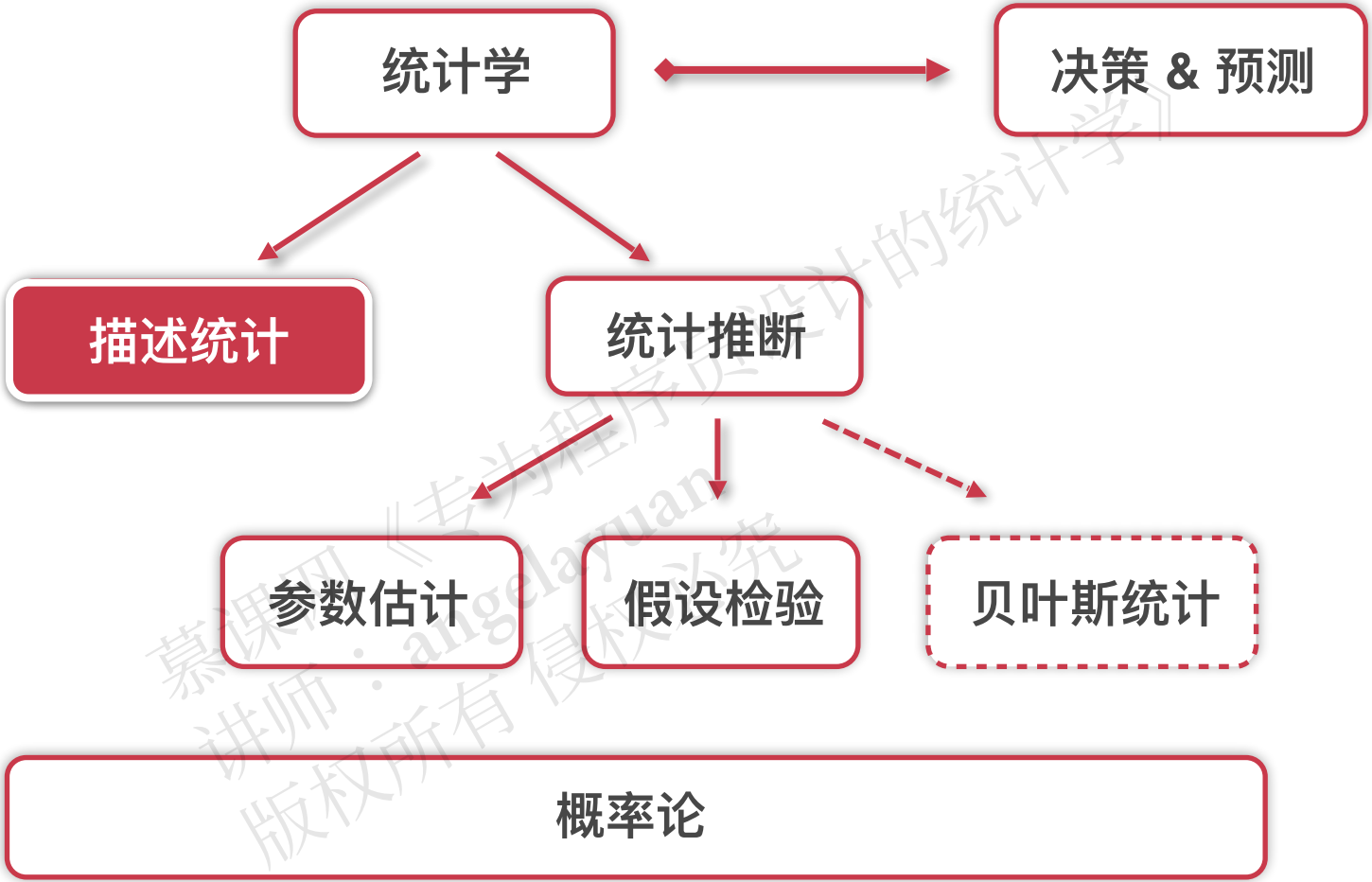
描述统计 (descriptive statistics)



什么是描述统计

描述统计是研究

- 如何取得反映客观现象的数据(数据的收集)
- 通过图表形式对数据进行加工处理和可视化
- 通过概括与分析得出反应客观现象的规律性数量特征

数据的可靠性和有效性

数据是否可靠(reliable)和有效(valid)?

- 可靠性(reliability): 多次测量得到的数据是否一致

对统计学的态度	某治疗的效果
早: 喜欢	评价者1: 有效
中: 不喜欢	评价者2: 无效
晚: 喜欢	

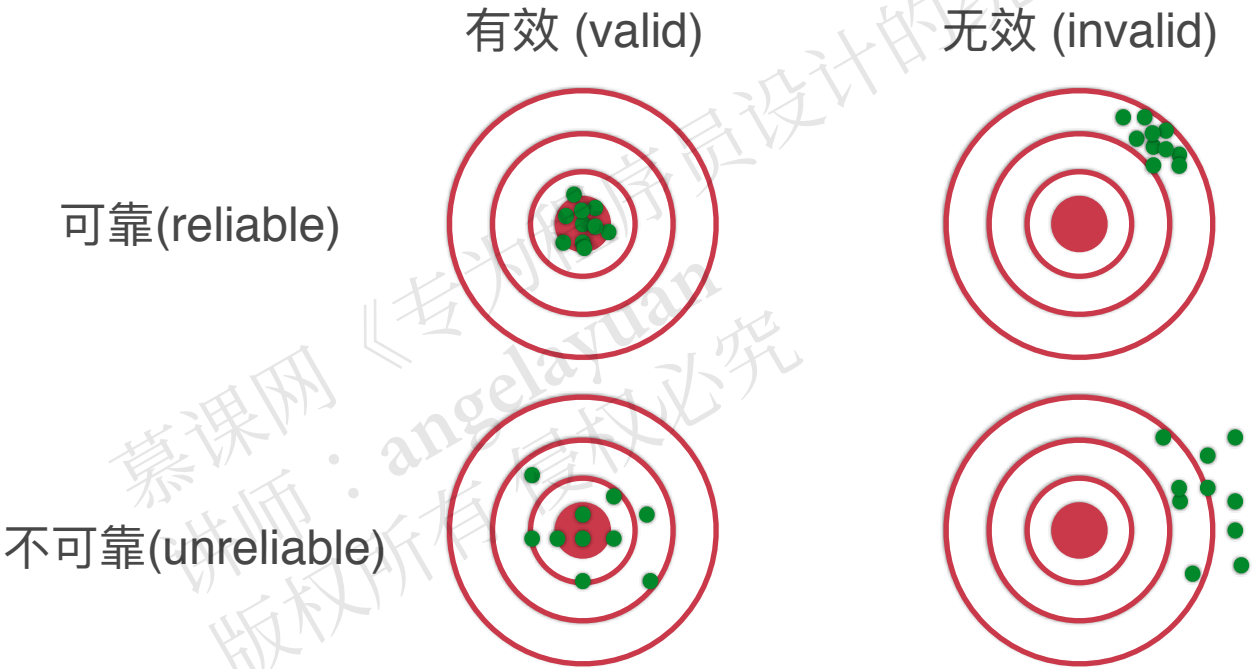
数据的可靠性和有效性

数据是否可靠(reliable)和有效(valid)?

- 有效性(validity): 实际测量的对象 = 认为/希望测量的对象

认为/希望测量: 身高	认为/希望测量: 身高
实际测量的是: 身高	实际测量的是: 体重

数据的可靠性和有效性



描述统计

- 通过图表形式对数据进行加工处理和可视化
- 通过概括与分析得出反应客观现象的规律性数量特征

尺度	举例	逻辑与数学运算	类别
名目	性别、颜色	$=, \neq$	定性/(无序)分类变量
次序	教育程度、评价	$=, \neq, >, <$	定性/(有序)分类变量
等距	温度、年份、时间	$=, \neq, >, <, +, -$	定量/数值变量
等比	身高、体重、年龄	$=, \neq, >, <, +, -, \times, \div$	定量/数值变量

一个分类变量的特征和可视化

无序分类变量

性别 (名目; =, ≠): 男，女

观测12个新生儿的性别 (n = 12)

女, 男, 女, 女, 男, 男, 男, 男, 女, 男, 男, 女

无序分类变量

女, 男, 女, 女, 男, 男, 男, 男, 女, 男, 男, 女

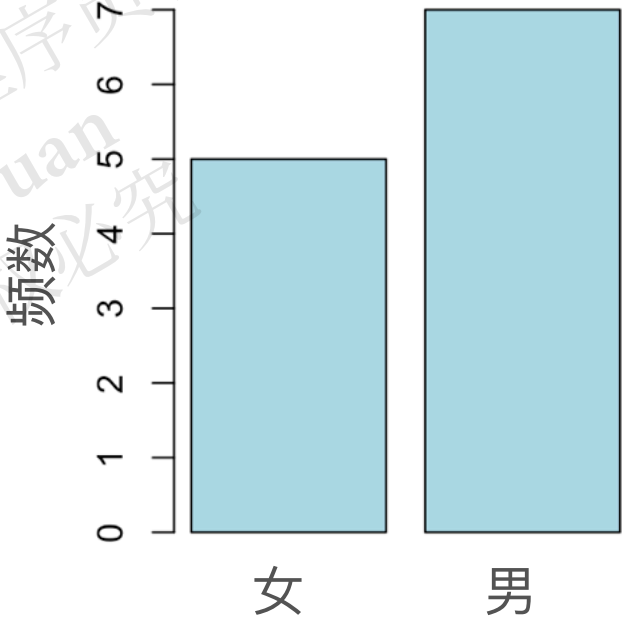
频率表(frequency table)

性别	频数 (Count)	频率 (Frequency)
女	5	5/12 = 41.7%
男	7	7/12 = 58.3%

无序分类变量

女, 男, 女, 女, 男, 男, 男, 男, 女, 男, 男, 女

条形图(bar plot)



无序分类变量

女, 男, 女, 女, 男, 男, 男, 男, 女, 男, 男, 女

集中趋势(central tendency):

一组观测值向其中心集中的倾向和程度

- 众数(mode): 一组观测值中出现次数最多的数

无序分类变量

- 众数(mode): 一组观测值中出现次数最多的数
- 可能存在多个众数，也可能不存在众数

颜色: 赤1, 橙1, 黄1, 绿1, 青1, 蓝1, 紫1 ➡ 不存在众数

颜色: 赤2, 橙6, 黄1, 绿10, 青3, 蓝10, 紫4 ➡ 存在多个众数

有序分类变量

教育程度 (次序; $=, \neq, >, <$):
小学(1), 初中(2), 高中(3), 本科(4), 研究生(5)

观测19个人的教育程度 ($n = 19$)

3, 3, 4, 1, 5, 4, 2, 1, 5, 4, 4, 4, 5, 3, 2, 1, 4, 5, 5

有序分类变量

3, 3, 4, 1, 5, 4, 2, 1, 5, 4, 4, 4, 5, 3, 2, 1, 4, 5, 5

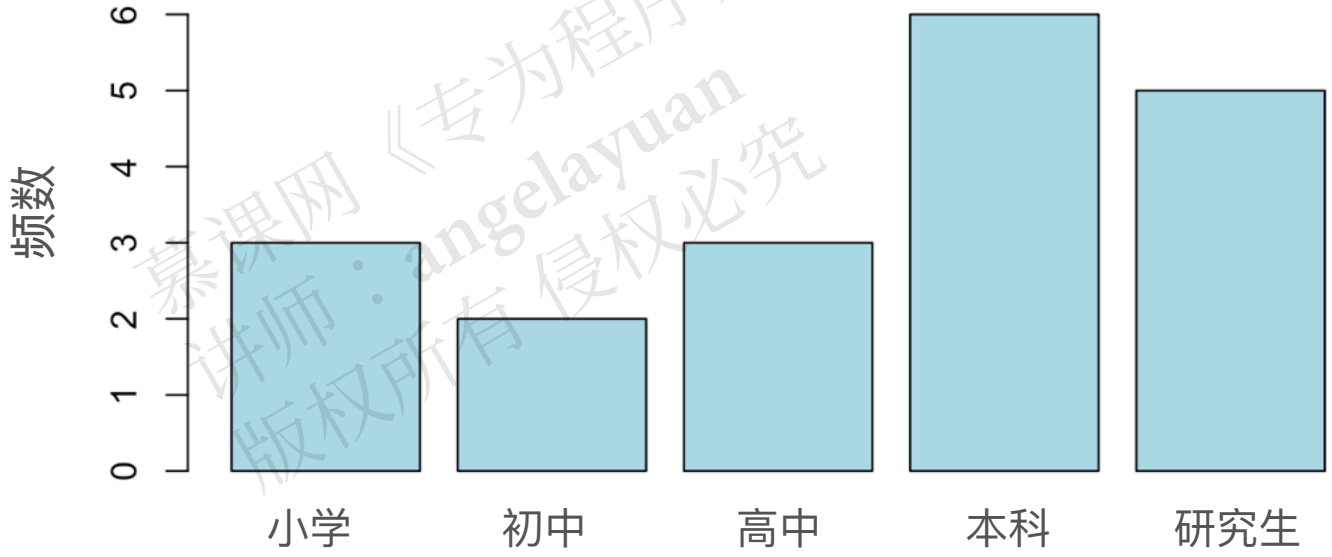
频率表

教育程度	频数	频率
小学 (1)	3	3/19 = 15.8%
初中 (2)	2	2/19 = 10.5%
高中 (3)	3	3/19 = 15.8%
本科 (4)	6	6/19 = 31.6%
研究生 (5)	5	5/19 = 26.3%

有序分类变量

3, 3, 4, 1, 5, 4, 2, 1, 5, 4, 4, 4, 5, 3, 2, 1, 4, 5, 5

条形图(bar plot)



有序分类变量

3, 3, 4, 1, 5, 4, 2, 1, 5, 4, 4, 4, 5, 3, 2, 1, 4, 5, 5

集中趋势: 众数 ➡ 本科

集中趋势: 中位数(median)

- 对于有限的数集，把所有观测值按大小排序后, 位于正中间的观测值即为中位数/中值

1, 1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5 (n = 19)

有序分类变量

1, 1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5 (n = 19)

1, 1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5 (n = 20)

偶数个观测值, 中位数 = (4+4)/2 = 4

1, 1, 1, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5 (n = 20)

偶数个观测值, 中位数 = (4+5)/2 = 4.5

小结

	无序分类变量	有序分类变量
表	频率表	频率表
图	条形图	条形图
集中趋势	众数	众数、中位数

一个数值变量的特征和可视化

数值变量

温度 (等距; $=, \neq, >, <, +, -$)

5月份前两周的温度

19, 22, 21, 17, 13, 19, 18, 17, 17, 21, 21, 21, 19, 20 (n = 14)

数值变量

19, 22, 21, 17, 13, 19, 18, 17, 17, 21, 21, 21, 19, 20 (n = 14)

频率表

温度	频数	频率
13	1	0.07
17	3	0.21
18	1	0.07
19	3	0.21
20	1	0.07
21	4	0.30
22	1	0.07

数值变量频率表

vs

分类变量频率表

数值变量

19, 22, 21, 17, 13, 19, 18, 17, 17, 21, 21, 21, 19, 20 (n = 14)

等距 分割小区间 $\Delta = 1$

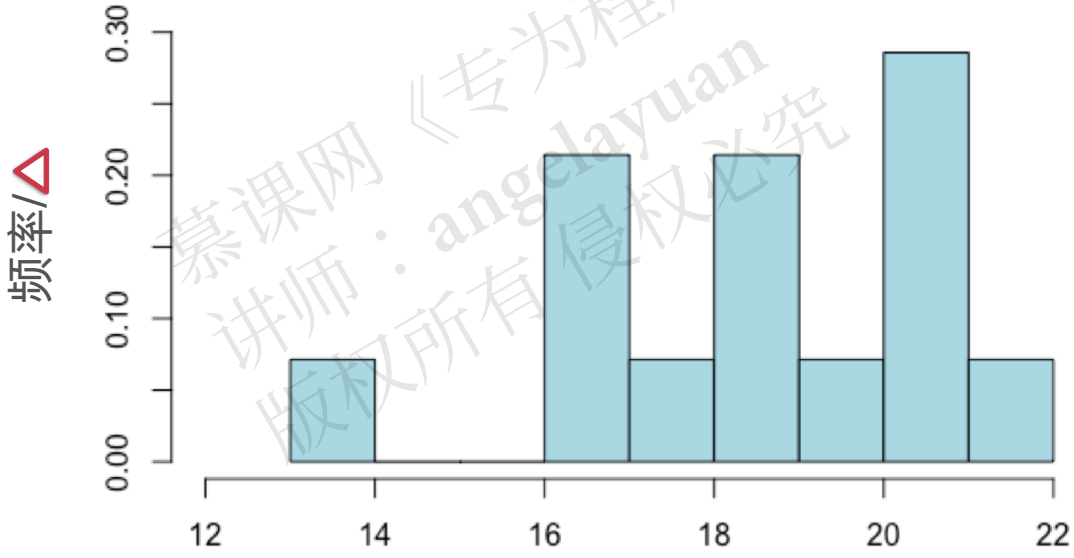
温度	频数	频率
(12, 13]	1	0.07
(13, 14]	0	0
(14, 15]	0	0
(15, 16]	0	0
(16, 17]	3	0.21
(17,18]	1	0.07
(18, 19]	3	0.21

温度	频数	频率
(19, 20]	1	0.07
(20, 21]	4	0.30
(21, 22]	1	0.07

数值变量

19, 22, 21, 17, 13, 19, 18, 17, 17, 21, 21, 21, 19, 20 (n = 14)

频率直方图(histogram)



蓝色覆盖的区域
面积为1

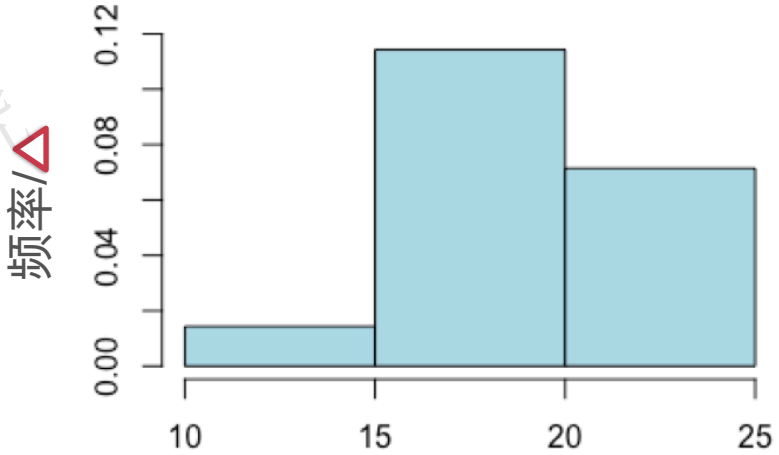
数值变量

19, 22, 21, 17, 13, 19, 18, 17, 17, 21, 21, 21, 19, 20 (n = 14)

△ = 5

温度	频数	频率
(10, 15]	1	0.07
(15, 20]	8	0.57
(20, 25]	5	0.36

频率直方图



蓝色覆盖的区域面积为1

数值变量

19, 22, 21, 17, 13, 19, 18, 17, 17, 21, 21, 21, 19, 20 (n = 14)

集中趋势: 众数、中位数

13, 17, 17, 17, 18, 19, 19, 19, 20, 21, 21, 21, 21, 22 (n = 14)

众数 = 21

中位数 = (19+19)/2 = 19

数值变量

19, 22, 21, 17, 13, 19, 18, 17, 17, 21, 21, 21, 19, 20 (n = 14)

集中趋势: 均值(mean)

- 在一组数据中，所有数据之和再除以这组数据的个数，所得即为这组数据的均值

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\frac{19+22+21+17+13+19+18+17+17+21+21+21+19+20}{14} = 265/14 = 18.93$$

数值变量

19, 22, 21, 17, 13, 19, 18, 17, 17, 21, 21, 21, 19, 20 (n = 14)

离散趋势(tendency of dispersion): 观测值偏离其中心的趋势

- 极差/全距 (Range): 最大值减去最小值，用于简单描述数据的范围大小

  13, 17, 17, 17, 18, 19, 19, 19, 20, 21, 21, 21, 21, 22

  极差 = 22 - 13 = 9

数值变量

19, 22, 21, 17, 13, 19, 18, 17, 17, 21, 21, 21, 19, 20 (n = 14)

离散趋势(tendency of dispersion): 观测值偏离其中心的趋势

- 分位数/分位点 (quantile): 把数据n等分的分割点

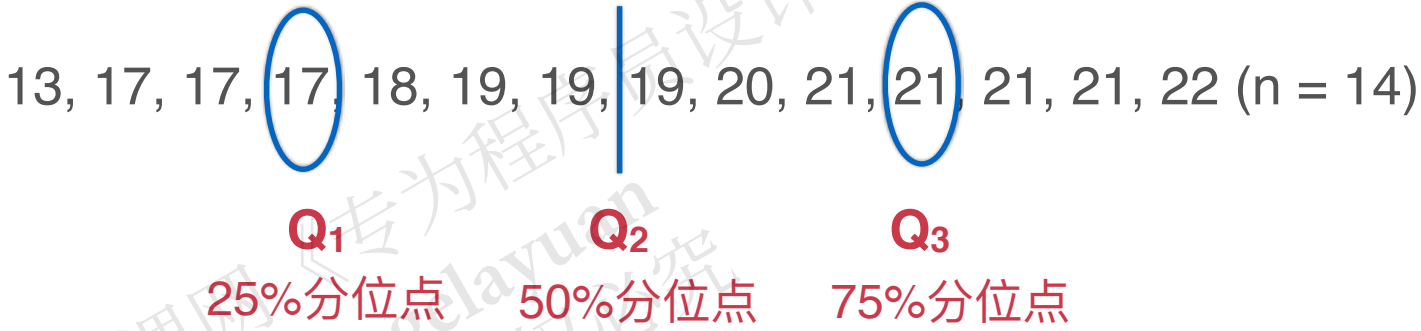
13, 17, 17, 17, 18, 19, 19, 19, 20, 21, 21, 21, 21, 22

中位数 = (19+19)/2 = 19

中位数把数据分成了数目相等的两部分，是二分位数/点

数值变量

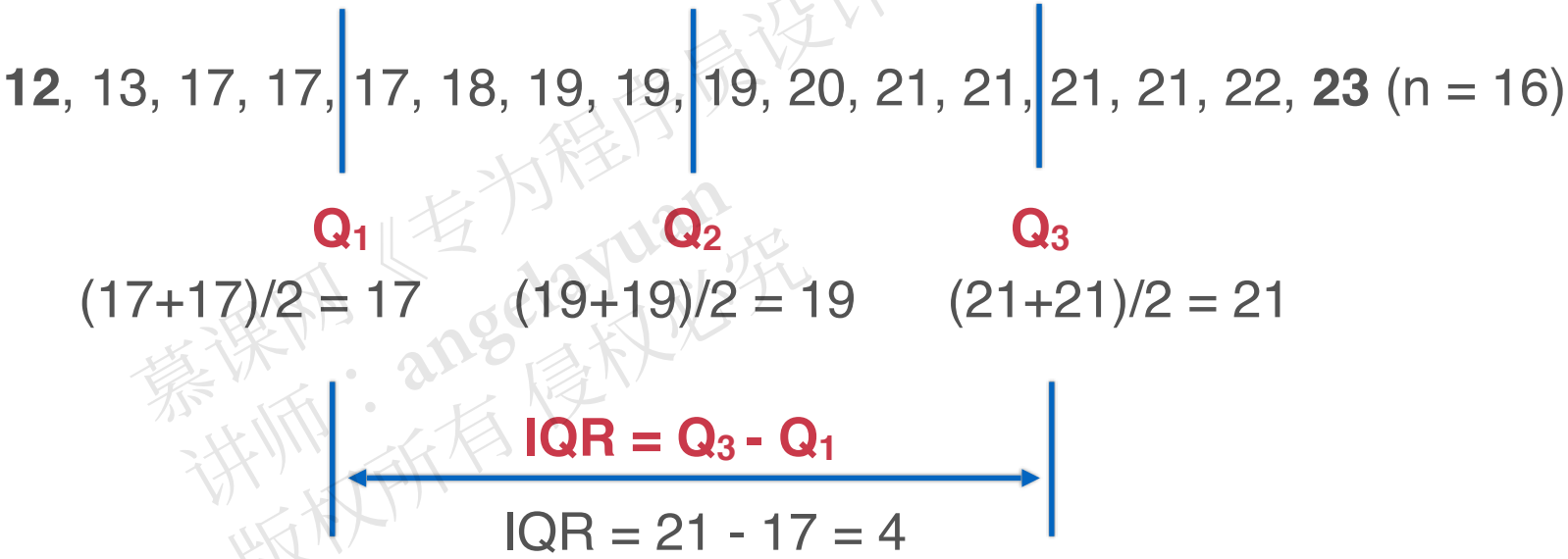
- 四分位数 (quartile)



- 四分位距 (interquartile range, IQR)

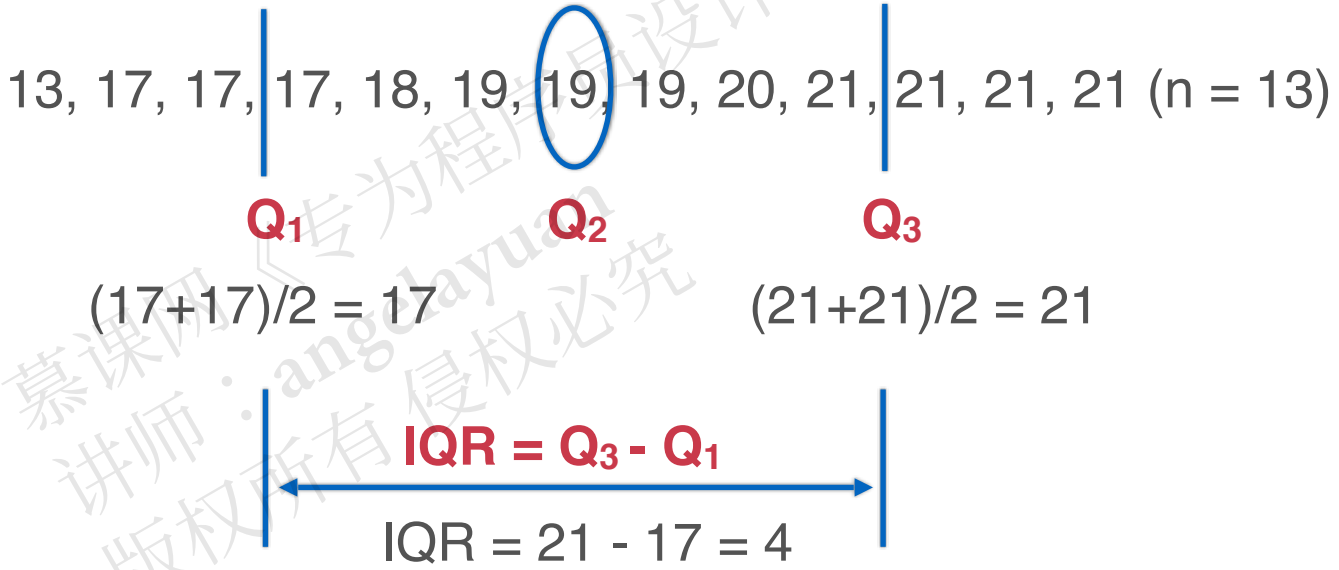
数值变量

- 四分位数



数值变量

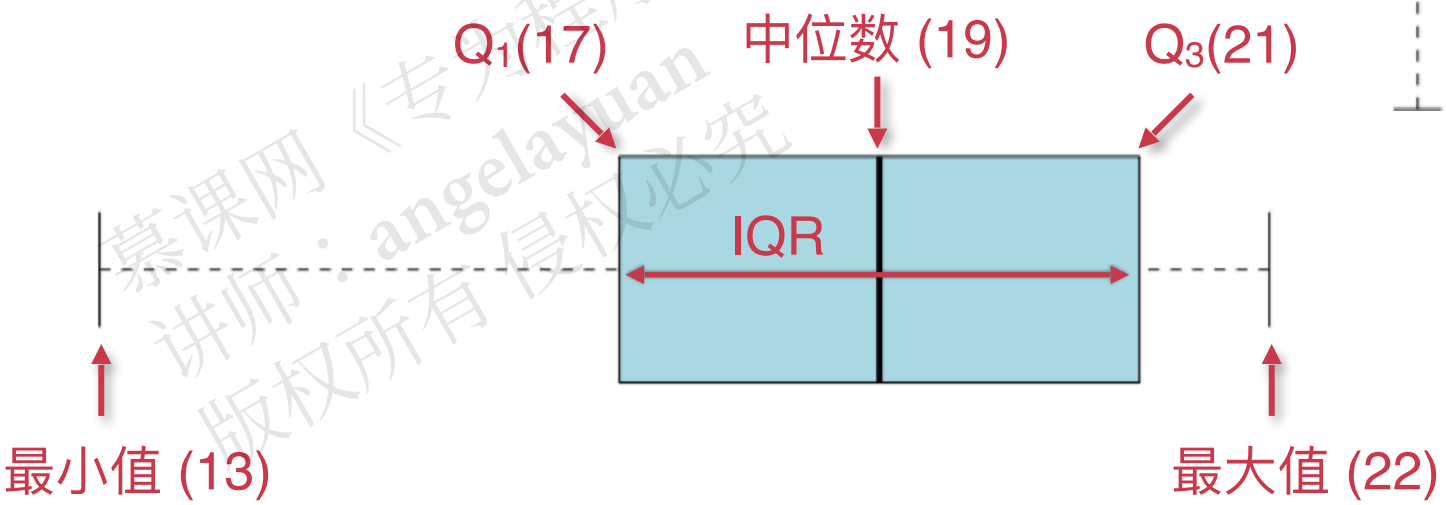
- 四分位数



数值变量

- 箱线图(box plot)

13, 17, 17, 17, 18, 19, 19, 19, 20, 21, 21, 21, 21, 22



小结

	(等距)数值变量
表	频率表
图	频率直方图、箱图
集中趋势	众数、中位数、均值
离散趋势	极差、分位数、四分位数

数值变量

鸢尾花花瓣的长度 (等比; =, ≠, >, <, +, −, ×, ÷)

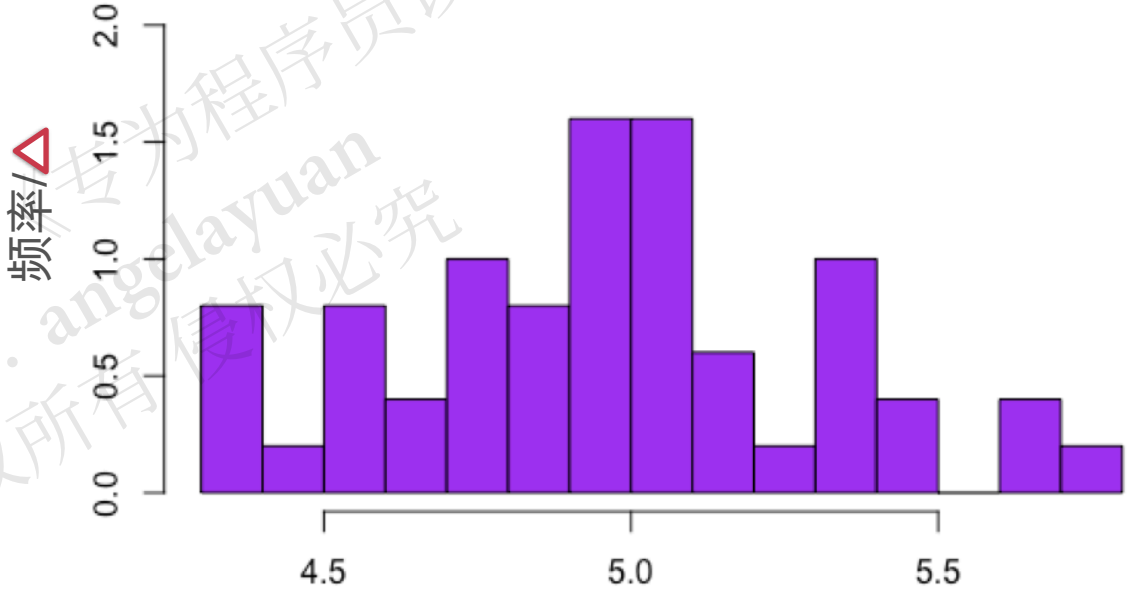
n = 50 (观测了50朵鸢尾花花瓣的长度)

4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
1	3	1	4	2	5	4	8

5.1	5.2	5.3	5.4	5.5	5.7	5.8
8	3	1	5	2	2	1

数值变量

鸢尾花花瓣的长度 (等比; $=, \neq, >, <, +, -, \times, \div$)



数值变量

鸢尾花花瓣的长度 (等比; =, ≠, >, <, +, −, ×, ÷)

4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
1	3	1	4	2	5	4	8

5.1	5.2	5.3	5.4	5.5	5.7	5.8
8	3	1	5	2	2	1

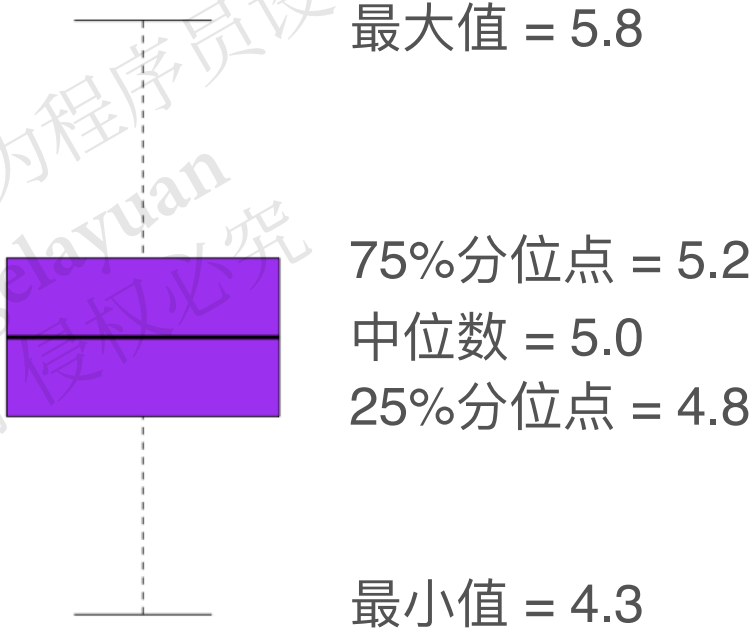
众数 = 5.0和5.1　　极差 = 5.8 - 4.3 = 1.5

均值 = (4.3*1 + 4.4*3 + … + 5.8*1)/50 = 5.0

中位数 = 5.0

数值变量

鸢尾花花瓣的长度 (等比; =, ≠, >, <, +, −, ×, ÷)



数值变量

鸢尾花花瓣的长度 (等比; $=, \neq, >, <, +, -, \times, \div$)

离散趋势: 方差(variance)和标准差(standard deviation)

- 方差: 每一个观测值与均值之间的差异的平方和的平均数

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

数值变量

鸢尾花花瓣的长度 (等比; =, ≠, >, <, +, −, ×, ÷)

4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
1	3	1	4	2	5	4	8

5.1	5.2	5.3	5.4	5.5	5.7	5.8
8	3	1	5	2	2	1

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

均值 = 5.0

方差 = $\frac{(4.3-5.0)^2+3*(4.4-5.0)^2+\cdots+2*(5.7-5.0)^2+(5.8-5.0)^2}{50}$

= 0.124

数值变量

鸢尾花花瓣的长度 (等比; =, ≠, >, <, +, −, ×, ÷)

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}}$$

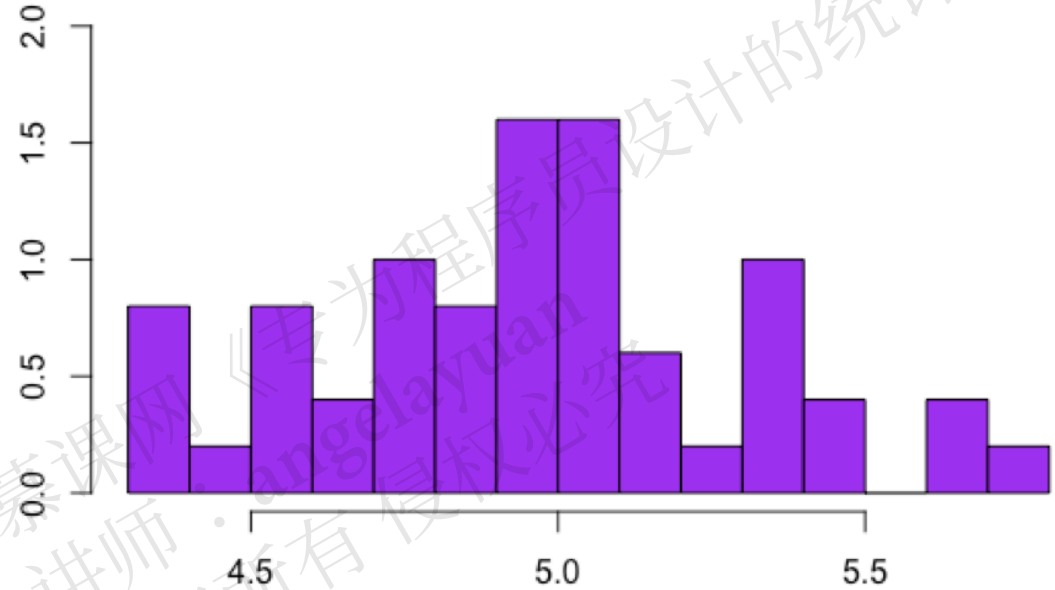
标准差 = 方差开根号 = 0.352

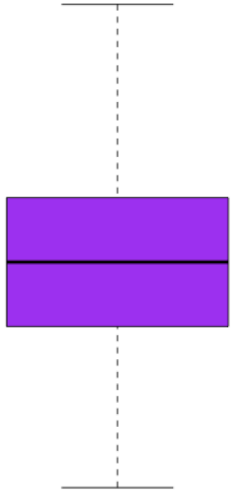
标准差与原观测值具有相同的单位

小结

	(等比)数值变量
表	频率表
图	频率直方图、箱图
集中趋势	众数、中位数、均值
离散趋势	极差、分位数、方差、标准差

分布的形状

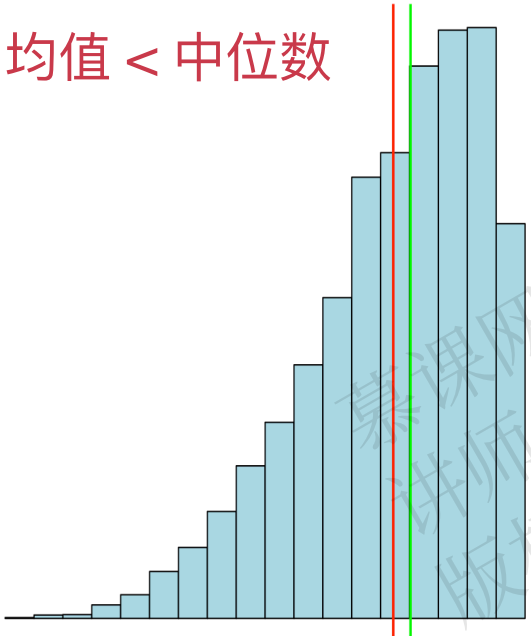




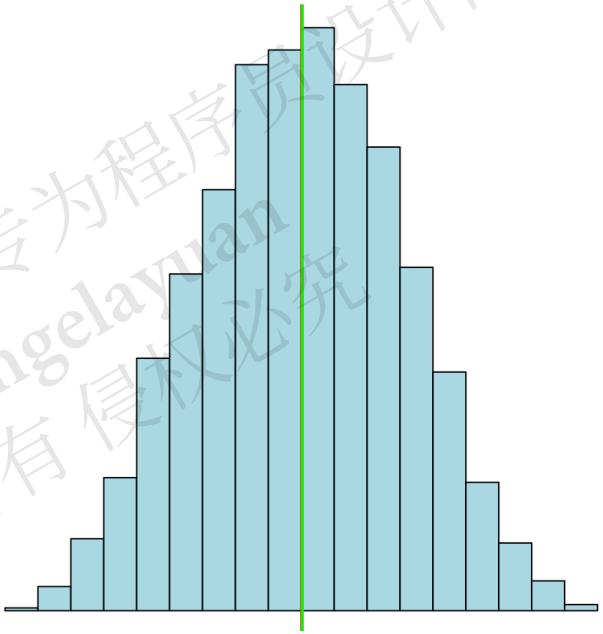
偏度 (skewness)

左偏

均值 < 中位数

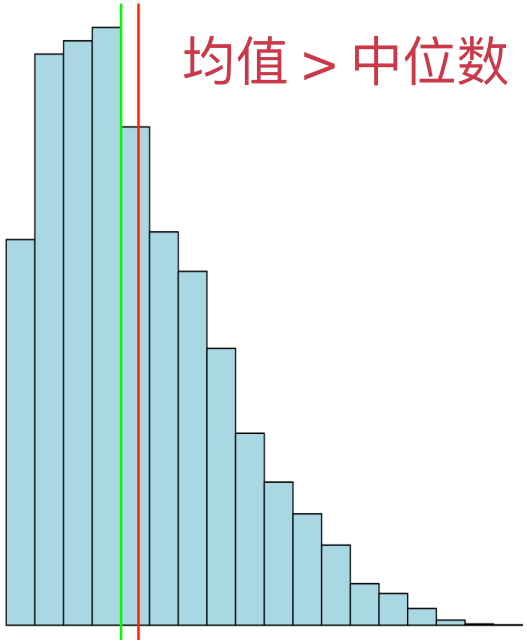


对称



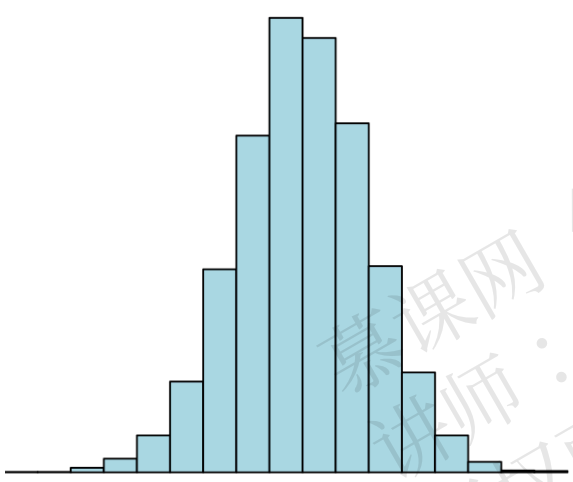
右偏

均值 > 中位数

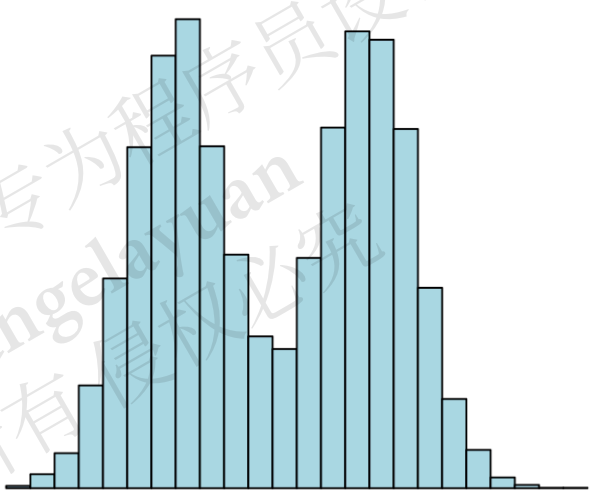


形态 (modality)

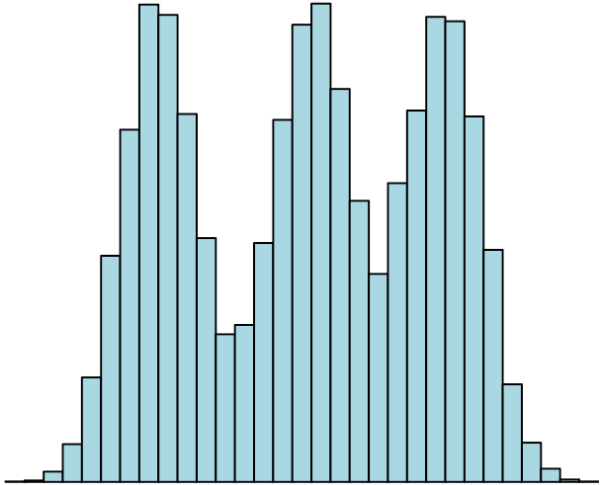
单峰 (unimodal)



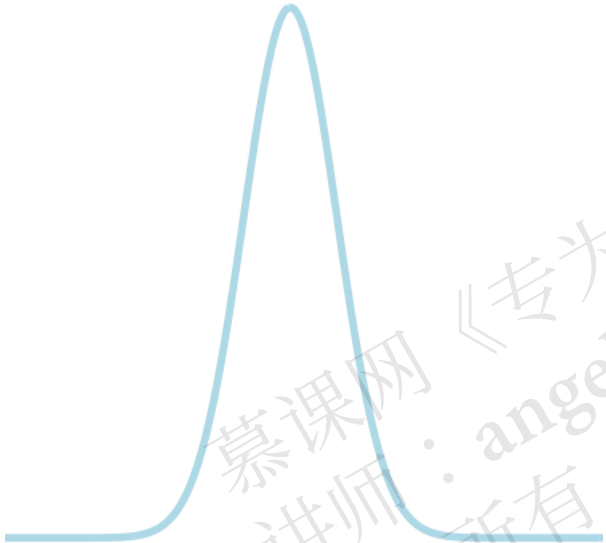
双峰 (bimodal)

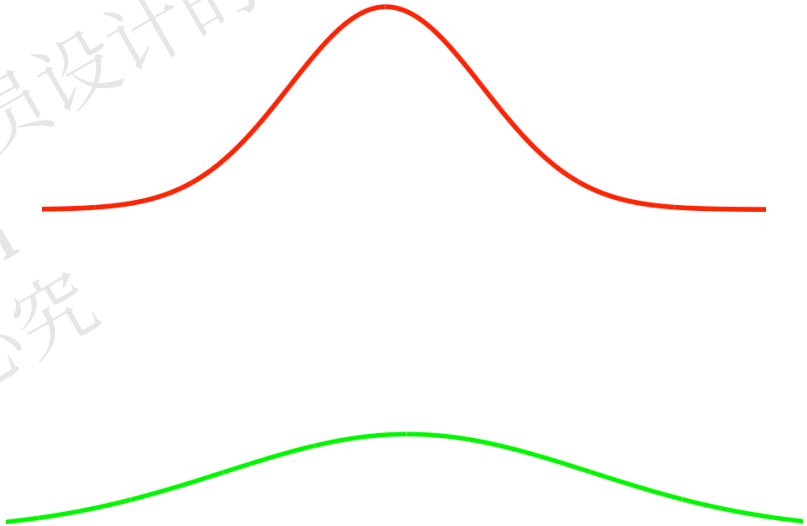


多峰 (multimodal)



峰度 (kurtosis)





峰尖, 尾平, 数据向中心聚拢程度高

扁平, 数据向中心聚拢程度低

变量间的关系

两个分类变量的关系

泰坦尼克号的325名乘客: 幸存(是/否); 年龄(儿童/成人)

编号	幸存	年龄
1	是	成人
2	否	儿童
…	…	…
325	否	儿童

两个分类变量的关系

年龄与幸存是否有关系?

- 关联表(contingency table)

相对频率表

0%　38%

100%　62%

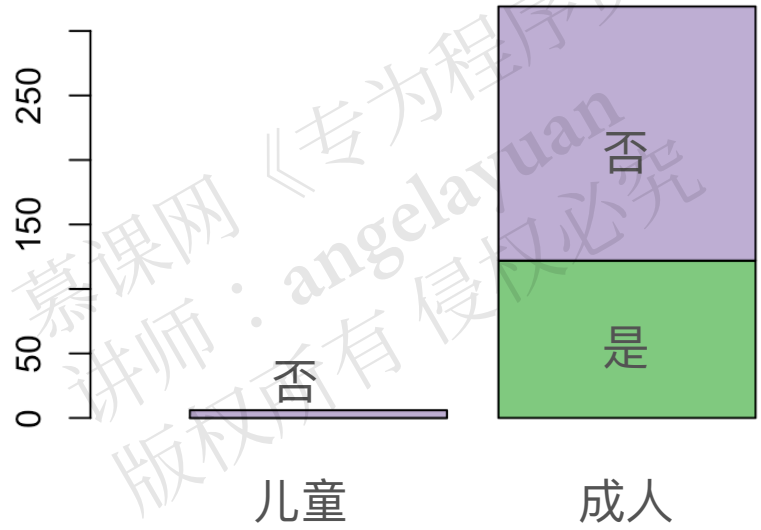
互依不等于因果

		年龄		
		儿童	成人	总数
幸存	是	0	122	122
	否	6	197	203
	总数	6	319	325

两个分类变量的关系

年龄与幸存是否有关系?

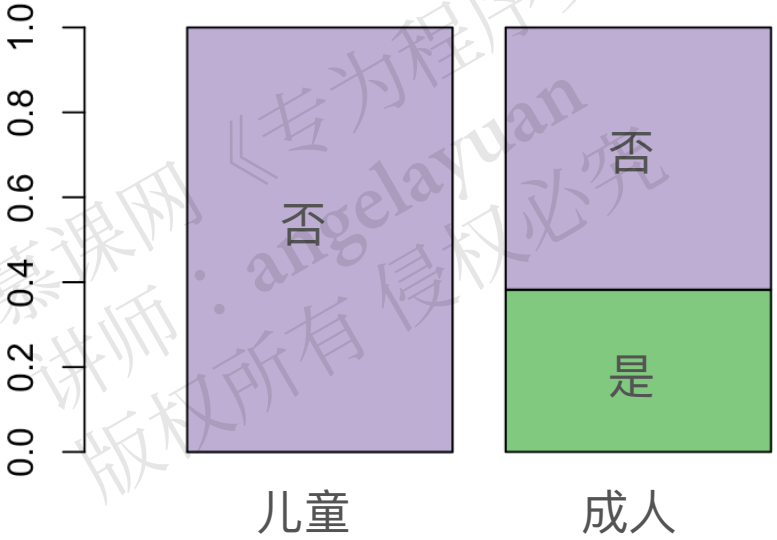
- 分段条形图



两个分类变量的关系

年龄与幸存是否有关系?

- 相对频率分段条形图



两个数值变量的关系

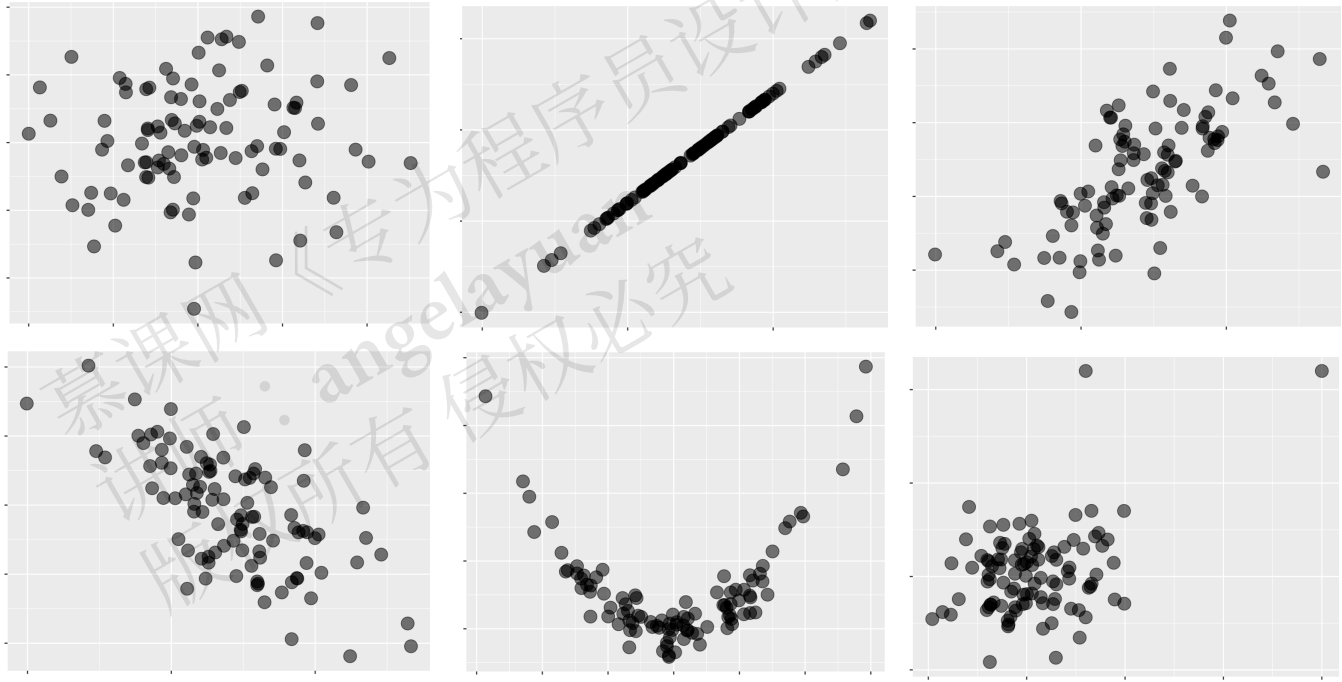
工资与入职时间是否有关？什么关系？关系强弱？

房价与到学校的距离是否有关？什么关系？关系强弱？

抽烟年数与寿命是否有关？什么关系？关系强弱？

两个数值变量的关系

- 散点图 (scatter plot): 方向、形状、强度、极端值



一个数值变量和一个分类变量的关系

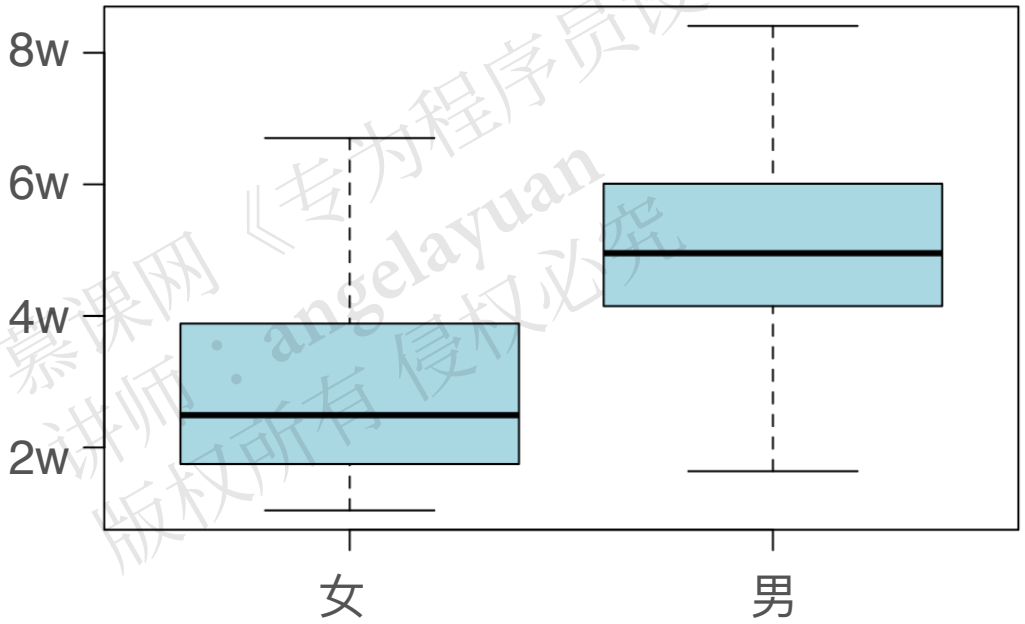
工资与性别是否有关?

房价与学区房是否有关?

抽烟年数与肺癌是否有关?

一个数值变量和一个分类变量的关系

- 并排箱图 (side-by-side box plot)



极端值和缺失值

极端值

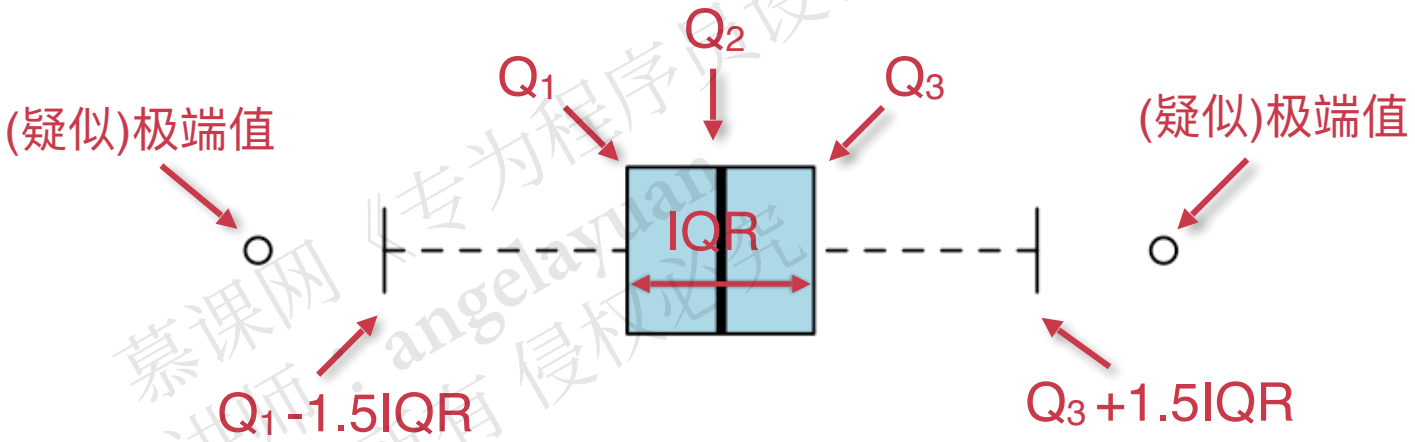
极端值/异常值 (outliers)

- 在一组数据中，小于Q1 - 1.5IQR 或者 大于Q3 + 1.5IQR的数据是疑似极端值

- 在一组数据中，小于Q1 - 3IQR 或者 大于Q3 + 3IQR的数据是极端值

极端值

修正箱线图



极端值

极端值的产生可能源于

- 数据的测量、记录或输入时的错误
- 数据来自不同的总体(例如: 病人 vs 健康人)
- 数据是正确的，但它只体现小概率事件

极端值

极端值可能产生的影响

- 某公司员工的收入水平

  1.2, 1.3, 1.4, 1.5, 1.6, 1.6, 1.8, 2.0, 2.2, 15

  均值: 2.96 vs 1.62　　极差: 13.8 vs 1

  中位数: 1.6　　标准差: 4.24 vs 0.33

  众数: 1.6　　IQR: 2.0 - 1.4 = 0.6 vs 1.9 - 1.35 = 0.55

极端值

如何处理极端值

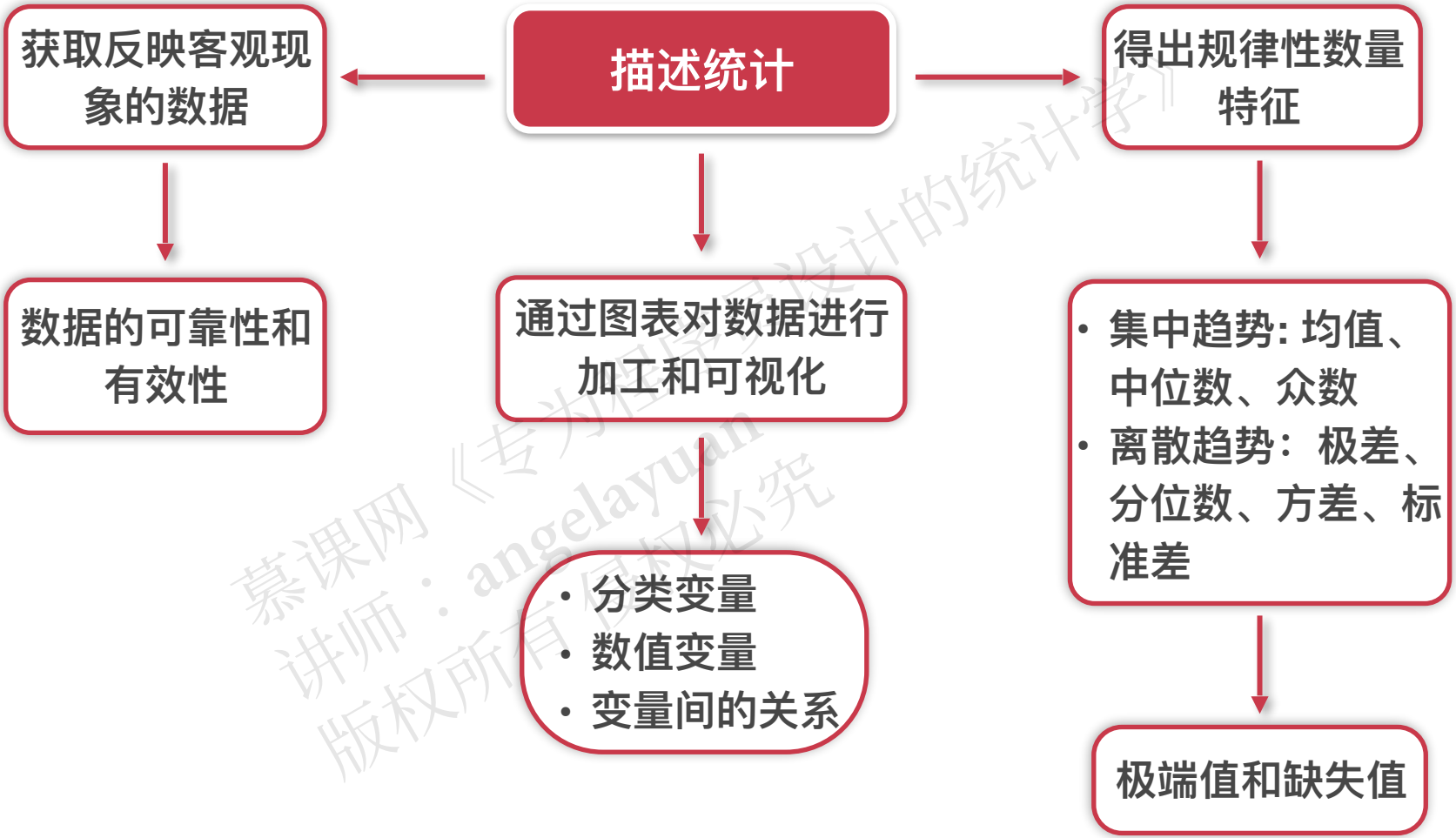
- 如果是由于测量或记录的错误，或其他明显的原因造成的，直接丢弃即可
- 如果极端值出现的原因无法解释，那么，丢弃或保留极端值则需要具体问题具体分析；尽量选用受极端值影响小的指标
- 可以通过对比保留极端值和丢弃极端值对结果的影响，来判断结果是否受到极端值的影响

缺失值

如何处理缺失值

- 如果含有缺失值的观测记录很少，而数据量很大，可以把含有缺失值的观测记录丢弃
- 如果含有缺失值的观测记录很多，需要分析原因，看是否能够把缺失的记录补全
- 如果含有缺失值的观测记录较少，可以使用均值/中位数/众数/最大值等进行替代

本章小结



描述统计的编程实现

小节列表

- 频数
- 频率
- 集中趋势: 众数、中位数、均值
- 离散趋势: 极差、四分位数、方差、标准差
- 折线图和散点图
- 条形图和频率直方图
- 箱线图和并排箱线图