数据科学导论
Introduction to Data Science

第三章 数据统计

刘 淇
Email: qiliuql@ustc.edu.cn
课程主页：
http://staff.ustc.edu.cn/~qiliuql/DS2017.html

假设检验

- 假设检验
  - 假设检验（hypothesis testing）是统计中一种应用广泛的统计推断方法。
  - 它既和参数估计有一定的联系，又有明显的区别。
    - 都利用样本对总体进行推断，采用的技术手段相似；
    - 推断的出发点不同，结果也不同
      - 参数估计是用样本的统计来来估计总体参数的推断方法，待估计的总体参数在估计前是未知的；
      - 在假设检验中，通常先对待估计的总体参数提出一个假设，再利用样本去检验该假设是否成立。

假设检验

- 假设与否定论证
  - 假设检验首先根据实际问题提出原假设H_0和备择假设H_1。
    - 当假设的数学形式同总体分布的某个参数有关时，原假设H_0可用所关心的总体参数等于某个特定值表示，备择假设H_1可用该参数与特定值不相等（或大于，或小于）来表示。
    - 当假设的数学形式为未知总体的分布情况时，原假设H_0可用问题所关心的总体分布为某特定已知分布表示，备择假设H_1用该分布于特定的已知分布不相同来表示。

假设检验

- 案例分析
  - 有报道称，随着电子商务的快速发展，35.6%的中国人在 2015 年有过网购经历，达到 4 亿人。
  - 如何利用假设检验判断参数 35.6%的真实性？
  - 原假设为: $H_0 : \pi = 35.6\%$。
  - 对备择假设H_1，只要求参数值不等于某个特定值（35.6%）。
    - (1) $H_1 : \pi < 35.6\%$;
    - (2) $H_1 : \pi \neq 35.6\%$;
    - (3) $H_1 : \pi > 35.6\%$
  - 思考：在本问题中备择假设(2)是否有意义？什么时候用备择假设(1)？什么时候用备择假设(3)？

假设检验

- 案例分析
  - 如果 35.6% 的中国人在 2015 年有过网购经历这个假设是真实的，那么不支持这一假设的小概率事件在一次实验中是几乎不可能发生的。
  - 如果在一次实验中，不支持这一假设的事件发生了，则有理由怀疑假设本身的真实性，拒绝这一假设。
  - 故在提出原假设和备择假设之后，需要构造一个适当的能度量观测值与原假设下的期望数之间差异程度的统计量——**检验统计量**。以此计算小概率事件是否发生。

假设检验

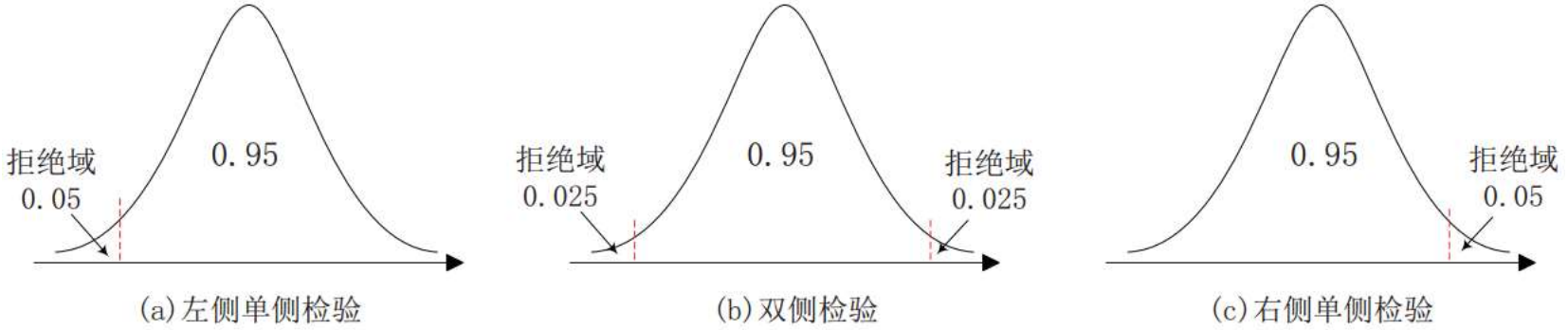
- 案例分析
  - 若样本中有网购经历的比例为 p，和假定的参数值(π_0) 进行对比，并构造检验统计量 Z:

$$Z = \frac{p - \pi_0}{se_0}, se_0 = \sqrt{\frac{\pi_0(1 - \pi_o)}{n}}$$

  - 其中 n 为样本大小。
  - 可通过计算样本中检验统计量 Z的值，与原假设理想分布下的值对比来判断是否发生了小概率事件。
  - 需要提前确定小概率事件的置信度，常用的置信度为 $\alpha = 0.05$，在这个置信度下，假设检验得到的结果出错概率为 5%。

假设检验

- 案例分析
  - 对于三种不同的备择假设 (1) $H_1: \pi < 35.6\%$；(2) $H_1: \pi \neq 35.6\%$；(3) $H_1: \pi > 35.6\%$，在显著性检验时的拒绝域是不同的，分别对应左侧单侧检验、双侧检验、和右侧单侧检验。



(a) 左侧单侧检验　(b) 双侧检验　(c) 右侧单侧检验

图中数字表示原假设 H_0 成立时，检验量 Z 落在该区间内的概率。

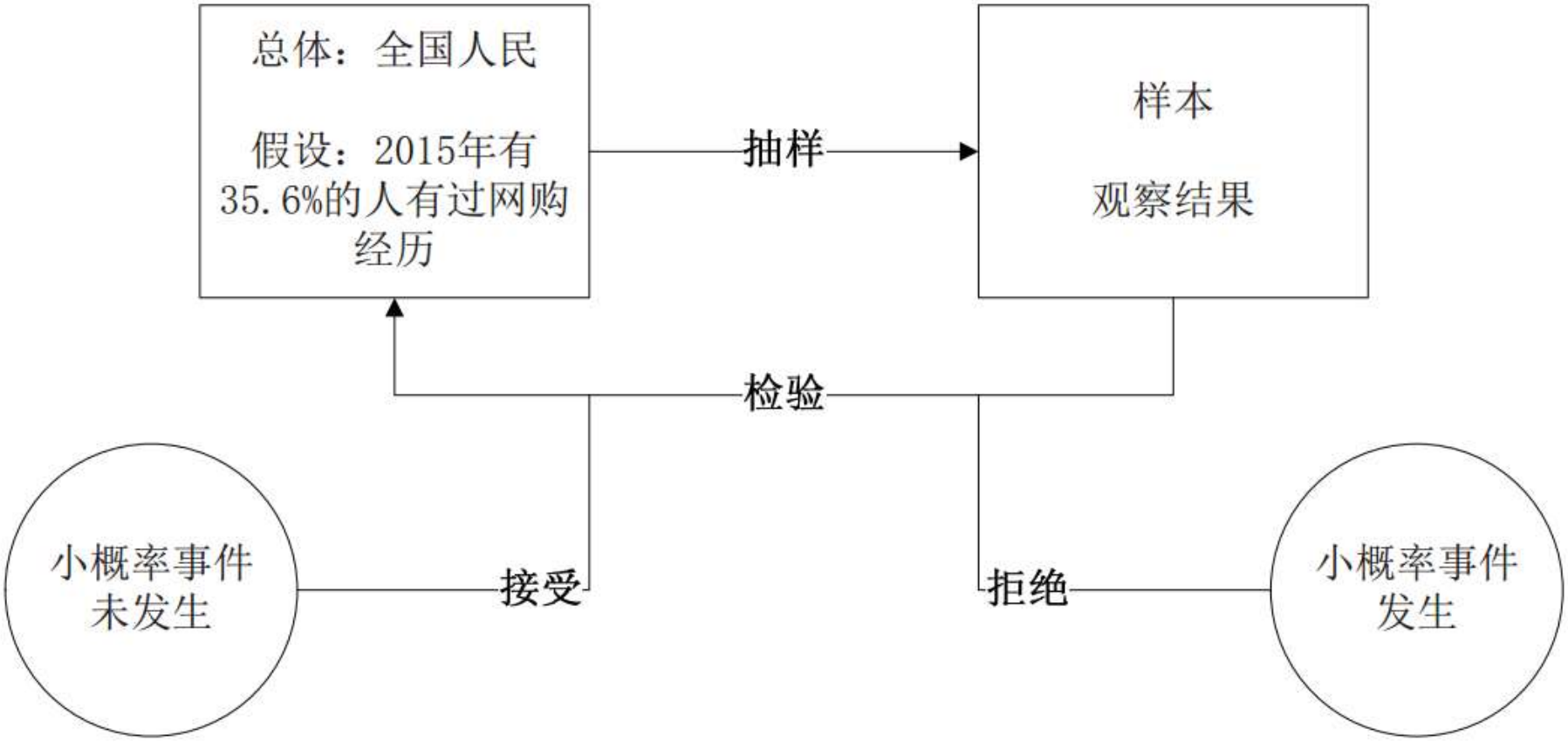
当由样本值计算出的统计量落入拒绝域则拒绝原假设H0，接受备择假设H1

假设检验

- 案例分析
  - 检验统计量落在拒绝域内是小概率事件。
  - 当这个小概率事件在某次检验中发生时，就认为其与实际推断相矛盾，拒绝原假设，接受备择假设；
  - 反之，若检验统计量为落在接受域，则接受原假设（但并不能说明原假设是正确的）。

假设检验

- 案例分析


总体：全国人民
假设：2015年有35.6%的人有过网购经历
抽样
样本
观察结果
检验
小概率事件未发生
接受
拒绝
小概率事件发生

假设检验

- 单总体假设检验

均值检验

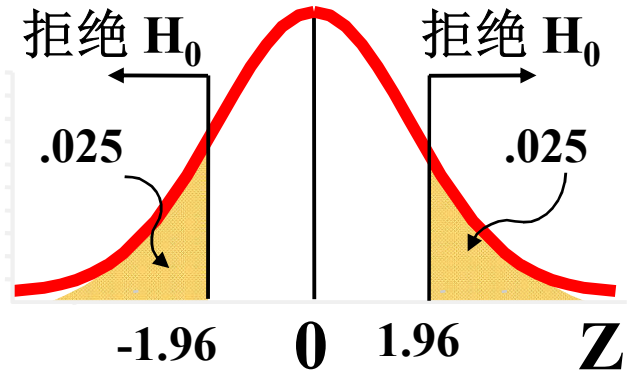
  - Z检验
    - 假定总体服从正态分布。
    - 即使总体不服从正态分布，由中心极限定理，当样本量足够大时(样本量n>30)，可用正态分布来近似。
  - Z-统计量
    - σ^2已知：$$Z=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\sim N(0,1)$$
    - σ^2未知：$$Z=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\sim N(0,1)$$

假设检验

- 例子
  - 某机床厂加工一种零件，根据经验知道，该厂加工零件的椭圆度近似服从正态分布，其总体均值为μ_0=0.081mm，总体标准差为σ= 0.025 。今换一种新机床进行加工，抽取n=200个零件进行检验，得到的椭圆度为0.076mm。试问新机床加工零件的椭圆度的均值与以前有无显著差异？（α＝0.05）

解：$H_0: \mu = 0.081$；$H_1: \mu \neq 0.081$

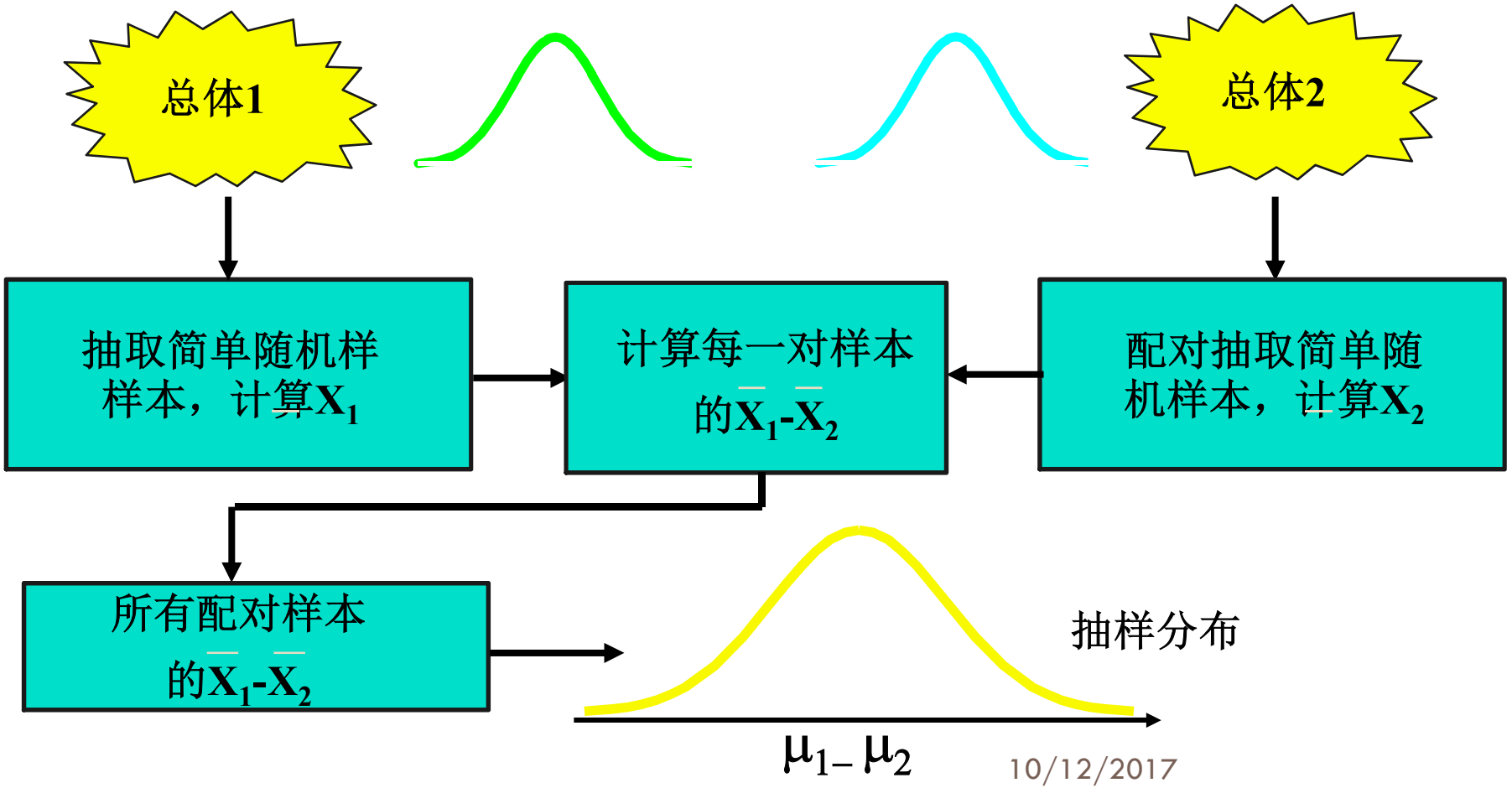
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{0.076 - 0.081}{0.025/\sqrt{200}} = -2.83$$



在 $\alpha = 0.05$的水平上拒绝H_0。新机床加工的零件的椭圆度与以前有显著差异。

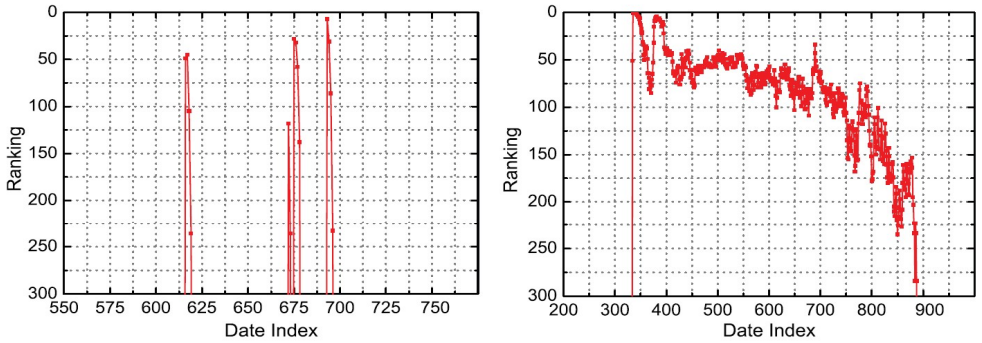
假设检验

- 双总体正态检验（例如，配对样本的 t 检验或z检验）



假设检验

- 应用举例---作为数据决策（分类、异常检测）特征



(a) Example 1 (b) Example 2

Figure 4: Two real-world examples of leading events.

▷ HYPOTHESIS 0: *The signature θ_s of leading session s is not useful for detecting ranking fraud.*

▷ HYPOTHESIS 1: *The signature θ_s of leading session s is significantly greater than expectation.*

Here, we propose to use the popular Gaussian approximation to compute the p-value with the above hypotheses.

- Hengshu Zhu, Hui Xiong, Yong Ge, and Enhong Chen, Discovery of Ranking Fraud for Mobile Apps, IEEE Transactions on Knowledge and Data Engineering (***IEEE TKDE***), 27(1): 74-87, January 2015.

假设检验

- 应用举例---验证推荐结果的有效性

User Study Ratings

	LUCF	LBSVD	TTER	TASTContent	Cocktail
Mean	3.22	3.30	3.46	3.20	**3.55**
SD	0.74	0.75	0.81	**0.94**	0.76

applying z-test, we find that the differences between the ratings obtained by Cocktail and the other algorithms are statistically significant with $|z| \geq 2.58$ and thus $p \leq 0.01$

- Qi Liu, Enhong Chen, Hui Xiong, Yong Ge, Zhongmou Li, Xiang Wu, A Cocktail Approach for Travel Package Recommendation, *IEEE Transactions on Knowledge and Data Engineering* ***(IEEE TKDE)***, 26(2): 278-293, 2014.

假设检验

- 课外阅读（不做强制要求）
  - 更多关于均值的检验
    - 双总体均值之差的Z检验（非配对）
  - 关于比例的检验
    - 单总体比例的Z检验
    - 双总体比例之差的Z检验
  - 关于方差的检验
    - 单总体方差的χ^2检验
    - 双总体方差比的F检验

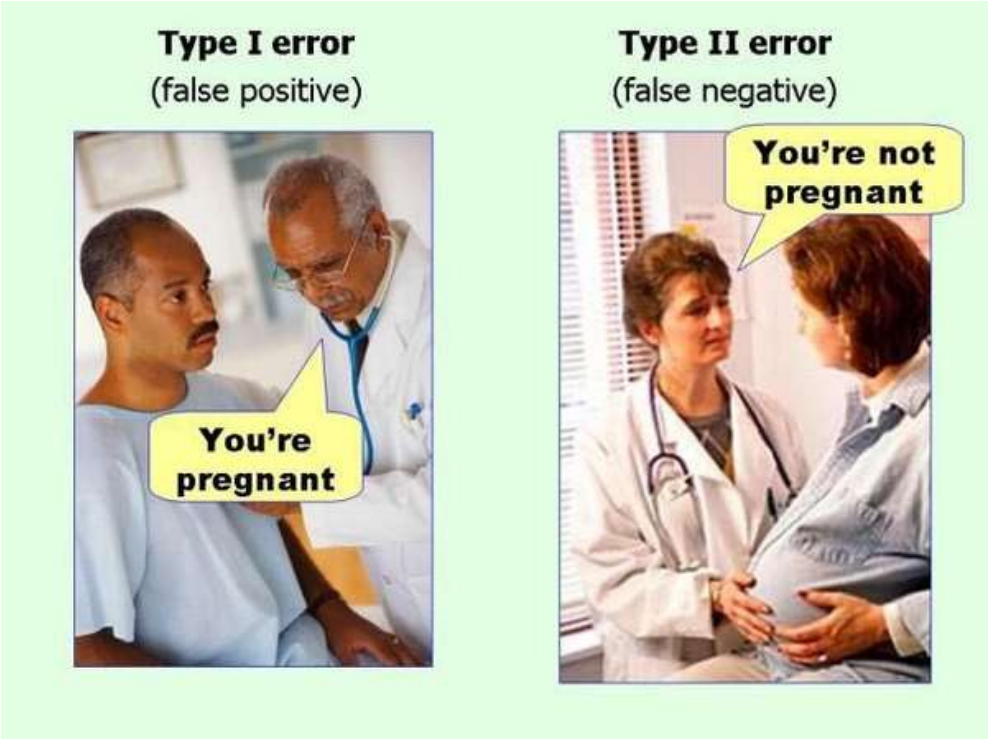
假设检验

- 两类错误
  - 假设检验需要对原假设提出的命题做出“拒绝原假设”或“不拒绝原假设”的判断，并且这种判断来自于样本的信息，是由部分推断总体。
  - 原假设是正确的但检验结果拒绝了原假设，称为弃真错误 α。
  - 原假设是错误的但检验结果没拒绝原假设，称为取伪错误 β。

项目	没有拒绝 H_0	拒绝 H_0
H_0为真	$1-\alpha$（正确决策）	α（弃真错误）
H_0为伪	β（取伪错误）	$1-\beta$（正确决策）

假设检验

- 两类错误
  - 两类错误造成的后果可能是不一样严重的。



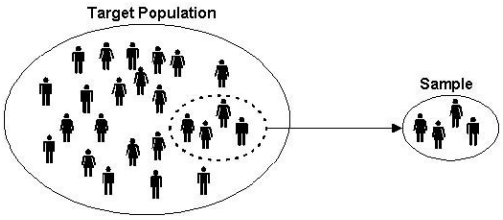
假设检验

- 两类错误
  - 假设检验的过程希望判断的结果犯错率越低越好。
  - 但对于一定量的样本 n，一个类型错误的错误率降低伴随着的是另一个类型错误犯错率的增加。
  - 哪一类错误所造成的后果更严重，在假设检验中就应当把哪一类错误作为首要的控制目标。
    - α 错误的犯错率即为置信度，降低置信度就可以降低这一类错误的犯错率；
    - β错误则是由很多客观因素造成的，难以明确表示。

增大样本量可以使得两类错误同时减小！

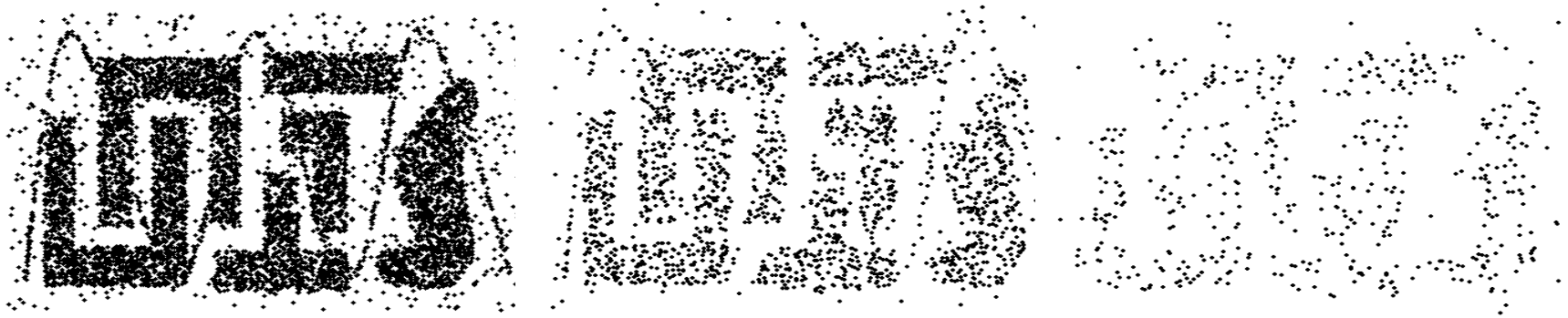
抽样方法

- 前面介绍了对数据的一些统计方法。但在分析数据之前，还有一个很重要的部分就是采集数据。
- 由于人力物力的限制，抽样的数量是有限的，好的抽样方法可以通过较少的样本数量反映正确的总体信息。
- 抽样结果是否具有代表性，决定了通过数据得到对总体的认知是否是合适的。



抽样方法

- 抽样
  - 抽样是通过抽取总体中的部分个体，收集这些个体的信息，从而对总体进行推断的一种手段。



8000 points　　2000 Points　　500 Points

- 常见抽样方法：
  - 非概率抽样
  - 等概率抽样
  - 不等概率抽样

抽样方法

- 非概率抽样
  - 抽取样本时不是依据随机原则，而是根据研究目的对数据的要求，采用某种方式从总体中抽出部分单位对其实施调查

- 常见非概率抽样方法：
  - 随意抽样——随便选择抽样对象
  - 判断抽样——由抽样人制定抽样对象
  - 志愿抽样——以志愿者为对象抽样
  - 滚雪球抽样——由被抽样对象推荐其他被抽样对象。

抽样方法

- 非概率抽样缺点：
  - 抽取样本有主观性，使结果有偏差；
  - 不可能计算各个元素的入样概率，无法得到可靠的估计值及抽样误差估计值，不能推断总体。

抽样方法

- 等概率抽样
  - 每一个单元的入样概率均相等；
  - 等概率抽样的基本出发点是将总体（或层）中的每一个单元看作是平等的，不“偏向”也不“疏远”某些特定的单元。
  - 如果总体单元差异不大，这种方式既简单也合理

- 常见等概率抽样
  - 简单随机抽样
  - 分层抽样
  - 整群抽样
  - 系统抽样

抽样方法

- 简单随机抽样
  - 一般地，设一个总体的个体数为N。如果通过逐个抽取的方法从中抽取一个样本，且每次抽取时各个个体被抽到的概率相等，就称这样的抽样为简单随机抽样。
    - 要求被抽取样本的总体的个体数有限；
    - 是从总体中逐个进行抽取；
    - 是不放回抽样。
    - 如果抽取一个容量为n的样本，那么每个个体被抽取的概率等于$\frac{n}{N}$。
  - 抽取方法：抽签法、随机数表法

抽样方法

- 抽签法
  - 将总体中的所有个体（共N个）编号（号码可以从1到N）。
  - 利用小球、卡片、纸条等手段随机抽取。

简单随机抽样并不是随意或随便抽取，因为随意或随便抽取都会带有主观或客观的影响因素！

抽样方法

- 随机数表法
  - 随机数表是统计工作者用计算机生成的随机数，并保证表中的每个位置上的数字是近似等可能出现的。
  - 随机数表并不是唯一的，因此可以任选一个数作为开始，读数的方向可以向左，也可以向右、向上、向下等等。

- 随机数表示例

16 22 77 94 39 49 54 43 54 82 17 37 93 23 78 87 35 20 96 43 84 26 34 91 64
84 42 17 53 31 57 24 55 06 88 77 04 74 47 67 21 76 33 50 25 83 92 12 06 76
63 01 63 78 59 16 95 55 67 19 98 10 50 71 75 12 86 73 58 07 44 39 52 38 79
33 21 12 34 29 78 64 56 07 82 52 42 07 44 38 15 51 00 13 42 99 66 02 79 54
57 60 86 32 44 09 47 27 96 54 49 17 46 09 62 90 52 84 77 27 08 02 73 43 28

抽样方法

- 分层抽样
  - 当已知总体由差异明显的几部分组成时，为了使样本充分地反映总体的情况，常将总体分成几部分，然后按照各部分所占的比例进行抽样。其中所分成的各部分叫做层。
    - 分层抽样的一个重要问题是一个总体如何分层。
    - 分层抽样中分多少层，要视具体情况而定。
    - 总的原则是：层内样本的差异要小，而层与层之间的差异尽可能地大，否则将失去分层的意义。
    - 既可以对总体参数进行估计，也可以对各层的目标量进行估计

抽样方法

- 例子
  - 一个单位的职工有500人，其中不到35岁的有125人，35～49岁的有280人，50岁以上的有95人。为了了解该单位职工身体状况的有关指标，从中抽取100名职工作为样本，应该怎样抽取？
  - 解：抽取人数与职工总数的比是100：500＝1：5，则各年龄段（层）的职工人数依次是125：280：95＝25：56：19，然后分别在各年龄段（层）运用简单随机抽样方法抽取。

抽样方法

- 整群抽样
  - 将总体全部单位分为许多个“群”，然后随机抽取若干“群”，对被抽中的各“群”内的所有单位登记调查。
    - 抽样时只需抽取群即可，操作简单；
    - 当总体单位自然成群时，抽样简单；
    - 当群内单位差异大，群间差异小时，效率更高；
    - 无法提前知道总样本量；

抽样方法

- 整群抽样到多阶段抽样
  - 先抽取群，但并不是调查群内的所有单位，而是再进行一步抽样，从选中的群中抽取出若干个单位进行调查
    - 群是初级抽样单位，第二阶段抽取的是最终抽样单位。将该方法推广，使抽样的段数增多，就称为多阶段抽样；
    - 具有整群抽样的优点，保证样本相对集中，节约调查费用；
    - 在大规模的抽样调查中，经常被采用的方法；

抽样方法

- 系统抽样
  - 当总体的个数较多时，采用简单随机抽样太麻烦，这时将总体分成均衡的部分，然后按照预先定出的规则，从每一部分中抽取1个个体，得到所需要的样本，这种抽样称为系统抽样。
    - 将总体中的个体均分后的每一段进行抽样时，采用简单随机抽样
    - 如总体的个体数不能被样本容量整除时，可以先用简单随机抽样从总体中剔除几个个体；
    - 整个抽样过程中每个个体被抽到的概率仍然相等；

抽样方法

- 例子
  - 一个礼堂有30排座位，每排40个座位。一次报告会礼堂坐满了听众。会后为听取意见留下了座位号为20的30名听众。
    - 由于每排的座位有40个，各排每个号码被抽取的概率都是$\frac{1}{40}$。
    - 因而第1排被抽取前，其他各排中各号码被抽取概率也是$\frac{1}{40}$，也就是说所有个体被抽取的概率都是$\frac{1}{40}$。
    - 第1排用简单随机抽样确定起始号码20，后面的每排都可以按照这个规则抽取。

抽样方法

- 不等概率抽样
  - 如果总体单元相差较大，等概率抽样效果不一定好
  - 例子：估计合肥市商业零售总额，大型商场、中型超市和小型商店的差别非常明显，平等对待显然不合理。
    - 分层抽样：按规模分层，大型抽样比高、小型抽样比低
    - 目录抽样：少数大单元普查而大多数小单元进行抽样
    - 不等概率抽样

抽样方法

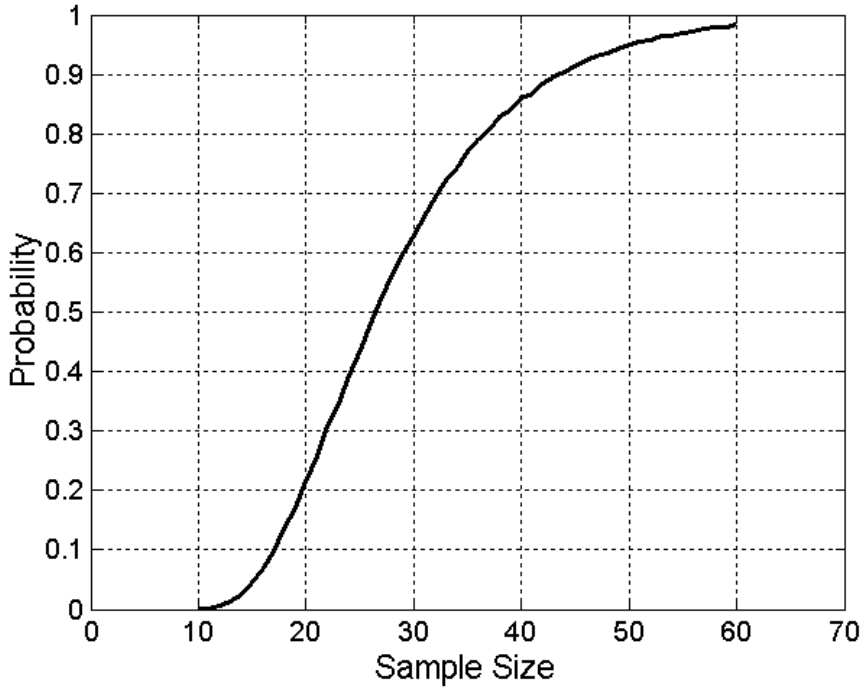
- 不等概率抽样
  - 不等概率抽样（sampling with unequal probability）是指在抽取样本之前给总体中的每一个单元赋予一定的入样概率，从而保证大的（重要的）单元抽到的概率大，而小的（不重要的）的单元抽到的概率小。这里每个单元被赋予的入样概率通常与某个辅助变量有关（比如单元规模等）
  - 必要的约束条件
    - 对总体的每一个单元，都要已知一个辅助变量用于确定其入样概率或两个单元同时入样的概率

抽样方法

- 不等概率抽样适用情况：
  - 需要估计总体总量但总体单元规模相差很大的情况
  - 抽样审计，注册会计师对某类交易或账户余额中低于百分之百的项目实施审计程序，使所有抽样单元都有被选取的机会。
  - 在不能直接对基本的较小单元抽样的情形下，与其它抽样结合，完成对大的单元的抽样。

抽样方法

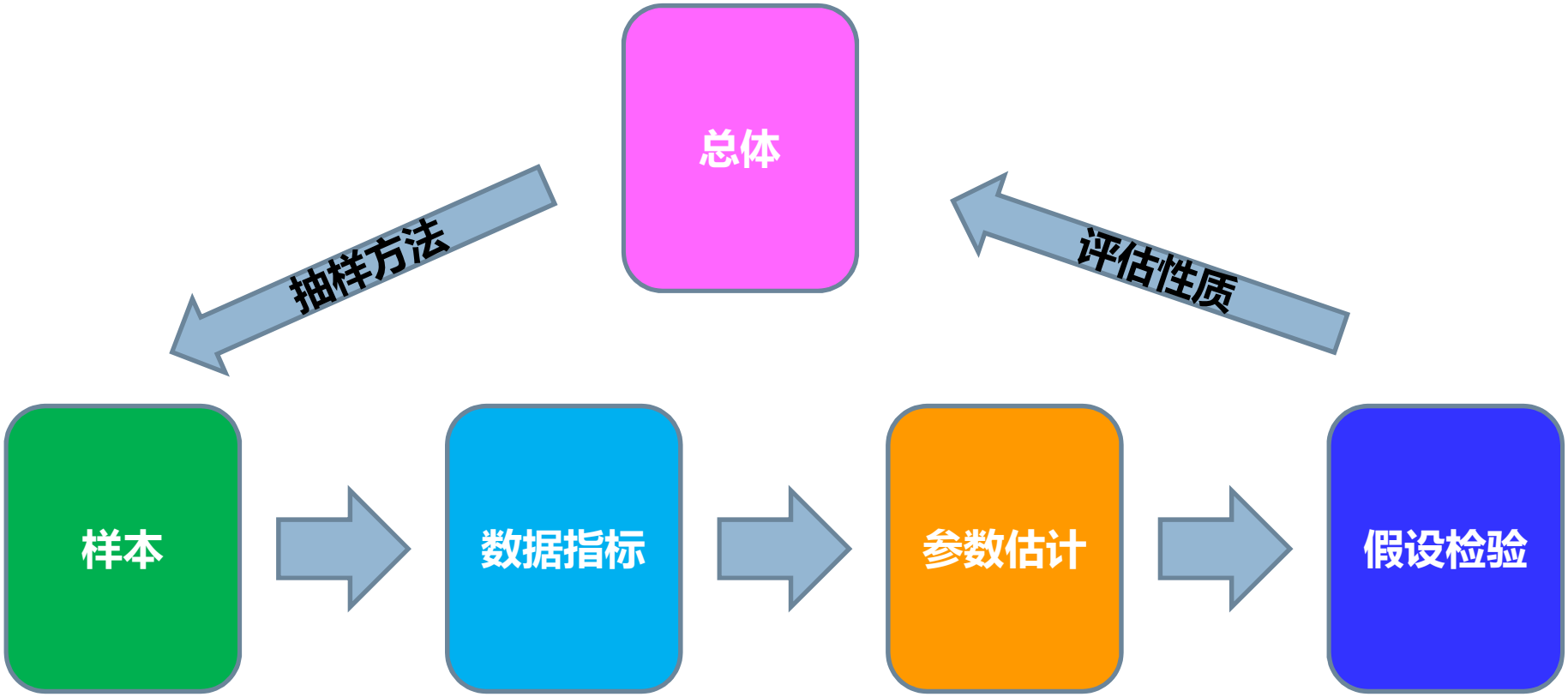
- 抽样规模
  - 假定有**10个组**（组的大小大致相当）的数据样本，如果从这些数据样本中进行抽样，**需要抽样多少次**才能保证抽样到的数据样本里**至少包含了每个组**的一个样本？



总结:大数据为什么还需要抽样

- 提高效率，节省时间成本与计算资源
  - 无法得到样本整体
  - 或者得到样本整体的成本太高
- 通过抽样来解决样本不均衡问题
  - 欠抽样、过抽样…
- 定性分析的工作需要
  - 通常不需要定量分析时的完整假设、精确数据和复杂统计分析过程，更多的是采用访问、观察和文献法收集资料并通过主观理解和定性分析找到问题答案
  - 主要依靠人自身的能力而非密集的计算机能力来完成研究工作

总结



数据统计

- 数据分布基本指标
- 参数估计
- 假设检验
- 抽样方法

- Tips:
  - 一项研究工作不一定使用到所有的数据统计
  - 数据统计量的使用方法必须结合实际场景的需求
    - 同一类数据统计不一定适用于不同的场景
      - 例如，使用哪一类假设检验方法
  - 有些统计量也可能在结果评估等场景中使用