数据科学导论

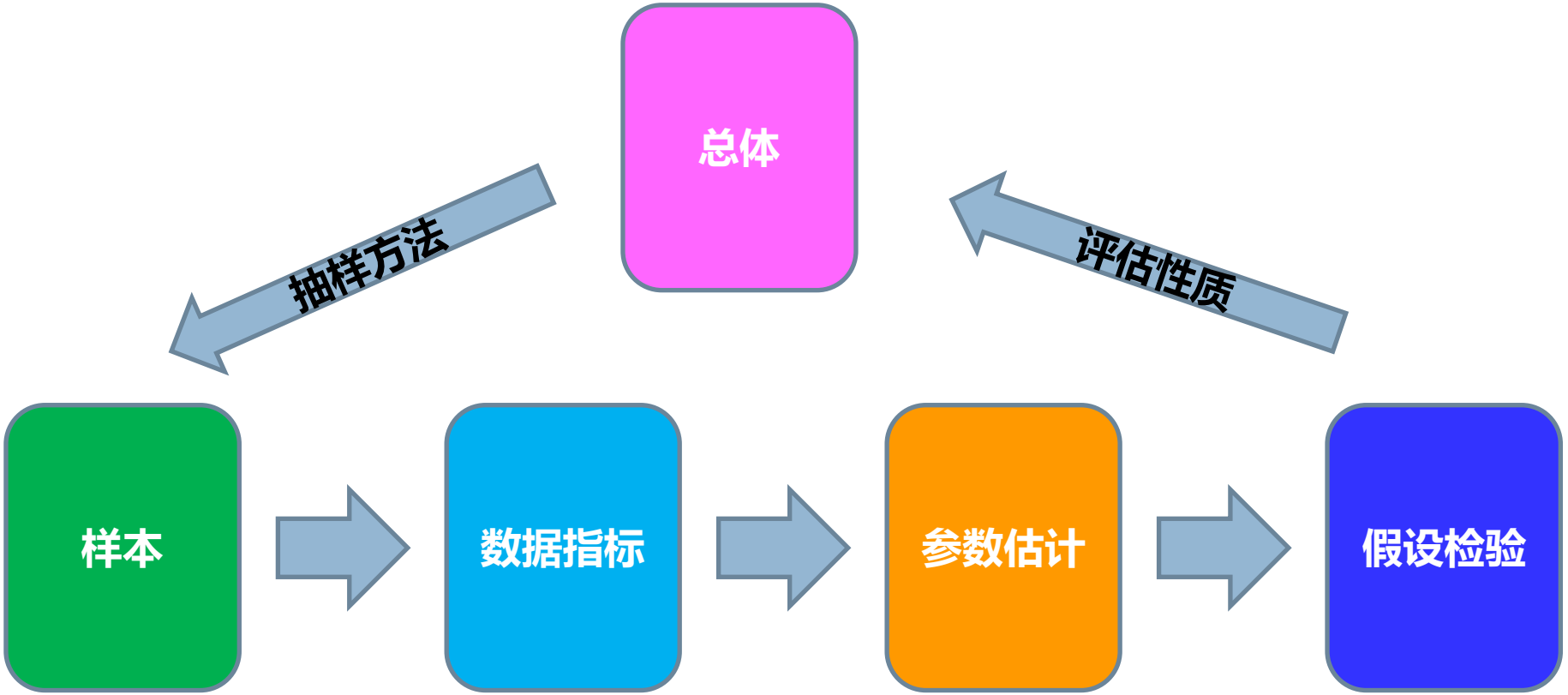
Introduction to Data Science

第三章 数据统计基础

黄振亚，陈恩红
Email: huangzhy@ustc.edu.cn, cheneh@ustc.edu.cn

课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/DS2025.html

回顾：数据统计



数据统计

- 数据分布基本指标
- 参数估计
- 假设检验
-

假设检验

- 假设检验
  - 假设检验（ hypothesis testing）是统计推断方法
    - 面对场景：对总体数据进行估计
    - 参数检验 vs 非参数检验
  - 假设检验与参数估计
    - 相同：都利用样本对总体进行推断，采用的技术手段相似；
    - 不同：推断的出发点不同，结果也不同
      - **参数估计**：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
      - **假设检验**：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
  - 大数据分析与假设检验
    - 大数据分析采用所有获得数据
    - 侧重分析变量之间的相关性

假设检验

树上的苹果都很甜

- 假设检验
  - 原理：小概率事件；逻辑：反证法
    - 给出假设，实验分析，观察数据，检验假设(下结论)
  - 假设：总体的参数：均值、方差、比例等
  - 两个假设定义
    - 原假设H0：想要拒绝的假设
    - 备择假设H1：想要接收的假设

H0和H1不一定是事件的全集

与研究目的相关，主观性

  - 两类错误
    - 第一类错误，弃真α：H0成立时，拒绝了H0
      - 原假设实际上真，但通过样本估计总体后，拒绝了原假设
    - 第二类错误，取伪β：H0错误时，接收了H0
      - 原假设实际上假，但通过样本估计总体后，接受了原假设

假设检验

- 两类错误
  - 第一类错误，弃真α：H0成立时，拒绝了H0
  - 第二类错误，取伪β：H0错误时(H1成立)，接收了H0

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）

  - 假设检验目标：拒绝H0
  - 假设检验的过程：希望判断的结果犯错率越低越好
    - 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

假设检验

- 两类错误
  - 取伪、弃真两类错误造成的后果可能是不一样严重的
  - 假设检验中应当把哪一类错误作为首要的控制目标？即，哪一类错误所造成的后果更严重
    - α 错误的犯错率为置信度，降低置信度就可以降低α 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
    - 因此，首要降低α 错误
      - 原假设被拒绝，如果出错的话，只能犯弃真错误，而犯弃真错误的概率已经被规定的显著性水平所控制了。对统计者来说更容易控制

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）

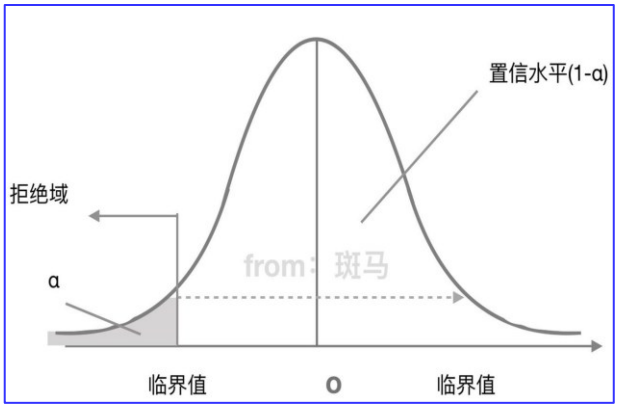
增大样本量可以使得两类错误同时减小

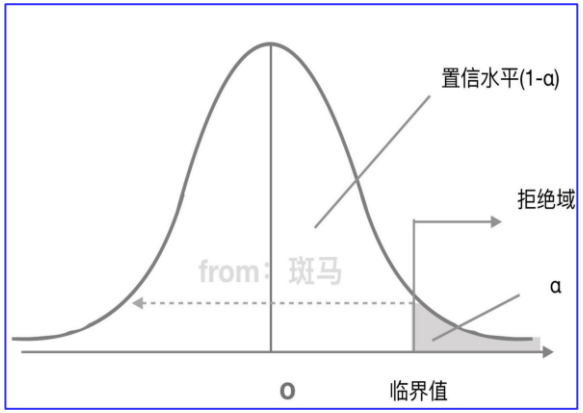
假设检验

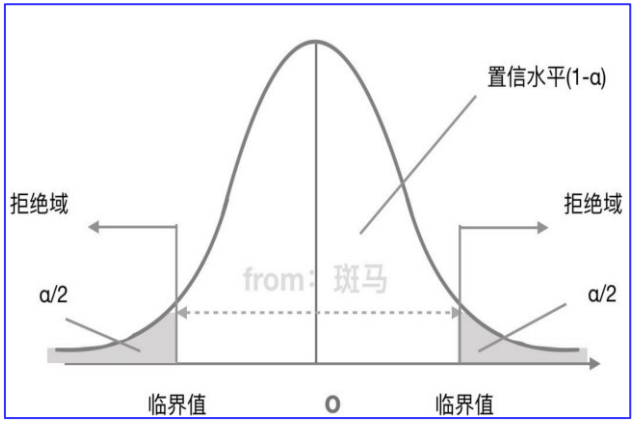
- 假设检验
  - 检验统计量
    - 对原假设和备择假设作出决策的某个样本统计量
    - 是否大样本(样本数量>30)？方差是否已知？
  - 显著性水平：指当原假设正确时(H0正确)，检验统计量落在拒绝域的概率，即，犯弃真错误α的概率
    - 小概率事件发生的概率（越小，越可信）
    - 显著性水平α越小，犯第I类错误的概率自然越小，一般取值：0.01、0.05、0.1等

假设检验

- 假设检验
  - 检验方式
    - （H1）单侧检验：左侧检验(<)，右侧检验(>)；双侧检验(≠)
  - 拒绝域：由显著性水平围成的区域
    - 判断假设检验是否拒绝原假设H0。
      - 若样本观测计算出来的检验统计量的数值落在拒绝域内，拒绝原假设，否则不拒绝原假设
      - 给定显著性水平α后，查表就可以得到具体临界值，将检验统计量与临界值进行比较，判断是否拒绝原假设



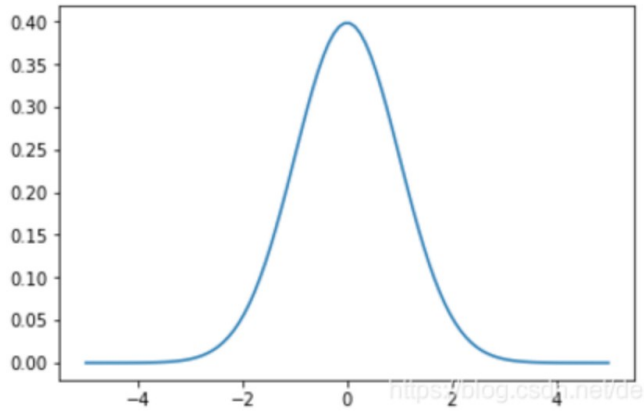




假设检验

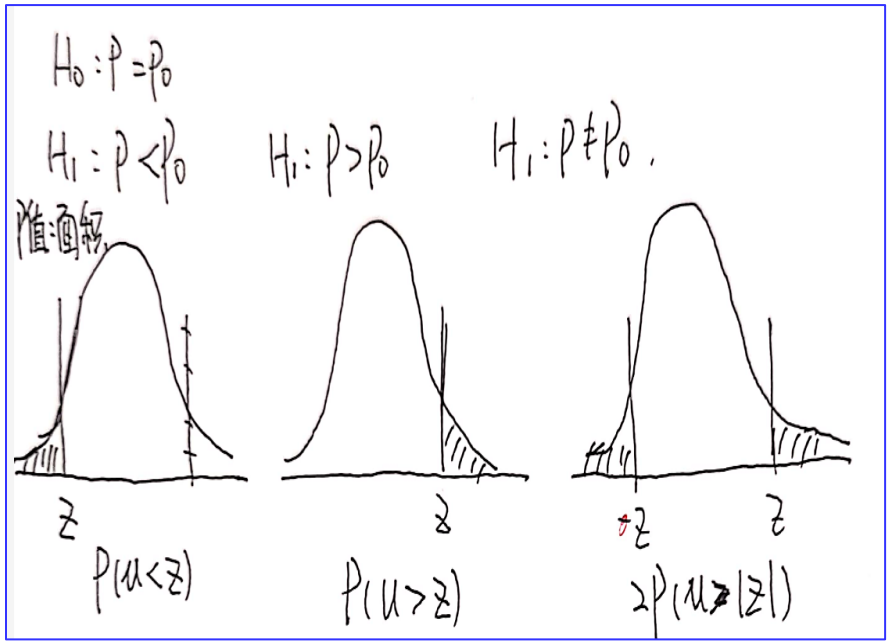
- 假设检验
  - 检验过程：利用标准正态分布查表
    - 提出原假设与备择假设
      - H0: p=p0(凭空假设，给出)，H1：p≠p0； p<p0；p>p0
      - 例如：投掷飞镖，平均5环
    - **构造小概率事件：**进行实验
    - 构造检验统计量
    - 样本试验归一化
      - 样本统计量 $\hat{p} = \frac{X}{N} \sim N\left(p, \frac{p(1-p)}{N}\right)$
      - 归一化：$Z = \frac{\hat{p}-p0}{\sqrt{\frac{p0(1-p0)}{N}}} \sim N(0,1)$
    - 检验思路：Z值是否靠近中心p0

Z-score标准化

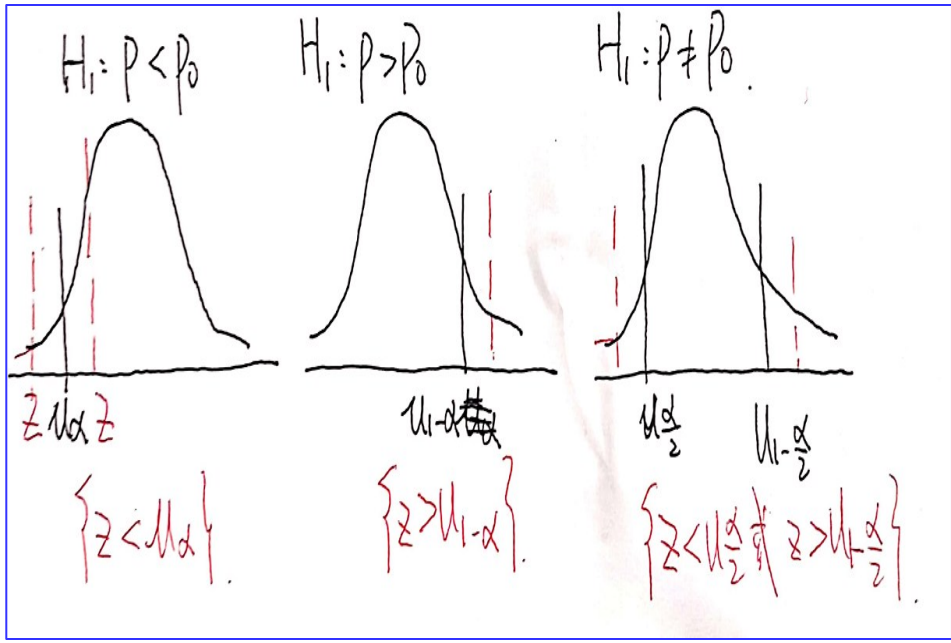


假设检验

- 假设检验
  - 检验过程：利用标准正态分布查表
    - 检验策略（与显著性水平α比较）
      - P值检验：出现与观察值极端或者更极端的情况的概率(越小越拒绝α)
      - **通过拒绝域：找出与接收α所要求的临界值**



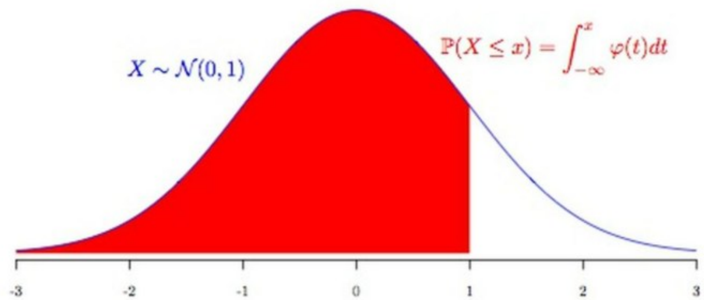
P值与α对比



拒绝域

假设

- 假设检验
  - 检验过程
  - 检验策略
    - P值检验 …… 越小越拒绝α)
    - 通过拒……
      - 给定
      - 给定
      - 给定 …… 5
      - 给定



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	[illegible]	[illegible]	[illegible]

假设检验

- 案例分析—提出假设
  - 有报道称，随着电子商务的快速发展，35.6%的中国人有过网购经历，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:$H_0: p0 = 35.6\%$
  - 对备择假设H_1，只要求参数值不等于某个特定值(35.6%)。
    - (1) $H_1: p0 < 35.6\%$
    - (2) $H_1: p0 = 35.6\%$
    - (3) $H_1: p0 > 35.6\%$
  - 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)？

如果原假设是保守估计（默认>35.6），这时选择（1），反之选择(3)

假设检验

- 案例分析—构造小概率事件
  - 理解小概率事件—为何目标是拒绝原假设H0？
    - 如果 35.6% 的中国人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几乎不可能发生
    - 小概率事件可以验证
  - 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
    - 提出原假设和备择假设后，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**。
    - 计算小概率事件是否发生。

假设检验

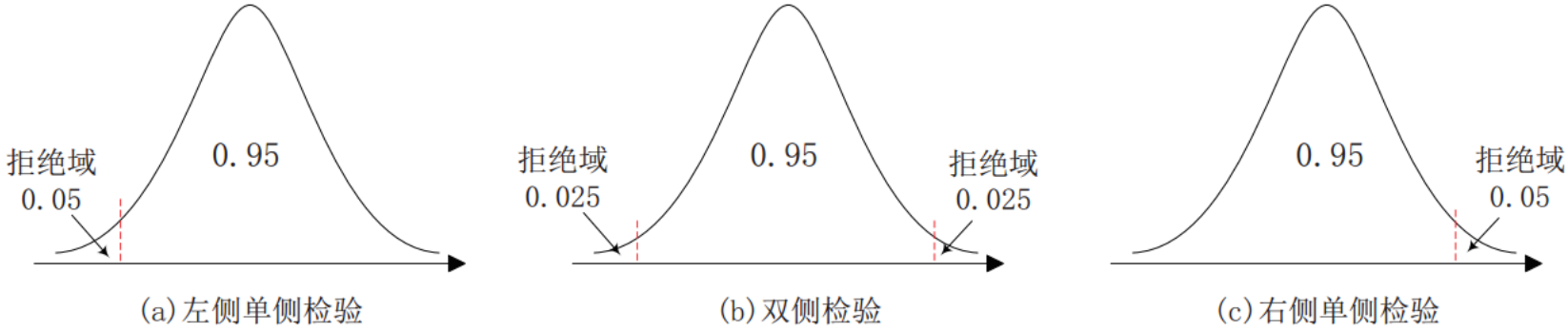
- 案例分析—构造样本统计量
  - 若样本(大小n)中由网购经历的比例为p，和假定的参数值(p_0)进行对比，并构造检验统计量Z：

$$Z = \frac{p - p_0}{s_0}, \qquad s_0 = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

  - 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
  - 确定小概率事件的置信度
    - 常用的置信度为$\alpha = 0.05$，表示：假设检验结果出错概率为5%

假设检验

- 案例分析
  - 对于三种不同的备择假设(1) $H_1: p0 < 35.6\%$；(2) $H_1: p0 \neq 35.6\%$；(3) $H_1: p0 > 35.6\%$，
  - 显著性检验时的**拒绝域**不同（以置信度0.05为例）
    - 单侧检验
    - 双侧检验



(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H_0，接受备择假设H_1

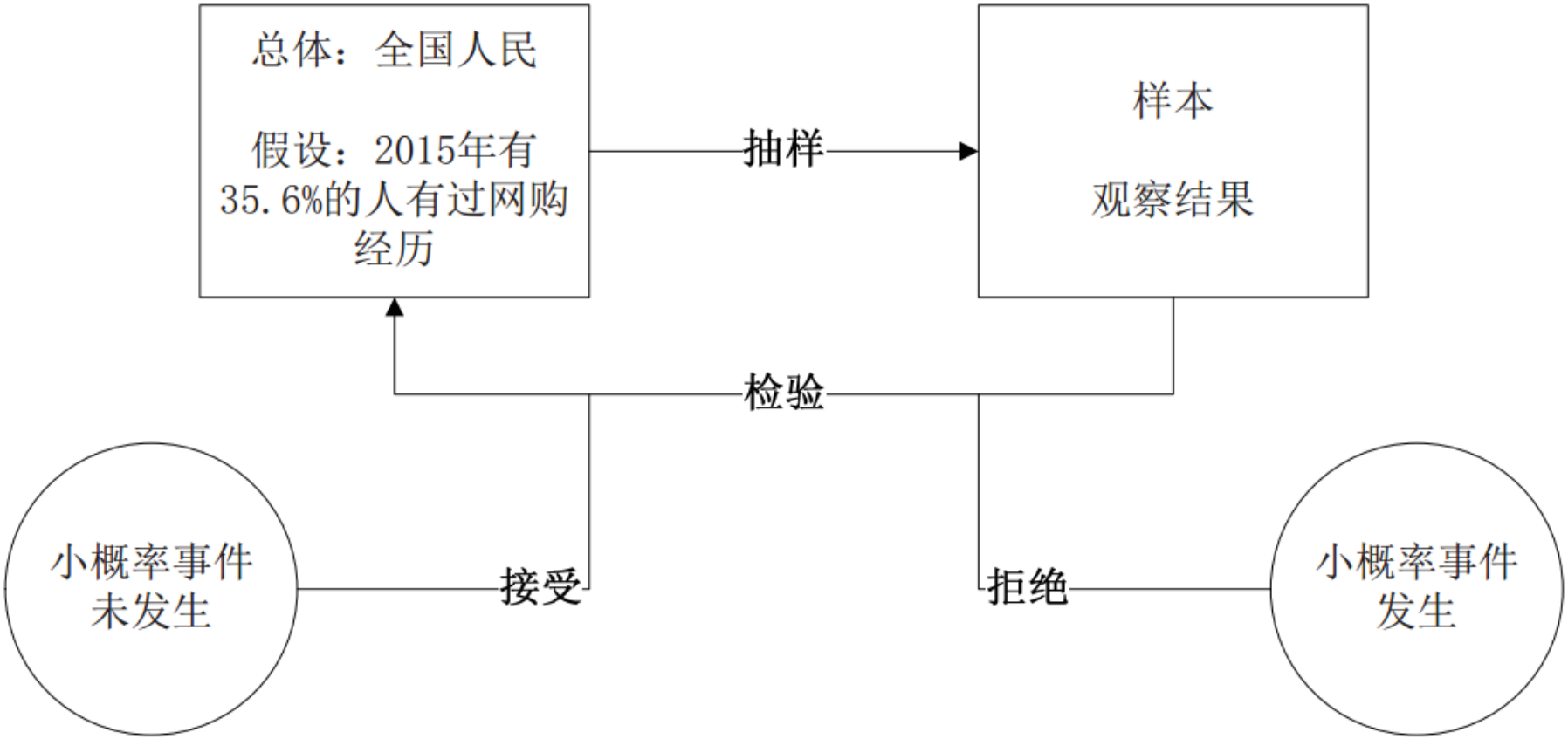
假设检验

- 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）。

接受原假设并不代表原假设本身是成立的，只能认为没有充分的理由（证据）否定原假设。

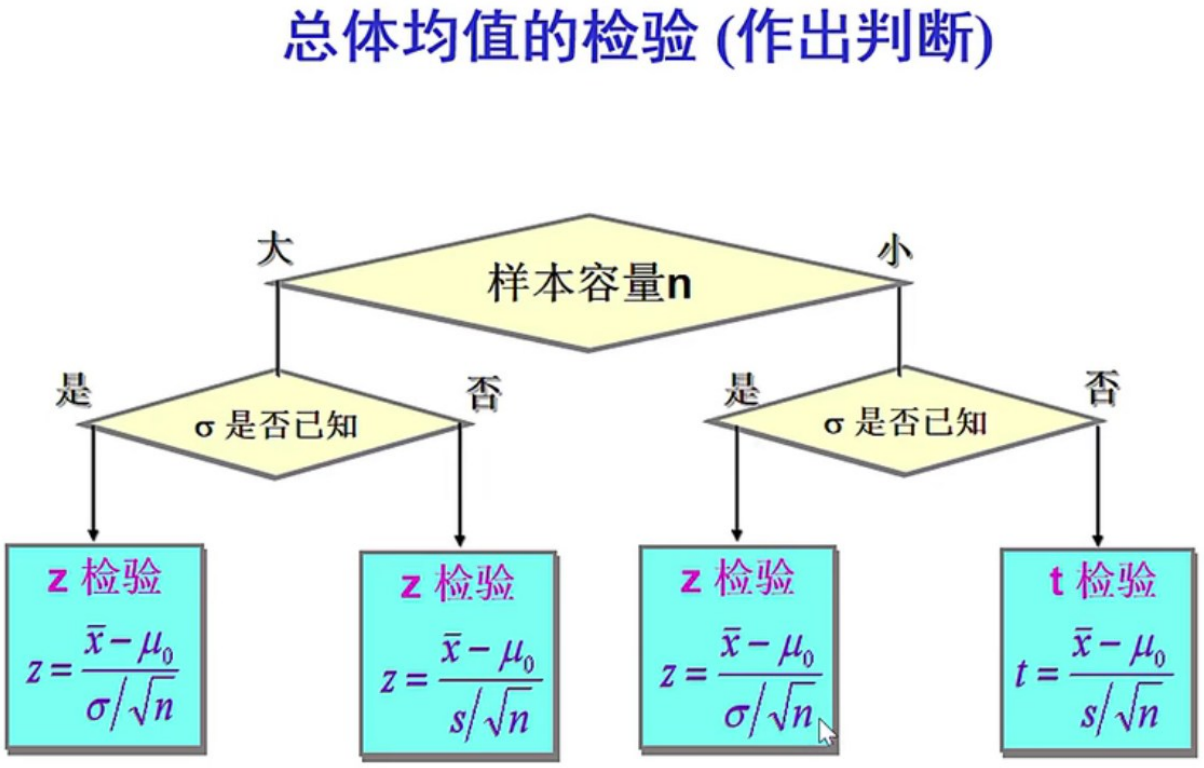
假设检验

- 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

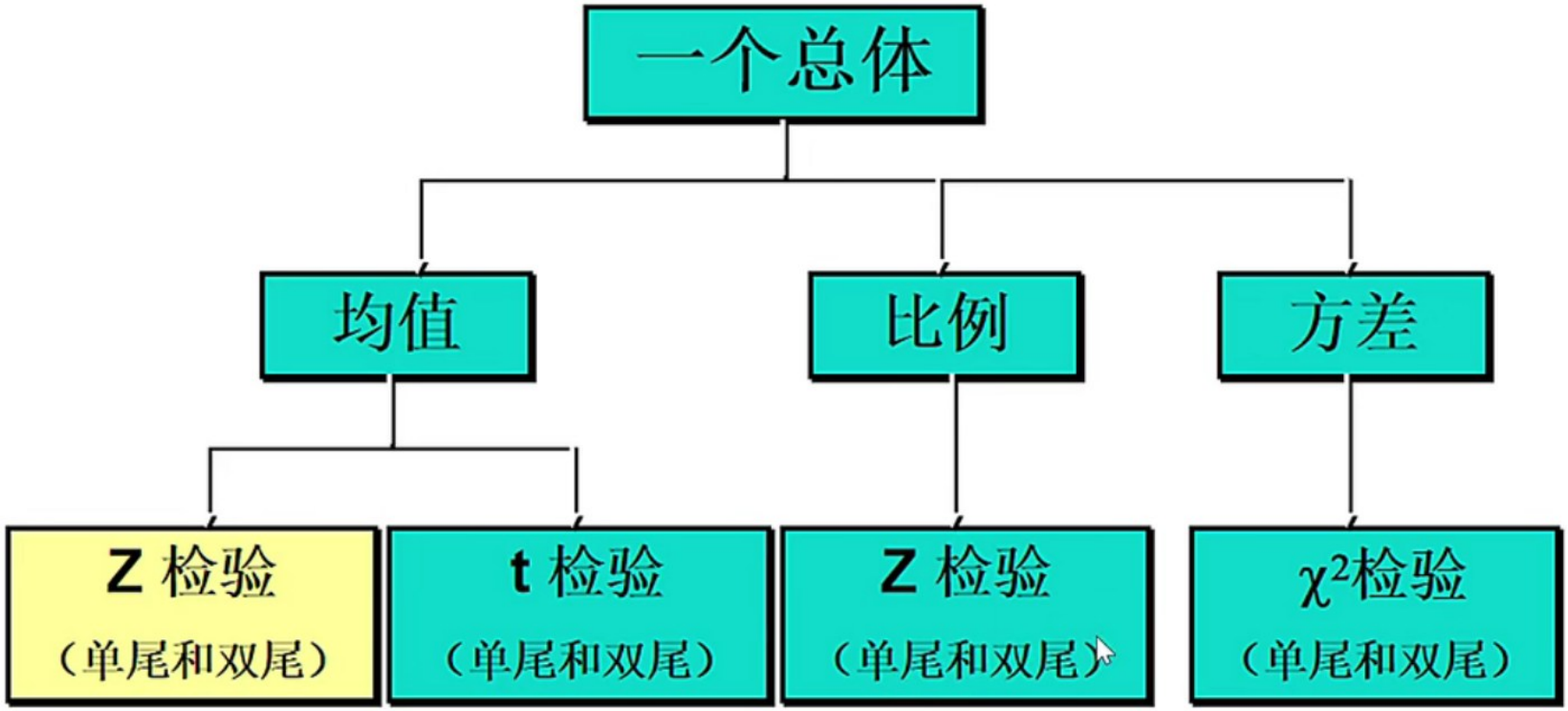
假设检验（几个常用的检验方法）

- 单总体假设检验
  - 均值检验



假设检验（几个常用的检验方法）

- 单总体假设检验

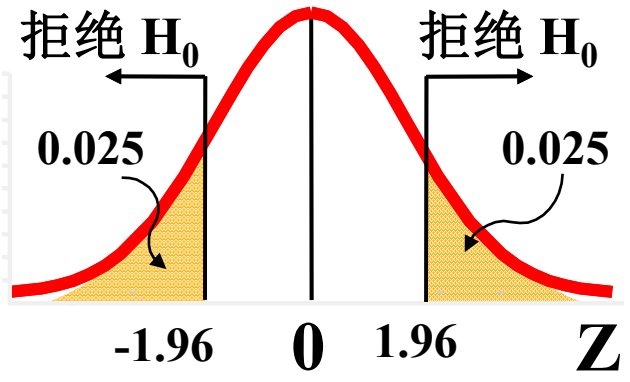


假设检验

- 例子　　大样本总体均值的双侧检验，σ 已知
  - 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: μ = 0.081；　H_1: μ ≠ 0.081

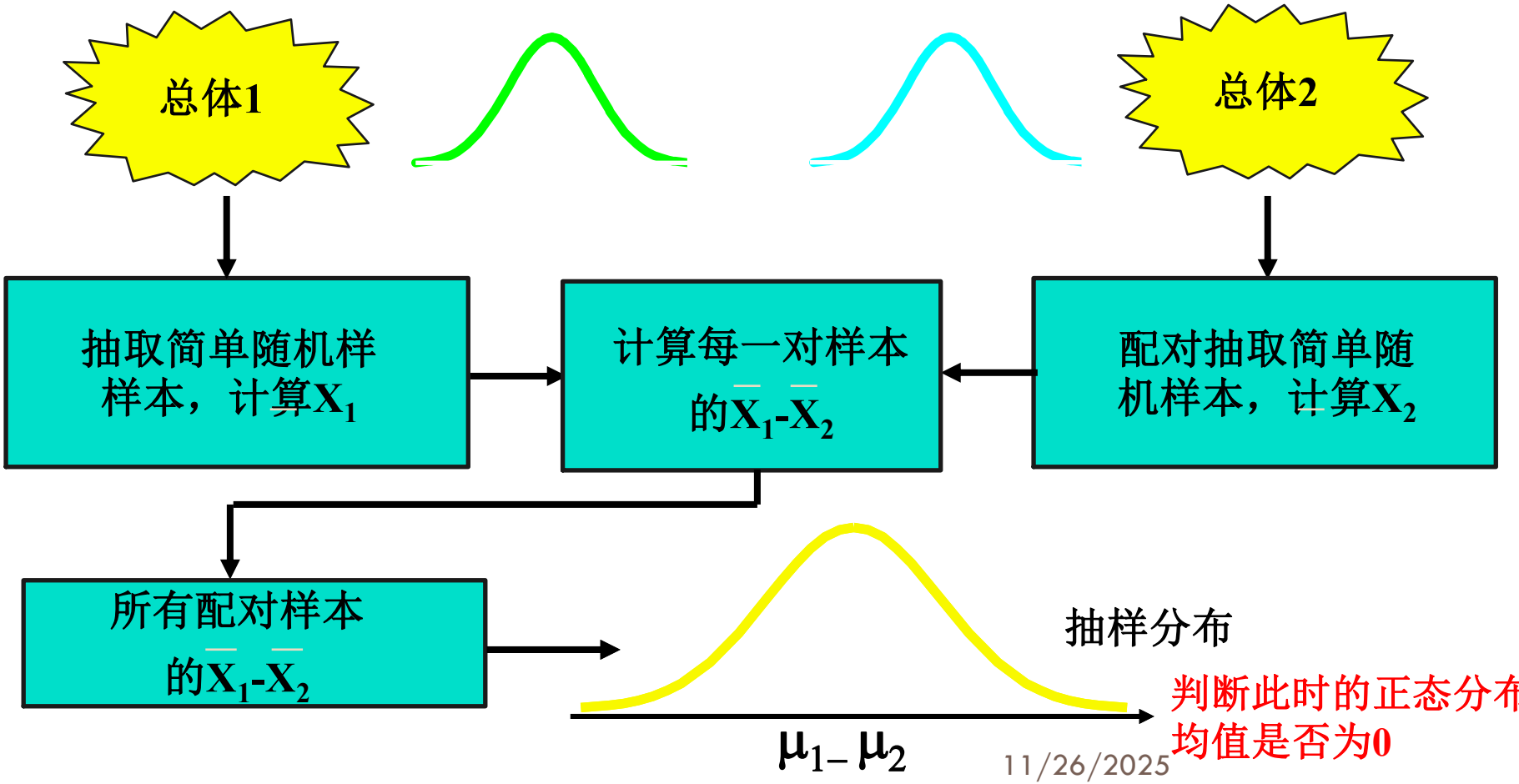
$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}=\frac{0.076-0.081}{0.025/\sqrt{200}}=-2.83$$



在 α = 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。

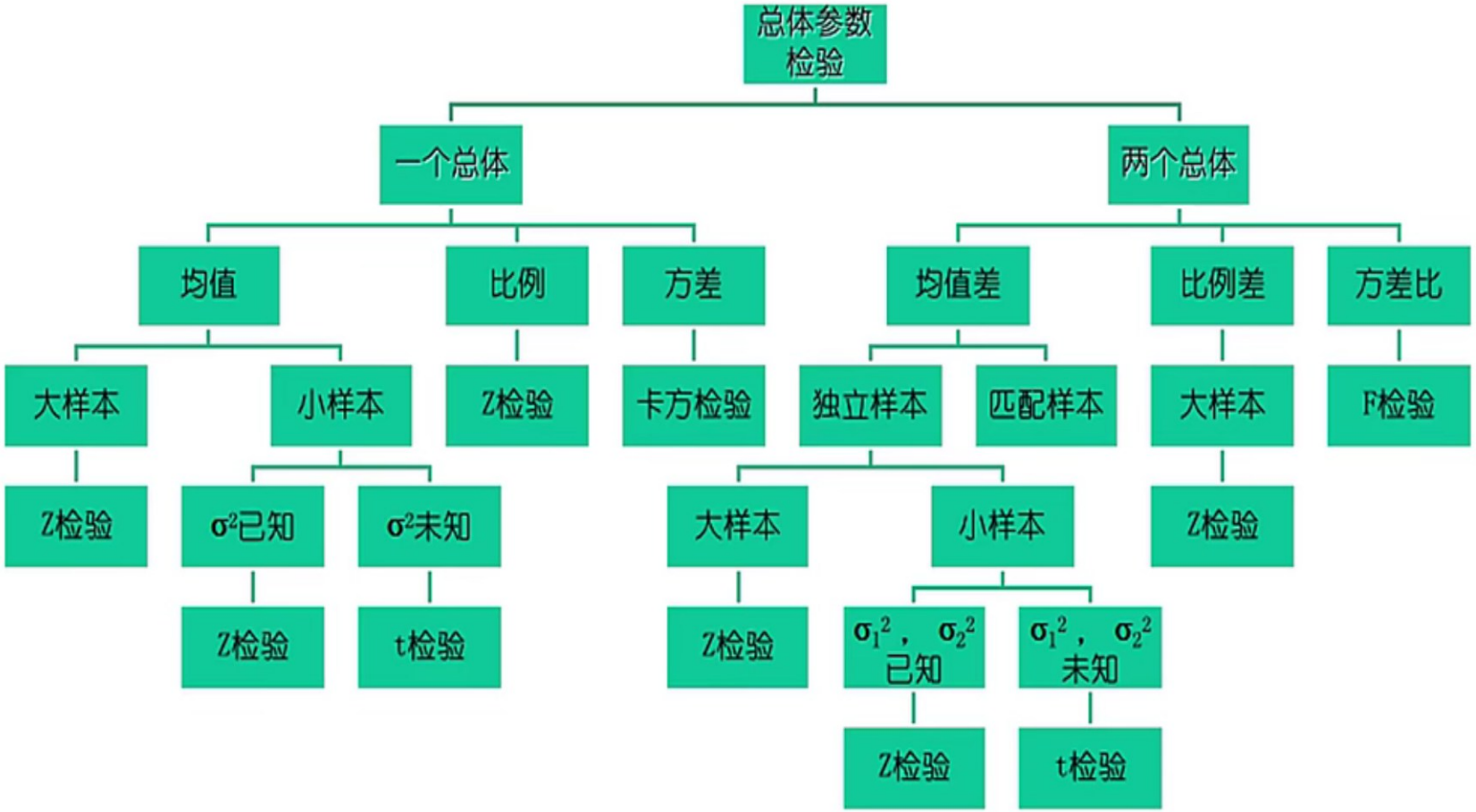
假设检验

- 双总体均值检验（例如，配对样本的 t 检验或z检验）



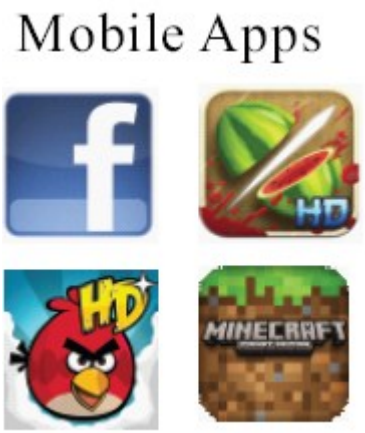
假设检验

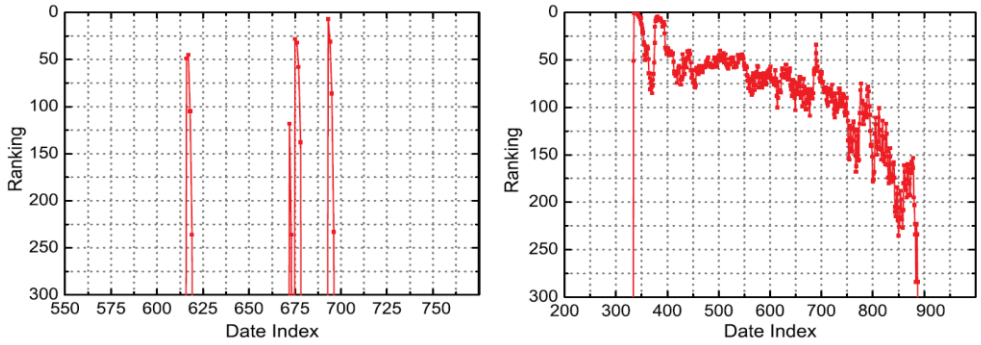
- 课外阅读(不做强制要求)



假设检验——应用

- 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 TKDE 27(1): 74-87, January 2015.

假设检验——应用

- 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 $p \leq 0.01$

非参数假设检验的例子：检验模型结果是否符合数据分布

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

假设检验——应用

- 应用举例---验证实验结果的可信度

	Math23K			MAWPS			SVAMP		
	ORI	LeAp	LeAp-EK	ORI	LeAp	LeAp-EK	ORI	LeAp	LeAp-EK
Seq2Seq	0.640	**0.660****	0.652**	0.797	0.803	**0.807***	0.200	**0.236*****	0.220***
Graph2Tree	0.774	0.779*	**0.782****	0.837	**0.852****	0.849**	0.319	**0.341*****	0.325*
HMS	0.761	**0.769**	0.765	0.803	**0.812***	0.805	0.179	**0.196****	0.191**
GTS	0.756	**0.772****	0.767**	0.826	**0.834****	0.830*	0.277	**0.285**	0.279
TSN-MD	0.774	**0.786****	0.778*	0.844	**0.853***	0.848	0.290†	**0.302****	0.294*
Multi-E/D	0.784	0.791*	**0.793****	/	/	/	/	/	/

Table 1: Answer accuracy ($***: p \leq 0.001, **: p \leq 0.01, *: p \leq 0.05$). †: implemented by MTPToolkit (Lan et al. 2022).

- From Table 1, we observe that LeAp improves the answer accuracy of all backbones, and by applying paired t-test, the improvements are statistically significant with $p \leq 0.001 \sim 0.05$.

Jiayu Liu, Zhenya Huang, Chengxiang Zhai, Qi Liu, Learning by Applying: A General Framework for Mathematical Reasoning via Enhancing Explicit Knowledge Learning, AAAI'2023