#### 第一章 导论

李德山

四川师范大学商学院

2022年2月26日

#### Contents

1 统计及其应用领域

2 统计数据的类型

3 统计中的几个基本概念

- 1. 身体超重 30% 会使寿命减少多少? 1300 天?
- 2. 身材高的父亲, 其子女的身高也较高吗?
- 3. 吸烟对健康是有害的,吸烟会减少寿命多少天? 2250 天? 每年世界约有330万人死亡。
- 4. 全国各省市旱灾分布情况是怎样的?
- 5. 上课坐在前面的学生平均考试分数比坐在后面的学生高?
- 6. 每年春节成都市或武汉市的人口是如何流动的? 流向哪些城市?
- 7. 辉瑞能治疗新冠肺炎吗? 什么是自然实验? 什么是准自然实验?

- 1. 身体超重 30% 会使寿命减少多少? 1300 天?
- 2. 身材高的父亲, 其子女的身高也较高吗?
- 3. 吸烟对健康是有害的, 吸烟会减少寿命多少天? 2250 天? 每年世
- 4. 全国各省市旱灾分布情况是怎样的?
- 5. 上课坐在前面的学生平均考试分数比坐在后面的学生高?
- 7. 辉瑞能治疗新冠肺炎吗? 什么是自然实验? 什么是准自然实验?

- 1. 身体超重 30% 会使寿命减少多少? 1300 天?
- 2. 身材高的父亲, 其子女的身高也较高吗?
- 3. 吸烟对健康是有害的,吸烟会减少寿命多少天? 2250 天? 每年世界约有 330 万人死亡。
- 4. 全国各省市旱灾分布情况是怎样的?
- 5. 上课坐在前面的学生平均考试分数比坐在后面的学生高?
- 6. 每年春节成都市或武汉市的人口是如何流动的? 流向哪些城市?
- 7. 辉瑞能治疗新冠肺炎吗? 什么是自然实验? 什么是准自然实验?

- 1. 身体超重 30% 会使寿命减少多少? 1300 天?
- 2. 身材高的父亲, 其子女的身高也较高吗?
- 3. 吸烟对健康是有害的,吸烟会减少寿命多少天? 2250 天? 每年世界约有 330 万人死亡。
- 4. 全国各省市旱灾分布情况是怎样的?
- 5. 上课坐在前面的学生平均考试分数比坐在后面的学生高?
- 6. 每年春节成都市或武汉市的人口是如何流动的? 流向哪些城市?
- 7. 辉瑞能治疗新冠肺炎吗? 什么是自然实验? 什么是准自然实验?

第一章 导论

- 1. 身体超重 30% 会使寿命减少多少? 1300 天?
- 2. 身材高的父亲, 其子女的身高也较高吗?
- 3. 吸烟对健康是有害的,吸烟会减少寿命多少天? 2250 天? 每年世界约有 330 万人死亡。
- 4. 全国各省市旱灾分布情况是怎样的?
- 5. 上课坐在前面的学生平均考试分数比坐在后面的学生高?
- 6. 每年春节成都市或武汉市的人口是如何流动的? 流向哪些城市?
- 7. 辉瑞能治疗新冠肺炎吗? 什么是自然实验? 什么是准自然实验?

- 1. 身体超重 30% 会使寿命减少多少? 1300 天?
- 2. 身材高的父亲, 其子女的身高也较高吗?
- 3. 吸烟对健康是有害的,吸烟会减少寿命多少天? 2250 天? 每年世界约有 330 万人死亡。
- 4. 全国各省市旱灾分布情况是怎样的?
- 5. 上课坐在前面的学生平均考试分数比坐在后面的学生高?
- 6. 每年春节成都市或武汉市的人口是如何流动的? 流向哪些城市?
- 7. 辉瑞能治疗新冠肺炎吗? 什么是自然实验? 什么是准自然实验?

- 1. 身体超重 30% 会使寿命减少多少? 1300 天?
- 2. 身材高的父亲, 其子女的身高也较高吗?
- 3. 吸烟对健康是有害的,吸烟会减少寿命多少天? 2250 天? 每年世界约有 330 万人死亡。
- 4. 全国各省市旱灾分布情况是怎样的?
- 5. 上课坐在前面的学生平均考试分数比坐在后面的学生高?
- 6. 每年春节成都市或武汉市的人口是如何流动的?流向哪些城市?
- 7. 辉瑞能治疗新冠肺炎吗? 什么是自然实验? 什么是准自然实验?

2 统计数据的类型

③ 统计中的几个基本概念



## 什么是统计学

• 统计学是一门收集、处理、分析、解释数据并从数据中得出结论的 科学

收集数据:取得数据

处理数据:图表展示

分析据数:利用统计方法分析数据

数据解释: 结果的说明

• 得结论结: 从数据分析中得出客观结论(检验)



统计在工商管理中的一些应用:企业发展战略、产品质量管理、市场研究、财务分析、经济预测、人力资源管理,等等。



Figure: 统计学主要应用领域

econometrics (经济计量学) education (教育学) election forecasting and projection (选举预测和策划) engineering (工程) epidemiology (流行病学) finance (金融) fisheries research (水产海业研究) gambling (赌博) genetics (遗传学) geography (地理学) geology (地质学) historical research (历史研究) human genetics (人类遗传学) hydrology (水文学) industry (工业) linguistics (语言学) literature (文学) manpower planning (劳动力计划)

management science (管理科学) marketing (市场营销学) medical diagnosis (医学诊断) meteorology (气象学) military science (军事科学) nuclear material safeguards (核材料安全管理) ophthalmology (眼科学) pharmaceutics (制药学) physics (物理学) political science (政治学) psychology (心理学) psychophysics (心理物理学) quality control (质量控制) religious studies (宗教研究) sociology (社会学) survey sampling (调查抽样) taxonomy (分类学) weather modification (气象改善)

Figure: 统计学主要应用领域

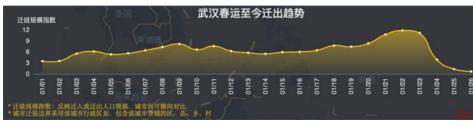


Figure: 武汉春运人口迁出趋势



Figure: 武汉人口迁出目的地地图

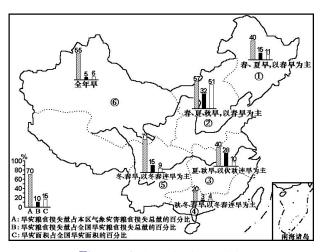


Figure: 中国不同地区的旱灾分布

#### 统计方法

• 数据分析是通过统计方法研究数据,其所用的方法可分为描述统计和推断统计

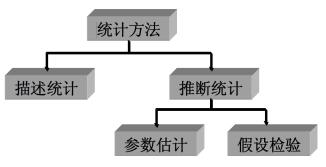


Figure: 统计方法

#### 描述性统计

描述性统计:研究数据收集、处理、汇总、图表描述、概括与分析等统计方法

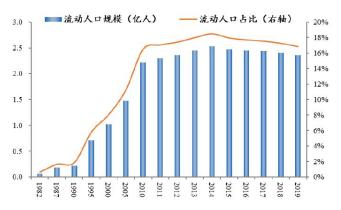


Figure: 中国流动人口规模变动趋势

#### 推断统计

• 推断统计: 研究如何利用样本数据来推断总体特征的统计方法

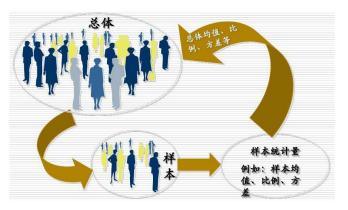


Figure: 推断统计

- 例 1: 1061 年《本草图经》: 如何判断一个人参的真假?

- 例 1: 1061 年《本草图经》: 如何判断一个人参的真假?
  - 两个人(其中一个人 A 口里含着人参), 地上走好几公里, 这样他们 必然会大喘气, 最后发现含着人参的那个人就不怎么大喘气



- 例 1: 1061 年《本草图经》: 如何判断一个人参的真假?
  - 两个人(其中一个人 A 口里含着人参), 地上走好几公里, 这样他们 必然会大喘气, 最后发现含着人参的那个人就不怎么大喘气
  - 但是这个对照实验还是有问题的, 比如 B 这个人身体素质本来就比 较弱, A 身体素质比较好

- 例 1: 1061 年《本草图经》: 如何判断一个人参的真假?
  - 两个人(其中一个人 A 口里含着人参). 地上走好几公里, 这样他们 必然会大喘气, 最后发现含着人参的那个人就不怎么大喘气
  - 但是这个对照实验还是有问题的, 比如 B 这个人身体素质本来就比 较弱, A 身体素质比较好
- 例 2: 1747 年, 英国, 詹姆斯. 林德 (海军军医)。治疗坏血病——酸物质

- 例 1: 1061 年《本草图经》:如何判断一个人参的真假?
  - 两个人(其中一个人 A 口里含着人参). 地上走好几公里, 这样他们 必然会大喘气, 最后发现含着人参的那个人就不怎么大喘气
  - 但是这个对照实验还是有问题的, 比如 B 这个人身体素质本来就比 较弱. A 身体素质比较好
- 例 2: 1747 年, 英国, 詹姆斯, 林德(海军军医)。治疗坏血病——酸物质
  - 12 个水手, 分成 6 组——不同酸的物质。苹果汁、低浓度的硫酸、 醋、海水、橙子和柠檬、大蒜和芥末

- 例 1: 1061 年《本草图经》:如何判断一个人参的真假?
  - 两个人(其中一个人 A 口里含着人参). 地上走好几公里, 这样他们 必然会大喘气, 最后发现含着人参的那个人就不怎么大喘气
  - 但是这个对照实验还是有问题的, 比如 B 这个人身体素质本来就比 较弱. A 身体素质比较好
- 例 2: 1747 年, 英国, 詹姆斯, 林德(海军军医)。治疗坏血病——酸物质
  - 12 个水手, 分成 6 组——不同酸的物质。苹果汁、低浓度的硫酸、 醋、海水、橙子和柠檬、大蒜和芥末
  - 新鲜的水果可以治疗坏死病



- 例 1: 1061 年《本草图经》:如何判断一个人参的真假?
  - 两个人(其中一个人 A 口里含着人参). 地上走好几公里, 这样他们 必然会大喘气, 最后发现含着人参的那个人就不怎么大喘气
  - 但是这个对照实验还是有问题的, 比如 B 这个人身体素质本来就比 较弱. A 身体素质比较好
- 例 2: 1747 年, 英国, 詹姆斯, 林德(海军军医)。治疗坏血病——酸物质
  - 12 个水手, 分成 6 组——不同酸的物质。苹果汁、低浓度的硫酸、 醋、海水、橙子和柠檬、大蒜和芥末
  - 新鲜的水果可以治疗坏死病
  - 但是他并不知道是因为缺乏维生素 C

## 统计学的应用

- 社会统计、环境统计(社会发展与评价;持续发展与环境保护;资源保护与利用,政府统计数据收集等)
- 生物统计、医学统计(分子生物学中的统计方法、高科技农业研究中的统计方法、流行病规律研究与探索的统计方法、人类染色体工程研究中的统计方法)、制药业(比对试验)、疾病诊断(Bayes 方法,图模型等)、病理分析、疾病的控制

● 统计及其应用领域

2 统计数据的类型

③ 统计中的几个基本概念



• 根据所采用的定义不同,统计数据的分类有多种类型

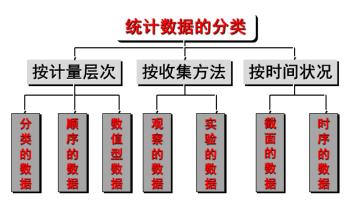
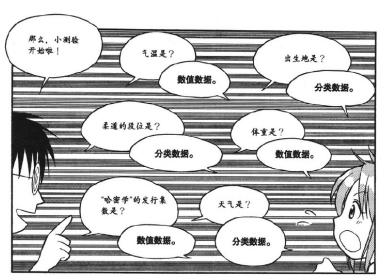


Figure: 统计数据的分类

- 1. 分类数据 (categorical data)
  - 只能归于某一类别的非数字型数据
  - 对事物进行分类的结果,数据表现为类别,用文字来表述
  - 例如, 人口按性别分为男、女两类
- 2. 顺序数据 (rank data)
  - 只能归于某一有序类别的非数字型数据
  - 对事物类别顺序的测度,数据表现为类别,用文字来表述
  - 例如, 产品分为一等品、二等品、三等品、次品等
- 3. 数值型数据 (metric data)
  - 按数字尺度测量的观察值
  - 结果表现为具体的数值, 对事物的精确测度
  - 例如: 身高为 175cm、168cm、183cm

- 1. 分类数据 (categorical data)
  - 只能归于某一类别的非数字型数据
  - 对事物进行分类的结果,数据表现为类别,用文字来表述
  - 例如, 人口按性别分为男、女两类
- 2. 顺序数据 (rank data)
  - 只能归于某一有序类别的非数字型数据
  - 对事物类别顺序的测度,数据表现为类别,用文字来表述
  - 例如,产品分为一等品、二等品、三等品、次品等
- 3. 数值型数据 (metric data)
  - 按数字尺度测量的观察值
  - 结果表现为具体的数值, 对事物的精确测度
  - 例如: 身高为 175cm、168cm、183cm

- 1. 分类数据 (categorical data)
  - 只能归于某一类别的非数字型数据
  - 对事物进行分类的结果,数据表现为类别,用文字来表述
  - 例如, 人口按性别分为男、女两类
- 2. 顺序数据 (rank data)
  - 只能归于某一有序类别的非数字型数据
  - 对事物类别顺序的测度,数据表现为类别,用文字来表述
  - 例如,产品分为一等品、二等品、三等品、次品等
- 3. **数值型数据** (metric data)
  - 按数字尺度测量的观察值
  - 结果表现为具体的数值, 对事物的精确测度
  - 例如: 身高为 175cm、168cm、183cm



- 1. 观测的数据 (observational data)
  - 通过调查或观测而收集到的数据
  - 在没有对事物人为控制的条件下而得到的
  - 有关社会经济现象的统计数据几乎都是观测数据
- 2. 实验的数据 (experimental data)
  - 在实验中控制实验对象而收集到的数据
  - 比如,对一种新药疗效的实验,对一种新的农作物品种的实验等
  - 自然科学领域的数据大多数都为实验数据

- 1. 截面数据 (cross-sectional data)
  - 在相同或近似相同的时间点上收集的数据
  - 描述现象在某一时刻的变化情况
  - 比如, 2020 年我国各地区的国内生产总值数据
- 2. 时间序列数据 (time series data)
  - 在不同时间上收集到的数据
  - 描述现象随时间变化的情况
  - 比如, 2000-2020 年国内生产总值数据

• 下面这种定义就是广义的统计数据分类

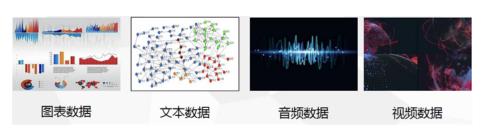


Figure: 统计数据的分类

● 统计及其应用领域

2 统计数据的类型

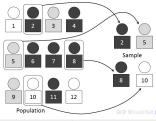
3 统计中的几个基本概念

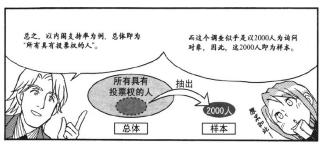
#### • 1. 总体 (population)

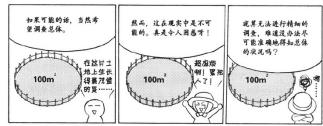
- 所研究的全部个体 (数据) 的集合, 其中的每一个个体也称为元素
- 分为有限总体和无限总体
- 特点: 大量性、同质性、差异性

#### • 2. 样本 (sample)

- 从总体中抽取的一部分元素的集合
- 构成样本的元素的数目称为样本容量或样本量







- 1. 总体是所要研究的对象,而样本则是所要观测的对象;
- 2. 样本是总体的代表和缩影
- 3. 样本是用来推断总体的;
- 4. 随着考察角度的改变,总体和样本的角色是可以改变的

- 1. 总体是所要研究的对象, 而样本则是所要观测的对象;
- 2. 样本是总体的代表和缩影
- 3. 样本是用来推断总体的;
- 4. 随着考察角度的改变,总体和样本的角色是可以改变的

- 1. 总体是所要研究的对象,而样本则是所要观测的对象;
- 2. 样本是总体的代表和缩影
- 3. 样本是用来推断总体的;
- 4. 随着考察角度的改变,总体和样本的角色是可以改变的

- 1. 总体是所要研究的对象,而样本则是所要观测的对象;
- 2. 样本是总体的代表和缩影
- 3. 样本是用来推断总体的;
- 4. 随着考察角度的改变, 总体和样本的角色是可以改变的

- 1. 总体是所要研究的对象, 而样本则是所要观测的对象;
- 2. 样本是总体的代表和缩影
- 3. 样本是用来推断总体的;
- 4. 随着考察角度的改变,总体和样本的角色是可以改变的

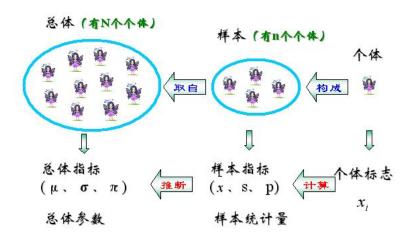


#### • 1. **参数** (parameter)

- 描述总体特征的概括性数字度量,是研究者想要了解的总体的某种特征值
- 所关心的参数主要有总体均值、标准差、总体比例等
- 总体参数通常用希腊字母表示

#### • 2. 统计量 (statistic)

- 用来描述样本特征的概括性数字度量,它是根据样本数据计算出来的 一些量,是样本的函数
- 所关心的样本统计量有样本均值、样本标准差、样本比例等
- 样本统计量通常用小写英文字母来表示



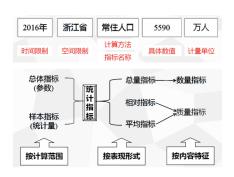
#### • 变量的分类

- 分类变量 (categorical variable): 说明事物类别的名称。定类变量
- 顺序变量 (rank variable): 说明事物有序类别的名称。定序变量
- 数值型变量 (metric variable): 说明事物数字特征的名称。它又可分为离散变量和连续变量
- 随机变量和非随机变量
- 经验变量和理论变量

定类变量		定序变量	定量变量	
城区	是否为学区房	楼层	面积	房价
朝阳	否	高	46.06	4.885
海淀	是	低	59.09	7.166
丰台	是	高	36.74	5.443
海淀	否	中	81.00	5.797

#### • 统计指标

- 反应总体综合数量特征的概念及其数值
- 构成要素: 名称、计算方法、时间、空间、数值和计量单位



### 相关资料

#### • 相关政府部门的数据资料

- 国家统计局: http://www.stats.gov.cn/; 中国历年统计数据
- 工业和信息化部: http://www.miit.gov.cn; 尤其是有关工业运行及信息化相关数据
- 中国人民银行: http://www.pbc.gov.cn/; 中国金融市场政策及运行相关数据
- 银监会: http://www.cbrc.gov.cn; 银行金融相关数据
- 中国海关: http://www.customs.gov.cn; 中国进出口相关数据
- 国家知识产权局: http://www.sipo.gov.cn; 专利相关查询
- .....

### 统计学的前世今生

- 统计学的发展大致划分为三个阶段:
  - 第一阶段称之为"城邦政情"阶段。
  - 第二阶段称之为"政治算数"阶段。
  - 第三阶段称之为"统计分析科学"阶段。
- 国势学派(亚里士多德)、政治算术学派(威廉. 配弟)、数理统计学派(凯特勒)、 社会统计学派(恩格尔)等



Figure: 统计史上的著名人物: 配第、凯特勒和皮尔逊

### 统计学的前世今生

- 统计学经过漫长的发展,尤其是计算机的大量应用,目前包括但不限于下面这些分支(或者交叉领域):
  - (1) 概率 (比如 Stochastic Process)、计算统计理论 (比如 Asymptotic Theory);
  - (2) 统计模型(比如 Egression Model、Stratification、Clustering、Classification 等) 以及 Test 的发展(比如 Likelihood Ratio Test、Wald Test、Permutation Test 等);
  - (3) 计算统计方法 (比如 Monte Carlo Simulation, Bootstrap 等);
  - (4) 数据采集(比如 Census, Survey, Clinical Trial 等);
  - (5) 机器学习 (Machine Learning);
  - (6) 数据挖掘 (Data Mining), 等等。

• SAS, SPSS, R, Stata, Minitab, Matlab, Mathematica, Python, Excel \( \begin{array}{c} \begin{array}{c} \epsilon \\ \epsilon \end{array} \end{array} \)

• 主流商业数据分析工具



















主流开源数据分析工具

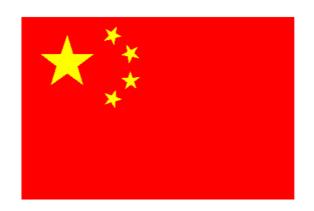




Figure: 常用统计分析软件

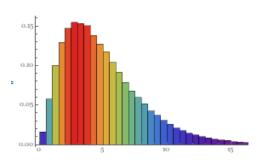
#### R 代码

```
1
     win.graph(width=5,height=3);plot.new()
 2
     par(mar=c(0,0,0,0))
 3
     plot.window(xlim = c(0, 1), ylim = c(0, 1),
 4
     xaxs="i",yaxs="i")
 5
     image(matrix(1:1000,10), col = rep('red',100),add=T, axes=F)
     symbols(0.2.0.8, stars=as.matrix(t(c(0.1.0.05.0.1.0.05,
6
 7
     0.1.0.05.0.1.0.05.0.1.0.05)/2)),add=T, fg='vellow',bg="vellow", lwd=1, inches=F)
     symbols(0.37,0.9, stars=as.matrix(t(c(0.1,0.05,0.1,0.05,
 8
9
     0.1,0.05,0.1,0.05,0.1,0.05)/4)),add=T,fg='yellow', bg="yellow", lwd=1, inches=F)
10
     symbols(0.39,0.75, stars=as.matrix(t(c(0.1,0.05,0.1,0.05,
11
     0.1,0.05,0.1,0.05,0.1,0.05)/4)),add=T, fg='yellow',bg="yellow", lwd=1, inches=F)
12
     symbols(0.35,0.62, stars=as.matrix(t(c(0.1,0.05,0.1,0.05,
13
     0.1,0.05,0.1,0.05,0.1,0.05)/4)),add=T,fg='yellow', bg="yellow", lwd=1, inches=F)
14
     symbols(0.25,0.56, stars=as.matrix(t(c(0.1,0.05,0.1,0.05,
15
     0.1,0.05,0.1,0.05,0.1,0.05)/4)),add=T,fg='yellow', bg="yellow", lwd=1, inches=F)
```



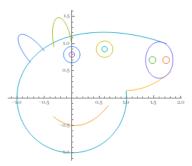
#### Mathematica 代码

- data = RandomVariate[ChiSquareDistribution [5], 10^5]
- Histogram[data, 40, "ProbabilityDensity", ColorFunction -> "Rainbow"]



导论

 $\begin{aligned} & \textbf{ParametricPlot}[\{(1+t^10) \ \{1 \ \text{Cos}[\backslash [Pi]/6-2 \ \backslash [Pi] \ (1/16+15/16 \ t)], \\ & \textbf{Sin}[\backslash [Pi]/6-2 \ \backslash [Pi] \ (1/16+15/16 \ t)] \\ & \textbf{Pi}, \textbf{t*4/5+1,1/3} \ \textbf{t^2+1/7}, \textbf{\{1/4} \ \textbf{Cos}[2 \ \backslash [Pi] \ \textbf{t}] + 8/5,1/3 \ \textbf{Sin}[2 \ \backslash [Pi] \ \textbf{t}] + 7/10\}, \textbf{\{1/16} \ \textbf{Cos}[2 \ \backslash [Pi] \ \textbf{t}] \\ & \textbf{P*5-1/8,1/16} \ \textbf{Sin}[2 \ \backslash [Pi] \ \textbf{t}] + 7/10\}, \textbf{\{1/16} \ \textbf{Cos}[2 \ \backslash [Pi] \ \textbf{t}] + 8/5+1/8,1/16 \ \textbf{Sin}[2 \ \backslash [Pi] \ \textbf{t}] \\ & \textbf{P*7/10}, 0.15 \ \textbf{\{Cos}[2 \ \backslash [Pi] \ \textbf{t}], \textbf{Sin}[2 \ \backslash [Pi] \ \textbf{t}] + 0.8/0.15\}, 0.15 \ \textbf{\{Cos}[2 \ \backslash [Pi] \ \textbf{t}] + 0.6/0.15, \textbf{Sin}[2 \ \backslash [Pi] \ \textbf{t}] \\ & \textbf{P*0.9/0.15}, 0.05 \ \textbf{\{Cos}[2 \ \backslash [Pi] \ \textbf{t}], \textbf{Sin}[2 \ \backslash [Pi] \ \textbf{t}] + 0.8/0.05\}, 0.05 \ \textbf{\{Cos}[2 \ \backslash [Pi] \ \textbf{t}] + 0.6/0.05, \textbf{Sin}[2 \ \backslash [Pi] \ \textbf{t}] \\ & \textbf{P*1} \ \textbf{t}] + 0.9/0.05\}, \textbf{\{t-1/3,(t-2/5)^2-1/2\}}, \textbf{\{(3/2-2 \ (t-1/2)^2) \ \textbf{Cos}[t/3+(2 \ \backslash [Pi])/3], (3/2-2 \ (t-1/2)^2) \ \textbf{Sin}[t/3+(2 \ \backslash [Pi]/3], \textbf{\{(3/2-2 \ (t-1/2)^2) \ \textbf{Cos}[t/3+(Pi]/2], (3/2-2 \ (t-1/2)^2) \ \textbf{Sin}[t/3+(Pi]/2]), \textbf{\{(1/2)^2, (1/2)$ 



39 / 41

# 参考资料

- 贾俊平. 《统计学》(第八版) [M]. 北京: 中国人民大学出版社, 2021.
- 管于华. 《统计学》[M]. 北京, 高等教育出版社, 2013 年。

# Q&A THANK YOU