

数据科学基础期末试卷回忆

选择

计算

简答

只记了印象深刻的题。其实绝大部分是PPT原题。

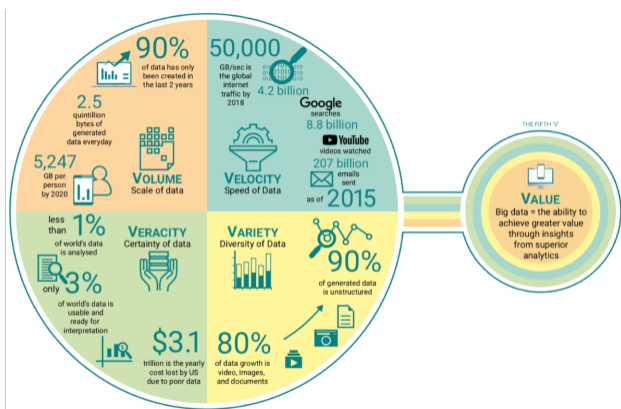
选择

1. 不属于非抽样误差的是：回答误差、无回答误差、调查员误差、**系统误差**

非抽样误差 (non-sampling error)

- 相对于抽样误差而言
- 除抽样误差之外的，由于其他原因造成的样本观察结果与总体真值之间的差异
- 存在于所有的调查之中
 - 非抽样误差的产生贯穿在调查的每一个环节，任何一个环节出错都有可能导致非抽样误差增加而使数据失真。
 - 平时说的控制误差主要指的就是控制非抽样误差
- 有抽样框误差、回答误差、无回答误差、调查员误差、测量误差等

2. 不属于大数据“5V”的是：**volume**、**velocity**、**veracity**、**valid**



3. 求秩和统计量。

计算

1. 画出帕累托图
2. 计算矩估计和最大似然估计
3. 计算p的置信区间

例1 设从一大批产品的100个样品中,得一级品60个,求这批产品的一级品率 p 的置信水平为0.95的置信区间.

4. 手撕方差分析 (不需要p值)

The score for those students on the second exam was recorded

- Front: 82, 83, 97, 93, 55, 67, 53
- Middle: 83, 78, 68, 61, 77, 54, 69, 51, 63
- Back: 38, 59, 55, 66, 45, 52, 52, 61

Source	SS	df	MS	F	p
Between	1902	2	951.0	5.9	
Within	3386	21	161.2		
Total	5288	23			

简答

1. 评选估计量的标准?

无偏、有效、相合 (一致)

估计量的优良准则

- 一. 无偏性
- 二. 有效性
- 三. 一致性

2. 举例说明常见的数据形态。

