别表示什么含义
  - 图表中圆点聚在一起和比较分散分别表示什么含义
  - 放眼世界，你能从可视化中看出什么趋势
  - 聚焦中国，你能从可视化中观察到哪些趋势或特点

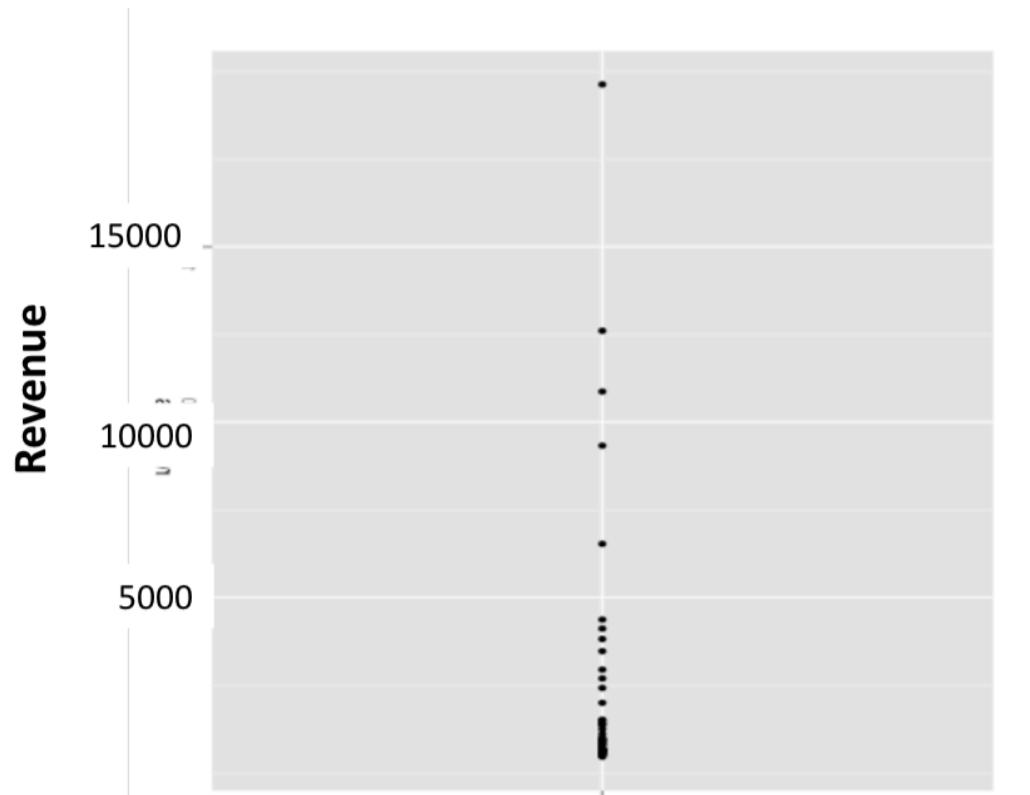
# 可视化图表类型

- 单变量
- 典型图表
  - 点阵图 - Dot plot
  - 直方图 - Histogram
  - 累积分布函数 - Cumulative distribution function
  - 箱线图 - Box plot

# 散点图

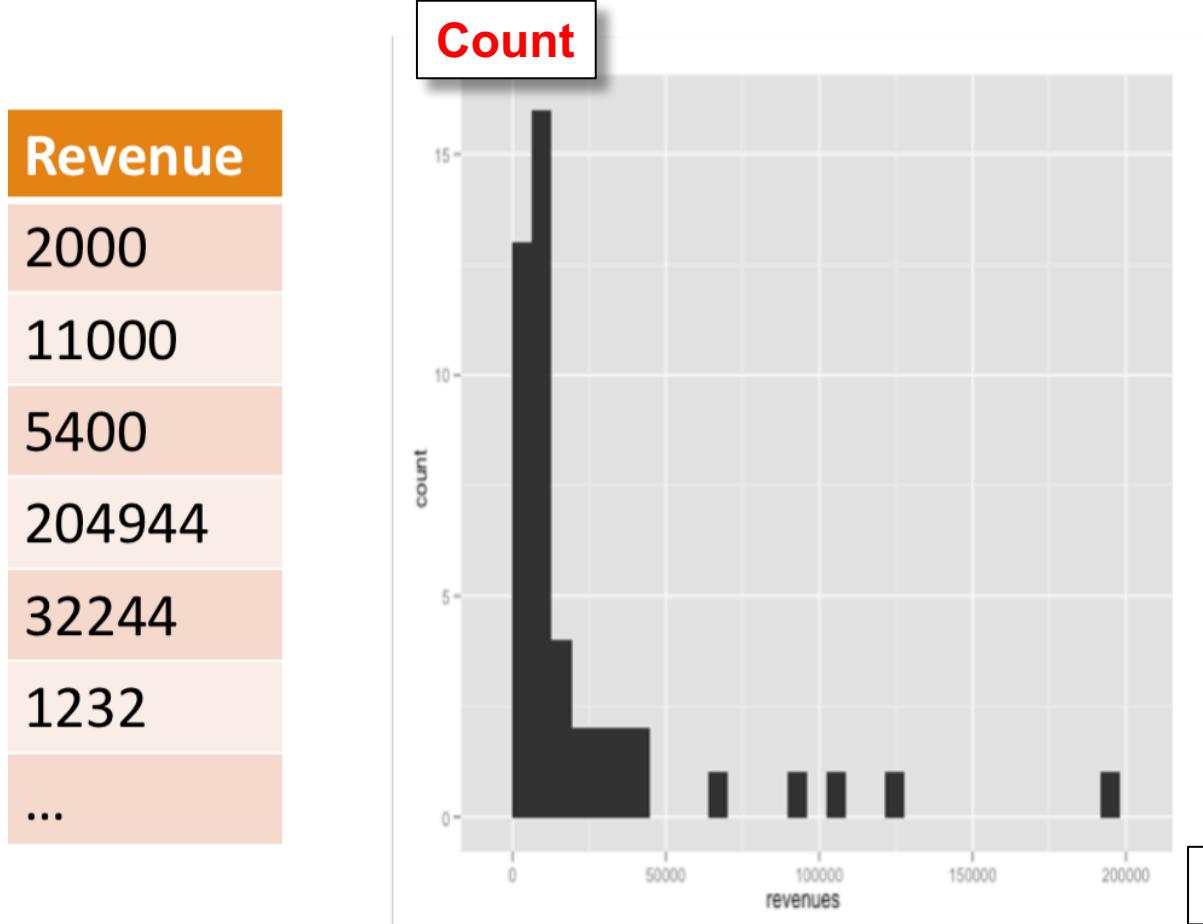
- 散点图将所有的数据以点的形式展现在直角坐标系上

Revenue
2000
11000
5400
204944
32244
1232
...



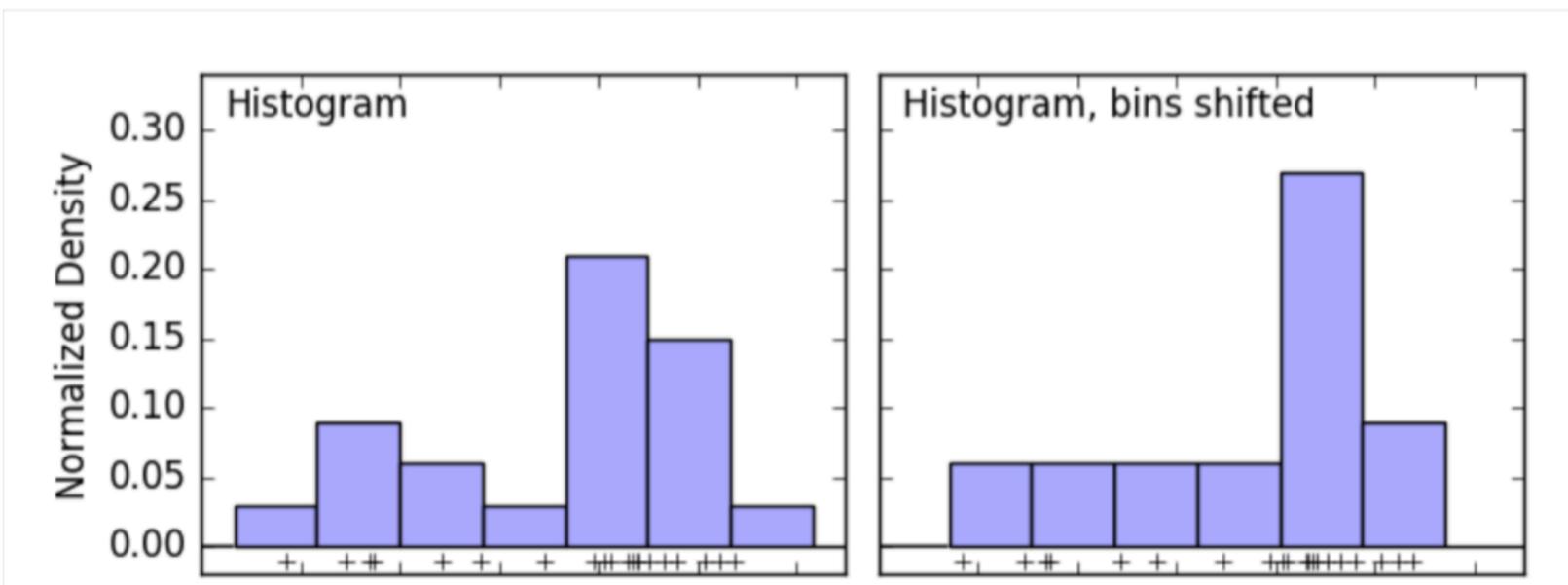
# 直方图

- 直方图表示了数据的分布情况



# 直方图的局限性

- 如何分桶（binning）会带来很大的视觉差异



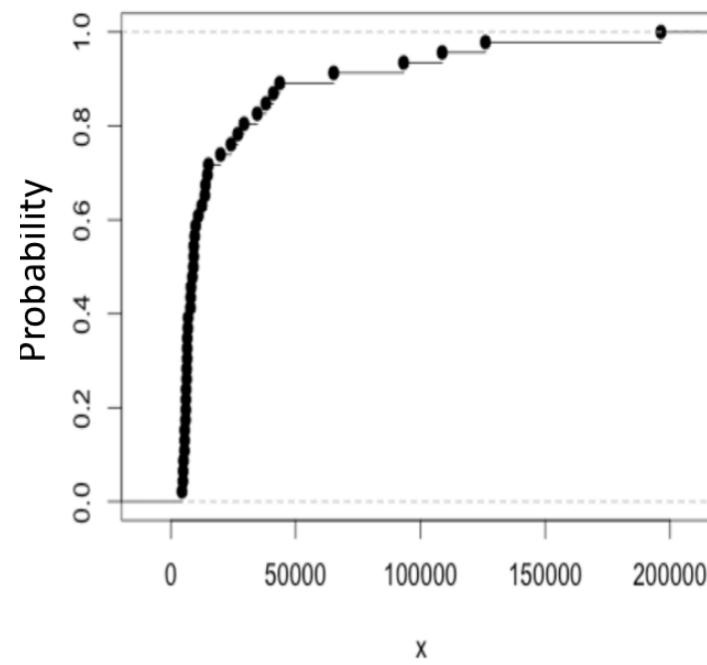
累积分布函数定义如下：

# 累积分布函数

- 对于连续变量 $x$ , 累积分布函数定义如下:  
$$F_X(x) = P(X \leq x).$$
- 离散变量: 所有小于等于 $a$ 的值出现概率的和

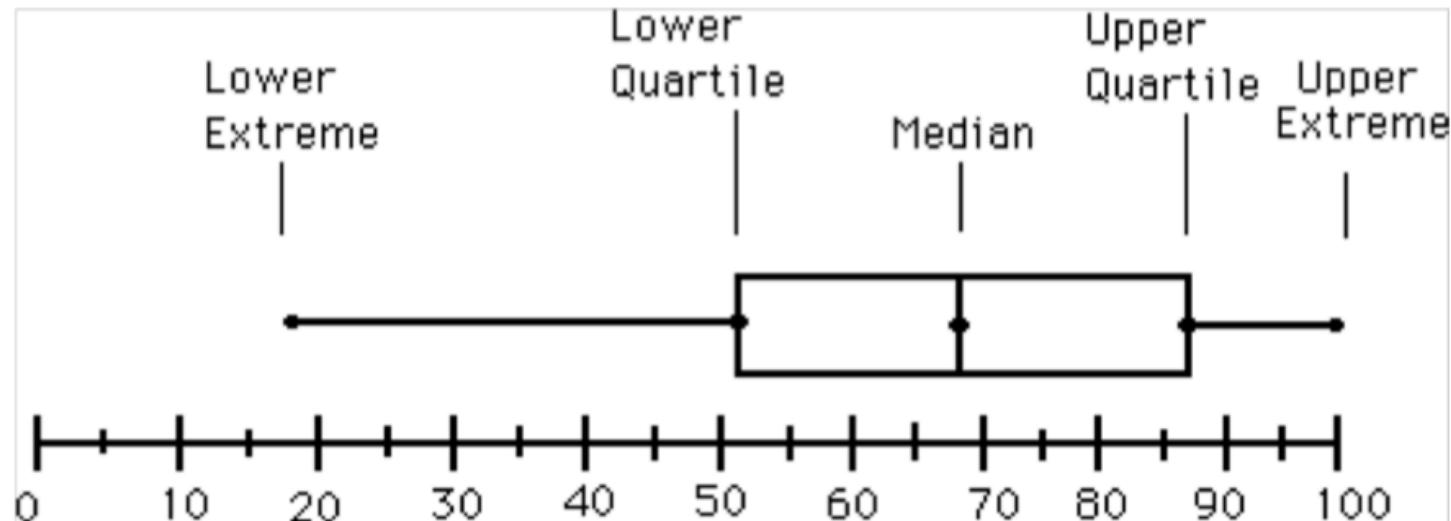
Revenue

2000
11000
5400
204944
32244
1232
...



# 箱线图

- 通过图形的方式表示数据的
  - Min, 25% Quartile, Median, 75% Quartile, Max



# 箱线图

- Find the median, lower quartile and upper quartile of the following numbers.
  - The **median** divides the data into a lower half and an upper half. The **lower quartile** is the middle value of the lower half. The **upper quartile** is the middle value of the upper half.

12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25

5, 7, 12, 14, 15, 22, 25, 30, 36, 42, 53  
↑                   ↑                   ↑  
lower quartile      median      upper quartile

# 课堂练习

排名	上年排名	公司名称	营业收入 (百万美元)	利润 (百万美元)	国家
1	1	沃尔玛 (WALMART)	514,405	6,670	美国
2	3	中国石油化工集团公司 (SINOPEC GROUP)	414,649.90	5,845	中国
3	5	荷兰皇家壳牌石油公司 (ROYAL DUTCH SHELL)	396,556	23,352	荷兰
4	4	中国石油天然气集团公司 (CHINA NATIONAL PETROLEUM)	392,976.60	2,270.50	中国
5	2	国家电网公司 (STATE GRID)	387,056	8,174.80	中国
6	--	沙特阿美公司 (SAUDI ARAMCO)	355,905	110,974.50	沙特阿拉伯
7	8	英国石油公司 (BP)	303,738	9,383	英国
8	9	埃克森美孚 (EXXON MOBIL)	290,212	20,840	美国
9	7	大众公司 (VOLKSWAGEN)	278,341.50	14,322.50	德国
10	6	丰田汽车公司 (TOYOTA MOTOR)	272,612	16,982	日本

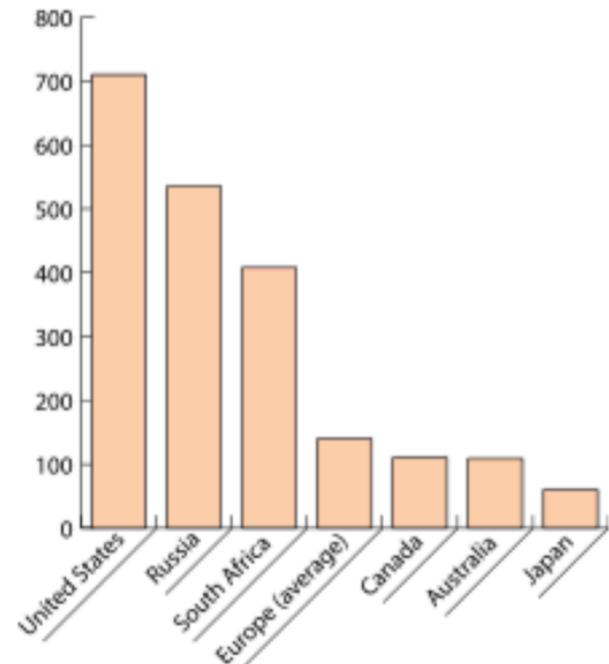
# 可视化图表类型

- 多变量
- 典型图表
  - 柱状图 - Bar chart
  - 散布图 – Scatter plot
  - 线条图 - Line plot

# 柱状图

- 一个变量为范畴型 (Categorical)

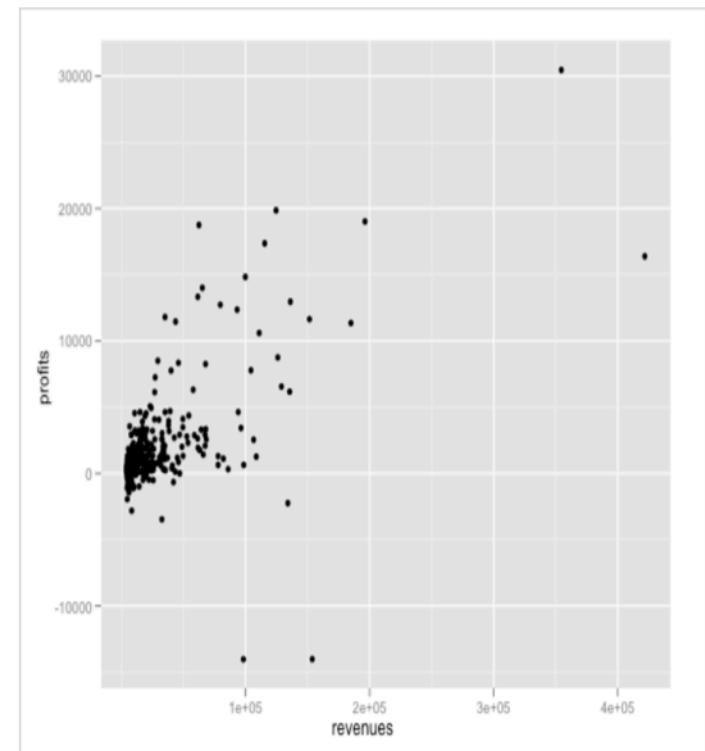
Region	Revenue
US	720
Russian	540
South Africa	400
Canada	120
	...



# 散布图

- 两个变量均为数值型

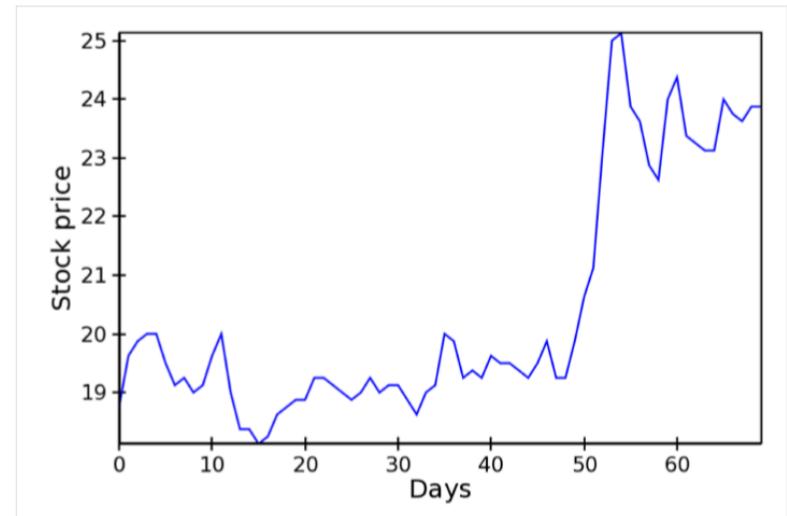
Revenue	Profit
20000	1000
45000	450
50234	-200
34522	900
	...



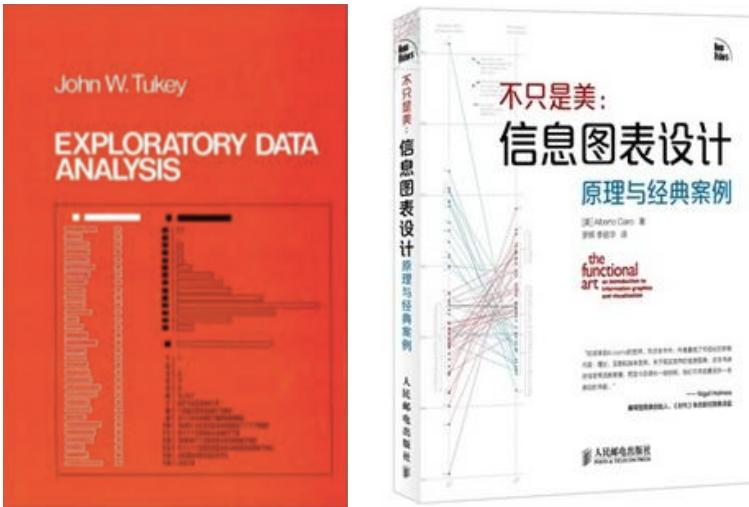
# 线条图

- 一个变量为数值型 - 看趋势

Days	Price
1	15.34
2	17.12
3	18.56
4	19.21
...	...

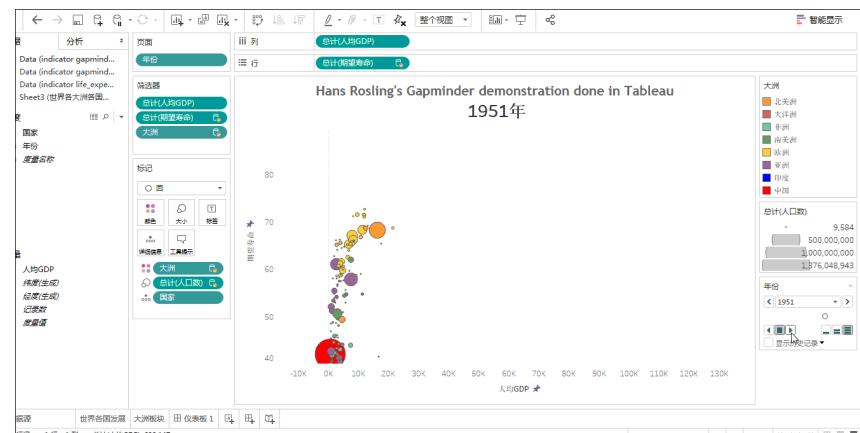


# 扩展阅读



- John W. Tukey. (1977). *Exploratory data analysis*. Pearson Modern Classic.
- Alberto Cairo (2015). 不只是美：信息图表设计原理与经典案例. 人民邮电出版社

- 实用工具
  - 探索数据分析：Tableau
  - 可视化
    - Highcharts
    - ECharts

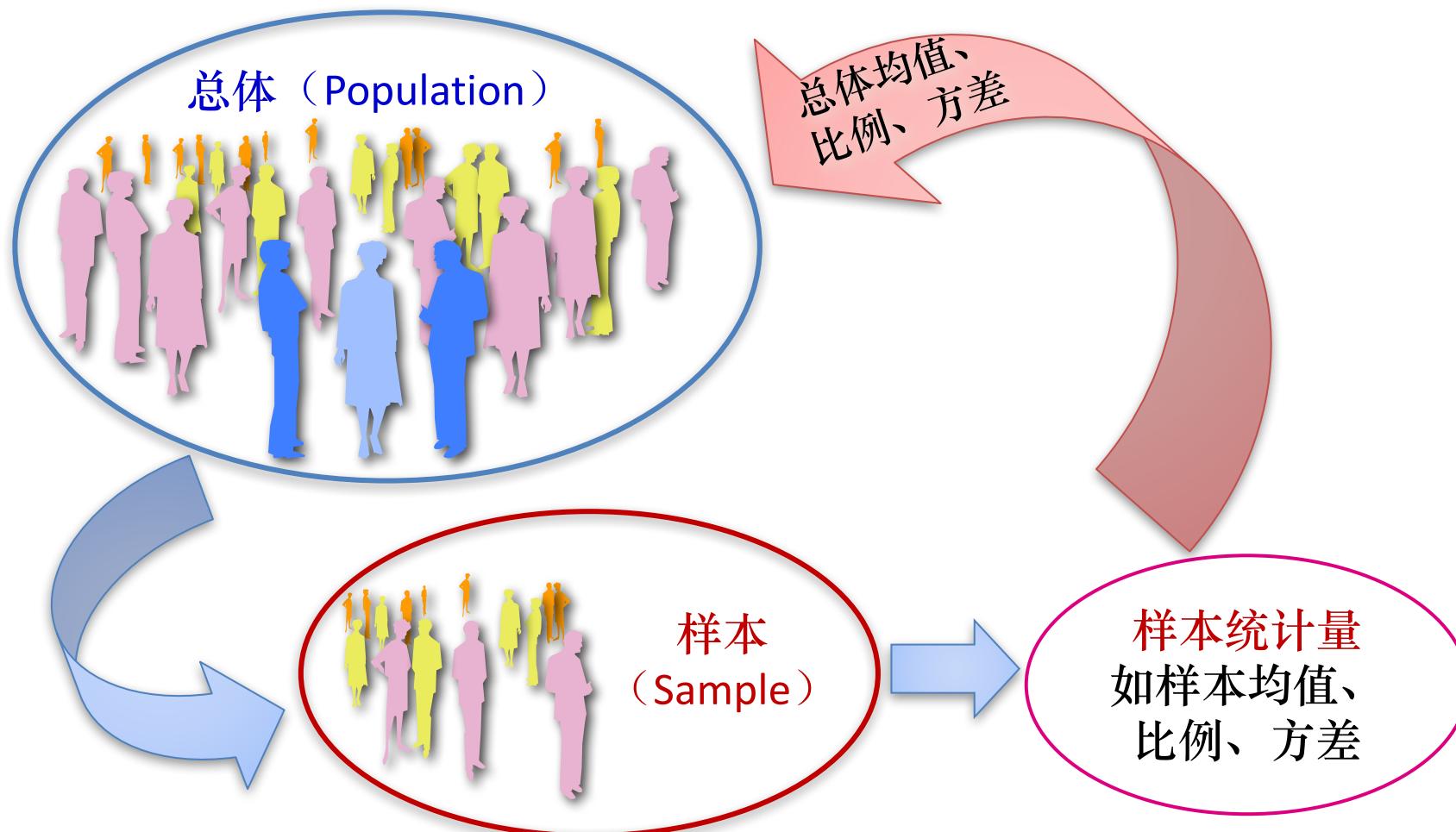


# 推论统计

- 推论统计 (Inferential Statistics)
  - 举例：选取100名同学估计学院的平均绩点
  - 目标：使用数据（样本）**学习**总体的特性
  - 方法：点估计、区间估计、假设检验
- 主要内容<sup>1</sup>
  - 总体 (Population) 与样本 (Sample)
  - 参数估计 (Estimation)
  - 假设检验 (Hypothesis Testing)
  - 回归分析 (Regression)

<sup>1</sup>说明：课上只介绍基本想法，具体方法细节请学习概率论与数理统计的相关教材。

# 推论统计的基本思路



# 回顾之前的例子

- 谁会赢得美国大选，希拉里还是特朗普？
  - 选前民调：希拉里支持率52%、特朗普支持率48%
  - 总体：所有在大选当天投票的选民
  - 样本：参与民调的1000人



# 图像分类的例子

- 图片中的动物是喵星人还是汪星人?
  - 从Web上收集了1000张图片，训练出分类模型，其分类的准确率为95%
  - 总体：所有在Web上的图片
  - 样本：从Web上收集到的1000张照片
- 思考：机器学习为什么要分训练集/测试集？



# 基本概念

- 总体 (Population)
  - 调查研究的事物或现象的全体
- 个体 (Item unit)
  - 组成总体的每个元素
- 样本 (Sample)
  - 从总体中所抽取的部分个体
- 样本容量 (Sample size)
  - 样本中所含个体的数量

# 抽样方法

- 概率抽样：根据已知的概率选取样本
  - 简单随机抽样：完全随机地抽选样本
  - 分层抽样：总体分成不同的“层”，在每一层内进行抽样
  - 整群抽样：将一组被调查者（群）作为一个抽样单位
  - 等距抽样：在样本框中每隔一定距离抽选一个被调查者
- 非概率抽样：不是完全按随机原则选取样本
  - 非随机抽样：由调查人员自由选取被调查者
  - 判断抽样：通过某些条件过滤来选择被调查者
- 配额抽样：选择一群特定数目、满足特定条件的被调查者

# 抽样分布

- 所有样本指标（如均值、比例、方差等）所形成的分布称为抽样分布
- 是一种理论上的概率分布
- 随机变量是 样本统计量
  - 样本均值, 样本比例等结果
  - 如何理解样本统计量, 如均值是个随机变量?
- 来自容量相同的所有可能样本

# 样本均值的抽样分布

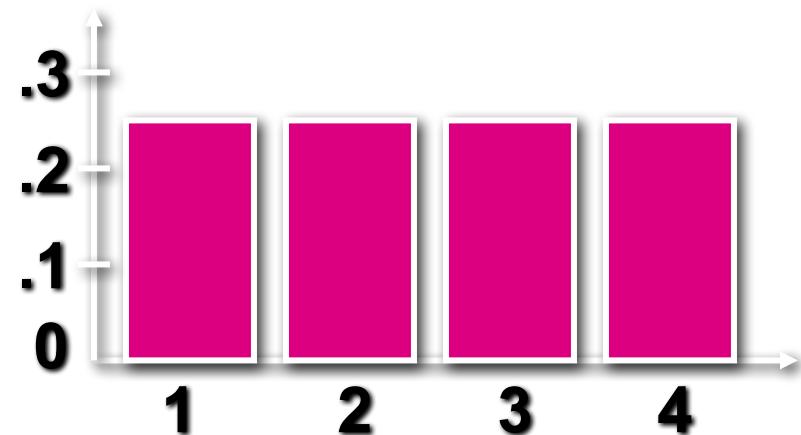
- 设一个总体，含有4个元素（个体），即总体单位数N=4。4个个体分别为X<sub>1</sub>=1、X<sub>2</sub>=2、X<sub>3</sub>=3、X<sub>4</sub>=4。总体的均值、方差及分布如下

均值和方差

$$\mu = \frac{\sum_{i=1}^N X_i}{N} = 2.5$$

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N} = 1.25$$

总体分布



# 样本均值的抽样分布

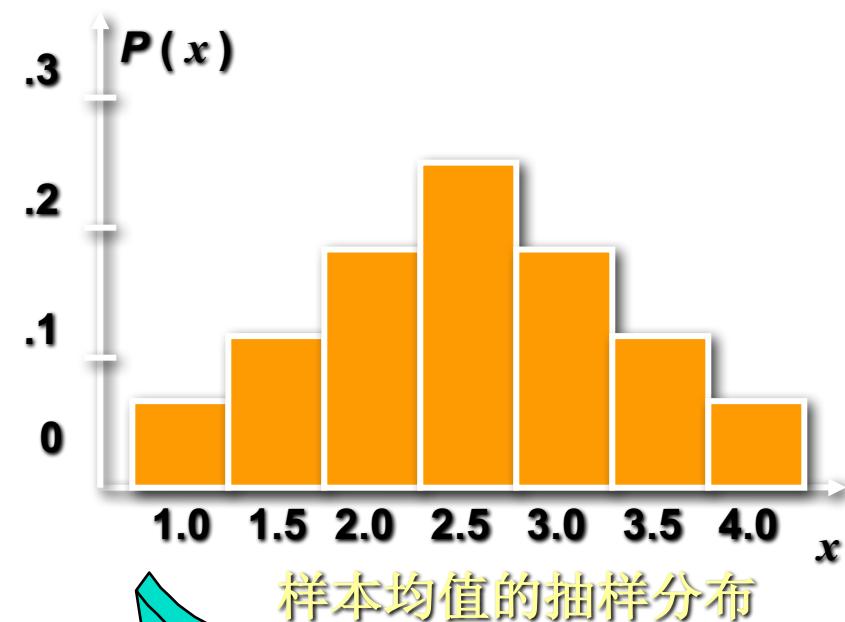
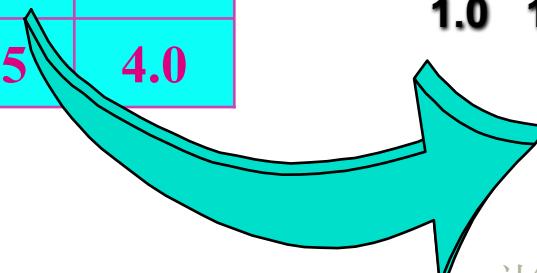
- 现从总体中抽取 $n=2$ 的简单随机样本，在重  
复抽样条件下，共有 $4^2=16$ 个样本。所有样  
本的结果如下表

所有可能的 $n = 2$ 的样本（共16个）				
第一个 观察值	第二个观察值			
	1	2	3	4
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,4
3	3,1	3,2	3,3	3,4
4	4,1	4,2	4,3	4,4

# 样本均值的抽样分布

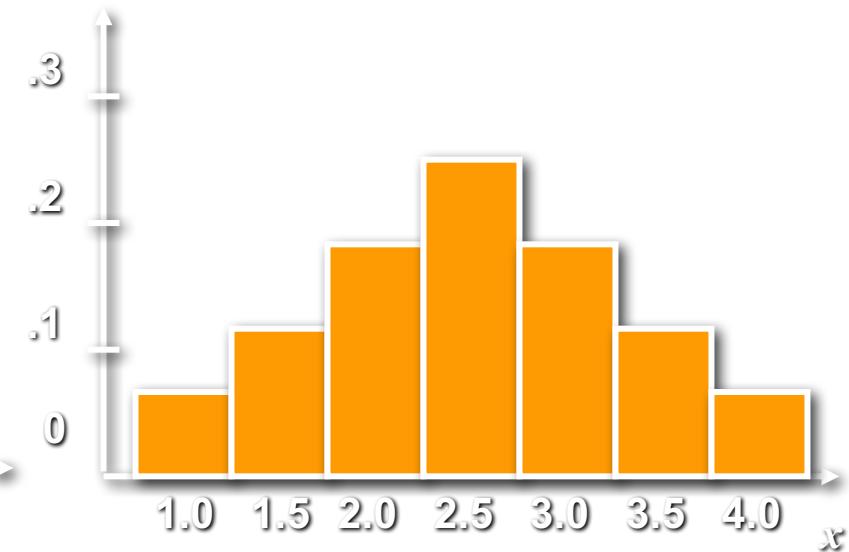
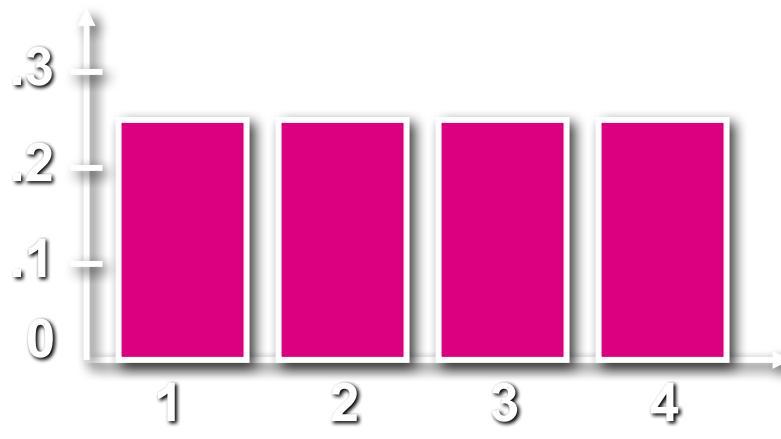
- 计算出各样的均值，如下表。并给出样本均值的抽样分布

16个样本的均值 ( $\bar{x}$ )				
第一个 观察值	第二个观察值			
	1	2	3	4
1	1.0	1.5	2.0	2.5
2	1.5	2.0	2.5	3.0
3	2.0	2.5	3.0	3.5
4	2.5	3.0	3.5	4.0



# 总体分布 vs. 抽样分布

- 比较



$$\mu = 2.5$$

$$\sigma^2 = 1.25$$

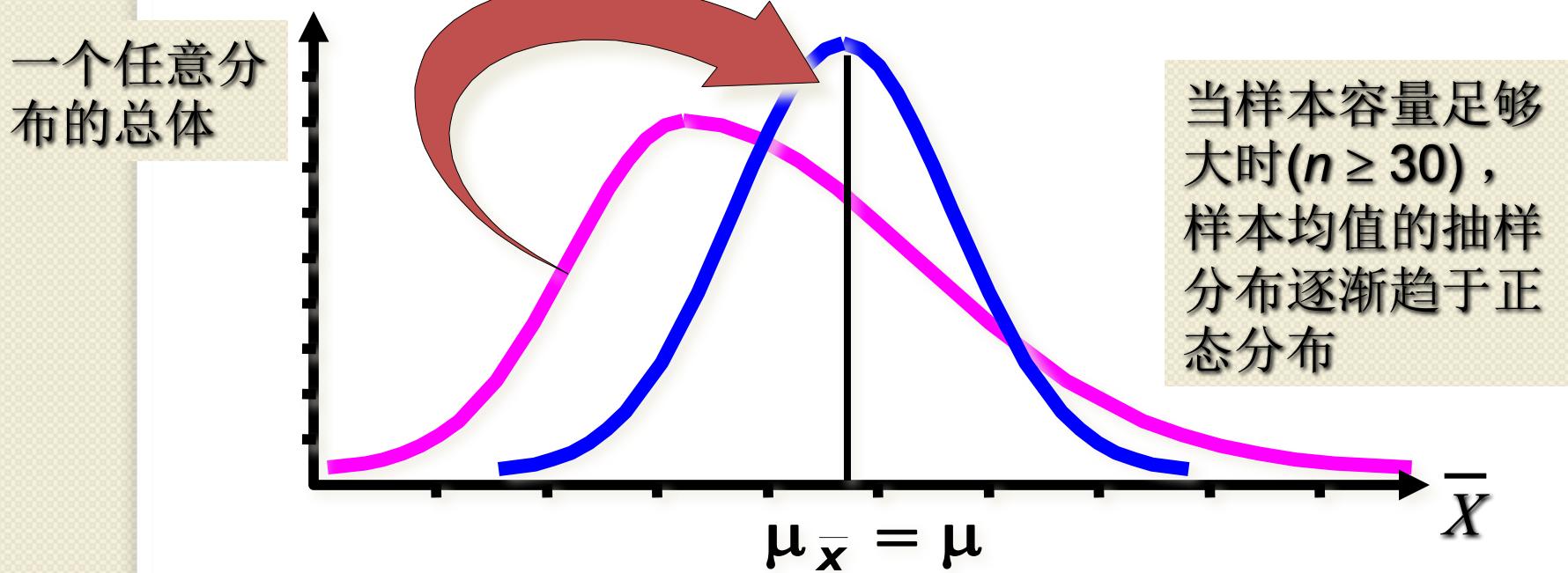
$$\mu_{\bar{x}} = \frac{\sum_{i=1}^n \bar{x}_i}{M} = \frac{1.0 + 1.5 + \dots + 4.0}{16} = 2.5 = \mu$$

$$\sigma_{\bar{x}}^2 = \frac{\sum_{i=1}^n (\bar{x}_i - \mu_{\bar{x}})^2}{M}$$

$$= \frac{(1.0 - 2.5)^2 + \dots + (4.0 - 2.5)^2}{16} = 0.625 = \frac{\sigma^2}{n}$$

# 中心极限定理

- 中心极限定理：设从均值为 $\mu$ ，方差为 $\sigma^2$ 的一个任意总体中抽取容量为n的样本，当n充分大时，样本均值的抽样分布近似服从均值为 $\mu$ 、方差为 $\sigma^2/n$ 的正态分布



# 点估计

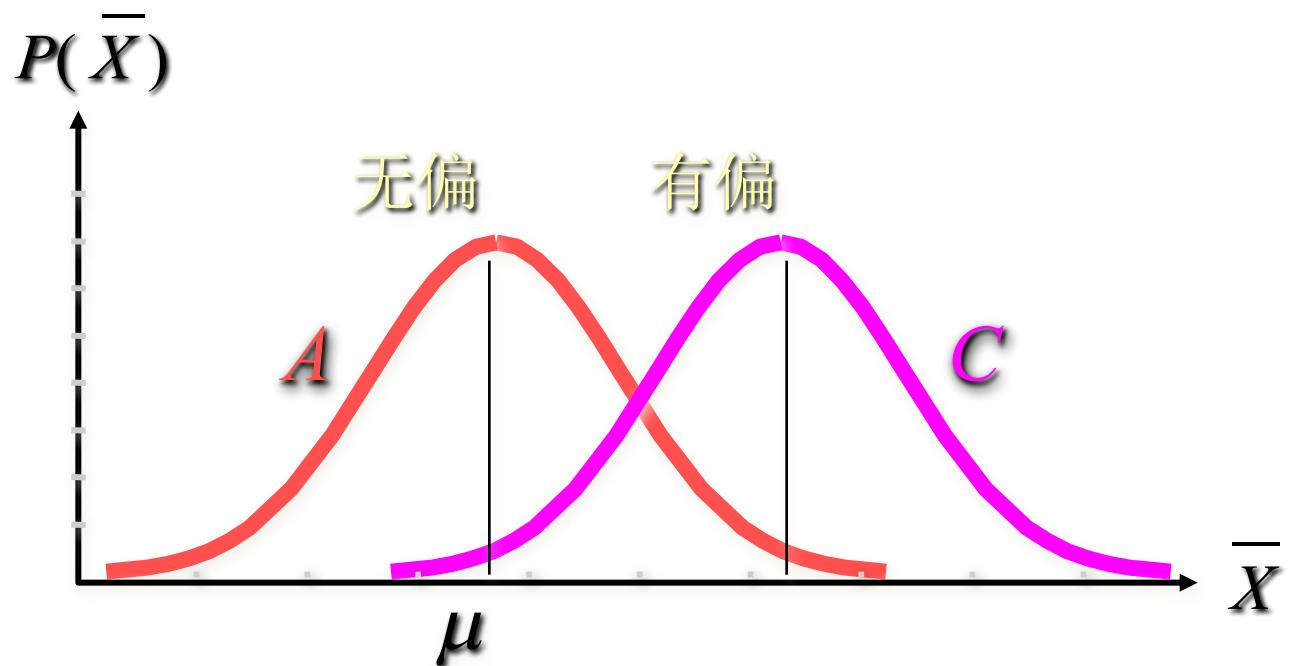
- 从总体中抽取一个样本，根据该样本的统计量对总体的未知参数作出一个数值点的估计
  - 例：用样本均值作为总体未知均值的估计值就是一个点估计
- 点估计没有给出估计值接近总体未知参数程度的信息
- 点估计的方法有矩估计法、顺序统计量法、最大似然法、最小二乘法等

# 点估计

- 估计量：用于估计总体某一参数的随机变量
  - 如样本均值，样本比例、样本中位数等
- 例如：样本均值就是总体均值 $\mu$ 的一个估计量
  - 如果样本均值 $x=3$ ，则 3 就是  $\mu$  的估计值
- 理论基础是抽样分布

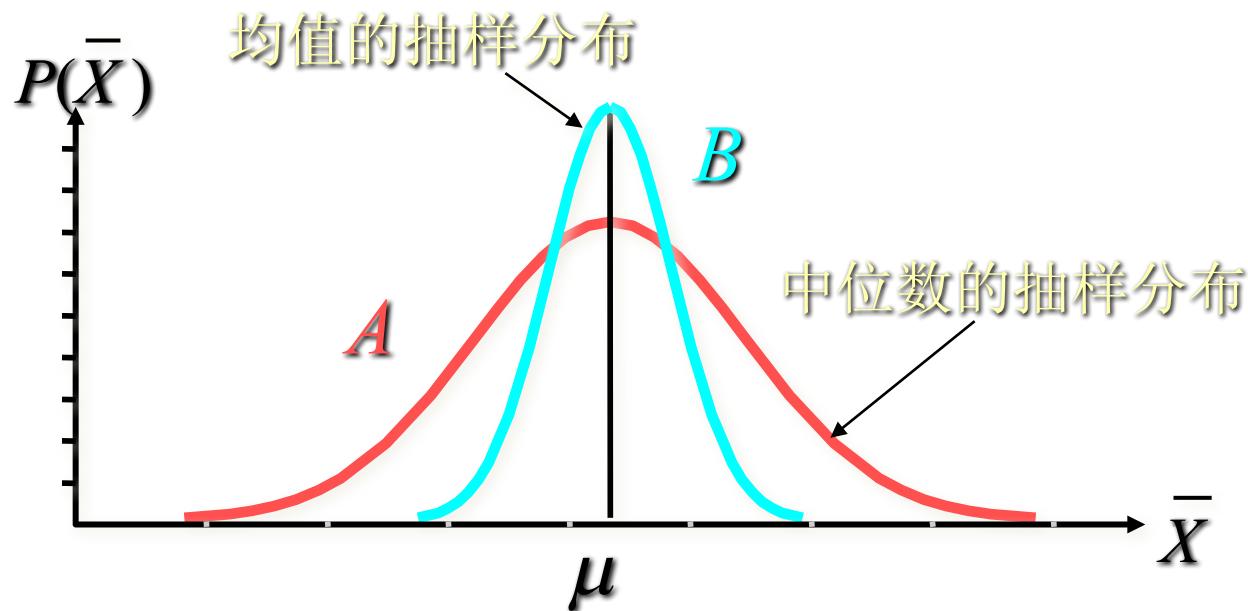
# 如何判断估计量的好坏

- 无偏性：估计量的数学期望等于被估计的总体参数



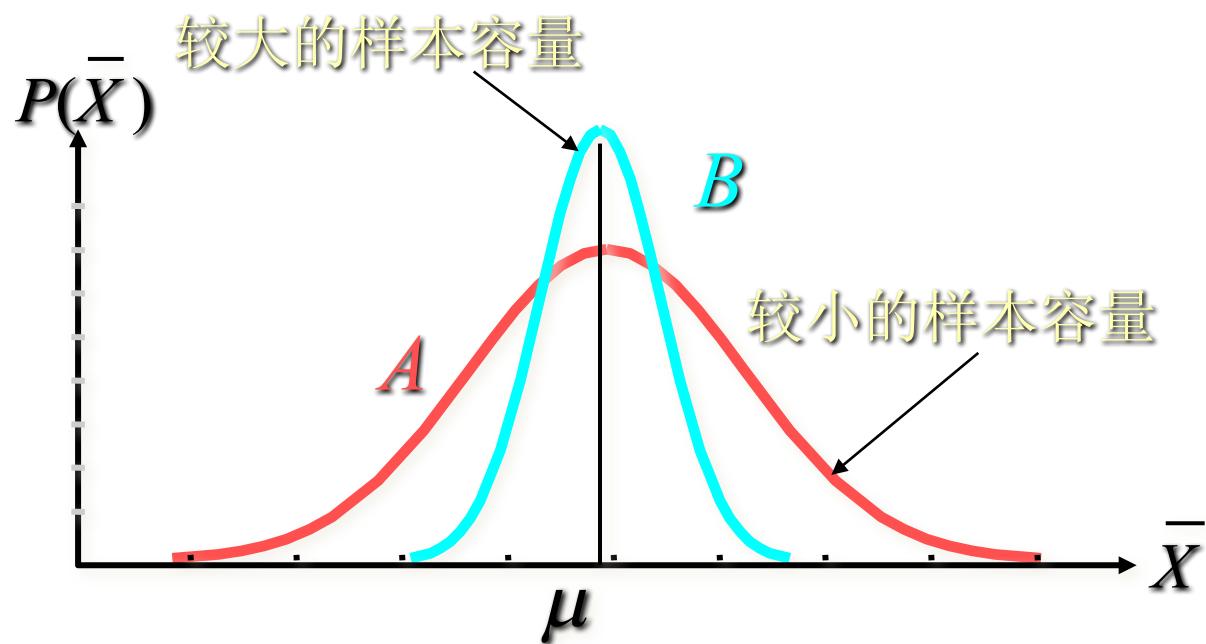
# 如何判断估计量的好坏

- 有效性：一个方差较小的无偏估计量称为一个更有效的估计量。如，与其他估计量相比，样本均值是一个更有效的估计量



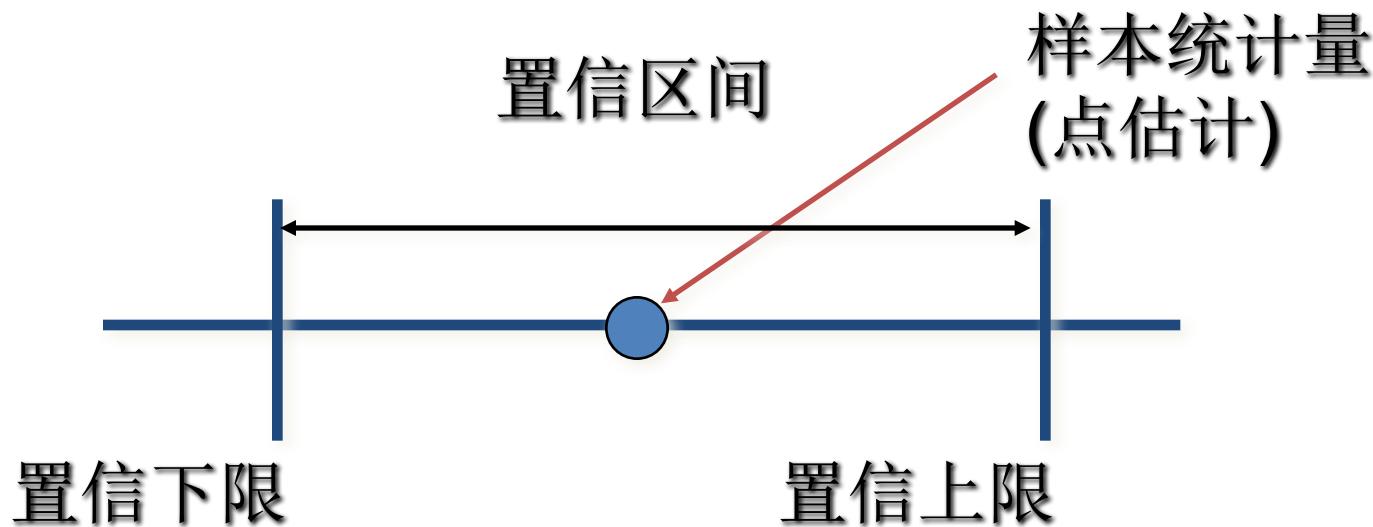
# 如何判断估计量的好坏

- 一致性：随着样本容量的增大，估计量越来越接近被估计的总体参数



# 区间估计

- 根据一个样本的观察值给出总体参数的估计范围
- 给出总体参数落在这一个区间的概率
  - 例如：总体均值落在50~70之间，置信度为 95%



# 总结

- 描述统计 (Descriptive Statistics)
  - 目标: 更好的理解数据
  - 方法: 数据汇总、可视化、探索数据分析
- 推论统计 (Inferential Statistics)
  - 目标: 使用数据 (样本) 学习总体的特性
  - 方法: 点估计、区间估计、假设检验
- 学习要点
  - 要求掌握: 基本原理与编程实现 (第一次作业)
  - 基本了解: 数学公式们.....