# 数据世界探秘

## 第三章　数据分析的道与术

# 授课要点

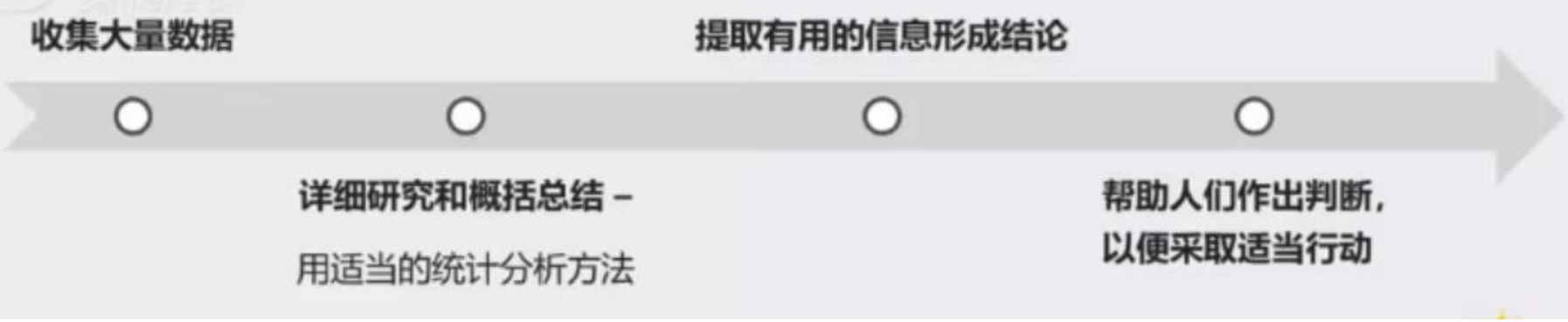
1. 数据分析定义与价值
2. 统计建模工具
3. 探索性数据分析框架
4. 统计建模技术

# 主要内容

- 什么是数据分析（道）
  - ✓ 数据分析是什么

# 数据分析的定义

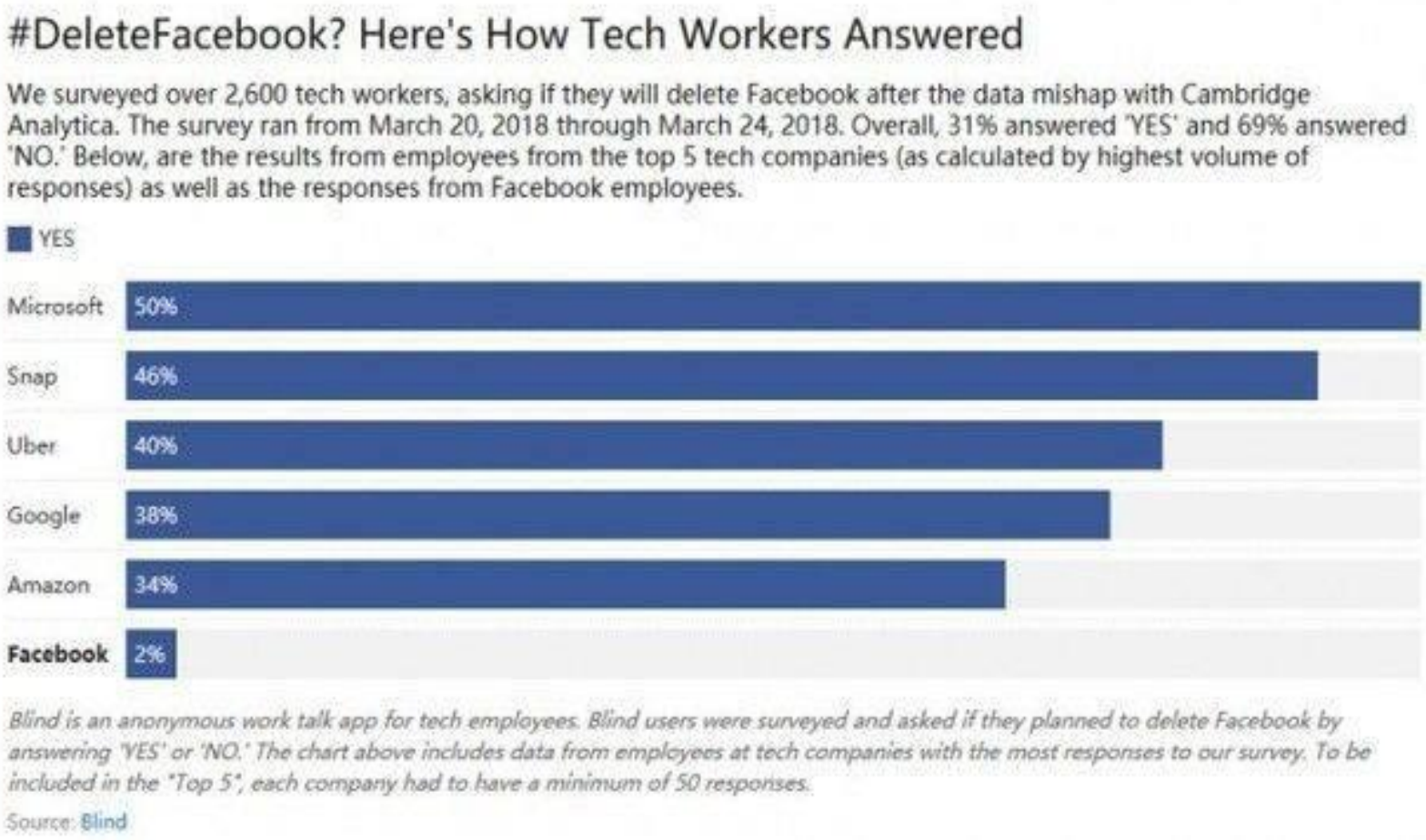
- 数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，提取有用的信息形成结论并对数据加以详细研究和概括总结的过程。在实际应用中，数据分析可以帮助人们作出判断，以便采取适当行动。



# FACEBOOK的数据泄露与政治斗争

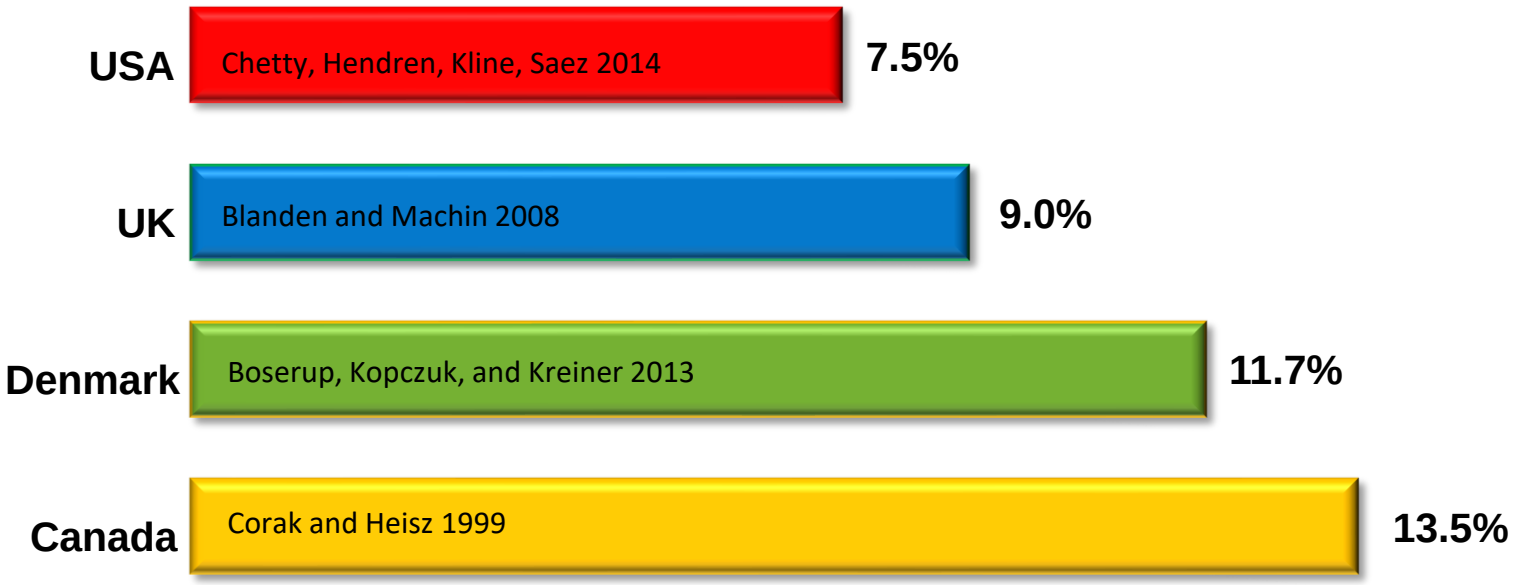
- **背景**
  - 2018年3月17日当地时间，美国《纽约时报》和英国《观察者报》共同发布了深度报道。声称英国一家基于数据分析的政治咨询公司剑桥分析被控利用Facebook的信息管理不力，窃取了高达5000万名Facebook用户的个人资料，在2016年美国大选期间帮助共和党候选人、现任总统特朗普投放针对性的政治广告，可能影响到大选结果。这篇报道在世界范围内引发了轩然大波。
- **影响**
  - #删除Facebook账号#迅速成为了Twitter的热门话题。连WhatsApp联合创始人布莱恩·阿克顿(Brian Acton)也不失时机地呼吁删除Facebook账号。(在把WhatsApp作价190亿美元出售给Facebook三年后，他已经离开Facebook。)

# FACEBOOK的数据泄露与政治斗争



# 美国梦?

- Chance that a child born to parents in the bottom fifth of the income distribution reaches the top fifth:




- 在加拿大实现“美国梦”的机会几乎是在美国的两倍。

# 为什么美国的概率较小？

- 仅从国家层面数据维度很难回答这个问题：
  - 不同国家之间的众多差异使得在不同的解释之间进行检验变得困难；
  - **问题**：数据点较少。
- 至今，社会科学家还没有足够的数据来研究这样的政策问题，因此社会科学一直是一个理论领域：
  - **问题**：对于一个给定的问题，五个经济学家通常有五个不同的答案。
- 如今，由于数据的日益普及，社会科学正成为一个更加经验性的领域：
  - 使用真实数据测试和改进理论。

# 假如上海地铁延长运营1小时


在晚上20点至22点之间，地铁与出租车的客流波动情况有明显的同步性。但是在22:00-22:10之间，由于地铁的停运，导致出租车的客流需求有个明显的小高峰。

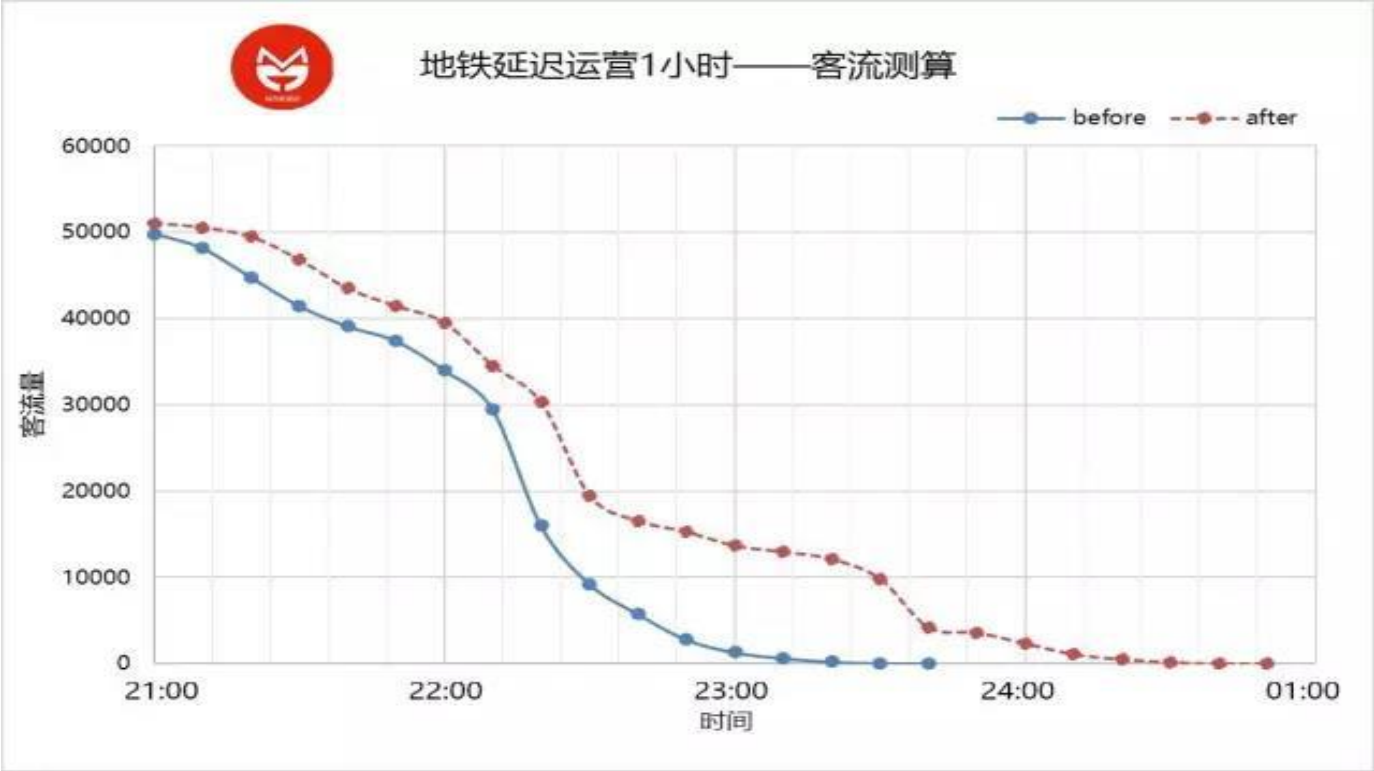
# 假如上海地铁延长运营1小时


在停运前的半小时之内（22：50-23：20），陆家嘴地区的出租车需求有了非常明显的上升，而且其上升的顶峰正是停运的23点20分。

# 假如上海地铁延长运营1小时


轨交站点所服务地区的市民，在地铁停运以后，仍然有对公共交通的需求。随着地铁停运，地铁站点附近地区会有大量的公共交通需求转移到了出租车等非公共交通方式上。这种现象存在于商业区、办公区以及交通枢纽区。

# 假如上海地铁延长运营1小时




若上海地铁系统延运1小时，上海轨道交通的客流量约增加13万人/次。

# 假如上海地铁延长运营1小时


## 4月28日起上海地铁多条线路延时运营

环球网
17-04-25 14:16

央广网上海4月25日消息（记者韩晓余傅闻捷）为进一步提升上海地铁网络服务水平，特别是进一步满足市民乘客周末、节假日夜间出行需求，上海地铁全网络多条线路自2017年4月28日起延长运营时间：10号线和16号线全线，分别常态延时运营25分钟和30分钟；1、2、7、8、9、10号线六条线路在周五、周六延时运营至零点，“五一”、“十一”及元旦节前最后一个工作日也将延时运营。届时，上海中心城区（外环线内）延时线路的地铁车站周末夜间将运行至零点。

10、16号线常态延时运营


天运营，
的地铁
而且晚
分别运
上海

# 数据分析是什么

- 字面拆解 数据+分析
- 有骨有肉方成一个人
  - 分析是骨架（主）
  - 数据是血肉（附）
- **常见错误**
  - 只有数据（机器报表不行吗）
  - 只有分析（你是瞎猜的吧）

# 主要内容

- 什么是数据分析（道）
  - ✓ 数据分析有什么价值

# 案例：菜鸟物流助力阿里巴巴双十一


菜鸟网络通过**分析海量历史数据**，对热卖品在不同城市的销量做出预测，并据此建立前置仓，提前将商品布局在离消费者最近的仓库。

# 案例：依靠数据点石成金的CAPITAL ONE

20世纪80年代的美国信用卡行业
现象：不同客户申请到信用卡几乎完全一样-无差别对待

1990年，理查德.费尔班克斯和纳杰尔.莫里斯意识到，当时的数据处理已经**足够支持更加复杂的客户和信用卡产品匹配策略。**

Signet银行采纳了两人意见，基于丰富数据分析，制定了针对不同客户人群的**精细化运营策略**，降低坏账率，大大提升利润率！

信用卡中心独立出来成立Capital One，成为全球Top 5信用卡公司。

# 什么是做好数据分析的关键?

- “黛玉教香菱学诗”—学诗要先学立意（格局）而不是辞藻（技巧）

写诗的立意与辞藻

同理，做好数据分析与学写诗文一样，真正的关键在技巧（统计技术）之外，而在于对业务的观察、思考与感悟，即**分析的思路**

# 什么是做好数据分析的关键?

- **创新思考：**广阔的知识面和积极的思考力，是分析思路的源泉
- **学：**跨领域的知识面

**经济学**：“价格歧视”原理等

➢ “产品优惠券”案例：很多快餐店经常会发放一大张纸的优惠券汇总，可以撕成一张张小的优惠券，对应不同套餐和优惠。

➢ “杀熟”案例：一位网友曾说，自己经常通过某平台订某个特定酒店的房间，长年价格在380元到400元左右。偶然一次，他用朋友的账号查询后发现是300元；但用自己的账号去查，还是380元。相同的酒店，对于老客户，价格却要高出一大截！

**心理学**：心理需求的满足等

➢张小龙在微信产品观中提到，微信产品设计的原则是“心理需求满足”大于“功能性实现”。

➢用户使用打车应用是为了快速叫到出租车；使用团购应用是为了更便宜的消费

➢微信针对人对人的好奇心，研发了摇一摇，附近的人，漂流瓶等产品；针对人在小圈子中存在感需求，研发了朋友圈

# 什么是做好数据分析的关键?

- **创新思考：**广阔的知识面和积极的思考力，是分析思路的源泉
- **学：**跨领域的知识面

  经济学：商业数据（财务报表分析）、产品优惠券等

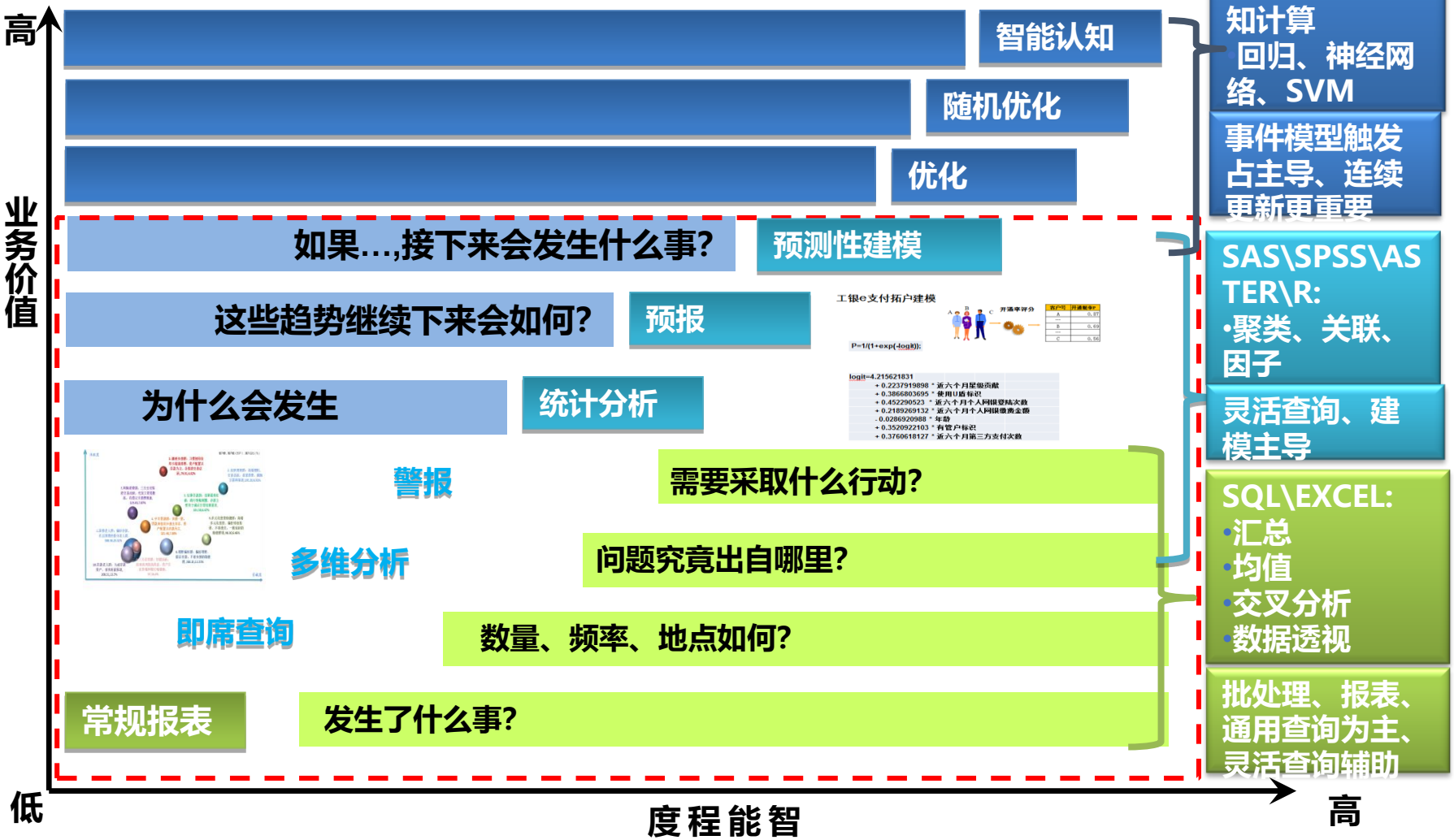
  心理学：心理需求的满足等

  **统计学**：聚类分析等，如客户关系管理中的RFM理论对客户评级：有价值、重要、潜力、无用客户等
- **思：**思考的习惯

  在职场，“选对路”往往比“开快车”有效得多

# 什么是做好数据分析的关键

分析能力的十个等级



# 什么是做好数据分析的关键?

- 数据分析人才
  - 同样的数据，仁者见仁智者见智，人才的不可复制性
  - 做好数据分析的人不一定能做老大，至少可以做军师

数据科学家

匠人

学者

艺术家

“数据贯穿业务，数据驱动业务”

**业务能力**
理解行业动态和发展趋势、业务的需求、行为、流程

数据分析人员

**数据分析能力**
掌握大数据分析方法，从数据中发现有价值的信息并转化为应用

**数据处理能力**
掌握大数据处理工具，能够对数据进行采集、整合和清理

# “数据科学家”能力发展路线图

# 授课要点

## 2. 统计建模工具

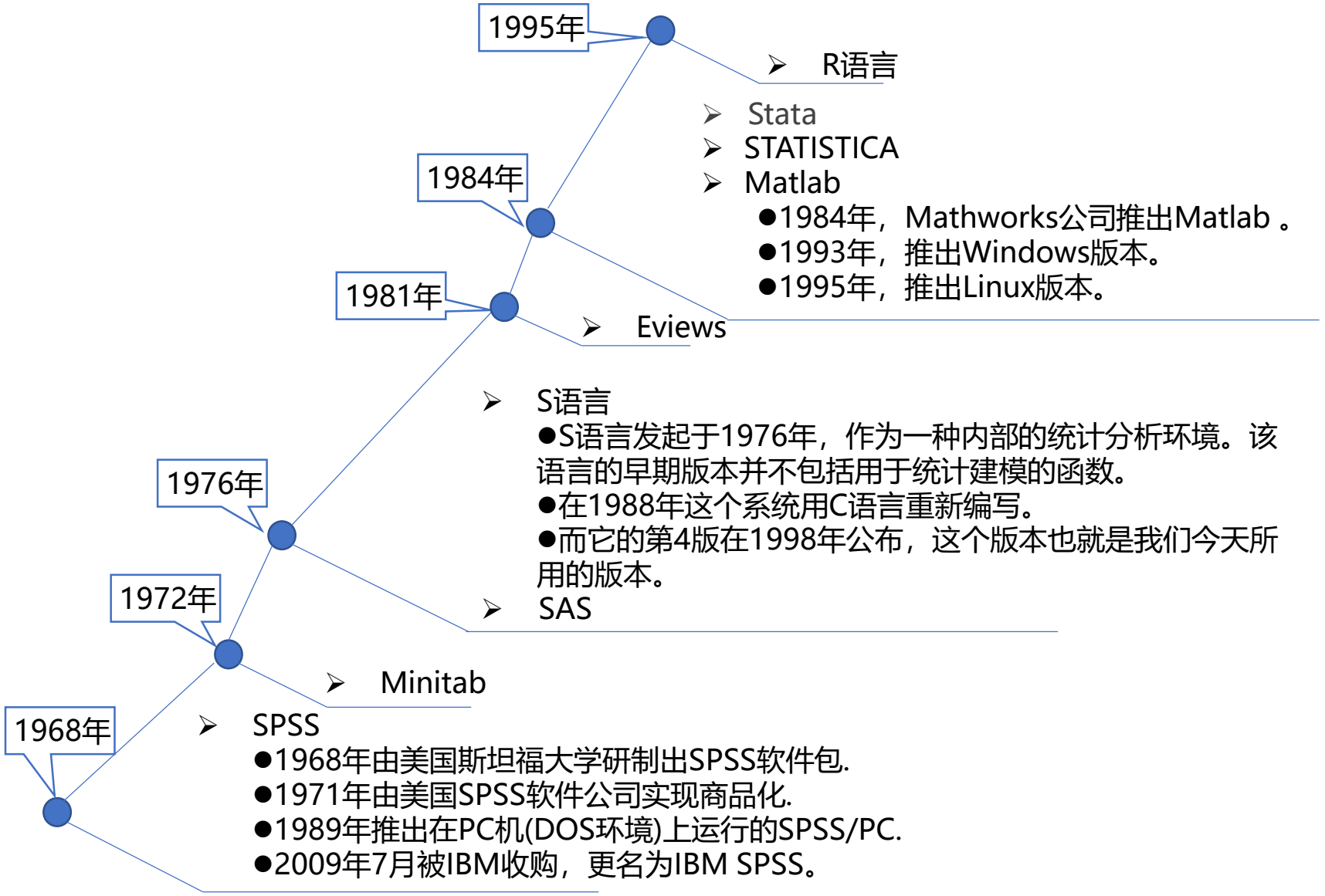
### 2.1. SPSS

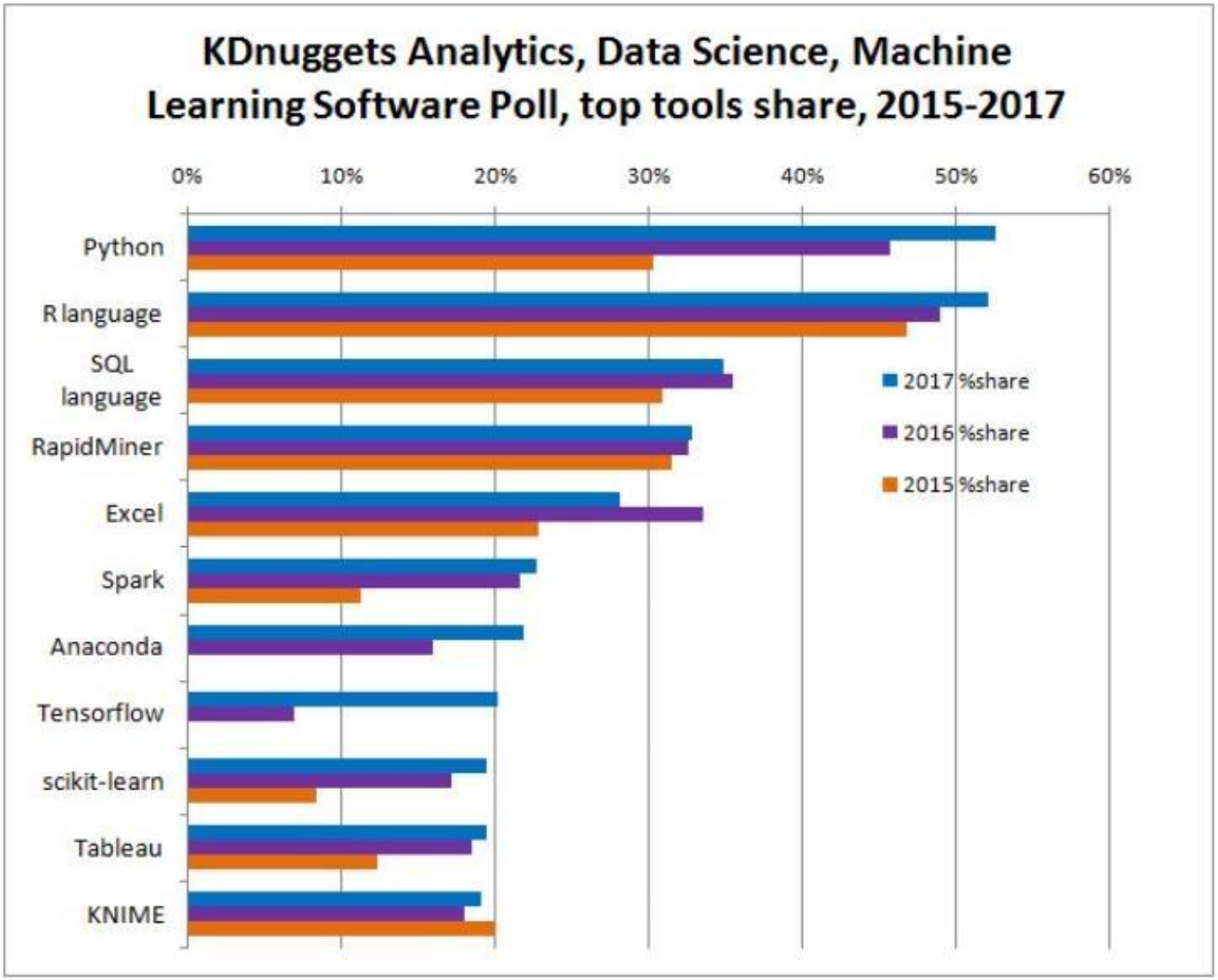
### 2.2. SAS

### 2.3. R语言

### 2.4. Python

***工欲善其事必先利其器！***


1995年
R语言
Stata
STATISTICA
Matlab
●1984年，Mathworks公司推出Matlab 。
●1993年，推出Windows版本。
●1995年，推出Linux版本。
1984年
1981年
Eviews
S语言
●S语言发起于1976年，作为一种内部的统计分析环境。该语言的早期版本并不包括用于统计建模的函数。
●在1988年这个系统用C语言重新编写。
●而它的第4版在1998年公布，这个版本也就是我们今天所用的版本。
1976年
SAS
1972年
Minitab
1968年
SPSS
●1968年由美国斯坦福大学研制出SPSS软件包.
●1971年由美国SPSS软件公司实现商品化.
●1989年推出在PC机(DOS环境)上运行的SPSS/PC.
●2009年7月被IBM收购，更名为IBM SPSS。


KDnuggets Analytics, Data Science, Machine Learning Software Poll, top tools share, 2015-2017
0%
10%
20%
30%
40%
50%
60%
Python
R language
SQL language
RapidMiner
Excel
Spark
Anaconda
Tensorflow
scikit-learn
Tableau
KNIME
2017 %share
2016 %share
2015 %share

# 人大经济论坛——计量经济学与统计论坛

| | |
|---|---|
| 计量经济学与统计软件 | Stata专版 stata上传下载区 EViews专版 Gauss专版<br>LISREL、AMOS等结构方程模型分析软件 IRT理论相关软件<br>winbugs及其他软件专版 HLM专版 LATEX论坛 统计从业与统计师考试<br>经管代码库 统计软件培训班VIP答疑区 |
| 商业数据分析 | 数据分析与数据挖掘 SPSS论坛 SAS专版 SAS上传下载区<br>R语言论坛 Excel MATLAB等数学软件专版 Clementine&Modeler版<br>JMP论坛 数据分析师（CDA）专版 每天一个数据分析师 |
| 大数据技术 | Hadoop论坛 python论坛 数据可视化 SQL及关系型数据库数据分析<br>Oracle数据库及大数据解决方案 mahout论坛 数据仓库技术<br>spark高速集群计算平台 nosql论坛 openstack云平台<br>storm实时数据分析平台 行业应用案例 |
| 机器学习技术 | 人工智能(自然语言处理/机器学习/智能设备与机器人) 人工智能论文版 |
| IT基础 | Scala及其他JVM语言 Linux操作系统 C与C++编程 JAVA语言开发技术 |
| 其他版块 | 原创成果区 |

# 2.1、SPSS

## （一）SPSS的前世

- SPSS是世界上最早的统计分析软件。
- SPSS由美国斯坦福大学的三位研究生Norman Nie、Dale Bent和Hadlai (Tex) Hull于1968年研究开发成功，并同时成立了SPSS公司。
- 1975年成立法人组织，在芝加哥组建了SPSS总部。
- 2009年7月28日，IBM公司宣布用12亿美元收购SPSS公司。2010年1月，产品更名为IBM SPSS Statistics。

# 2.1、SPSS


## （二）SPSS的今生

### 1. SPSS的定义

- SPSS（Statistical Product and Service Solutions），“统计产品与服务解决方案”软件。最初软件全称为“社会科学统计软件包”（Solutions Statistical Package for the Social Sciences），但是随着SPSS产品服务领域的扩大和服务深度的增加，SPSS公司已于2000年正式将英文全称更改为“统计产品与服务解决方案”，这标志着SPSS的战略方向正在做出重大调整。SPSS为IBM公司推出的一系列用于统计学分析运算、数据挖掘、预测分析和决策支持任务的软件产品及相关服务的总称，有Windows和Mac OS X等版本。

——百度百科

- 官网链接：中文：www.ibm.com/analytics/cn/zh/technology/spss/
  英文：www.ibm.com/analytics/us/en/technology/spss/

# 2.1、SPSS

2. SPSS的特点

- 操作简便。界面非常友好，大多数操作可通过鼠标拖曳、点击“菜单”、“按钮”和“对话框”来完成。
- spss编程方便。既可以采用程序运行方式，根据自己的分析需要手工编写程序；又可以采用混合运行方法，在使用菜单的同时编辑SPSS程序。
- spss功能强大。具有完整的数据输入、编辑、统计分析、报表、图形制作等功能。自带11种类型136个函数。
- spss数据接口。能够读取及输出多种格式的文件。
- spss模块组合。软件分为若干功能模块。用户可以根据自己的分析需要灵活选择。
- spss针对性强。SPSS针对初学者、熟练者及精通者都比较适用。

——百度百科

# 2.2、SAS

## （一）SAS的前世

- 全称：STATISTICAL ANALYSIS SYSTEM。
- 最早由北卡罗来纳大学的两位生物统计学研究生编制，并于1976年成立了SAS软件研究所，正式推出了SAS软件。
- 在数据处理和统计分析领域，SAS系统被誉为国际上的标准软件系统，并在1996～1997年度被评选为建立数据库的首选产品。
- 全世界120多个国家和地区的近三万家机构所采用，直接用户则超过三百万人，遍及金融、医药卫生、生产、运输、通讯、政府和教育科研等领域。
- 1997年，推出SAS 6.12版；2000年，推出SAS 8.0版；目前最新版本为9.4版。版本9.0+支持中文操作界面。

# 2.2、SAS

## （二）SAS的今生

1. SAS的定义

- SAS（previously “Statistical Analysis System”），是一款用于先进分析，多元分析，商业智能，数据管理以及预测分析的软件。它是由SAS软件研究所维护、开发。

——维基百科

- 官网链接：中文：https://www.sas.com/zh_cn/home.html

2. SAS的特点

- 功能强大，统计方法齐、全、新。
- 使用简便，操作灵活。
- 提供联机帮助功能。

# 2.2、SAS

## 3. SAS功能模块

- 是一个模块化、集成化的大型应用软件系统。
- 它由数十个专用模块构成，功能包括数据访问、数据储存及管理、应用开发、图形处理、数据分析、报告编制、运筹学方法、计量经济学与预测等等。
- SAS系统基本上可以分为四大部分：SAS数据库部分；SAS分析核心；SAS开发呈现工具；SAS对分布处理模式的支持及其数据仓库设计。
- SAS系统主要完成以数据为中心的四大任务：数据访问；数据管理（SAS的数据管理功能并不很出色，而是数据分析能力强大所以常常用微软的产品管理数据，再导成SAS数据格式．要注意与其他软件的配套使用）；数据呈现；数据分析。

- ●数据库部分：BASE，FSP，ACCESS,..
- ●分析核心：STAT，ETS，QC，OR，INSIGHT，…
- ●开发呈现工具：AF，EIS，GRAPH，…
- ●分布处理与数据仓库：CONNECT，WA，…

# 2.2、SAS

## 4. SAS市场规模

- SAS已在全球100多个国家和地区拥有29000多个客户群，直接用户超过300万人。
- 在我国，国家信息中心，国家统计局，卫生部，中国科学院等都是SAS系统的大用户。SAS已被广泛应用于政府行政管理，科研，教育，生产和金融等不同领域，并且发挥着愈来愈重要的作用。
- 客户对SAS企业级智能平台和行业解决方案需求的不断增长，验证了SAS的智能化战略所取得的卓越成效，并在2005年创下了新的销售记录：总收入从2004年的15.3亿美元增长到2005年的16.8亿美元，增幅为10%，这标志着SAS连续29年保持收入增长和盈利。
- 从全球收入细分从行业分布上看，SAS银行业解决方案的收入继续保持领先，增长率达10%，占SAS行业解决方案收入的28%。零售业解决方案则增长了20%、教育业16%、保险业12%、政府应用11%。

# 2.3、R语言

## （一）R的前世

- R语言由S语言演变而来，S语言于上世纪70年代诞生于AT&T贝尔实验室。
- 基于S语言开发的商业软件S-plus，在国外学术界应用很广。
- R语言由Auckland大学统计系的Robert Gentleman和Ross Ihaka于1995年编写而成（R命名取其两人名字的首字母）。
- R很快得到广泛用户的欢迎，目前它是由R核心发展团队维持，它是一个由志愿者组成的工作努力的国际团队。

Ross Ihaka和Robert Gentleman

# 2.3、R语言

## （二）R的今生

### 1. R语言定义

- R是用于统计分析、绘图、数据挖掘的语言和操作环境。R是属于GNU系统的一个自由的、免费、源代码开放的软件，它是一个用于统计计算和统计绘图的优秀工具。

——维基百科

- 官网：https://www.r-project.org/

### 2. R的特点

- R是自由软件。
- R是一种可编程的语言。
- 所有R的函数和数据集是保存在程序包里面的。
- R具有很强的互动性。
- 如果加入R的帮助邮件列表一，每天都可能会收到几十份关于R的邮件资讯。可以和全球一流的统计计算方面的专家讨论各种问题，可以说是全世界最大、最前沿的统计学家思维的聚集地。

# 2.3、R语言

3. R的优势

- R使用成本低。
- R扩展性强。
- R使用简单。

**R优秀的扩展包弥补了性能问题！！！**

（截至2019年3月18日，已经收录了各类包13903个）

4. R的劣势

- R初始设计完全基于单线程和纯粹的内存计算，处理大数据受到限制。
- R非“傻瓜”软件，需要一定的编程基础，需要足够的统计知识。
- R的技能核定并没有官方或者机构标准，企业想招到R相关人才也不那么简单；
- R的迁移成本高：对于大量工作已由其他软件实现（比如用SAS）的公司来讲，迁移成本很高。

# 2.4、Python

## （一）Python的前世

- Python是一种面向对象的解释型计算机程序设计语言，由荷兰人Guido van Rossum于1989年发明，第一个公开发行版发行于1991年。
- 1989年圣诞节期间，在阿姆斯特丹，Guido为了打发圣诞节的无趣，决心开发一个新的脚本解释程序，做为ABC语言的一种继承。之所以选中Python（大蟒蛇的意思）作为该编程语言的名字，是因为他是一个叫Monty Python的喜剧团体的爱好者。
- Python是从ABC语言发展起来，主要受到了Modula-3（另一种相当优美且强大的语言，为小型团体所设计的）的影响，并且结合了Unix shell和C的习惯。
- Python已经成为最受欢迎的编程语言之一。 自2003年以来，来Python已经3次被知名的全球编程语言流行度排行榜网站TIOBE评为年度编程语言，分别是2007年、2010年和2018年。
- 自从2004年以后，python的使用率呈线性增长。

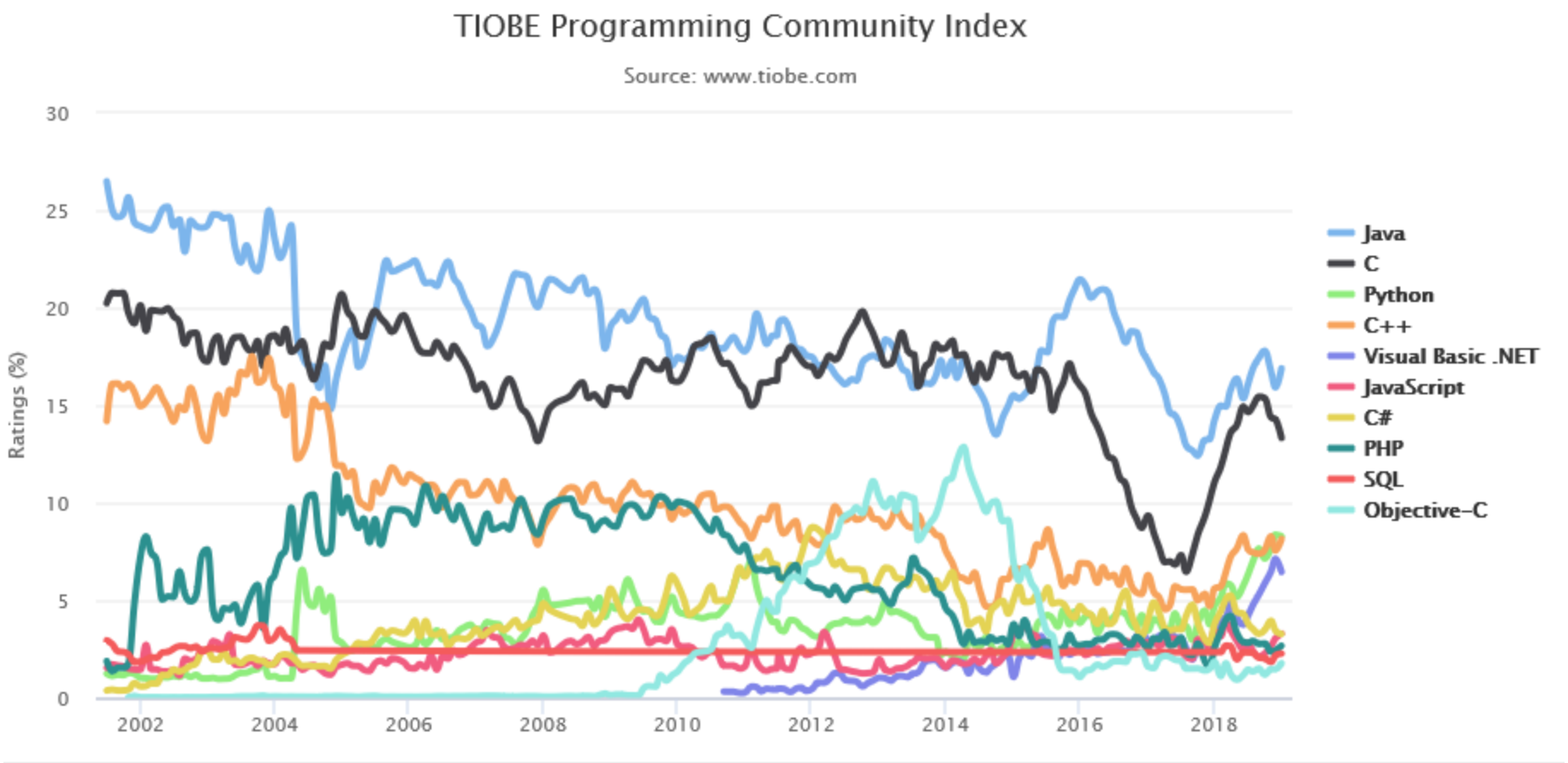
# 2.4、Python

2003年至2018年， TIOBE年度最佳编程语言

| Year | Winner |
| --- | --- |
| 2018 | Python |
| 2017 | C |
| 2016 | Go |
| 2015 | Java |
| 2014 | JavaScript |
| 2013 | Transact-SQL |
| 2012 | Objective-C |
| 2011 | Objective-C |
| 2010 | Python |
| 2009 | Go |
| 2008 | C |
| 2007 | Python |
| 2006 | Ruby |
| 2005 | Java |
| 2004 | PHP |
| 2003 | C++ |

资料来源：https://www.tiobe.com/tiobe-index/

# 2.4、Python




资料来源：https://www.tiobe.com/tiobe-index/

# 2.4、Python


## （二）Python的今生

### 1. Python的定义

- Python is an interpreted, high-level, general-purpose programming language.

——Wikipedia

- 官网：https://www.python.org/

### 2. Python的特点

- Python是一种面向对象、解释型计算机程序设计语言。
- Python是纯粹的自由软件，它的语法简洁、易读以及有很强可扩展性。一些知名大学已经采用Python教授程序设计课程。例如卡耐基梅隆大学的编程基础、麻省理工学院的计算机科学及编程导论。
- Python具有丰富和强大的库，能够把用其他语言制作的各种模块（尤其是C/C++）很轻松地联结在一起。有专门用于科学计算的Numpy、Pandas、Matplotlib等模块，有很多集成软件如Anaconda，以及数据挖掘的SK-learn。

# 2.4、Python

## 3. Python的优势

- 简单：Python是一种代表简单主义思想的语言。
- 易学：Python极其容易上手，因为Python有极其简单的说明文档 。
- 速度快：Python的底层是用C语言写的，很多标准库和第三方库也都是用C写的，运行速度非常快。
- 免费、开源：Python是FLOSS（自由/开放源码软件）之一。
- 高层语言：用Python语言编写程序的时候无需考虑诸如如何管理你的程序使用的内存一类的底层细节。
- 可移植性：由于它的开源本质，Python已经被移植在许多平台上（经过改动使它能够工作在不同平台上）。
- 解释性：一个用编译性语言比如C或C++写的程序可以从源文件（即C或C++语言）转换到一个你的计算机使用的语言（二进制代码，即0和1）。这个过程通过编译器和不同的标记、选项完成。
- 面向对象：Python既支持面向过程的编程也支持面向对象的编程。
- 可扩展性：如果需要一段关键代码运行得更快或者希望某些算法不公开，可以部分程序用C或C++编写，然后在Python程序中使用它们。
- 可嵌入性：可以把Python嵌入C/C++程序，从而向程序用户提供脚本功能。
- 丰富的库：Python标准库确实很庞大。除了标准库以外，还有许多其他高质量的库，如wxPython、Twisted和Python图像库等等。

# 2.4、Python

4. Python的劣势

- 单行语句和命令行输出问题：很多时候不能将程序连写成一行。
- 独特的语法：这也许不应该被称为局限，但是它用缩进来区分语句关系的方式还是给很多初学者带来了困惑。即便是很有经验的Python程序员，也可能陷入陷阱当中。
- 运行速度慢：这里是指与C和C++相比。

# 统计软件的未来

统计软件向商业化发展

企业大数据解决方案中
增加对统计软件支持

统计软件大数据
处理能力提升

# 授课要点

## 3.探索性数据分析框架

### 3.1.数值描述分析

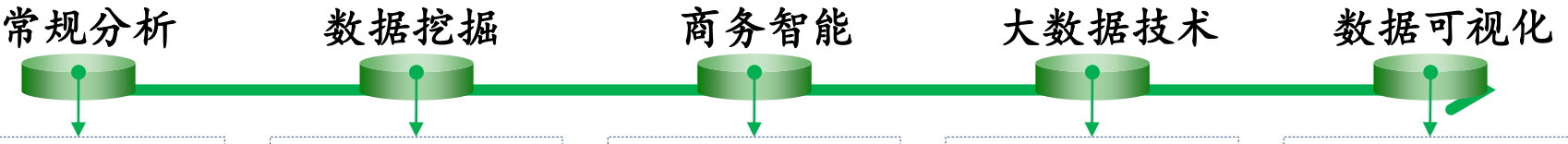
### 3.2.数据可视化

### 3.3.数据清洗和转换

### 3.4.实例分析

# 数据分析

随着计算机科学的进步，数据挖掘、商务智能、大数据等概念的出现，数据分析的手段和方法更加丰富




- 揭示数据之间的静态关系
- 分析过程滞后
- 对数据质量要求高

结构分析
分组分析
预警分析
杜邦分析
……

- 统计学和计算机技术等多学科的结合
- 揭示数据之间隐藏的关系
- 将数据分析的范围从“已知”扩展到“未知”，从“过去”推向“将来”

- 一系列以事实为支持，辅助商业决策的技术和方法，曾用名包括专家系统、智能决策等
- 一般由数据仓库、联机分析处理、数据挖掘、数据备份和恢复等部分组成
- 对数据分析的体系化管理，数据分析的主体依然是数据挖掘

- 从多种类型的数据中，快速获取知识的能力
- 数据挖掘技术的衍生

- 大数据时代，展示数据可以更好辅助理解数据、演绎数据

# 数据分析

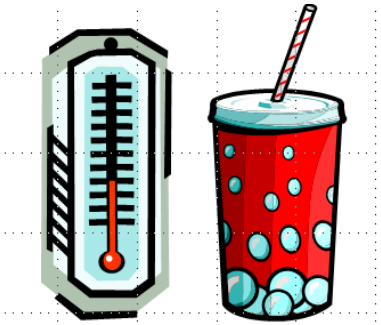
通过数据分析，初步发现数据特征、规律，为后续数据建模提供输入依据，常见的数据探索方法有数据特征描述、相关性分析、主成分分析等。

**数据分析要遵循由浅入深、由易到难的步骤**

# 变量分类

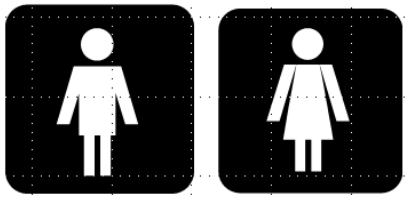
- **连续变量：**在一定区间内可以任意取值的变量叫连续变量，其数值 是连续不断的，相邻两个数值可作无限分割，即可取无限个数值。

例如：温度



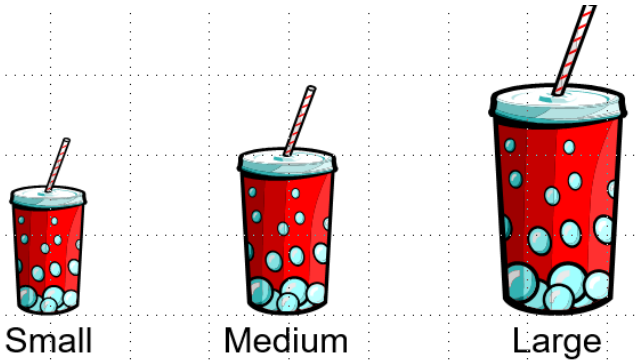
- **分类变量：**是说明事物类别的一个名称，其取值是分类数据。

例如：性别



# 变量分类

- **有序变量**：有序变量是指分类数大于等于3，且类别之间存在序次关系的响应变量。在对此类资料进行统计分析的过程中，我们发现，有序变量的“类间距”并不相等，也就是各类型之间的稀疏程度并不是均匀的。




- **定距变量**：定距变量（又叫连续性变量或者定量变量）与有序变量又有点儿像，但是定距变量可以确切地测量同一类别各个水平高低、大小次序之间的距离，因而可以做加减法。

# 数据分析

- 描述性统计：

运用制表和分类，图形以及计算概括性数据来描述数据特征的各项活动。主要包括：

- 数值展示：频数分析、集中趋势、离中趋势等。
- 图表展示：

# 3.1、数值描述分析

- 频数统计

•频数表是数理统计中由于所观测的数据较多，为简化计算，将这些数据按等间隔分组，然后按选举唱票法数出落在每个组内观测值的个数，称为(组)频数。这样得到的表称“频数表”或“频数分布表”。

•分析频数分布的目的是要根据子样中各个变值的频率分布情况来推测母体中各个变值的频率分布情况。

# 3.1、数值描述分析

- 频数表编制

① 求全距（range）：找出观察值中的最大值与最小值，其差值即为全距（或极差），用R表示。

② 确定组段和组距：根据样本含量的大小确定“组段”数，一般设8-15个组段。

- 第一组段应包括全部观察值中的最小值，最末组段应包括全部观察值中的最大值，并且同时写出其下限与上限。
- 各组段的起点和终点分别称为下限和上限，某组段包含下限，但不包含上限，其组中值为该组段的（下限+上限）/2。
- 相邻两组段的下限之差称为组距。

# 3.1、数值描述分析

- 频数表编制

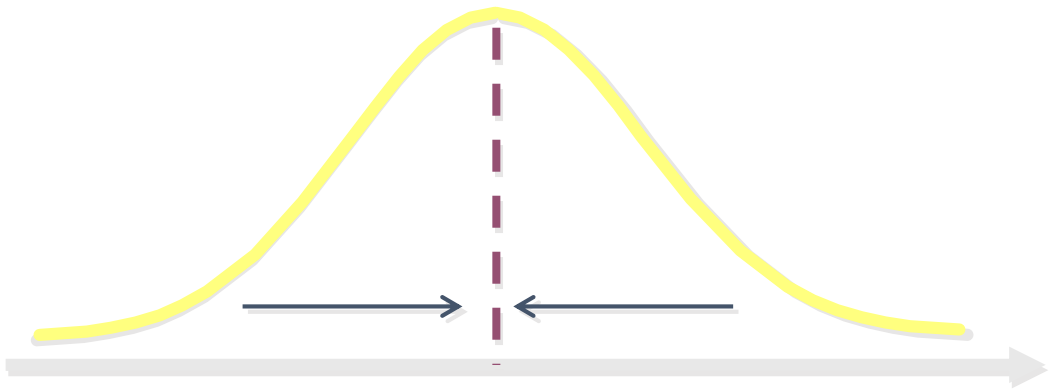
③ 列表划记：确定组段界限，得出各组段的观察例数，即频数，表中的第（1）、（3）栏即所需的频数表。

表　某地110名18岁男大学生身高（cm）均数的频数表

| 身高组段 | 划记 | 频数，$f$ | 组中值，$x$ |
|---|---|---|---|
| （1） | （2） | （3） | （4） |
| 108~ | 一 | 1 | 109 |
| 110~ | 下 | 3 | 111 |
| 112~ | 正丅 | 9 | 113 |
| 114~ | 正丅 | 9 | 115 |
| 116~ | 正正正 | 15 | 117 |
| 118~ | 正正正下 | 18 | 119 |
| 120~ | 正正正正一 | 21 | 121 |
| 122~ | 正正丅 | 14 | 123 |
| 124~ | 正正 | 10 | 125 |
| 126~ | 丅 | 4 | 127 |
| 128~ | 下 | 3 | 129 |
| 130~ | 丅 | 2 | 131 |
| 132~134 | 一 | 1 | 133 |
| 合计 | | 110 | |

# 3.1、数值描述分析

- 集中趋势
  - 一组数据向其中心值靠拢的倾向和程度
  - 测度集中趋势就是寻找数据水平的代表值或中心值
  - 不同类型的数据用不同的集中趋势测度值
- 集中趋势的度量
  - 众数
  - 中位数和分位数
  - 平均数(均值)



# 描述统计——常见统计量

- **众数(不唯一)**
  - 一组数据中出现次数最多的变量值
  - 不受极端值的影响
  - 一组数据可能没有众数或有几个众数

- **中位数**
  - 排序后处于中间位置上的值


  - 不受极端值的影响
  - 变量值与中位数的离差绝对值之和最小

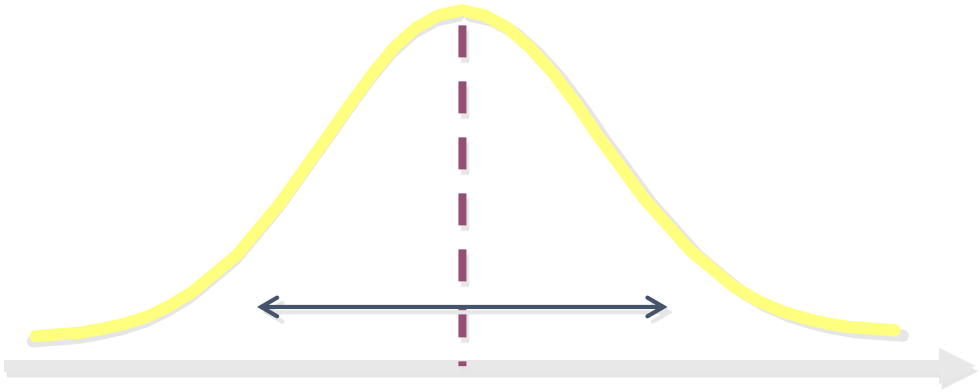
- **分位数**
  - 排序后处于25%和75%位置上的值


  - 不受极端值的影响

- **平均数**
  - 一组数据的均衡点所在
  - 易受极端值的影响
  - 集中趋势的最常用测度值
  - 简单平均数、加权平均数、几何平均数

# 3.1、数值描述分析

- 离散趋势
  - 反映各变量值远离其中心值的程度(离散程度)
  - 从另一个侧面说明了集中趋势测度值的代表程度
  - 不同类型的数据用不同的离散趋势测度值
- 离散趋势的度量
  - 异众比率
  - 极差
  - 四分位差
  - 方差和标准查
  - 离散系数



# 描述统计——常见统计量

## 极差

- 一组数据的最大值与最小值之差
- 离散程度的最简单测度值
- 易受极端值影响
- 未考虑数据的分布

## 方差和标准差

- 数据离散程度的最常用测度值
- 反映了各变量值与均值的平均差异
- 样本方差$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- 样本标准差$S = \sqrt{S^2}$

# 描述统计——常见统计量

## 分布形态

- **偏度**：数据分布偏斜程度的测度

$$SK = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)(n-2)S^3}$$

SK=0为对称分布

SK> 0为右偏分布

SK< 0为左偏分布

- **峰度**：数据分布扁平程度的测度

$$K = \frac{n(n+1)\sum_{i=1}^{n}(x_i - \bar{x})^4 - 3(n-1)[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2}{(n-1)(n-2)(n-3)S^4}$$

K=0扁平峰度适中

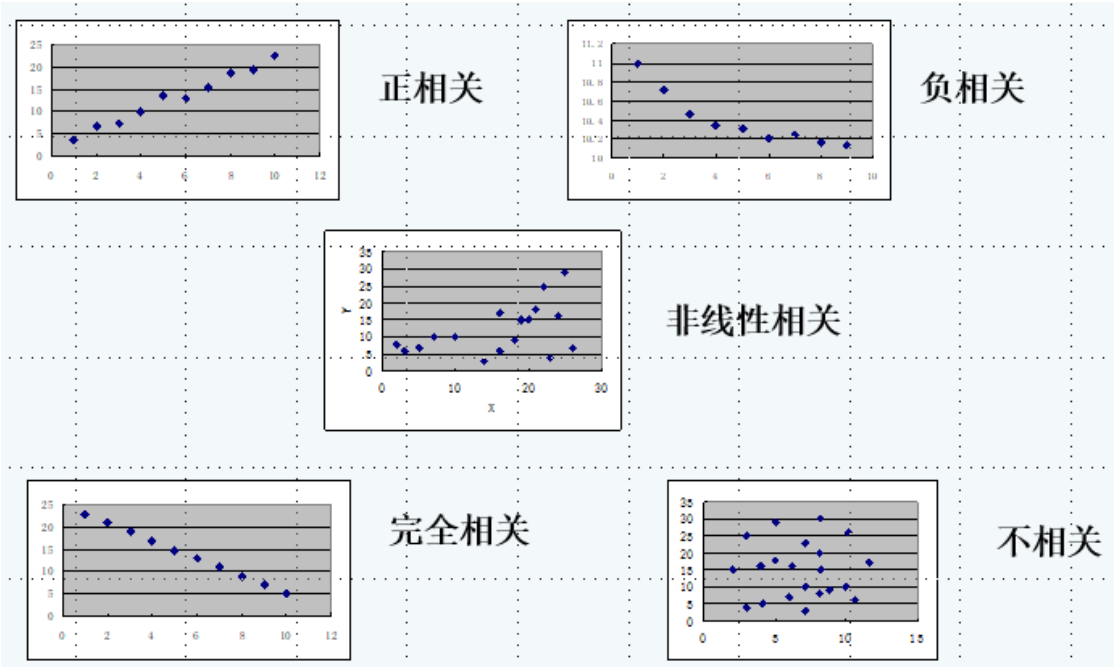
K<0为扁平分布

K>0为尖峰分布

# 3.1、数值描述分析

- 相关系数

  - 它反映现象之间确实存在的，但关系数值不固定的相互依存关系
  - 相关关系是指现象之间确实存在数量上的相互依存关系
  - 现象之间数量依存关系的具体关系值不是固定的



# 3.1、数值描述分析

- 相关关系与因果关系

例：

•2004年年底，美股走出了短线的“楔形向上”，预示了股市的下跌回调。五个交易日之后，当那时美股的“楔形向上”彻底完成之后，美股出现了中短线，急速的下跌回调。

•2004年12月26日，圣诞节后的第一天，印尼海啸发生。那时，那些恐惧，悲惨的照片，人们称是“人间炼狱”。

•研究表明，美股的涨跌和潮汐存在很大的相关性，但是两者之间不存在因果关系。

# 3.1、数值描述分析

- 协方差

  - 用于衡量两个变量的总体误差。而方差是协方差的一种特殊情况，即当两个变量是相同的情况。

  - 协方差表示的是两个变量的总体的误差，这与只表示一个变量误差的方差不同。

  - 如果两个变量的变化趋势一致，那么两个变量之间的协方差就是正值。如果两个变量的变化趋势相反，那么两个变量之间的协方差就是负值

$$Cov = \sum \frac{(x - \bar{x})(y - \bar{y})}{n}$$

# 3.1、数值描述分析

- 相关系数

  •说明变量之间在直线相关条件下相关关系密切程度和方向的统计指标

  •就参数统计而言，常用的是皮尔逊积矩相关系数(Pearson)，是没有量纲的、标准化的协方差

$$r = \frac{\sum \frac{(x-\bar{x})(y-\bar{y})}{n}}{\sqrt{\frac{\sum(x-\bar{x})^2}{n}}\sqrt{\frac{\sum(y-\bar{y})^2}{n}}}$$

# 3.1、数值描述分析

- 相关系数

  - 相关系数只度量变量间的线性关系，弱相关不一定表明变量间没有关系；
  - 极端值可能影响相关系数

例：两组数据x, y的相关系数

x: 0.1, 0.2, 0.3, 0.4, 0.5
y: -1, -3, 1, 3, 3

相关系数 $r = 0.85$

V. S.

x: 0.1, 0.2, 0.3, 0.4, 0.5
y: -1, -3, 1, 3, **-100**

相关系数 $r = -0.68$

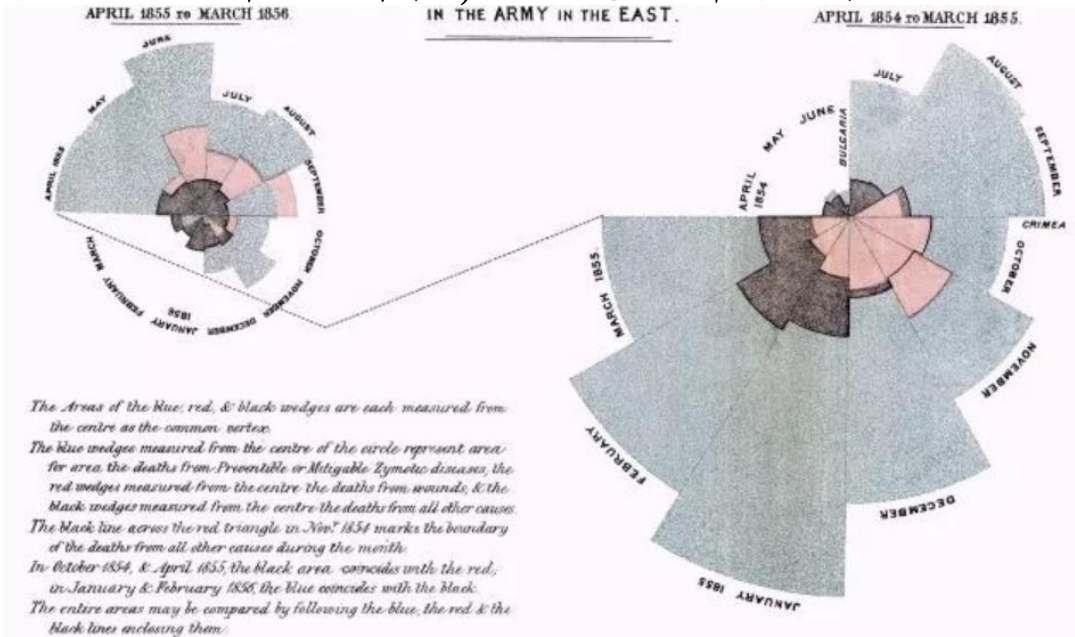
# 3.1、数值描述分析

- 列联表
  - 由两个以上的变量交叉分类的频数分布表
  - 行变量的类别用 $r$ 表示，$r_i$表示第 $i$ 个类别；列变量的类别用 $c$ 表示，$c_j$表示第 $j$ 个类别；每种组合的观察频数用 $f_{ij}$表示
  - 表中列出了行变量和列变量的所有可能的组合，所以称为列联表

| | 列($c_j$) | | | 合计 |
|---|---|---|---|---|
| | $j$ =1 | $j$ = 2 | … | |
| $i$ =1 | $f_{11}$ | $f_{12}$ | … | $r_1$ |
| $i$ = 2 | $f_{21}$ | $f_{22}$ | … | $r_2$ |
| : | : | : | : | : |
| 合计 | $c_1$ | $c_2$ | … | $n$ |

# 3.2、数据的可视化分析

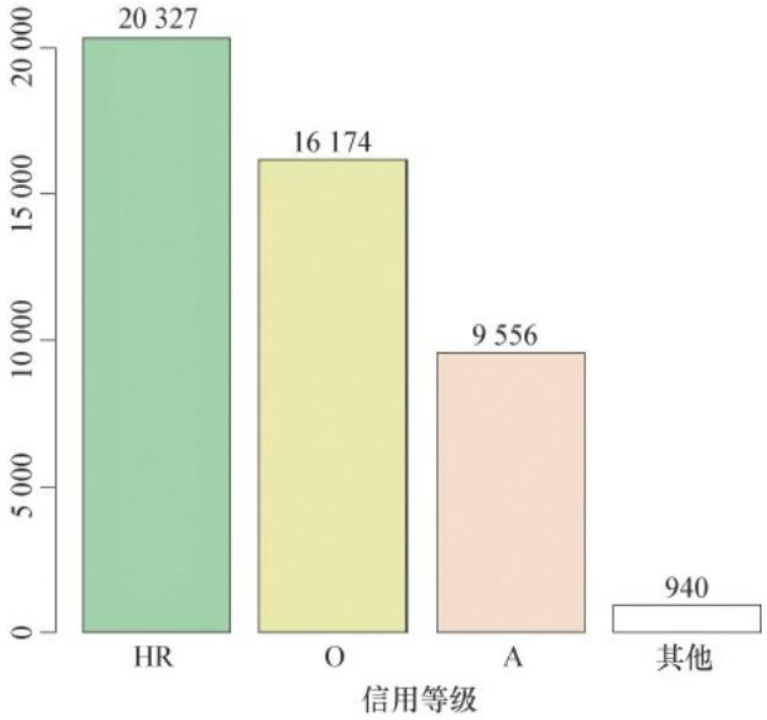
有这样一幅图，它出自克里米亚战争，一名叫南丁格尔的护士利用一幅扇状的玫瑰饼图展示了她所管理的野战医院不同季节中死于不用病因的病人数变化，直观地让英国政府看到：每年死于感染的士兵数（蓝色区域）比死于战场（红色区域）和其他原因（黑色区域）的要多得多，这才使得政府开始改善战地士兵的卫生条件，因此这幅图被称为拯救生命的图表，也是较早使用统计图形传达信息的例子。



# 3.2、数据的可视化分析

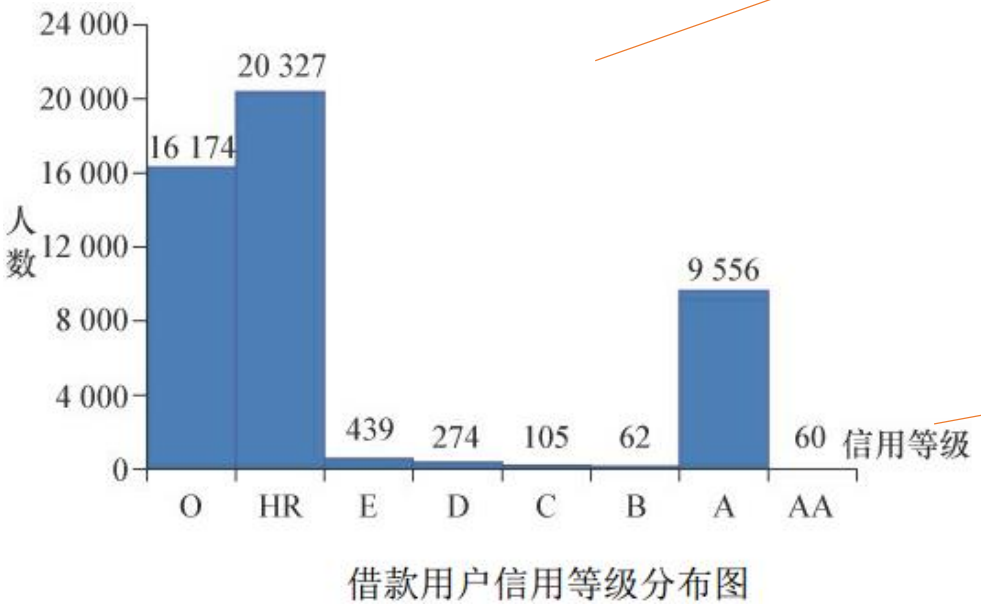
- 柱状图

•柱状图是针对**分类数据**所作的统计图。每根柱子代表一个类别，柱子的高度是这个类别的频数，有时也是百分比。



# 3.2、数据的可视化分析

- 柱状图之错误使用




借款用户信用等级分布图

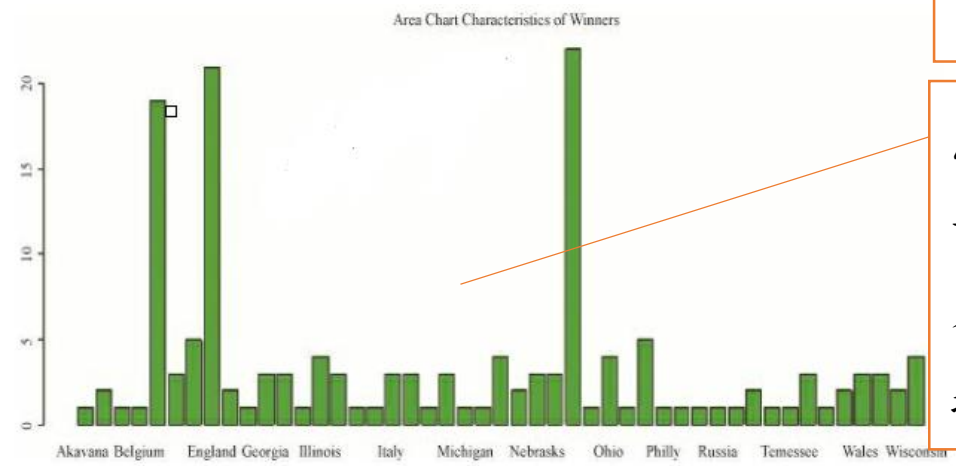
点评1：这不是在画统计图，而是在画诗，这幅图画的是《题西林壁》 中的“远近高低各不同”。 最高的柱子高2万多，最矮的柱子才60。

点评2：美观问题。人都说距离产生美，柱子之间需要留出空隙，让人喘口气。横坐标“信用等级”也体现了自己无处安放的青春，非要跟频数60挤在一起才有安全感吗？

点评3：是图的标题。图的大名叫“柱状图”，你起个绰号叫“分布图”？

# 3.2、数据的可视化分析

- 柱状图之错误使用




获奖者地区分布频数图

点评2：图的标题出现了两次。图的上方标注了一次标题(更多时候是统计软件默认的标题，而作者没有修改或者去掉)，然后图的下方又写了一遍。

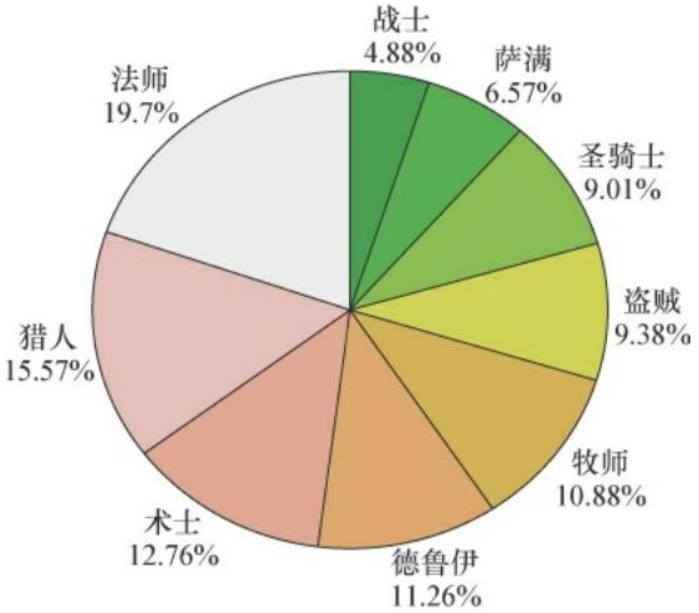
点评1：洋洋洒洒几十根柱子，精心排列得奇丑无比。而且由于柱子数太多，很多标签无法显示，根本无法知道每根柱子对应哪个地区，相当于柱状图没有传递任何信息！

点评3：图的标题和纵轴标题大名叫“柱状图”，就不要再给起个“频数图”或者“分布图”这种名字了。另外，这个图缺少纵轴标题，可以标注“频数”或者“人数”。

# 3.2、数据的可视化分析

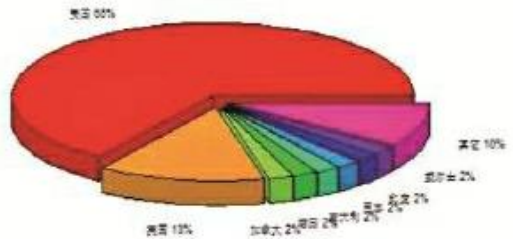
- 饼图

  •针对**分类数据**的统计图。 柱状图多用于展示频数， 饼图多用于展示频率(也就是比例) 。



# 3.2、数据的可视化分析

- 饼图之错误使用



点评1：饼的块数过多(如果只有两类也不适合画饼图)。

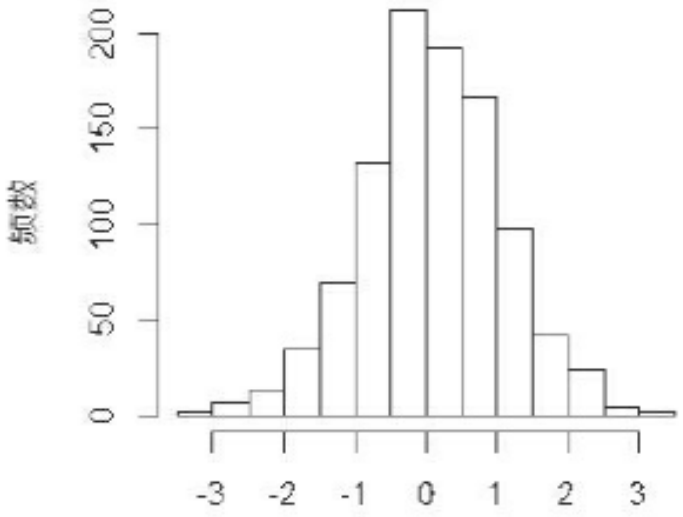
点评2：饼的标签单独打在旁边的时候，对应起来很费劲，比如右下角的饼图： 这个饼分了9块，右侧的标签只有8个。另外一个34.53%的饼对应的标签呢？

点评3：饼的标签，一般只标注百分比，很少标注频数或者两者都标注。 左下角的饼图就同时标注了频数和百分比，异常混乱。

# 3.2、数据的可视化分析

- 直方图

  - 针对**连续型变量**所作的统计图。
  - 直方图的横轴是实数轴， 纵轴代表频数或密度。
  - 一般纵轴使用密度较为合适，特别是在不等距情况下。
  - 直方图最大的用处是观察数据分布的形态。



# 3.2、数据的可视化分析

- 茎叶图
  - 是一种针对**连续型变量**的统计图
  - 茎叶图可以同时展示原始数据和分布的形状。
  - 图形由“茎”和“叶”两部分组成。 通常以数据的高位数字作为树茎， 低位数字作为树叶。

# 3.2、数据的可视化分析

- 茎叶图

- 例1：球员进球时间

| | A | B |
|---|---|---|
| 1 | 进球时间 | 时间段 |
| 2 | 57 | 下半场 |
| 3 | 65 | 下半场 |
| 4 | 89 | 下半场 |
| 5 | 5 | 上半场 |
| 6 | 10 | 上半场 |
| 7 | 61 | 下半场 |
| 8 | 81 | 下半场 |
| 9 | 73 | 下半场 |
| 10 | 92 | 伤停补时 |
| 11 | 41 | 上半场 |

```
0 | 5
1 | 089
2 |
3 | 122477
4 | 125889
5 | 016779
6 | 125
7 | 135
8 | 017789
9 | 00122266
```

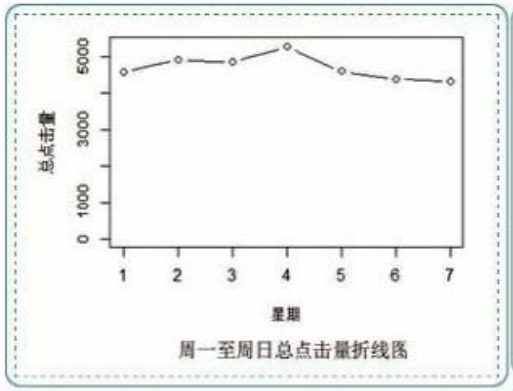
- 例2：球员得分对比（两组数据对比）

| 甲 | | 乙 |
|---|---|---|
| 97 | 0 | 78 |
| 6331 | 1 | 0579 |
| 83 | 2 | 13 |

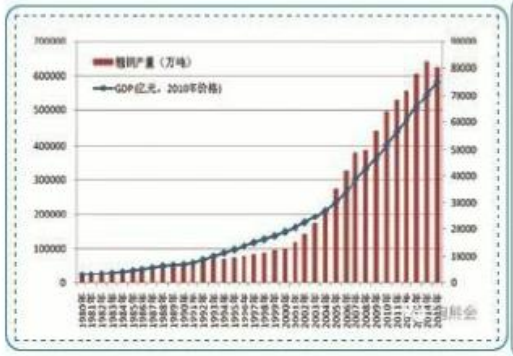
甲乙两名球员八场比赛得分对比

# 3.2、数据的可视化分析

- 折线图之错误使用







点评1：左上图：一根线飘在空中，让人不明所以。不妨对纵轴展示范围进行调整。

点评2：右上图：三根折线两个纵轴，让人难以比较。

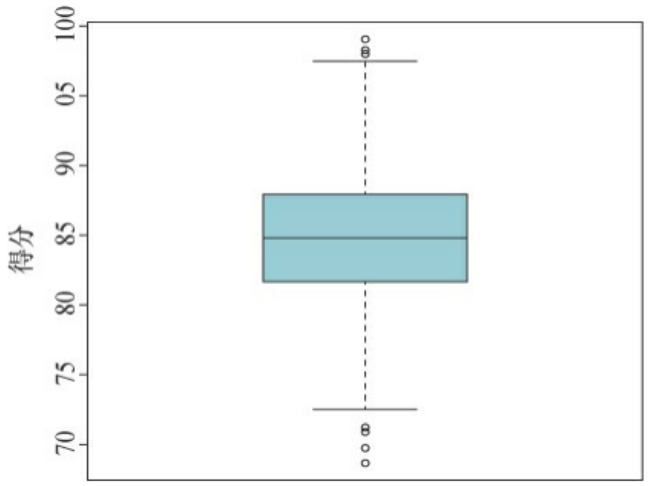
点评3：左下图：少了纵轴标题，横轴标签过于密集。

点评4：右下图：只能用一个词来表达：一团乱麻。 如果有太多的信息想要表达，而且非要在一个图中，就是这个效果。

# 3.2、数据的可视化分析

- 箱线图

  - 是一种针对**分类变量和连续型变量并存**的统计图。
  - 箱子的中间一条线是数据的中位数，代表了样本数据的平均水平。
  - 箱子的上下限分别是数上四分位数和下四分位数，意味着箱子包含50%的数据。箱子的高度在一定程度上反映数据的波动程度。
  - 在箱子的上方和下方，又各有一条线。超出判为“异常值”。

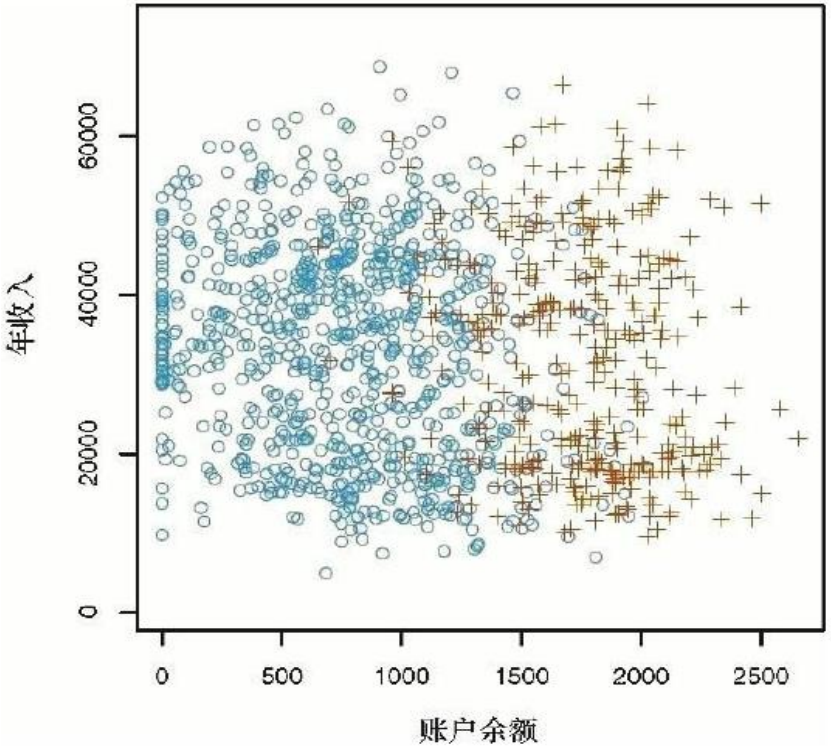


# 3.2、数据的可视化分析

- 散点图

  •散点图是用于展示**两个连续型变量**的一种常用统计图。

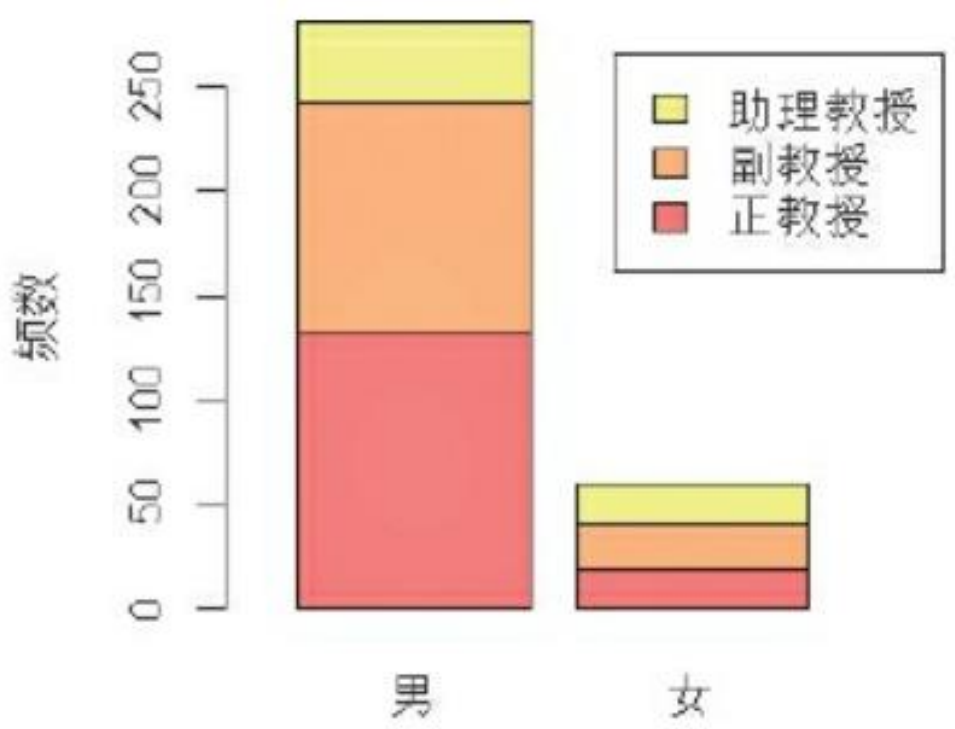
  •除了已知的两个变量，当数据中还有其他变量信息时，可以通过改变“点”的颜色、形状和大小来传递更多的信息。



# 3.2、数据的可视化分析

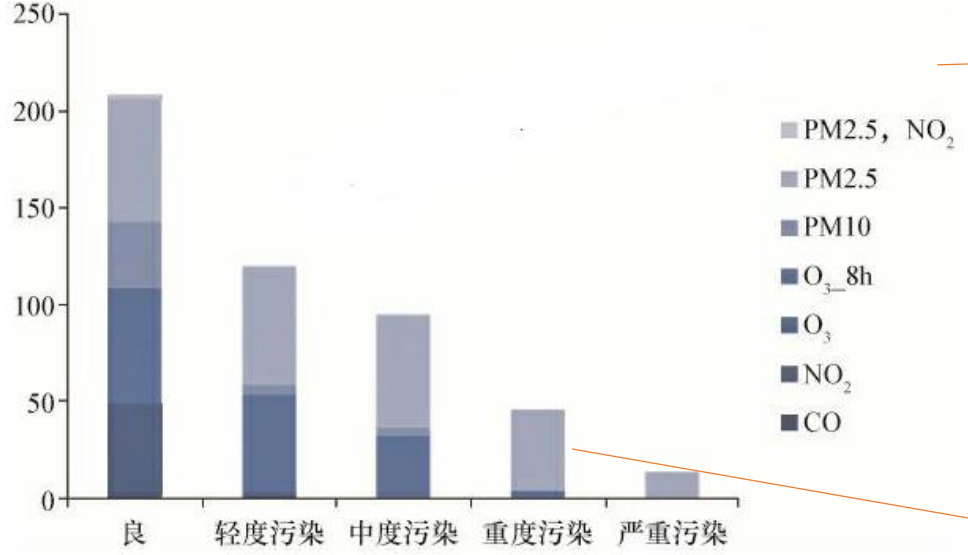
- 堆积柱状图

  - 堆积柱状图和柱状图的本质一样，都是在展示频数。
  - 涉及**两个离散型分类**变量。



# 3.2、数据的可视化分析

- 柱状图之错误使用




北京市不同空气质量指数类别下首要污染物分布图

点评2：这些柱子上面最多出现了4种颜色，然而标签却显示出7种物质。看原始数据才发现，CO或者O3频数太低，根本显示不出来。

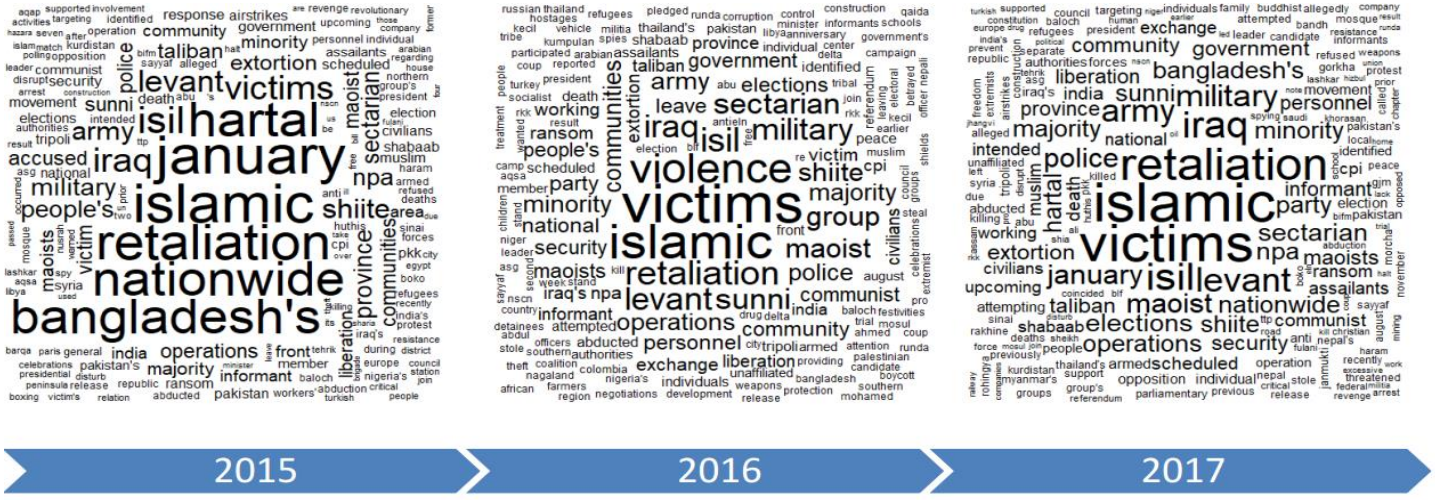
点评1：这是在对读者进行色弱测试吗？很难看出，哪段是PM2.5，哪段是PM10。注意，但凡类别较多，需要画堆积柱状图的时候，应选择区分度比较强的配色，让人能识别出每段柱子都是哪个类别。

# 3.2、数据的可视化分析

- 词云图

•词云图，也叫文字云，是对文本中出现频率较高的“关键词”予以视觉化的展现。

•词云图过滤掉大量的低频低质的文本信息，使得浏览者只要一眼扫过文本就可领略文本的主旨。




2015-2017恐怖袭击动机

# 3.2、数据的可视化分析

- 词云图

  - 例：2018年国内高校相关课程及公司技能需求核心词云图对比

# 3.2、数据的可视化分析

- 地理热力图

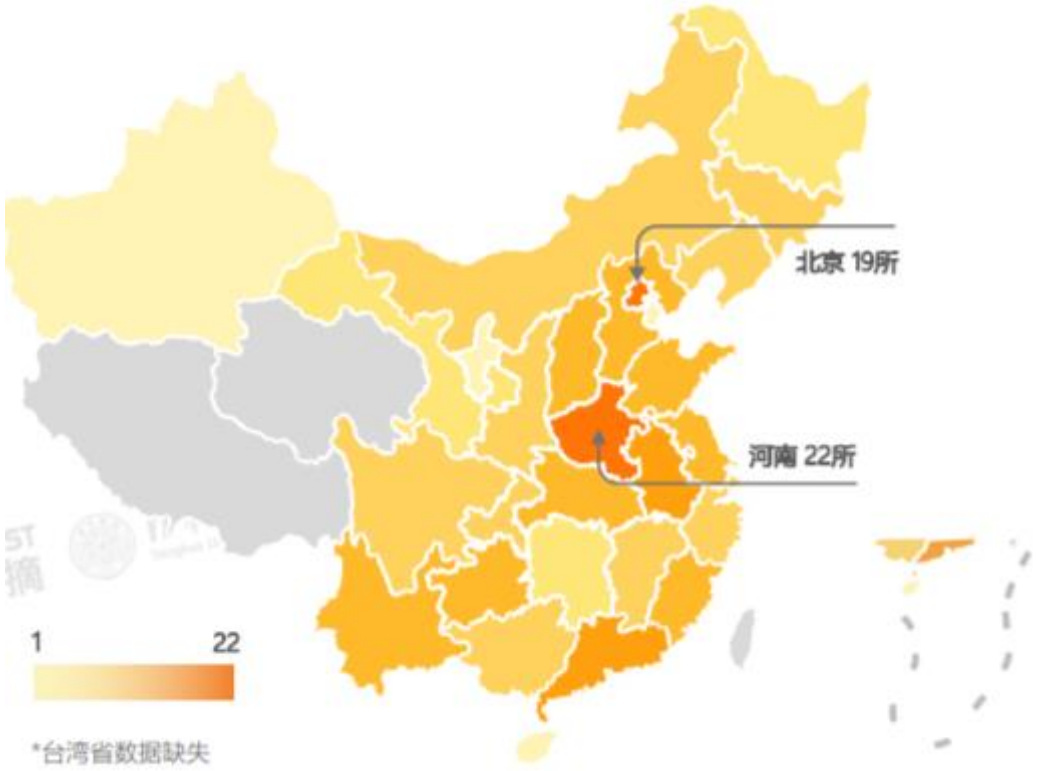
  - 若数据与地理位置意义对应，可视化这类数据的最佳方式是依据地图绘制热力图。
  - 热力图通过颜色的深浅反应对应地区某变量取值的大小。
  - 热力图的关键是依据某个变量的取值设置地图中各地区的填充颜色。


中国恐怖袭击事件地域分布图

# 3.2、数据的可视化分析

- 地理热力图

  - 2018年全国283所高校数据科学与大数据技术新专业分布地图



## 大数据专业迅猛发展

### 全国474所高校共设立了488个“数据科学与大数据技术”专业分布地图

# 3.3、数据的清洗与转换

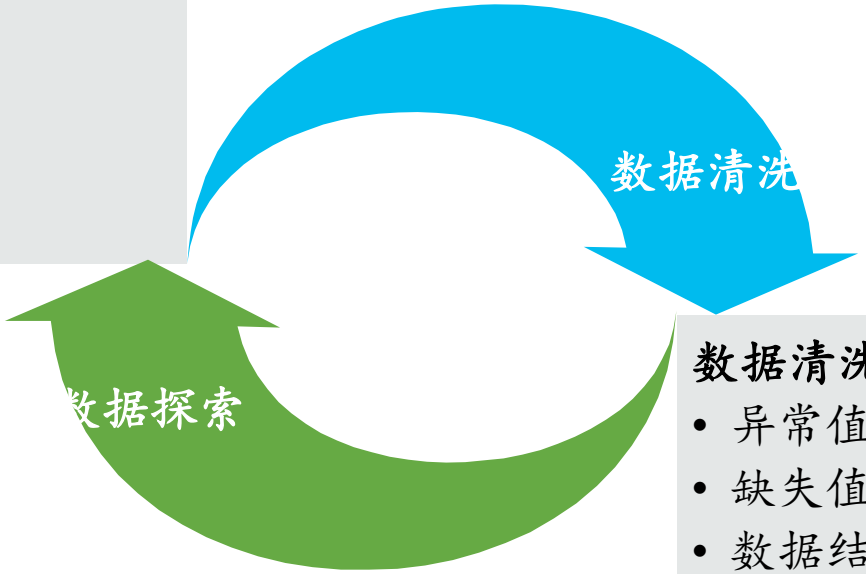
数据清洗&数据探索

数据收集的方法多种多样，本文不再详述。在对收集的数据进行分析前，要明确数据类型、规模，对数据有初步理解，同时要对数据中的“噪声”进行处理，以支持后续数据建模。

# 3.3、数据的清洗与转换

## 数据清洗&数据探索

**数据探索**
- 特征描述
- 分布推断
- 结构优化




**数据清洗**
- 异常值判别
- 缺失值处理
- 数据结构统一

- 数据清洗和数据探索通常交互进行
- 数据探索有助于选择数据清洗方法
- 数据清洗后可以更有效的进行数据探索

# 3.3、数据的清洗与转换

- 数据清洗：异常值识别

数据清洗的第一步是识别会影响分析结果的“异常”数据，然后判断是否剔除。目前常用的识别异常数据的方法有物理判别法和统计判别法

**物理判别法**

- 根据人们对客观事物、业务等已有的认识，判别由于外界干扰、人为误差等原因造成实测数据偏离正常结果，判断异常值。
- 比较困难

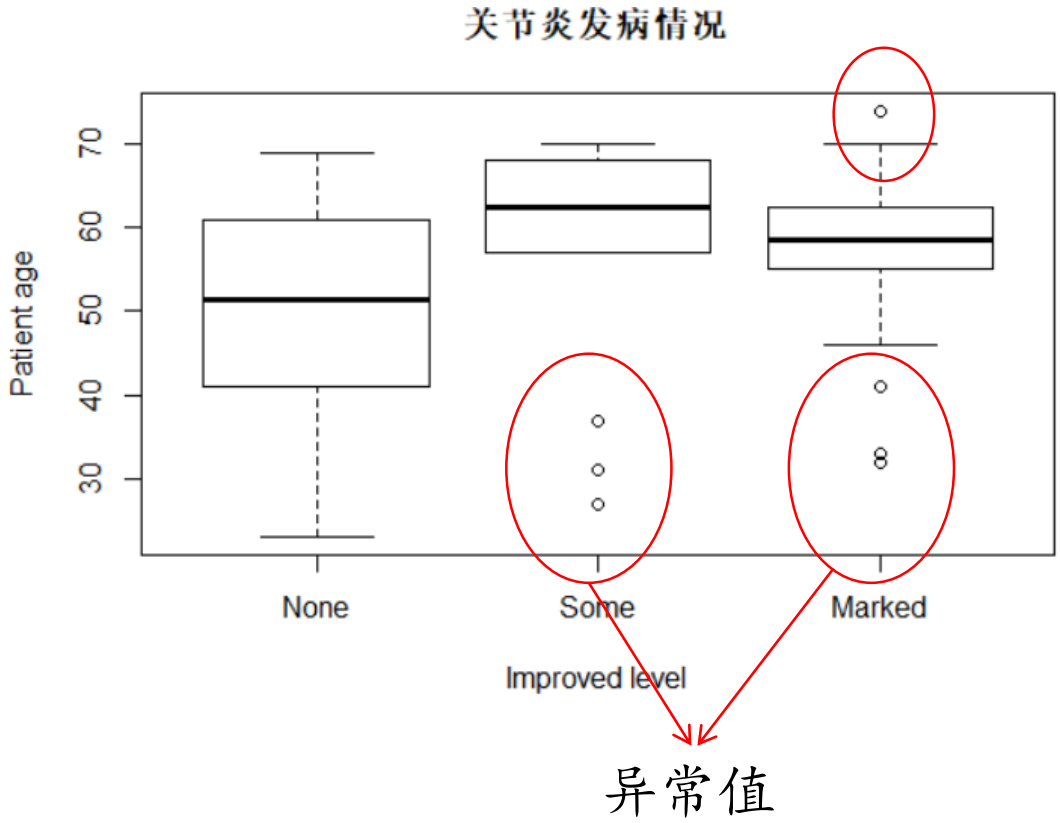
**统计判别法**

- 给定一个置信概率，并确定一个置信限，凡超过此限的误差，就认为它不属于随机误差范围，将其视为异常值。
- 常用的方法（数据来源于同一分布，且是正态的）：拉依达准则、肖维勒准则、格拉布斯准则、狄克逊准则、t检验。

# 3.3、数据的清洗与转换

- 数据清洗：异常值识别

  例：利用箱线图识别



# 3.3、数据的清洗与转换

- 数据清洗：异常值识别

  例：利用常识识别

Table 1: The samples of the data collected by the three sensors

| Time \ Sensor | BMA 06-3D | | | BMA EI-2 | | | BMA EI-3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | X | Y | Z | X | Y | Z | X | Y | Z |
| Thu Jul 10 08:00:00 2014 | 206 | 4048 | -16 | -904 | -3870 | -53 | -131 | 4049 | -135 |
| Thu Jul 10 08:00:01 2014 | 224 | 4031 | **5** | -897 | -3862 | -28 | -131 | 4039 | -137 |
| Thu Jul 10 08:00:02 2014 | 215 | 4048 | -11 | -896 | -3876 | -37 | -122 | 4053 | -146 |
| Thu Jul 10 08:00:03 2014 | 222 | 4046 | -25 | -907 | -3865 | -26 | -124 | 4050 | -137 |
| Thu Jul 10 08:00:04 2014 | 212 | 4060 | -12 | -912 | -3869 | -29 | -129 | 4051 | -130 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| Thu Jul 10 13:24:01 2014 | | | | -886 | -3865 | -50 | | | |

异常值

# 3.3、数据的清洗与转换

- 数据清洗：异常值识别


1. 慎重对待删除异常值：为减少犯错误的概率，可多种统计判别法结合使用，并尽力寻找异常值出现的原因；若有多个异常值，应逐个删除，即删除一个异常值后，需再行检验后方可再删除另一个异常值

2. 检验方法以正态分布为前提，若数据偏离正态分布或样本较小时，则检验结果未必可靠，校验是否正态分布可借助W检验、D检验

# 3.3、数据的清洗与转换

- 数据清洗：缺失值处理

1. 在数据缺失严重时，会对分析结果造成较大影响，因此对剔除的异常值以及缺失值，要采用合理的方法进行填补，常见的方法有平均值填充、K最近距离法、回归法、极大似线估计法等。

2. 随着数据量的增大，异常值和缺失值对整体分析结果的影响会逐渐变小，因此在“大数据”模式下，数据清洗可忽略异常值和缺失值的影响，而侧重对数据结构合理性的分析。

# 3.3、数据的清洗与转换

- 数据清洗：缺失值处理

| 平均值填充 | K最近距离法 | 回归 | 极大似然估计 | 多重插补法 |
| --- | --- | --- | --- | --- |
| 取所有对象（或与该对象具有相同决策属性值的对象）的平均值来填充该缺失的属性值 | 先根据欧式距离或相关分析确定距离缺失数据样本最近的K个样本，将这K个值加权平均来估计缺失数据值 | 基于完整的数据集，建立回归方程，对于包含空值的对象，将已知属性值代入方程来估计未知属性值，以此估计值来进行填充；但当变量不是线性相关或预测变量高度相关时会导致估计偏差 | 在给定完全数据和前一次迭代所得到的参数估计的情况下计算完全数据对应的对数似然函数的条件期望（E步），后用极大化对数似然函数以确定参数的值，并用于下步的迭代（M步） | 由包含m个插补值的向量代替每一个缺失值，然后对新产生的m个数据集使用相同的方法处理，得到处理结果后，综合结果，最终得到对目标变量的估计 |

# 3.3、数据的清洗与转换

- 数据清洗：缺失值处理

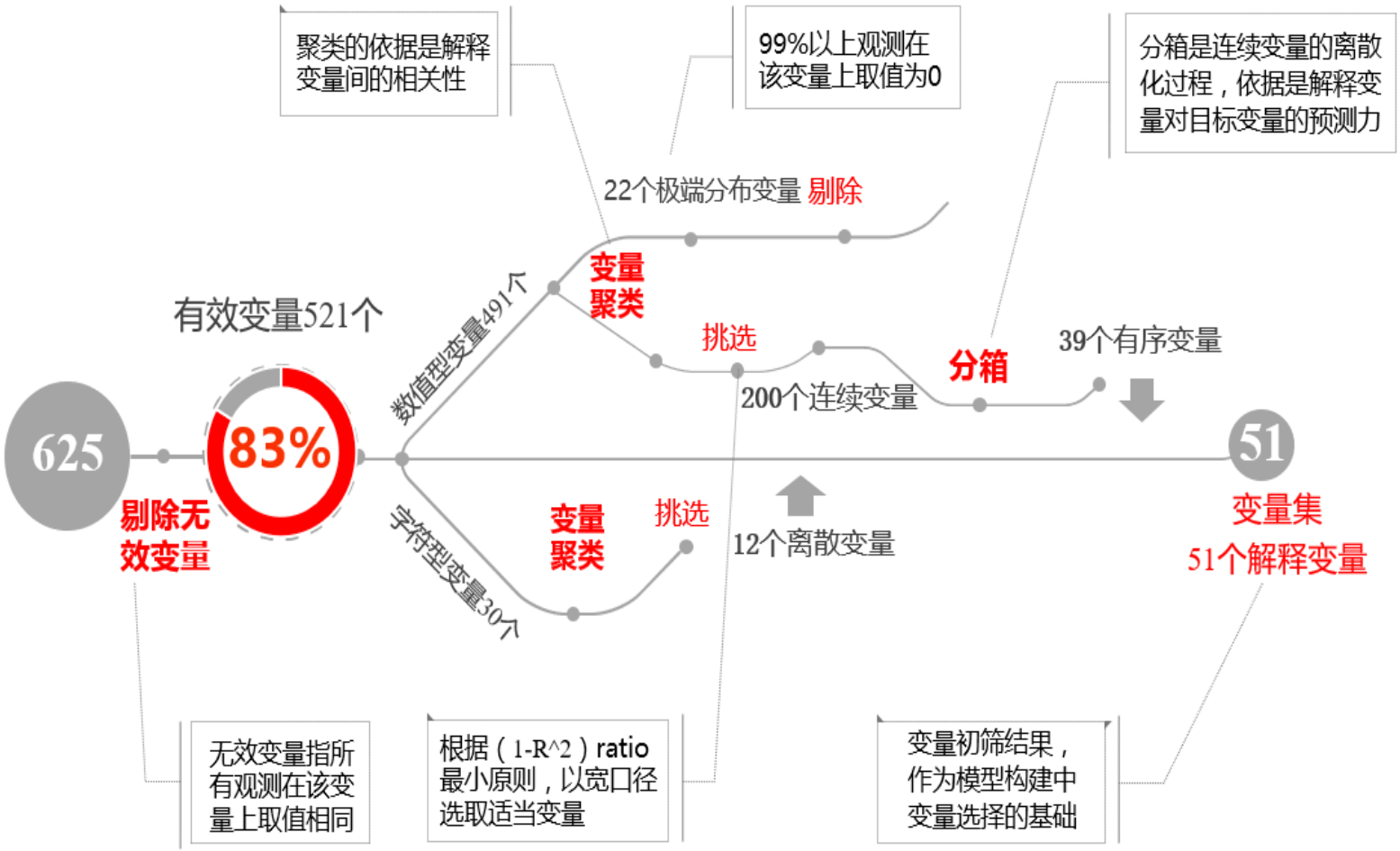
  - 例：对于2013年缺失的国内生产总值进行填补处理

| 年份 | 国内生产总值（亿元） | 均值填充 | 回归填充 | 前一个值填充 | 后一个值填充 |
|---|---|---|---|---|---|
| 2017 | 827121.7 | | | | |
| 2016 | 743585.5 | | | | |
| 2015 | 689052.1 | | | | |
| 2014 | 643974 | | | | |
| 2013 | | 557225.4 | 592170.7 | 643974 | 540367.4 |
| 2012 | 540367.4 | | | | |
| 2011 | 489300.6 | | | | |
| 2010 | 413030.3 | | | | |
| 2009 | 349081.4 | | | | |
| 2008 | 319515.5 | | | | |

数据来源：国家统计局

# 3.3、数据的清洗与转换

- 数据清洗和变量选择



# 3.3、数据的清洗与转换

- 数据转换

数据转换或统一成适合于挖掘的形式，通常的做法有数据泛化、标准化、属性构造等，本文详细介绍数据标准化的方法，即统一数据的量纲及数量级，将数据处理为统一的基准的方法。

① 基期标准化法

选择基期作为参照，

各期标准化数据＝各期数据/基期数据

② 直线法

极值法：$x'_i = \frac{x_i}{\max(x_i)}, x'_i = \frac{\max(x_i) - x_i}{\max(x_i)}, x'_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$

Z-score：$x'_i = \frac{x_i - \bar{x}}{s}$，其中$s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$

# 3.3、数据的清洗与转换

- 数据转换

③ 折线法

某些数据在不同值范围，采用不同的标准化方法，通常用于综合评价

$$x_i' = \begin{cases} 0(x_i < a) \\ \dfrac{x_i - a}{b - a}(a \le x_i < b) \\ 1(x_i \ge b) \end{cases}$$

④ 曲线法

Log函数法：$x_i' = \log(x_i) / \log(\max(x_i))$

Arctan函数法：$x_i' = \arctan(x_i) \times \frac{2}{\pi}$

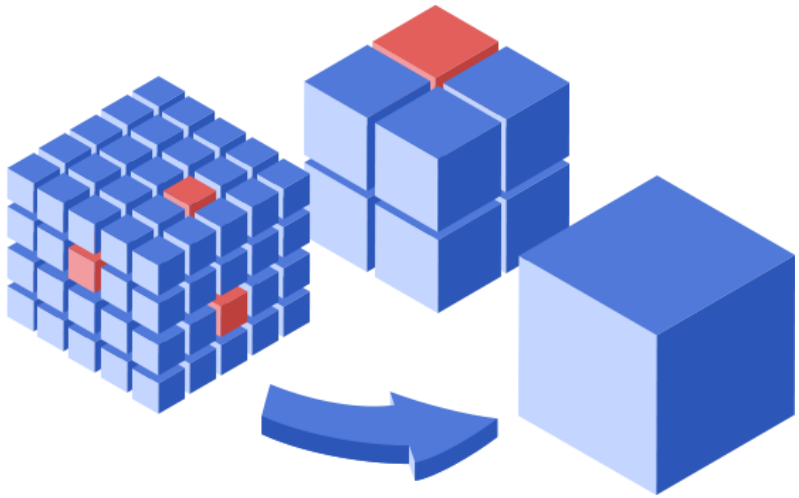
对数函数法、模糊量化模式等

# 3.3、数据的清洗与转换

- 数据转换

  - 各方法都有缺点，要根据客观事物的特征及所选用的分析方法来确定，如聚类分析、关联分析等常用直线法，且聚类分析必须满足无量纲标准；而综合评价则折线和曲线方法用得较多

  - 能简就简，能用直线尽量不用曲线。



# 3.4、实例分析

- 上海市民消费需求分析

  - ✓上海市某政府部门需要了解上海市民消费需求特点，从而为拉动内需提出相应的对策建议。
  - ✓2010年度针对上海市民进行了消费需求的抽样调查。
  - ✓每户家庭仅调查一位，18岁以下市民及全日制在校大学本科生和专科生不调查。
  - ✓分析上海市民消费需求特点；制约上海市民消费的因素；通过对具体项目消费意愿的分析寻求上海市民未来消费的热点。

# 3.4、实例分析

- 上海市民消费需求分析

  - 问卷展示，相关数据文件为问卷调查案例.word

| 基本信息 |
| --- |

Q1. 您的年龄（　　）

①18～20 岁　②21～30 岁　③31～40 岁

④41～50 岁　⑤51～60 岁　⑥60 岁以上

Q2. 您的受教育程度（　　）

①初中及以下　②高中或中专　③大专　④大学本科　⑤研究生及以上

Q3. 您家庭（指统一考虑收支并且居住在一起的人）有（　　）

①1 人　②2 人　③3 人　④4 人　⑤5 人及以上

Q4. 您所在工作部门的性质是（　　）

①政府部门　②事业单位　③国有企业　④外资企业　⑤中外合资企业

⑥民营企业　⑦个体户　⑧其他（请注明）_____

# 3.4、实例分析

- 上海市民消费需求分析

  - 问卷展示

**收入、消费支出的基本情况**

Q5. 您家庭的人均月收入（包括所有收入）为（　　）

①1000 元及以下　②1001～2000 元　③2001～3000 元　④3001～5000 元

⑤5001～10000 元　⑥10001～15000 元　⑦15001～20000 元　⑧20000 元以上

Q6. 未来 1 年内，预计您家庭的总收入将（　　）

①大幅度增加　②有所增加　③基本不变　④有所减少　⑤大幅度减少

**如果 Q6 回答为“③基本不变”，请跳至 Q8 回答。**

Q7. 未来 1 年内，预计您家庭总收入增加（或减少）的主要原因为（　　）（可多选）

①就业人口增加（或减少）　②工资奖金增加（或减少）　③养老金增加（或减少）

④经营收入增加（或减少）　⑤投资收益增加（或减少）　⑥其他（请注明）_____

Q8. 您家庭的人均月消费支出为（　　）

①1000 元及以下　②1001～2000 元　③2001～3000 元　④3001～5000 元

⑤5001～10000 元　⑥10001～15000 元　⑦15001～20000 元　⑧20000 元及以上

Q9. 您家庭的实际消费中，消费金额最多的消费项目为（　　），第二多的消费项目为（　　），第三多的消费项目为（　　）。

①衣着消费　②食品消费　③家庭设备用品消费　④医疗保健消费

# 3.4、实例分析

## ➢ 上海市民消费需求分析

### •问卷展示

**消费支出的制约因素**

Q14. 请您根据以下因素对您家庭消费支出的制约作用在 0 分～5 分之间进行打分，其中“0 分”表示“没有制约作用”，“5 分”表示“制约作用非常大”。在对应分值的[ ]内打勾选择。

| 制约因素 | 0<br>没有制约作用 | 1 | 2 | 3 | 4 | 5<br>制约作用非常大 |
|---|---|---|---|---|---|---|
| 收入水平 | [ ] | [ ] | [ ] | [ ] | [ ] | [ ] |
| 物价水平 | [ ] | [ ] | [ ] | [ ] | [ ] | [ ] |
| 存钱意愿 | [ ] | [ ] | [ ] | [ ] | [ ] | [ ] |
| 投资意愿 | [ ] | [ ] | [ ] | [ ] | [ ] | [ ] |
| 市场现有消费品满足需要的情况 | [ ] | [ ] | [ ] | [ ] | [ ] | [ ] |
| 用于消费的时间多少 | [ ] | [ ] | [ ] | [ ] | [ ] | [ ] |
| 是否愿意多花钱消费 | [ ] | [ ] | [ ] | [ ] | [ ] | [ ] |

Q15. 您家庭当期储蓄（银行储蓄存款）占当期收入的比例为（ ）

①0%～10%　②11%～30%　③31%～50%　④51%以上

Q16. 您家庭储蓄（银行储蓄存款）的主要目的为（ ）（可多选）

①防病　②养老　③子女教育　④防失业　⑤其他（请注明）_____

# 3.4、实例分析

## ➢ 上海市民消费需求分析

### •问卷展示

**具体项目的消费意愿**

Q22. 未来 1 年内，您家庭计划购买的轿车价格为（　　）

①暂无购车计划　②10 万及以下　③10 万～20 万（含 20 万）

④20 万～30 万（含 30 万）　⑤30 万～50 万（含 50 万）　⑥50 万～100 万（含 100 万）

⑦100 万以上

**如果 Q22 回答为“①暂无购车计划”，请跳至 Q26 回答。**

Q23. 您家庭购车受到 1.6 升及以下排量乘用车（小排量汽车）购置税优惠政策的影响为（　　）

①不了解该政策　②影响很大　③影响较大　④不确定　⑤影响较小　⑥没有影响

Q24. 在现有汽车以旧换新政策下，您家庭愿意汽车以旧换新的可能性为（　　）

①不了解该政策　②无换车条件　③非常可能　④有可能

⑤不确定　⑥不太可能　⑦不可能

Q25. 继中央五部委推出《关于开展私人购买新能源汽车补贴试点的通知》，明确国家对新能源汽车的各项补贴和扶持政策之后，上海正在积极筹划地方补贴政策。您家庭购车受到新能源汽车补贴政策的影响为（　　）

①不了解该政策　②影响很大　③影响较大　④不确定　⑤影响较小　⑥没有影响

**请跳至 Q27 回答。**

# 3.4、实例分析

- 上海市民消费需求分析

  - 问题：

    ①数据的类型是什么？ 离散的横截面数据

    ②多变量数据还是单变量数据？ 多变量的数据

    ③可以进行哪些数值分析？ 众数、中位数、列联表分析等

    ④可以进行哪些可视化化展示？ 柱状图、饼图、散点图等

# 3.4、实例分析

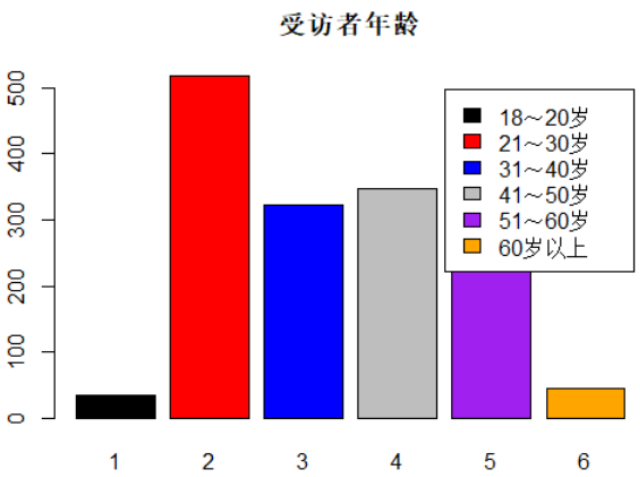
- 上海市民消费需求描述性分析

| | ID | Q1 | Q2 | Q3 | Q4 | Q4_8W | Q5 | Q6 | Q7_1 | Q7_2 | Q7_3 | Q7_4 | Q7_5 | Q7_6 | Q7_6W | Q8 | Q9_1 | Q9_2 | Q9_3 | Q9_8W | Q10_1 | Q10_2 | Q10_3 | Q10_4 | Q10_5 | Q10_6 | Q10_7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 4 | 3 | 3 | 3 | | 5 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | | 3 | 6 | 2 | 5 | | 3 | 4 | 4 | 3 | 4 | 4 | 4 |
| 2 | 2 | 2 | 4 | 1 | 4 | | 4 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | | 3 | 7 | 5 | 1 | | 3 | 2 | 3 | 2 | 2 | 4 | 3 |
| 3 | 3 | 1 | 2 | 3 | 7 | | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | | 2 | 3 | 4 | 7 | | 2 | 3 | 1 | 3 | 3 | 2 | 1 |
| 4 | 4 | 2 | 4 | 3 | 4 | | 5 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | | 4 | 2 | 3 | 6 | | 1 | 5 | 5 | 1 | 5 | 3 | 3 |
| 5 | 5 | 2 | 3 | 3 | 2 | | 8 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | | 6 | 1 | 5 | 3 | | 1 | 1 | 1 | 3 | 4 | 4 | 4 |
| 6 | 6 | 2 | 4 | 4 | 2 | | 5 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | | 5 | 2 | 4 | 6 | | 3 | 3 | 3 | 3 | 2 | 2 | 2 |
| 7 | 7 | 4 | 4 | 3 | 6 | | 7 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | | 1 | 2 | 6 | 4 | | 1 | 2 | 3 | 2 | 3 | 2 | 2 |
| 8 | 8 | 5 | 2 | 3 | 3 | | 3 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | | 2 | 6 | 4 | 2 | | 2 | 3 | 3 | 2 | 3 | 3 | 3 |
| 9 | 9 | 4 | 3 | 3 | 3 | | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | | 2 | 1 | 2 | 4 | | 3 | 3 | 3 | 2 | 3 | 3 | 3 |
| 10 | 10 | 4 | 2 | 2 | 7 | | 5 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | | 4 | 1 | 2 | 7 | | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 11 | 11 | 2 | 4 | 1 | 3 | | 4 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | | 2 | 2 | 6 | 5 | | 2 | 2 | 3 | 2 | 2 | 2 | 3 |
| 12 | 12 | 3 | 3 | 2 | 6 | | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | | 2 | 3 | 2 | 5 | | 3 | 3 | 3 | 3 | 2 | 3 | 3 |

1. 共有1496份调查数据，154个变量。

# 3.4、实例分析

- 上海市民消费需求描述性分析

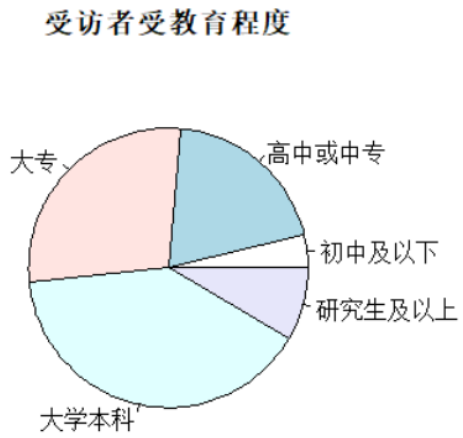


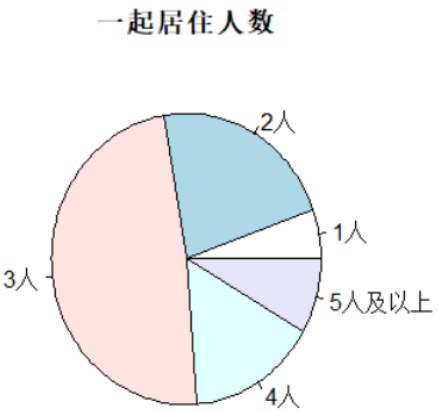

1. 本次调查受访者年龄基本集中于21~60岁之间，21~30岁之间人群最多，他们也是消费的主力军。

2. 受访者年龄的中位数与均值堵在31~40岁之间，较为符合社会结构。

# 3.4、实例分析

- 上海市民消费需求描述性分析



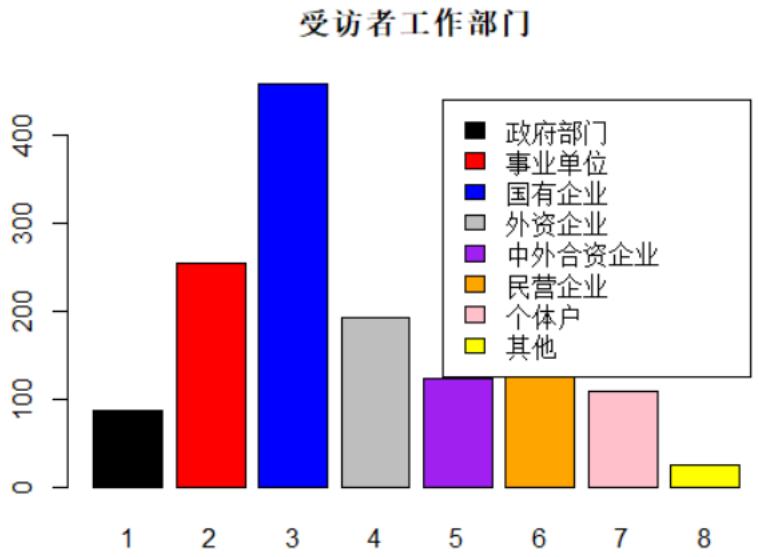




1. 受访者的受教育程度以大学本科和专科为主，极端值初中及以下意见研究生及以上水平的人数较少，比较符合社会现有结构。

2. 受访者基本以家庭为主，主要是三口之家居多，或者是2人活私人一起居住。独居者较少，符合现状。

# 3.4、实例分析

- 上海市民消费需求描述性分析




1. 受访者工作木门主要以国有企业居多，其他类型较为平均，主要反映了受访者不同工作性质的配合度。

2. 结合以上分析可得，受访者的样本与社会结构基本吻合，故该样本存在偏差较少，可以有效反应所有研究内容。

# 3.4、实例分析

- 上海市民消费需求描述性分析

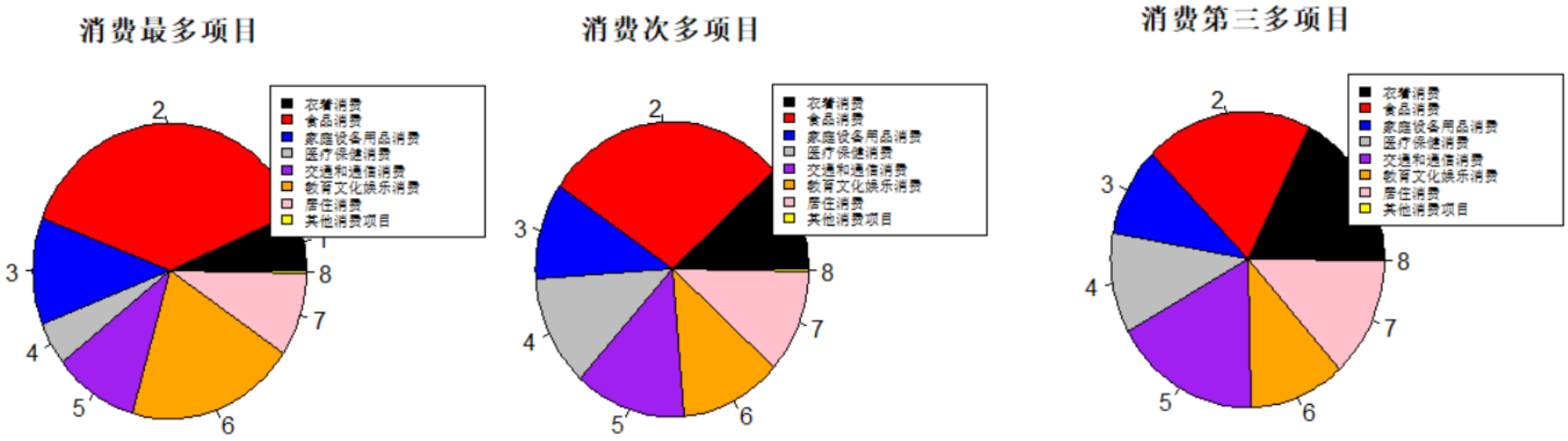
```
> mytable #显示列联表
       data$Q8
data$Q5   1   2   3   4   5   6   7   8
      1   9   2   0   0   0   0   0   0
      2  64  62   1   0   0   0   0   0
      3 103 243  51   2   0   0   0   0
      4  34 236 181  26   1   0   0   0
      5   8  61 132  75  16   0   0   0
      6   1   7  23  37  28   4   0   0
      7   1   1   6  12  17   5   2   0
      8   0   3   4  10  20   6   1   1
```

1. 明显地，数据集中于对角线上以及列联表的下三角部分，表明绝大多数人的消费与收入存在明显的正向关系。

```
       data$Q8
data$Q5           1           2           3           4           5           6           7           8
      1 0.040909091 0.003252033 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000
      2 0.290909091 0.100813008 0.002512563 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000
      3 0.468181818 0.395121951 0.128140704 0.012345679 0.000000000 0.000000000 0.000000000 0.000000000
      4 0.154545455 0.383739837 0.454773869 0.160493827 0.012195122 0.000000000 0.000000000 0.000000000
      5 0.036363636 0.099186992 0.331658291 0.462962963 0.195121951 0.000000000 0.000000000 0.000000000
      6 0.004545455 0.011382114 0.057788945 0.228395062 0.341463415 0.266666667 0.000000000 0.000000000
      7 0.004545455 0.001626016 0.015075377 0.074074074 0.207317073 0.333333333 0.666666667 0.000000000
      8 0.000000000 0.004878049 0.010050251 0.061728395 0.243902439 0.400000000 0.333333333 1.000000000
```

# 3.4、实例分析

- 上海市民消费需求描述性分析

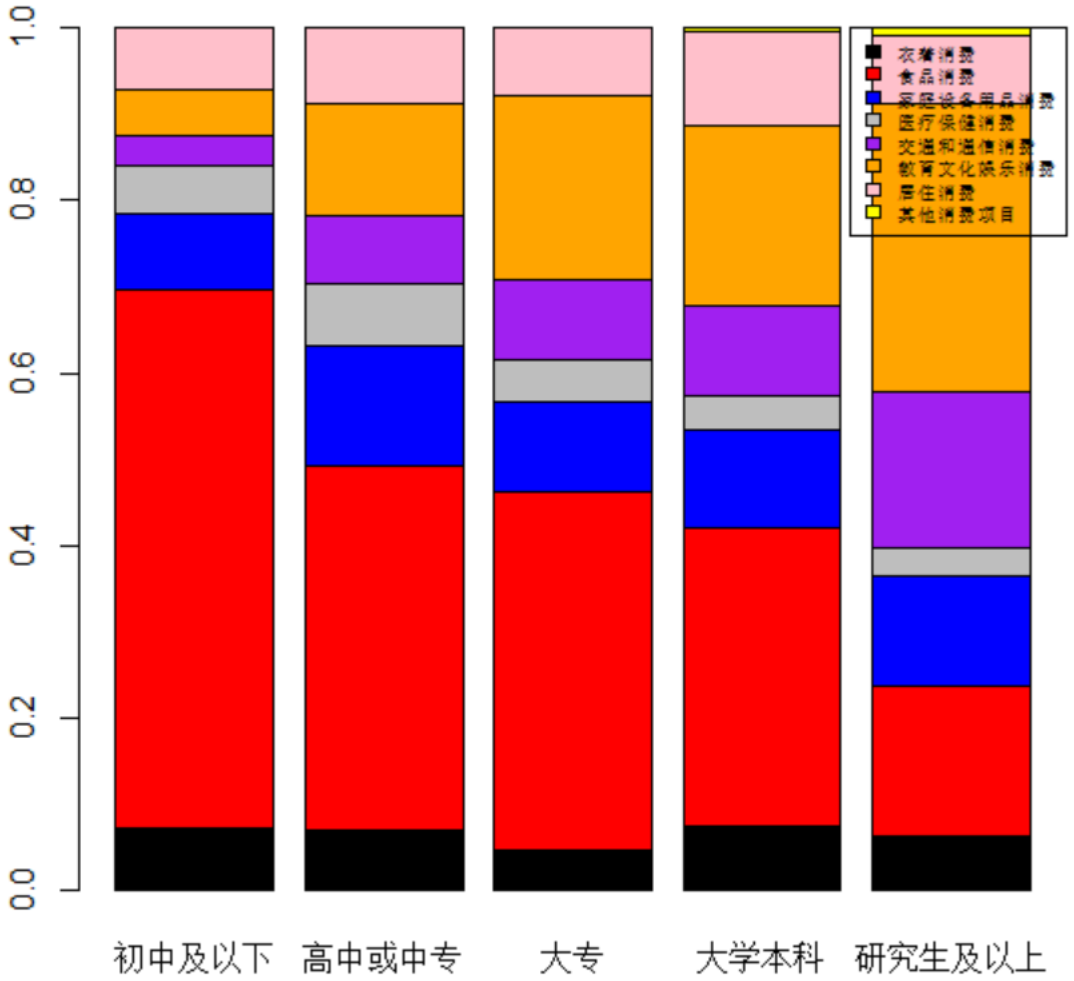



1. 总体来说，明显地，消费需求最大的还是食品消费。

2. 衣食住行占据着消费的很大比例。说明上海消费的Gini系数较高。

# 3.4、实例分析

- 上海市民消费需求描述性分析

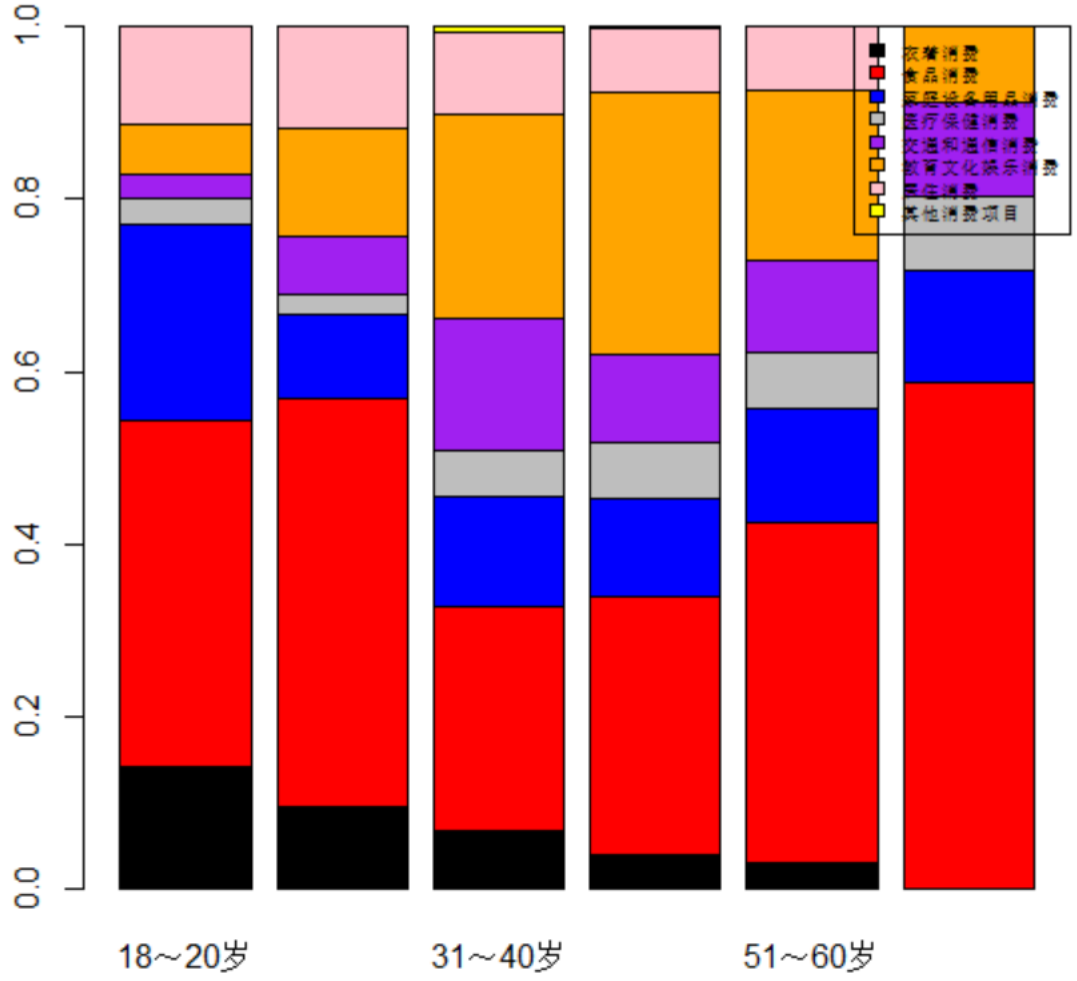



1. 随着学历的增加，对于食品的消费所占比例越来越小，而对于教育的支出则越来越大。

2. 随着学历的增加，对于居住的消费也有小微的增加。

# 3.4、实例分析

- 上海市民消费需求描述性分析



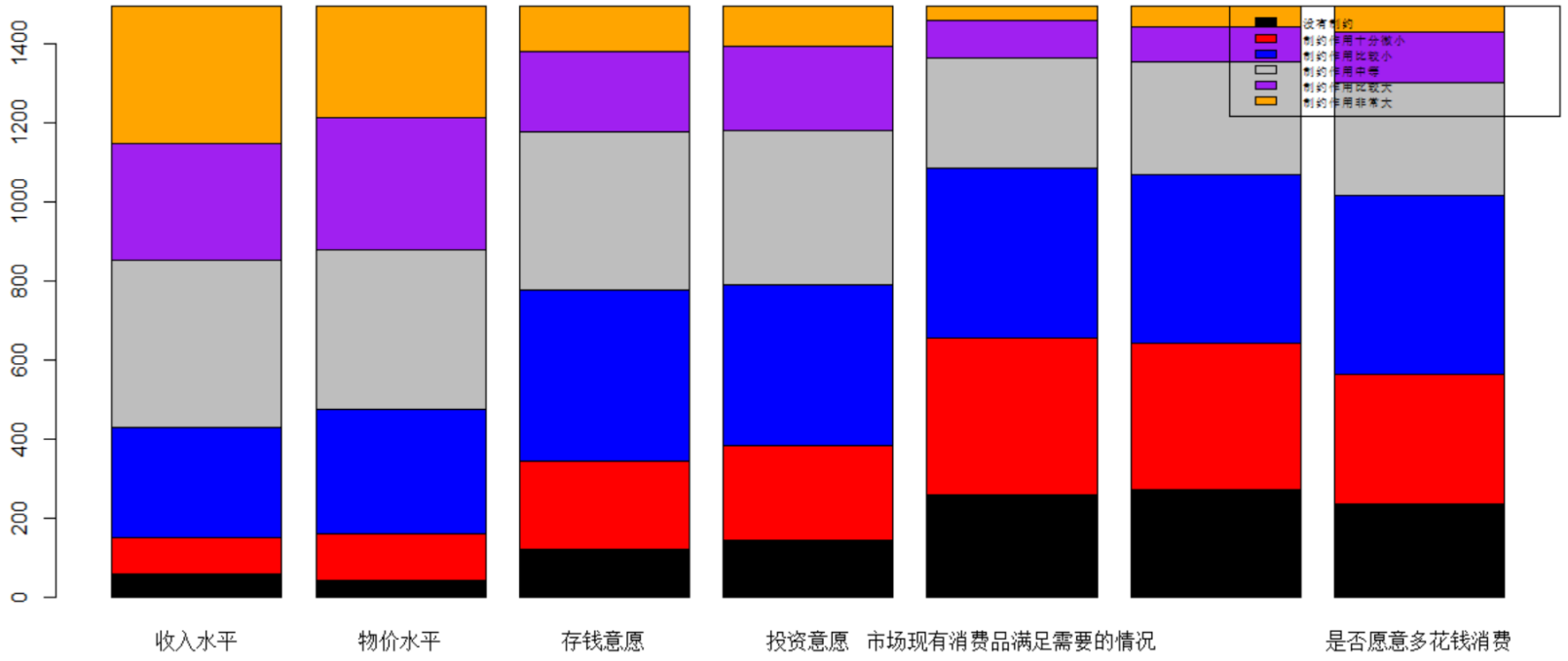

1. 食品消费这一块的支出与年龄的关系呈现“V”字型，即在31~40岁左右这一块消费比例较少，而年纪较小和年纪较大时，这块比例较大。

2. 与之相反的是，教育支出与年龄呈现倒“V”的关系。

3. 衣着支出则随着年龄增大而变少。

# 3.4、实例分析

- 上海市民消费需求描述性分析




1. 总体来说制约作用较大的市收入水平以及物价水平，而市场行为与个人意愿表现出的制约力相比较而言不是很大。

# 3.4、实例分析

- 上海市民消费需求描述性分析

  - 总结：

    1. 对于问卷调查的数据，大多数是分类变量，故在探究时需要根据问题寻找相关变量进行描述性分析，对于单个变量的分析较为常用的是中位数、众数、饼图、柱状图等。

    2. 不但要注意单个变量的分析，更要联合其他变量一起探究，这时候比较常用的是列联表、堆积柱状图、柱状图等。

# 3.4、实例分析

- 消费者信心指数

  - 消费者信心指数由**消费者评价指数**和**消费者预期指数**两个分类指数构成；

    - ✓消费者评价指数又由经济形势评价指数、收入评价指数、就业形势评价指数和耐用品购买意愿指数四个指数构成；

    - ✓消费者预期指数又由经济形势预期指数、收入预期指数、就业预期指数和耐用品购买预期指数四个指数构成。

# 3.4、实例分析

- 消费者信心指数

  - 消费者信心指数的取值范围在0~200之间

    - ✓当指数值大于100时，说明消费者信心整体偏向乐观
    - ✓指数值小于100时，表示消费者信心整体偏向悲观
    - ✓指数值越高，表示消费者整体信心越强

# 3.4、实例分析

- 消费者信心指数

  - 调查方式

    ✓上海财经大学上海市消费者信心指数采用电话调查方式。每次调查的样本量为1000个。

    ✓调查对象为上海中心城区和近郊共13个区县的常住居民，以及来沪工作或居住一年以上、年龄在20~69岁的务工者和居民，在校学生和外籍人士除外。

# 3.4、实例分析

- 消费者信心指数

  - 调查方式

    - ✓在每季度最后一个月的1日至10日实施调查，11日至30日为数据处理和分析阶段。本指数在每个季度的第一个月通过多家主流媒体对外发布，自2007年第三季度首次发布以来形成了广泛的影响力。

# 3.4、实例分析

- 消费者信心指数

•部分数据展示，相关数据文件为消费者信心指数.xlsx

| | A | AK | AL | AM | AN | AO | AP | AQ | AR | AS | AT | AU | AV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 日期 | 2016-3-20 | 2016-6-20 | 2016-9-20 | 2016-12-20 | 2017-3-20 | 2017-6-20 | 2017-9-20 | 2017-12-20 | 2018-3-20 | 2018-6-20 | 2018-9-20 | 2018-12-20 |
| 2 | 消费者信心指数ICS | 117.66405 | 108.983932 | 116.216675 | 111.9329758 | 119.042877 | 117.56025 | 121.36007 | 123.86657 | 118.1462 | 120.51463 | 117.77933 | 119.660294 |
| 3 | 消费者评价指数ICC | 120.064 | 112.497688 | 117.935936 | 114.1533508 | 121.260788 | 120.60604 | 122.32906 | 124.863548 | 119.38734 | 121.25392 | 119.88033 | 121.43367 |
| 4 | 消费者预期指数ICE | 115.26411 | 105.470146 | 114.497437 | 109.7125778 | 116.824944 | 114.51443 | 120.39109 | 122.869606 | 116.90509 | 119.77534 | 115.67834 | 117.886902 |

•问题：

①数据的类型是什么？ 连续型的时间序列数据

②多变量数据还是单变量数据？ 多变量的数据

③可以进行哪些数值分析？ 均值、方差、相关系数等

④可以进行哪些可视化化展示？ 折线图、箱线图等

# 3.4、实例分析

- 消费者信心指数描述性统计分析

预处理前与预处理后的对比展示：

| | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | X11 | X12 | X13 | X14 | X1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 日期 | 39253.0000 | 39345.0000 | 39436.0000 | 39527.0000 | 39619.00000 | 39711.00000 | 39802.00000 | 39892.00000 | 39984.0000 | 40076.0000 | 40167.0000 | 40257.0000 | 40349.0000 | 40 |
| 消费者信心指数ICS | 108.6158 | 109.1909 | 111.5794 | 103.6770 | 100.47018 | 103.13718 | 94.14720 | 98.88175 | 111.1575 | 109.3784 | 115.8972 | 109.0322 | 105.1590 | |
| 消费者评价指数ICC | 108.4895 | 109.7646 | 111.7408 | 103.5808 | 105.07249 | 107.72281 | 95.84207 | 97.13017 | 106.6544 | 106.7400 | 113.5891 | 106.4960 | 104.5480 | |
| 消费者预期指数ICE | 108.7421 | 108.6172 | 111.4181 | 103.7732 | 95.86786 | 98.55159 | 92.45234 | 100.63335 | 115.6606 | 112.0168 | 118.2053 | 111.5685 | 105.7701 | |

| | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 |
|---|---|---|---|---|---|---|---|---|---|
| 日期 | 2007-06-20 | 2007-09-20 | 2007-12-20 | 2008-03-20 | 2008-06-20 | 2008-09-20 | 2008-12-20 | 2009-03-20 | 2009-06-20 |
| 消费者信心指数ICS | 108.615814208984 | 109.190887451172 | 111.579444885254 | 103.67699432373 | 100.470184326172 | 103.137184143066 | 94.1472015380859 | 98.8817520141602 | 111.157501220703 |
| 消费者评价指数ICC | 108.489517211914 | 109.76456451416 | 111.740798950195 | 103.580757141113 | 105.072486877441 | 107.722808837891 | 95.8420715332031 | 97.1301727294922 | 106.654357910156 |
| 消费者预期指数ICE | 108.74210357666 | 108.617202758789 | 111.418113708496 | 103.773231506348 | 95.8678588867188 | 98.5515899658203 | 92.4523391723633 | 100.633346557617 | 115.660629272461 |

# 3.4、实例分析

- 消费者信心指数描述性统计分析

> 均值

[1] 110.2400 111.3588 109.1212

> 协方差

| | [,1] | [,2] | [,3] |
|---|---|---|---|
| [1,] | 52.56827 | 52.90387 | 52.23267 |
| [2,] | 52.90387 | 56.31587 | 49.49187 |
| [3,] | 52.23267 | 49.49187 | 54.97349 |

1. 三个指数的均值都超过100，说明总体来说，消费者信心整体偏向乐观。

2. 三个指数的方差分别为52.57，56.32，54.97。总体来说，三种指数的波定性较为一致，相较于均值来说，该程度的方差并不是十分巨大，所以说指数较为稳定。

# 3.4、实例分析

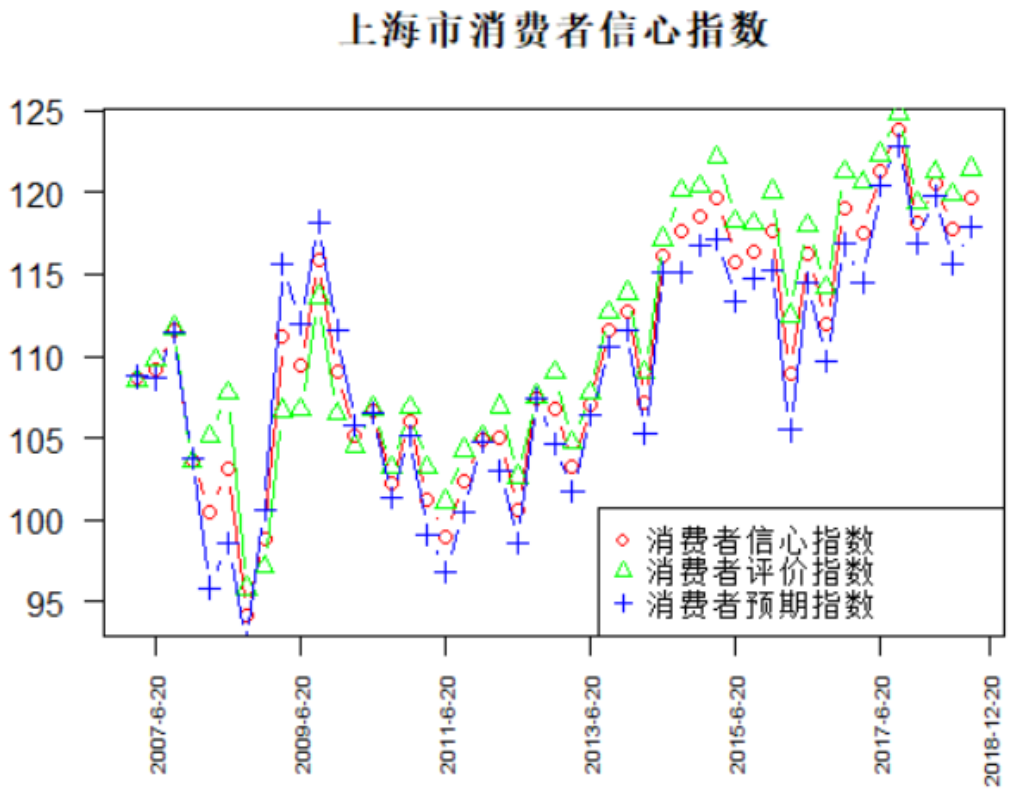
- 消费者信心指数描述性统计分析
  - 相关系数

| | [,1] | [,2] | [,3] |
|---|---|---|---|
| [1,] | 1.0000000 | 0.9723222 | 0.9716364 |
| [2,] | 0.9723222 | 1.0000000 | 0.8894915 |
| [3,] | 0.9716364 | 0.8894915 | 1.0000000 |

1. 三个指数俩俩之间的相关系数都达到了0.9左右，说明三个指数呈现明显的正向相关关系。

# 3.4、实例分析

- 消费者信心指数

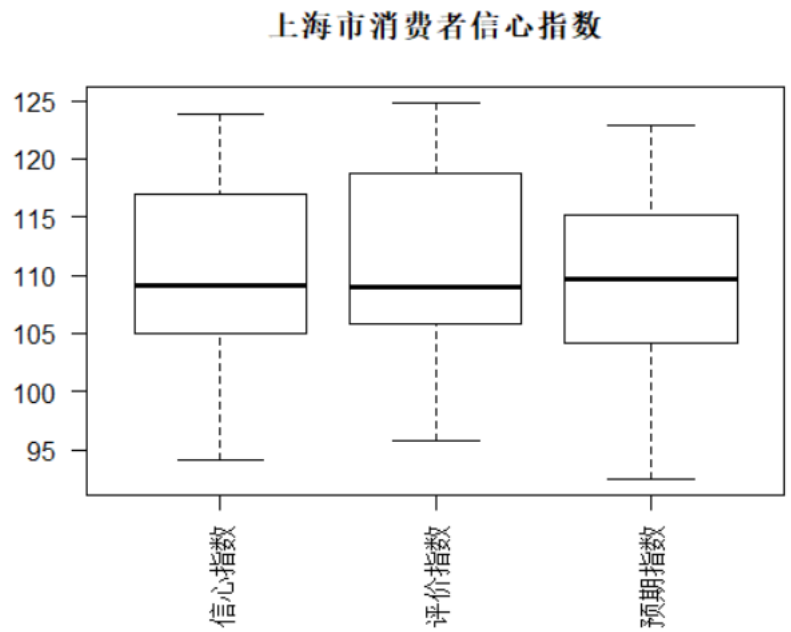



1. 各项指数在本季度总体回升，总指数及多项分类指数较上个季度有所上涨。纵观全年各项主要指数在本年度相对较为稳定，波动幅度不大。

2. 本年度国内经济形势和国际经济环境都面临了诸多的困难，各项指数依然能够在较高点位持续运行，说明消费者对上海的经济形势和预期都抱有较高的评价和直面挑战的信心。

# 3.4、实例分析

- 消费者信心指数




1. 信心指数、评价指数总体来说为右偏分布，而预期指数则相对较为对称。

2. 三大指数中位数较为接近，而前两个指数的波动性较大。

3. 数据不存在明显的异常点。

# 3.4、实例分析

- 上海市消费者信心指数处于较高点位原因

  - 近一年来，上海经济增速再次回升且较为平稳，转型升级前景乐观，使得消费者对经济发展的信心能够保持稳定。
  - 上个季度各项指数冲高回落，因此本季度的适度上涨也是对之前大幅下降的调整修正。
  - 中美贸易战暗潮涌动，资本市场乃至实体经济都面临着一定的挑战，消费者信心的稳定则是体现了对经济发展和面对各种挑战的信心，反映了我国的综合实力和抗危机能力逐步增强，因此本季度上海市消费者信心指数仍然在高位运行。

# 3.4、实例分析

## ➢ 本福特定律

- 法兰克·本福特本来是一个美国电气工程师，也是一名物理学家，在美国通用电气公司（GE）实验室里工作多年直到退休。
- 事实上，本福特定律的最早发现者并不是本福特，而是美国天文学家西蒙·纽康。
- 繁杂的天文计算经常需要用到对数表，但那个时代没有互联网，没有阿里云，对数表被印成书本，存于图书馆。细心的纽康发现一个奇怪的现象：**对数表中包含以1开头的数的那几页比其他页破烂得多，似乎表明计算所用的数值中，首位数是1。**

# 3.4、实例分析

## ➢ 本福特定律

- 本福特的发现便是如此：**以1开头的数字比较多，这也算是一个定律吗？**

- 本福特发现**这种现象不仅仅存在于对数表中，也存在于其它多种数据中**。于是，本福特检查了大量数据而证实了这点。

# 3.4、实例分析

- 本福特定律

•设想某银行有1000多个储存账户，金额不等。比如说，张本有存款23587元、老李1345元、小何35670元、刘红9000元、王军450元……等等

•奇怪的本福特定律不感兴趣存款金额本身，而感兴趣这些数值的开头第一位有效数字是什么，指的是这个数的第一个非零数字。所以，本福特定律也叫“首位数字定律”。

- ✓ 例如8.1、81、0.81的第一位有效数字都是8

# 3.4、实例分析

- 本福特定律

•一个数的第一位（非零）数字可能是1到9之间的任何一个。现在，如果我问，在刚才那个银行的上千个存款数据中，第一位数字是1的概率是多大？

•不需要经过很多思考，大部分人都会很快地回答：应该是1/9吧。因为从1–9，9个数字排在第一位的概率是相等的，每一个数字出现的概率都是1/9，大约11%左右。

•这听起来十分正常的思维方法却与许多自然得到的数据所遵循的规律不一样。人们发现，很多情况下，第一个数字是1的概率要比靠直觉预料的11%大得多。

# 3.4、实例分析

- 本福特定律

•数字越大，出现在第一位的概率就越小，数字9出现于第一位的概率只有4.5%左右。

•本福德和纽康都从数据中总结出首位数字为n的概率公式是：

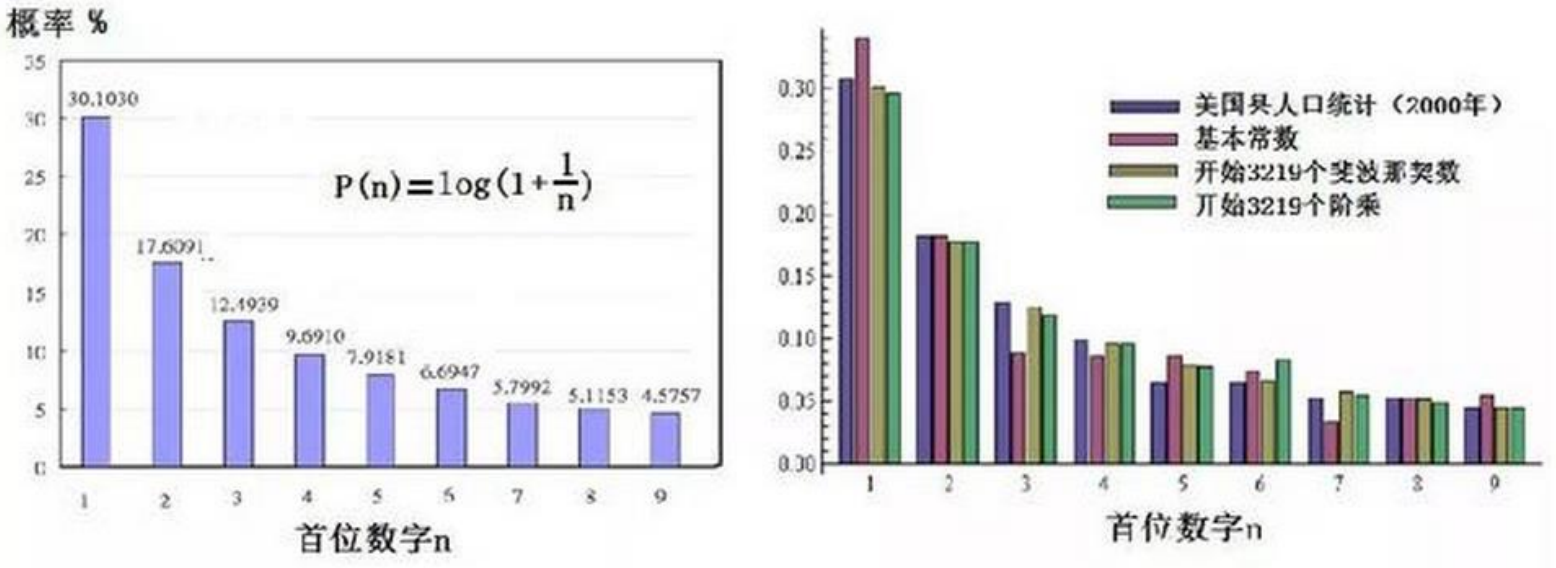
$$P(n) = \log_d(1 + \frac{1}{n})$$

其中d取决于数据使用的进位制，对十进制数据而言，d=10。

•根据本福德定律，首位数是1的概率最大，$\log_{10}2 = 0.301$，十成中占了三成；首位数是2的概率$\log_{10}3/2$=0.1761；然后逐次减小，首位数是9的概率最小，只有4.6%。

# 3.4、实例分析

- 本福特定律




本福特定律应用实例

# 3.4、实例分析

- 本福特定律

首位数字概率表

| 统计项目 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 河流面积 | 31.0 | 16.4 | 10.7 | 11.3 | 7.2 | 8.6 | 5.5 | 4.2 | 5.1 |
| 人口 | 33.9 | 20.4 | 14.2 | 8.1 | 7.2 | 6.2 | 4.1 | 3.7 | 2.2 |
| 常数 | 41.3 | 14.4 | 4.8 | 8.6 | 10.6 | 5.8 | 1.0 | 2.9 | 10.6 |
| 报纸 | 30.0 | 18.0 | 12.0 | 10.0 | 8.0 | 6.0 | 6.0 | 5.0 | 5.0 |
| 热量 | 24.0 | 18.4 | 16.2 | 14.6 | 10.6 | 4.1 | 3.2 | 4.8 | 4.1 |
| 压强 | 29.6 | 18.3 | 12.8 | 9.8 | 8.3 | 6.4 | 5.7 | 4.4 | 4.7 |
| 损失 | 30.0 | 18.4 | 11.9 | 10.8 | 8.1 | 7.0 | 5.1 | 5.1 | 3.6 |
| 分子量 | 26.7 | 25.2 | 15.4 | 10.8 | 6.7 | 5.1 | 4.1 | 2.8 | 3.2 |
| 下水道 | 27.1 | 23.9 | 13.8 | 12.6 | 8.2 | 5.0 | 5.0 | 2.5 | 1.9 |
| 原子量 | 47.2 | 18.7 | 5.5 | 4.4 | 6.6 | 4.4 | 3.3 | 4.4 | 5.5 |
| n−1, $\sqrt{n}$ | 25.7 | 20.3 | 9.7 | 6.8 | 6.6 | 6.8 | 7.2 | 8.0 | 8.9 |
| 设计 | 26.8 | 14.8 | 14.3 | 7.5 | 8.3 | 8.4 | 7.0 | 7.3 | 5.6 |
| 摘要 | 33.4 | 18.5 | 12.4 | 7.5 | 7.1 | 6.5 | 5.5 | 4.9 | 4.2 |
| 花费 | 32.4 | 18.8 | 10.1 | 10.1 | 9.8 | 5.5 | 4.7 | 5.5 | 3.1 |
| X射线 | 27.9 | 17.5 | 14.4 | 9.0 | 8.1 | 7.4 | 5.1 | 5.8 | 4.8 |
| 联盟 | 32.7 | 17.6 | 12.6 | 9.8 | 7.4 | 6.4 | 4.9 | 5.6 | 3.0 |
| 黑体 | 31.0 | 17.3 | 14.1 | 8.7 | 6.6 | 7.0 | 5.2 | 4.7 | 5.4 |
| 地址 | 28.9 | 19.2 | 12.6 | 8.8 | 8.5 | 6.4 | 5.6 | 5.0 | 5.0 |
| n, n2, . . . , n! | 25.3 | 16.0 | 12.0 | 10.0 | 8.5 | 8.8 | 6.8 | 7.1 | 5.5 |
| 死亡率 | 27.0 | 18.6 | 15.7 | 9.4 | 6.7 | 6.5 | 7.2 | 4.8 | 4.1 |
| 平均 | 30.6 | 18.5 | 12.4 | 9.4 | 8.0 | 6.4 | 5.1 | 4.9 | 4.7 |
| 本福德定律 | 30.1 | 17.6 | 12.5 | 9.7 | 7.9 | 6.7 | 5.8 | 5.1 | 4.6 |

# 3.4、实例分析

- 本福特定律

  - 由于大多数财务方面的数据，都满足本福德定律。因此，它可以用作检查财务数据是否造假。

  - 美国华盛顿州侦破过一个当时最大的投资诈骗案，金额高达1亿美元。

    ✓诈骗主谋凯文·劳伦斯及其同伙，以创办高技术含量的连锁健身俱乐部为名，向5000多个投资者筹集了大量资金。然后，他们挪用公款用作自身享乐，为他们自己买豪宅、豪华汽车、珠宝等。

# 3.4、实例分析

- 本福特定律

•美国华盛顿州侦破过一个当时最大的投资诈骗案，金额高达1亿美元（续）

✓为了掩饰他们的不法行为，他们将资金在海外公司和银行间进行频繁转账，并且人为做假账，给投资者造成生意兴隆的错觉。

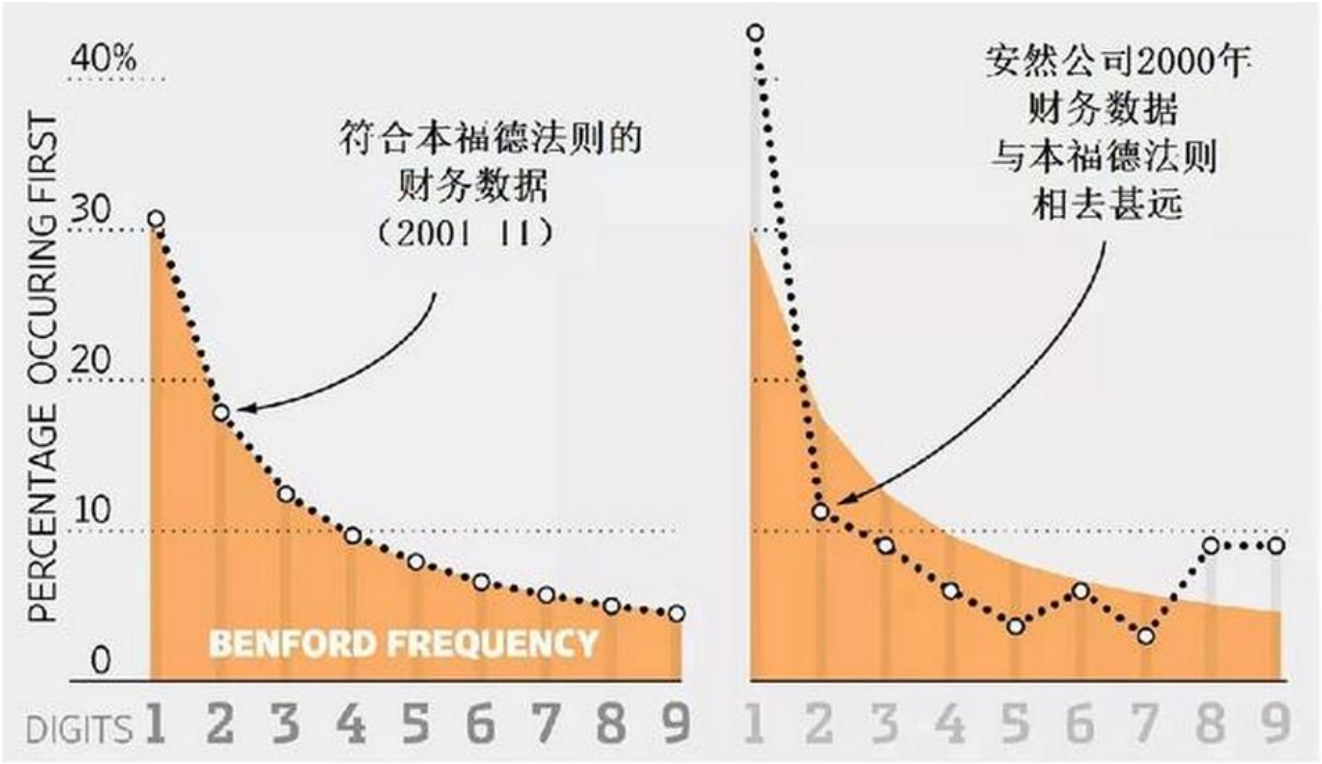
✓所幸有一位会计师（Darrell Dorrell）感觉不对，他将70000多个与支票和汇款有关的数据收集起来，将这些数据首位数字发生的概率与本福德定律相比较，发现这些数据通过不了第一数字法则的检验。

✓最后经过了3年的司法调查，终于拆穿了这个投资骗局。

# 3.4、实例分析

- 本福特定律

•据说安然高层改动过财务数据，因而他们所公布的2001-2002年每股盈利数据不符合本福特定律。



# 3.4、实例分析

- 本福特定律在国内某公司财报数据的应用

  预处理前与预处理后的对比展示：

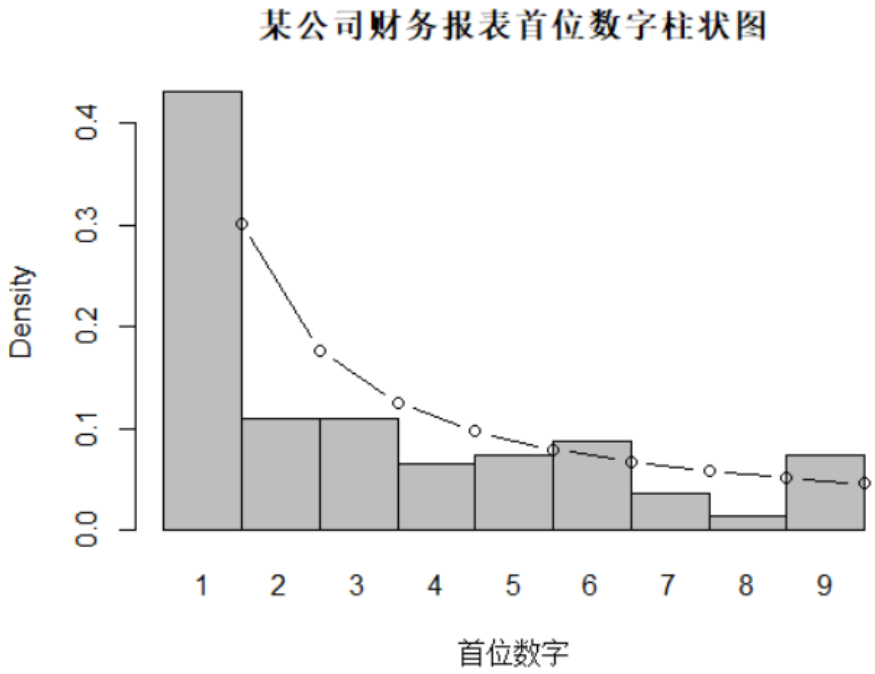
| | ths_commi_on_insurance_policy_stock | ths_financing_expenses_stock | ths_ii_from_jc_etc_stock | ths_noncurrent_asset_dispose_loss_stock | ths_rein_expenditure_stock | ths_fv_chg_income_stock |
|---|---|---|---|---|---|---|
| 1 | NA | 13730201 | NA | NA | NA | NA |

Filter

| | ths_financing_expenses_stock | ths_manage_fee_stock | ths_total_compre_income_atsopc_stock | ths_np_atoopc_stock | ths_basic_eps_stock | ths_np_stock | ths_total_profit_stock | ths_income_tax |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.37302e+16 | 3.031264e+16 | 1.876303e+17 | 1.876303e+17 | 1.06e+09 | 1.876303e+17 | 1.843787e+17 | 3.251668e+15 |

# 3.4、实例分析

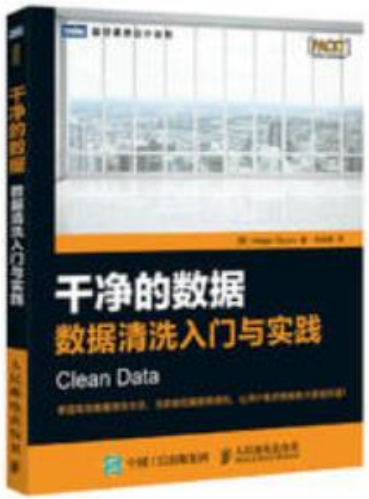
- 本福特定律在国内某公司财报数据的应用



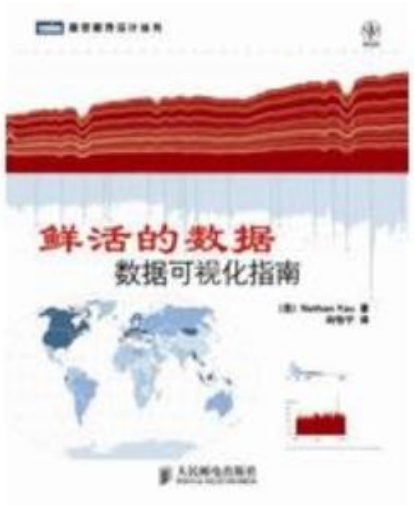

1. 该公司首位数字的分布于理论值存在较大差距，故有理由怀疑该公司的财务报表存在一定的粉饰。

# 推荐阅读


数据思维


干净的数据
数据清洗入门与实践
Clean Data


R数据可视化手册
O'REILLY


鲜活的数据
数据可视化指南

# 授课要点

## 4.统计建模技术

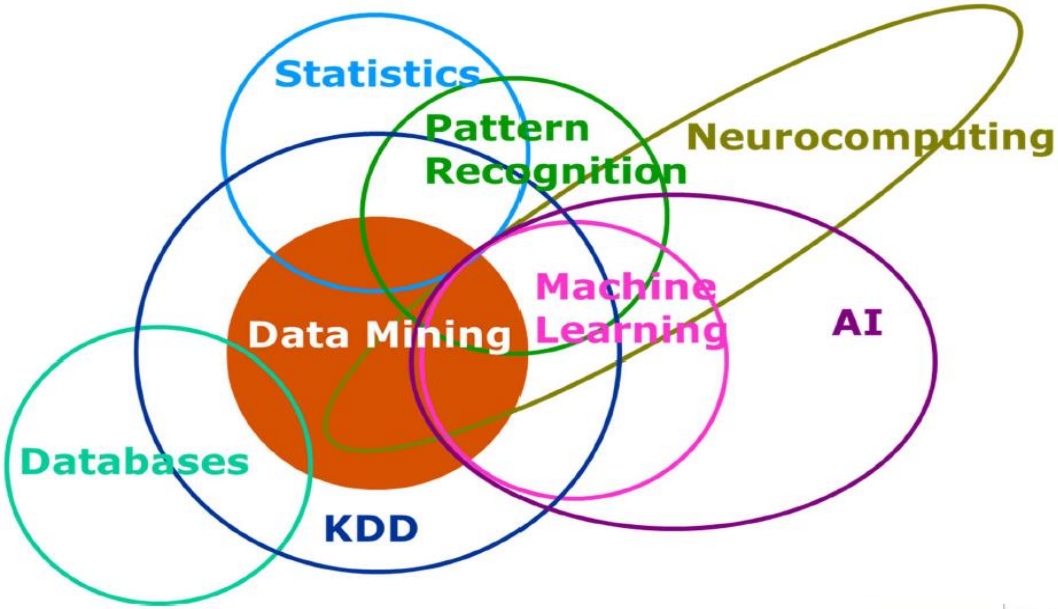
### 4.1.抽样分布

### 4.2.假设检验

### 4.3.预测分析

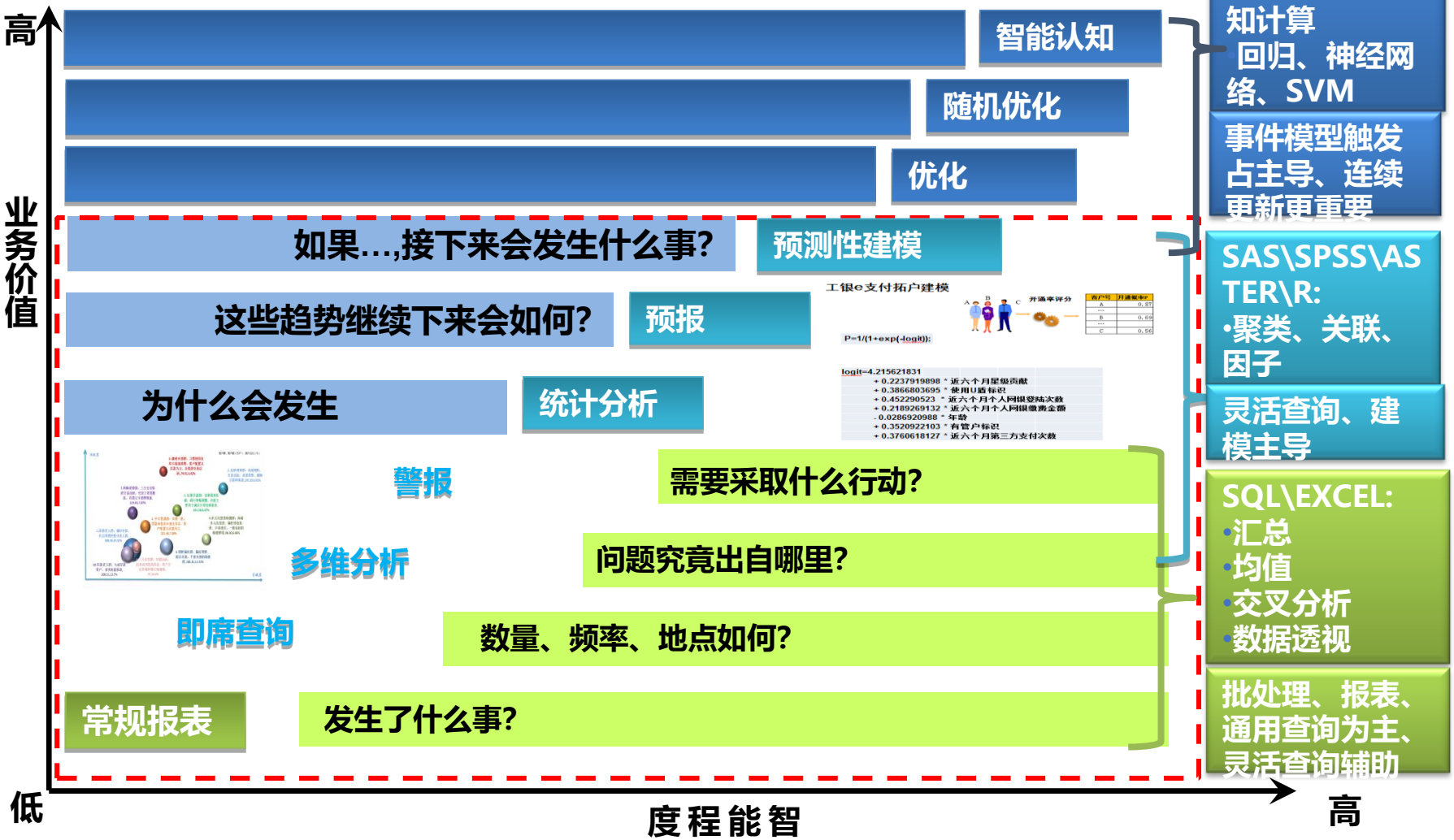
### 4.4.聚类分析

- 数据挖掘融合了数据库、人工智能、机器学习、**统计学**、高性能计算、模式识别、神经网络、数据可视化、信息检索和空间数据分析等多个领域的和技术，是21世纪初期对人类产生重大影响的十大新兴技术之一。



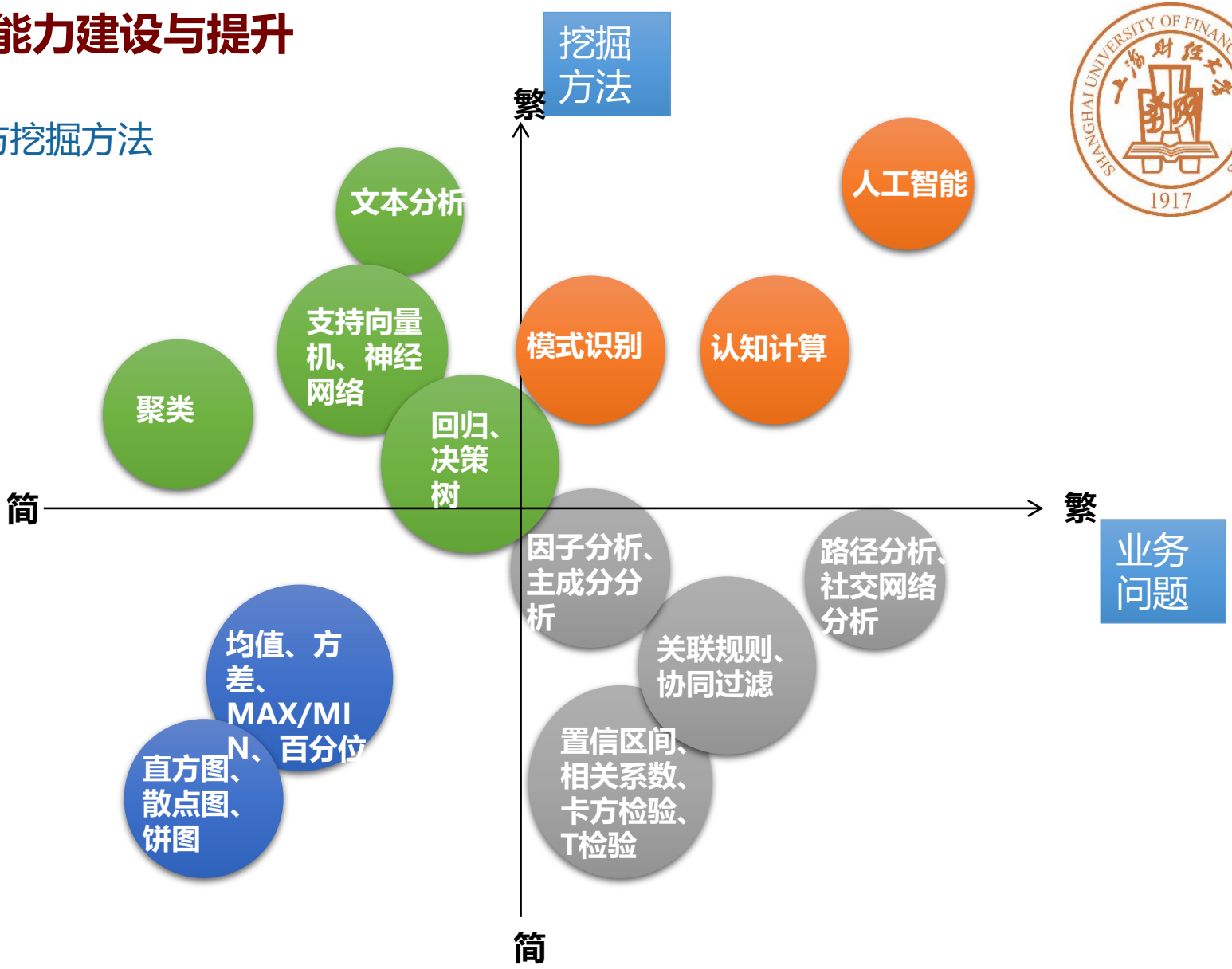
# 数据分析能力建设与提升

## 分析能力的十个等级



# 数据分析能力建设与提升

## 业务问题与挖掘方法

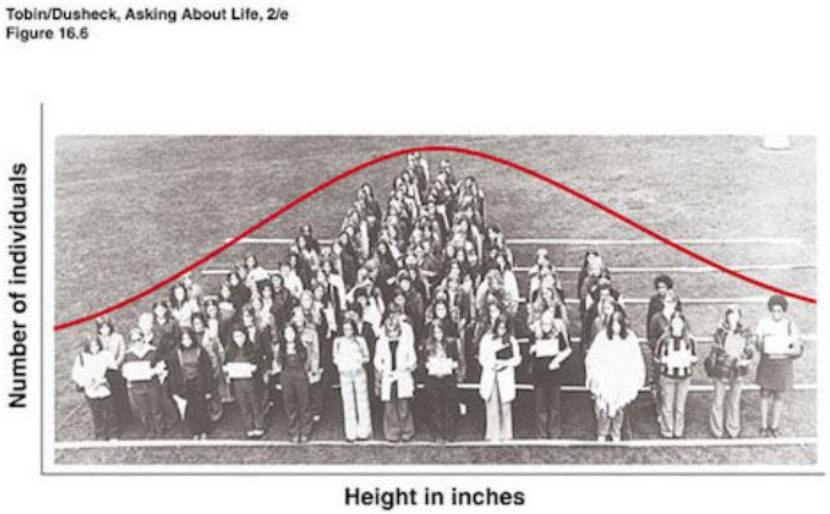


# 4.1、抽样分布

- **抽样分布是什么?**
  - 样本统计量的概率分布，是一种理论分布；
  - 在重复选取容量为$n$的样本时，由该统计量的所有可能取值形成的相对频数分布 ；
  - 结果来自容量相同的所有可能样本；
  - 提供了样本统计量长远而稳定的信息，是进行推断的理论基础，也是抽样推断科学性的重要依据。

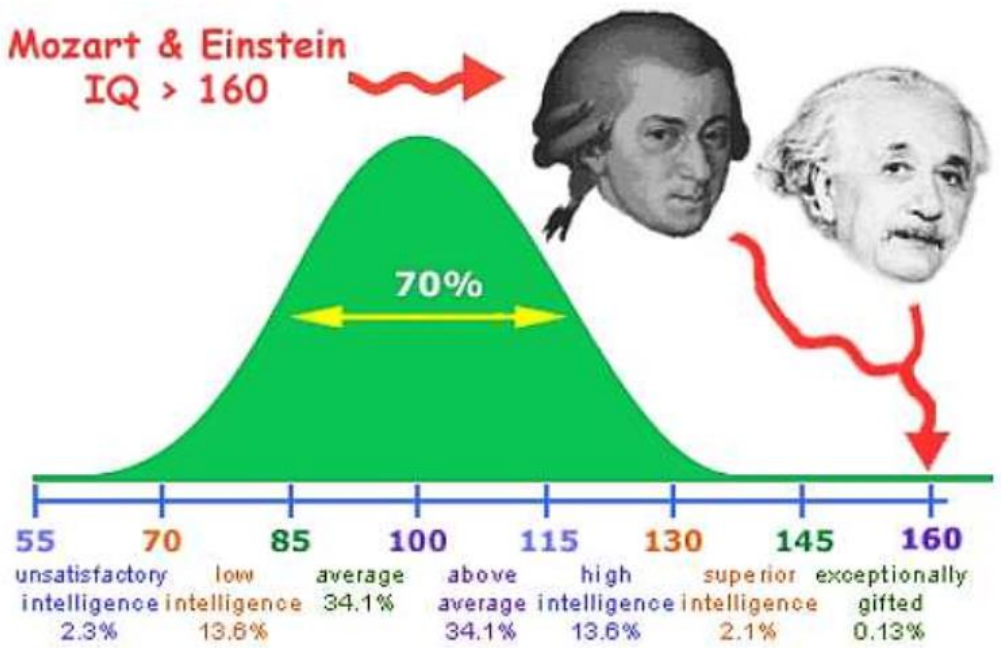
# 4.1、抽样分布

- 正态分布（Normal Distribution）
  - Normal意思是“常见的”，它能恰当代表多种多样的数据类型，如我们的考试成绩、身高的统计等。很多数据都符合正态分布，所以，它在数学、物理、工程领域有广泛的应用。



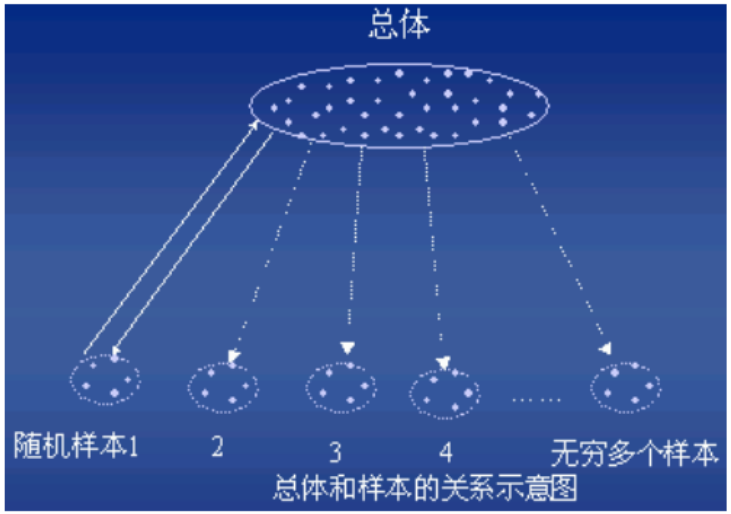
# 4.1、抽样分布

- 正态分布（Normal Distribution）
  - 正态分布的曲线呈钟型，两头低，中间高的形态，大部分数据集中在平均值，小部分在两端。
  - 例子：智商的分布



# 4.1、抽样分布

- 如果总体的分布类型已知（比如正态分布），从总体当中抽取大小为$N$的样本$X_1, \cdots, X_N$，利用这些样本可以算出样本的某些统计量$T$（比如均值等）。
- 统计量$T$服从的分布称为**抽样分布**。



# 4.1、抽样分布

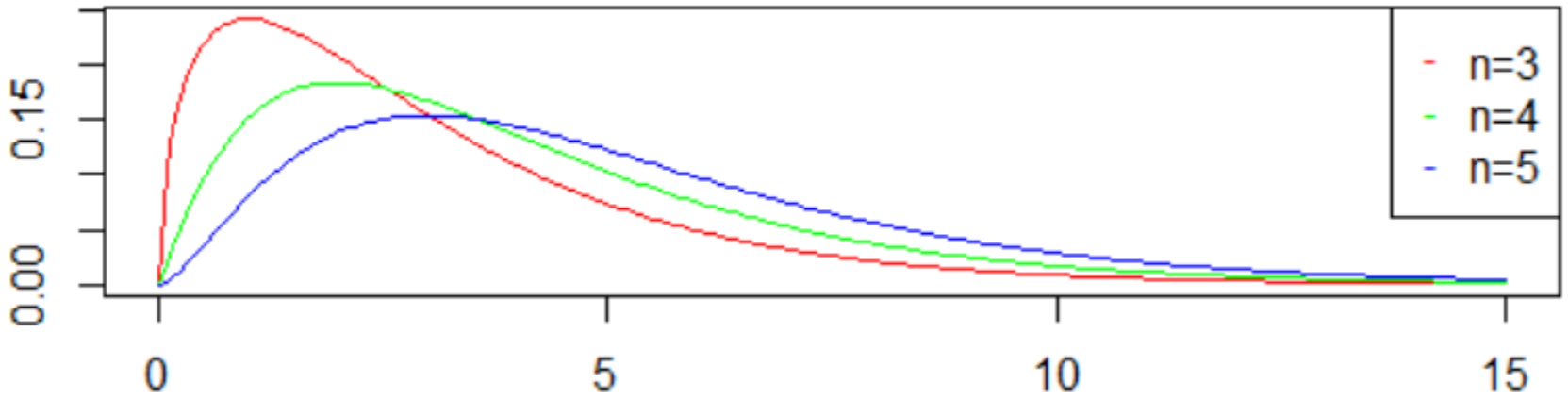
- 如果总体的分布类型已知（比如正态分布），从总体当中抽取大小为$N$的样本$X_1, \cdots, X_N$，利用这些样本可以算出样本的某些统计量$T$（比如均值等）。
- 统计量$T$服从的分布称为抽样分布。
- 在正态总体条件下，抽样分布主要有**$\chi^2$分布，$t$分布，$F$分布**
- 三大抽样分布在**假设检验**中有着重要作用

# 4.1、抽样分布

※ $\chi^2$分布

- 由阿贝(Abbe) 于1863年首先给出，后来由海尔墨特(Hermert)和卡·皮尔逊(K·Pearson) 分别于1875年和1900年推导出来
- 当总体 $X \sim N(\mu, \sigma^2)$，从中抽取容量为n的样本，则

$$\frac{\sum_i^n (x_i - \bar{x})^2}{\sigma^2} \sim \chi^2(n-1)$$

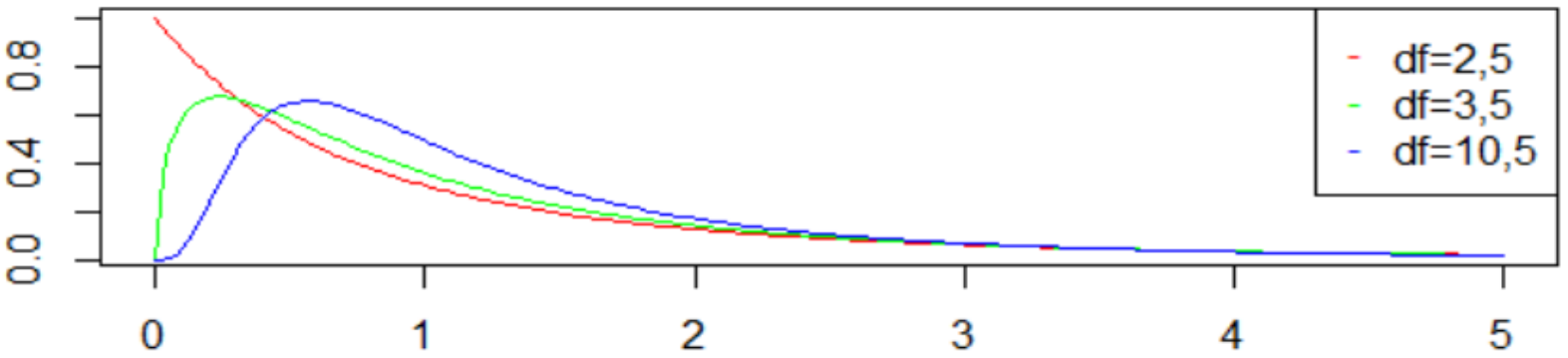


# 4.1、抽样分布

※ $F$分布

- 由统计学家费希尔(R.A.Fisher) 提出的，以其姓氏的第一个字母来命名
- 设若$U$为服从自由度为$n_1$的$\chi^2$分布，$V$为服从自由度为$n_2$的$\chi^2$分布，且$U$和$V$相互独立，则称$F$为服从自由度$n_1$和$n_2$的F分布，

$$F = \frac{U}{V} \sim F(n_1, n_2)$$

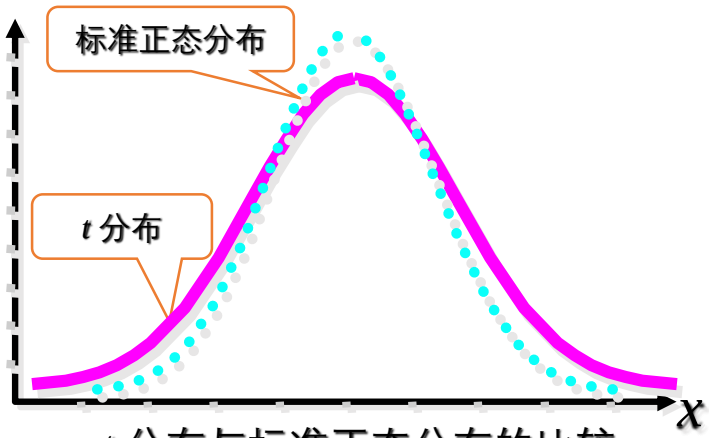


# 4.1、抽样分布

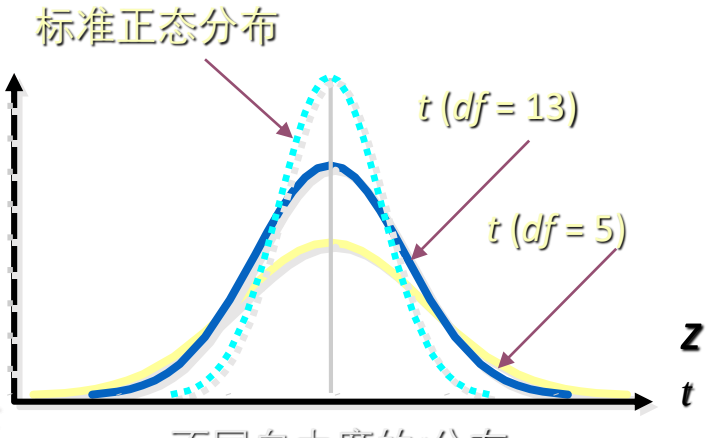
## ※ $t$分布

- 高塞特(W.S.Gosset)于1908年在一篇以“Student”（学生)为笔名的论文中首次提出
- 当总体 $X \sim N(\mu, \sigma^2)$，从中抽取容量为n的样本， $\bar{X}$ 与 $S^2$ 分别是该样本的样本均值与样本方差，则

$$T = \frac{\sqrt{n}(\bar{x} - \mu)}{S} \sim t(n-1)$$




t 分布与标准正态分布的比较




不同自由度的t分布

# 4.1、抽样分布

- 准确捕捉变量的**集中趋势**和**离散趋势**在统计中有极为重要的意义
  - 研究样本量的估计量更小。
  - 点估计更准确。如果我们需要根据一个小样本数据来估计学生的平均身高。那么使用正态分布来拟合，很容易就受到离群异常值的影响而得到错误的估计。
  - 回归中应用$t$分布，可以得到更稳健的估计量（β值），这也是我们实现“稳健回归”的一个重要手段。

# 4.2、假设检验

- 假设检验问题(Hypothesis Testing)：

在自然科学和社会科学等中，常常要对某些重要问题做出回答：是或否

- 如月球比地球早形成吗？
- 一种新药对某种病有效吗？
- 某种股票会涨吗？

为了回答这些问题，我们需要对感兴趣的问题进行试验或观察获得相关数据，根据这些数据决定是或否的过程称为**假设检验**

# 4.2、假设检验

- **假设检验**
  - 事先对总体参数或分布形式作出某种假设，然后利用样本信息来判断原假设是否成立
  - 采用逻辑上的反证法，依据统计上的小概率原理
- **原假设**
  - 待检验的假设，记为$H_0$
  - 研究者想收集证据予以反对的假设
- **备择假设**
  - 与原假设对立的假设，记为$H_1$
  - 研究者想收集证据予以支持的假设

# 4.2、假设检验

| | 真实情况 | |
|---|---|---|
| 决策 | $H_0$ 为真 | $H_0$ 为假 |
| 不能拒绝原假设 | 正确 | 第II类错误 |
| 拒绝原假设 | 第I类错误 | 正确 |

# 4.2、假设检验

假设检验的基本思想就是运用小概率事件原理的反证法，其表现如下：

1. 小概率事件在现实中是不可能发生的。

小概率是在一次实验中，一个几乎不可能发生事件的发生概率。在我们设定的原假设“假设方便面净含量为100g”前提下，如果在一次观察中小概率事件发生了，则认为原假设是不成立的；反之，如果小概率事件没有出现，我们没有理由否定原假设。

# 4.2、假设检验

2. 采用反证法。

要检验某个原假设是否成立，先假定它是正确的，例如，假定方便面净含量是100g，然后通过抽样，根据样本计算出的统计量信息判断由假设而得到的结果是否合理，从而确定对原假设的拒绝与否。

# 4.2、假设检验

**t检验**

由于个体差异的存在，在临床医学实践中充满了各种不确定性


**卡方检验**


医学统计，就是用数字去直观呈现这些不确定实践发生的概率

# 4.2、假设检验

**t检验**

由一名德国啤酒厂职员于1908年发明，他因没有大学教授身份而坚持自谦为“Student:”

发展史

比较两组正态分布数据均值的差异

用途

**卡方检验**

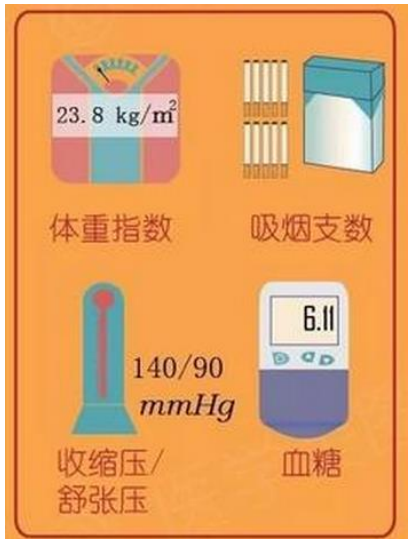
由英国数学家Karl Pearson于1900年提出理论基础，后经过多位统计学家的改良和发展

比较多组分类数据构成比的差异

# 4.2、假设检验

**t检验**

连续计量资料使用t检验




- 符合正态分布
- 不超过两组

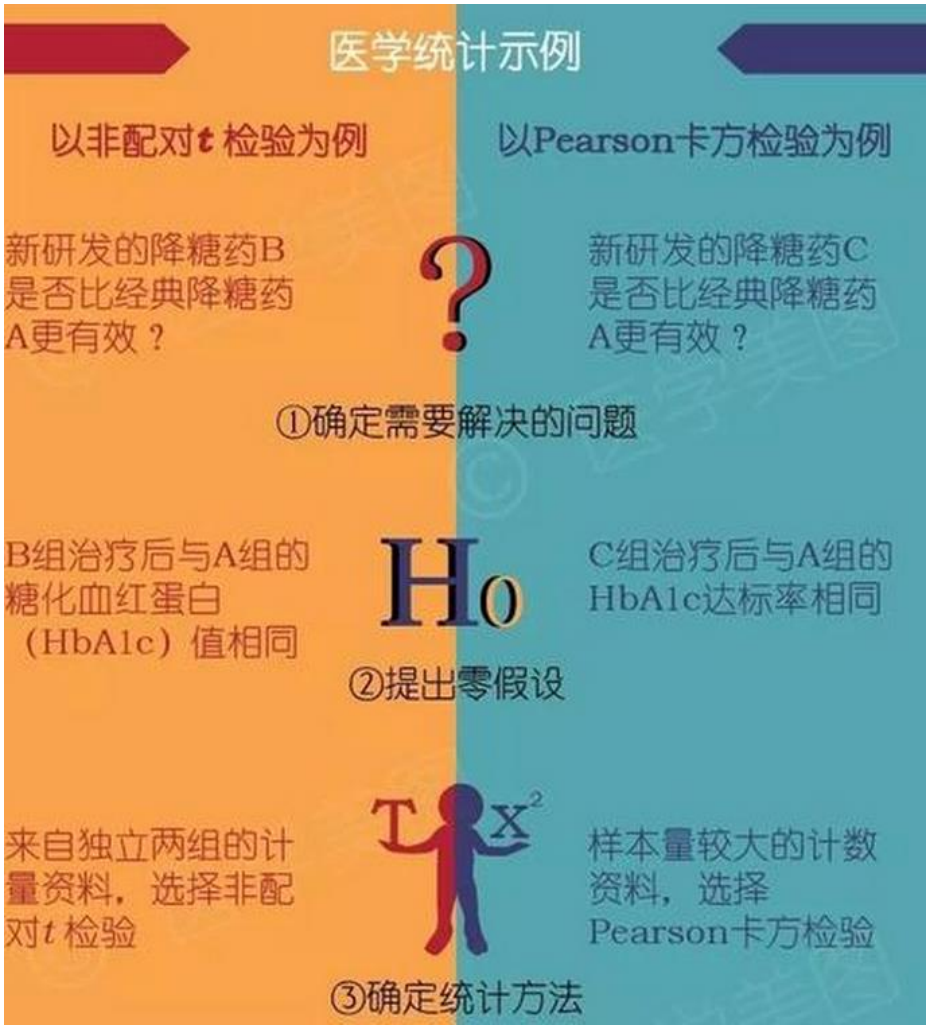
**卡方检验**

分类计数资料使用卡方检验


- 试验例数>10
- 可以有多组

# 4.2、假设检验

**t检验**



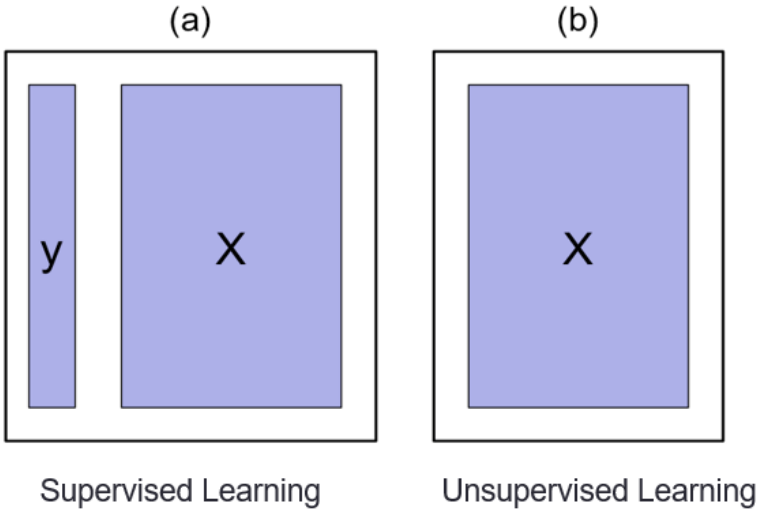

**卡方检验**

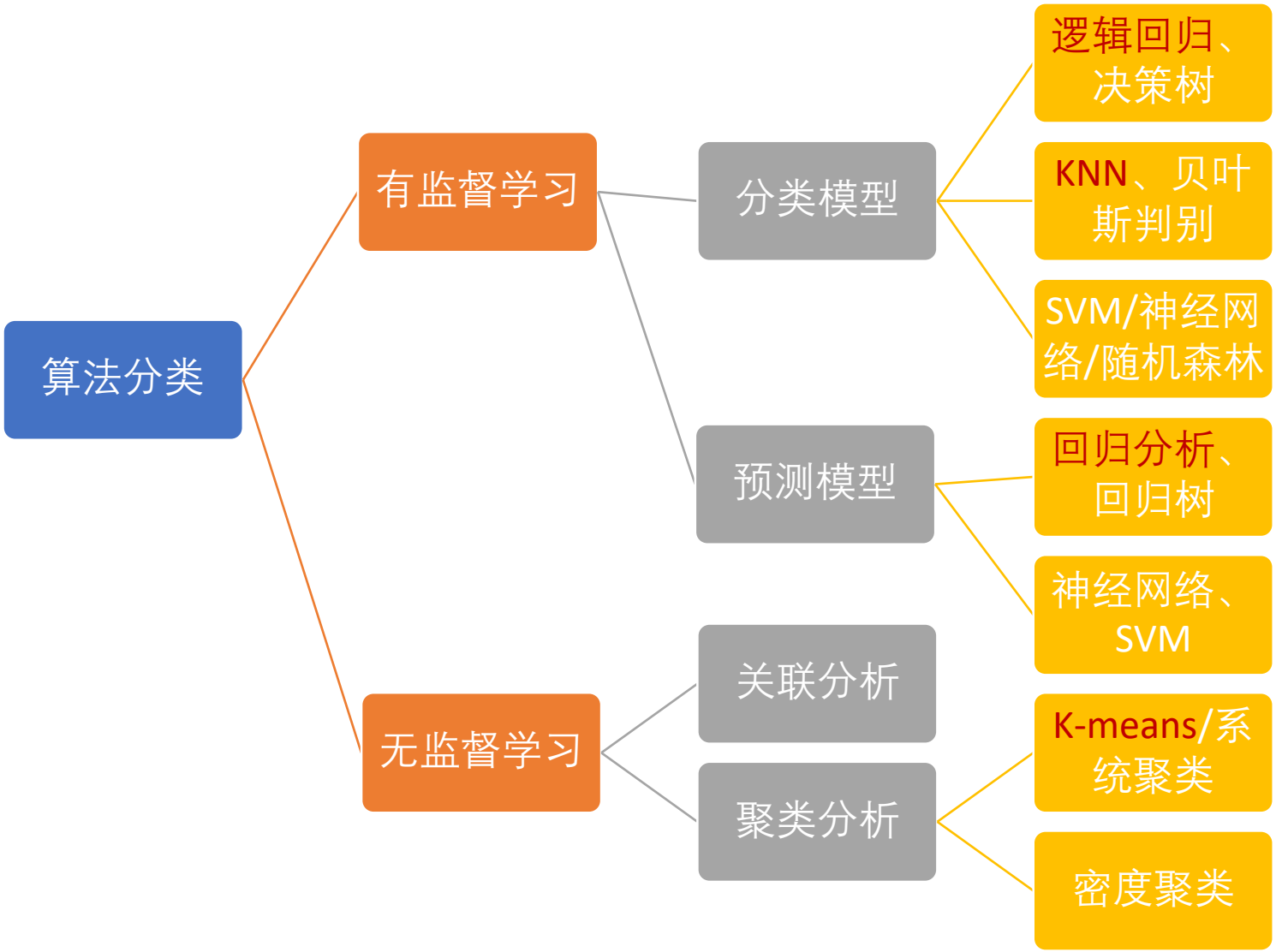
# 4.2、假设检验

**t检验**

**卡方检验**

- 如果我们得到一组样本，对我们感兴趣的问题应该怎样处理呢？
- 下面介绍几种处理数据的方法
  - 有响应变量$y$（有监督学习）
  - 没有响应变量$y$（无监督学习）




算法分类
有监督学习
无监督学习
分类模型
预测模型
关联分析
聚类分析
逻辑回归、决策树
KNN、贝叶斯判别
SVM/神经网络/随机森林
回归分析、回归树
神经网络、SVM
K-means/系统聚类
密度聚类

# 4.3、 预测分析

- 回归分析
  - 相关关系：一个变量的取值并不能由另一个变量唯一确定，当变量$x$的取某个值时，变量$y$的取值可能有几个。
  - 例子：某产品的销量（$y$）与广告投入（$x$）之间的关系
  - 思考：还有什么类似的例子？

# 4.3、 预测分析

- 回归分析

通过之前介绍的相关分析，可以知道变量之间有没有相关关系以及相关关系是否紧密，但并不知道变量之间究竟是什么样的一种关系。

在实际运用中，往往需要知道具有相关关系的变量之间的数量关系，这种数量关系是可以用函数表达式来表述的。

回归分析就是对具有相关关系的变量，用函数表达式来表述各个变量之间相关关系的研究过程。

# 4.3、 预测分析

## ➢ 一元线性回归模型

在人们的日常生活中，经常会有这样的现象：虽然某种事物的变化是众多因素作用是结果，但其中却有一个主要的因素，它往往是我们研究的首要对象。比如居民可支配收入的增加是其消费支出增加的首要因素；广告费用的增加是影响销售额增加的主要因素等等。

# 4.3.1 有监督学习-回归分析

- 回归方程和回归名称
  - 回归一词的英文是regression，其基本的思想和方法是由英国著名生物学家、统计学家F·高尔顿（F.Galton,1822-1911）在研究人类遗传问题时提出来的。
  - 为了研究父代与子代身高的关系，高尔顿和他的学生、现代统计学的奠基人之一K·皮尔逊（K.Pearson,1856-1936）在研究父母身高与其子女身高的遗传问题的时候，观察了1078对夫妇，以每对夫妇的平均身高最为$x$，而取他们的一个成年儿子身高作为y，将结果在平面直角坐标系上绘成散点图，发现趋势近乎一条直线。

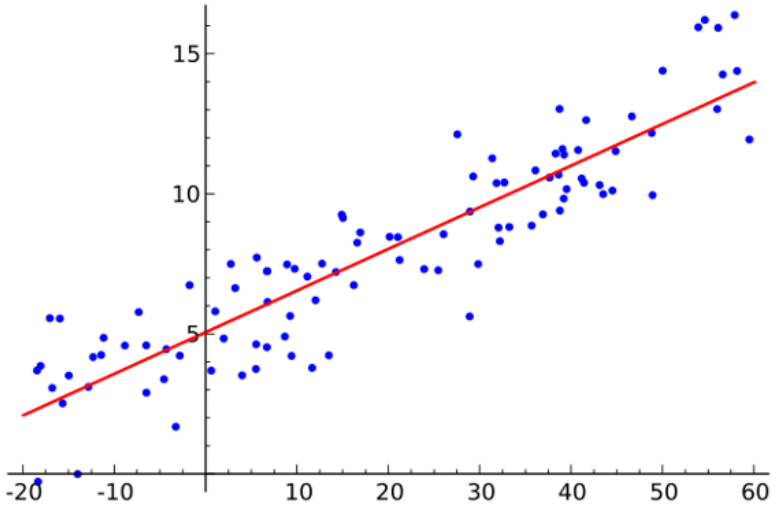
F·高尔顿和K·皮尔逊

# 4.3.1 有监督学习-回归分析

- 回归方程和回归名称
  - 计算出的回归直线方程为

$$成年儿子的身高 = 33.73 + 0.561 \times 父母平均身高$$

这种趋势以及回归方程总的表面父母平均身高$x$每增加一个单位时，其成年儿子的身高y也平均增加0.516个单位。




简单回归

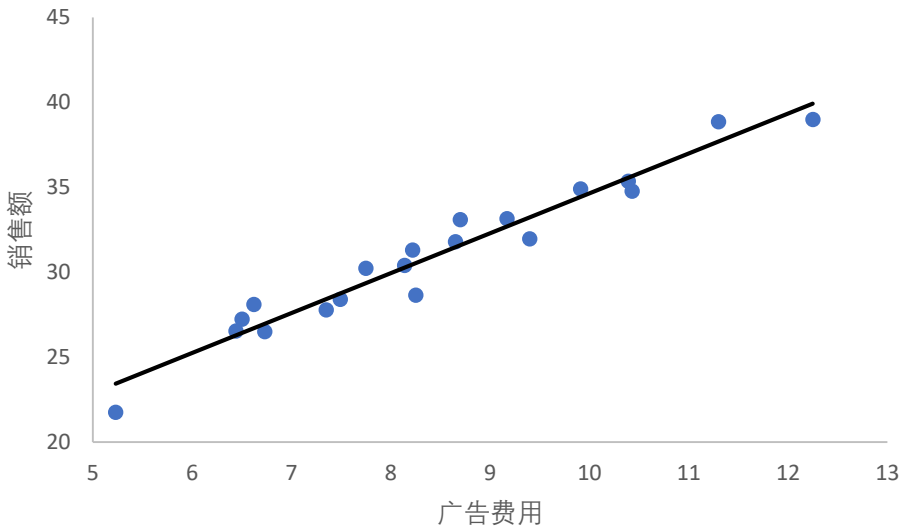
# 4.3.1 有监督学习-回归分析

## ➢ 一元线性回归模型

对具有相关关系的两个变量$x$和$y$，根据样本数据，确定两个变量之间相关关系的函数表达式就是一元回归分析。如果两个变量之间相关关系的是线性的，则称为一元线性回归。

右图展示了产品销售额与广告费用之间的散点图，可以看出样本数据点大致落在一条直线附近，

可以判断产品销售额与广告费用之间存在线性关系。

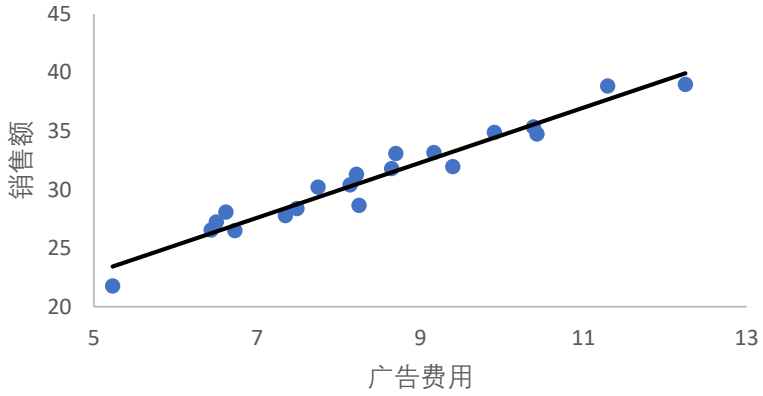


# 4.3.1 有监督学习-回归分析

- 一元线性回归模型

从图中还可以看到的一点是样本数据点并不完全落在直线上，也就是说产品销售额并不完全由广告费用所确定。

实际上，影响产品销售额的因素还有很多，比如人口总量、居民收入水平、居民消费偏好等都会对产品销售额产生影响，样本数据点与直线之间的差异可以看作是其他所有因素影响的结果。



# 4.3.1 有监督学习-回归分析

- 一元线性回归模型

再次把上例中的广告费用作为自变量$x$，产品销售额作为因变量$y$，变量$y$与变量$x$之间基本上是线性关系，将除广告费用外的其他一些因素作为随机因素处理。因此，可以假设变量$x$与变量$y$之间有下列关系：

$$y = \alpha + \beta x + \varepsilon$$

$$(\text{产品销售额} = \alpha + \beta \times \text{广告费用} + \varepsilon)$$

# 4.3.1 有监督学习-回归分析

- 一元线性回归模型

上述函数表达式是产品销售额与广告费用的一元线性模型，它表明了广告费用是决定产品销售额的主要因素，二者之间有密切关系，但密切的程度又没有达到由$x$唯一确定$y$的程度。

其中$\alpha$称为常数项，代表不受广告费用影响的产品销售额，$\beta$表明广告费用每增加一个单位时，产品销售额所增加的数量，而$\varepsilon$则表示影响产品销售额变化的其他因素，称为随机误差。

# 4.3.1 有监督学习-回归分析

- 多元线性回归模型

一元线性回归分析实际上是将实际的社会经济现象之间的关系简化了，因为影响一份经济变量发展的因素一般是多个的，在一元线性回归分析中，将这些因素看作在观察期内保持不变。这当然只是一种假定，并不准确地符合实际情况。

例如，对某种商品的需求$Q$进行研究，它可能与商品价格$P$、居民收入$R$等都有关系，此时，我们需要建立它们之间的多元回归分析。如果它们之间的关系是线性的，则可以表示为

$$Q = \beta_0 + \beta_1 P + \beta_2 R + \varepsilon$$

# 4.3.1 有监督学习-回归分析

- 多元线性回归模型

当影响变量$y$的因素有$k$个时，且$y$与$k$个变量为线性关系，则总体线性回归模型为：

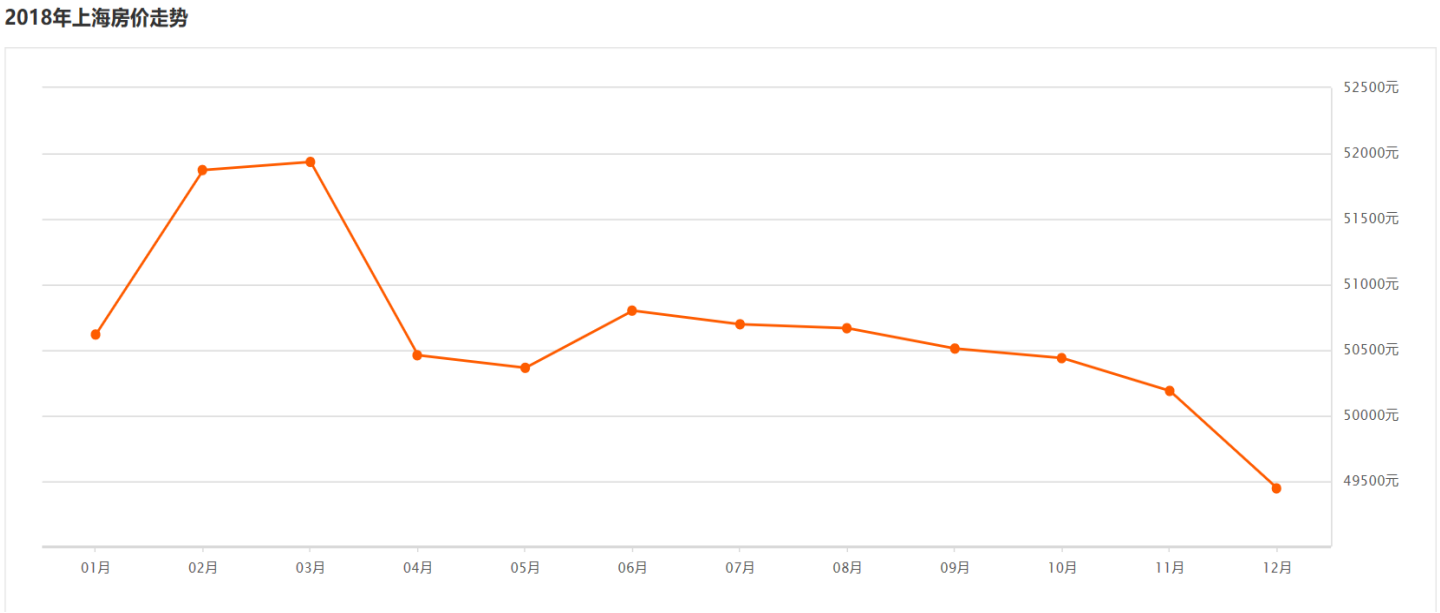
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

$\beta_0$，$\beta_1 \cdots \beta_k$为回归模型的参数，它们决定了因变量$y$与自变量$x$之间的线性关系的具体形式。其中某个回归系数$\beta_i$的含义是当控制其他变量不变时，第$i$个自变量对因变量均值的影响。$\varepsilon$为随机误差项。

# 4.3.1 有监督学习-回归分析

## ➢ 房价影响因素案例

住房价格对调节居民的生活水平有重要的功能和作用。住房价格高居民承受能力低，居住水平和居住质量会由此下降;反之住房价格水平低，能增强居民的购房能力，相应提高居民的居住水平和居住质量。因此，住房价格的高低成为关系到居民切身利益的重大经济问题和社会问题。



# 4.3.1 有监督学习-回归分析

- 房价影响因素案例

思考：那么有哪些因素会影响到房价呢？

# 4.3.1 有监督学习-回归分析

- 房价影响因素案例

我们以波士顿房价的数据为例，分析影响波士顿房价的因素，并预测房价。

波士顿数据集记录了来自506个社区房价的中位数（medv），为了预测社区的房价，同时收集了相关变量：平均房间个数（rm），房屋的房龄（age），社会经济地位低的家庭比例（lstat）等13个相关变量，具体见下表：

（数据文件详见5_1.csv）

# 4.3.1 有监督学习-回归分析

- 房价影响因素案例

| 变量 | 解释 |
| --- | --- |
| crim | 人均犯罪率。 |
| zn | 占地面积超过25,000平方英尺的住宅用地比例。 |
| indus | 每个城镇非零售业务的比例。 |
| chas | Charles River虚拟变量（如果管道限制河流则= 1;否则为0）。 |
| nox | 一氧化氮浓度（每千万份）。 |
| rm | 每栋住宅的平均房间数。 |
| age | 1940年以前建造的自住单位比例。 |
| dis | 到波士顿五个就业中心的加权距离。 |
| rad | 径向高速公路的可达性指数。 |
| tax | 每10,000美元的全额物业税率。 |
| ptratio | 城镇的学生与教师比例。 |
| black | 1000 （Bk - 0.63）* 2其中Bk是城镇黑人的比例。 |
| lstat | 社会经济地位低的家庭比例。 |
| medv | 房价中位数 |

# 4.3.1 有监督学习-回归分析

- 房价影响因素案例

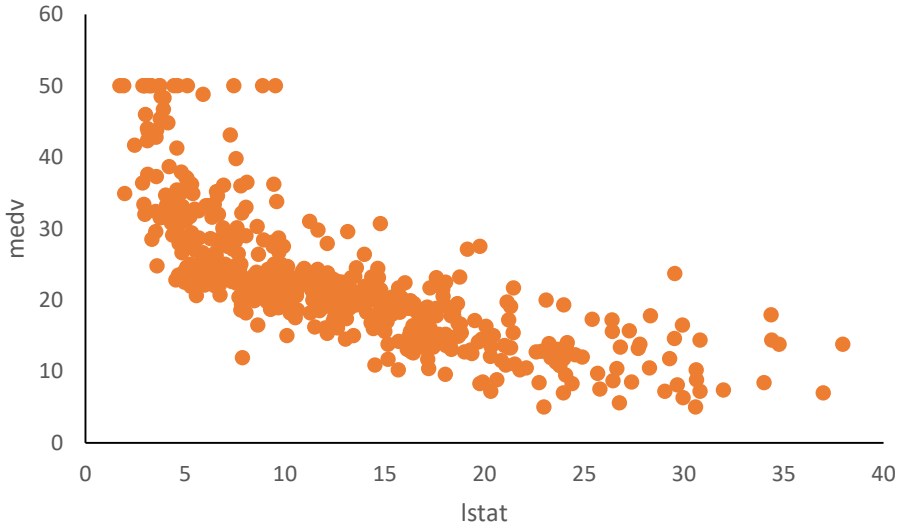
数据部分展示：

| | crim | zn | indus | chas | nox | rm | age | dis | rad | tax | ptratio | black | lstat | medv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.00632 | 18 | 2.31 | 0 | 0.538 | 6.575 | 65.2 | 4.09 | 1 | 296 | 15.3 | 396.9 | 4.98 | 24 |
| 2 | 0.02731 | 0 | 7.07 | 0 | 0.469 | 6.421 | 78.9 | 4.9671 | 2 | 242 | 17.8 | 396.9 | 9.14 | 21.6 |
| 3 | 0.02729 | 0 | 7.07 | 0 | 0.469 | 7.185 | 61.1 | 4.9671 | 2 | 242 | 17.8 | 392.83 | 4.03 | 34.7 |
| 4 | 0.03237 | 0 | 2.18 | 0 | 0.458 | 6.998 | 45.8 | 6.0622 | 3 | 222 | 18.7 | 394.63 | 2.94 | 33.4 |
| 5 | 0.06905 | 0 | 2.18 | 0 | 0.458 | 7.147 | 54.2 | 6.0622 | 3 | 222 | 18.7 | 396.9 | 5.33 | 36.2 |

- 我们首先考虑**单变量**的回归方程

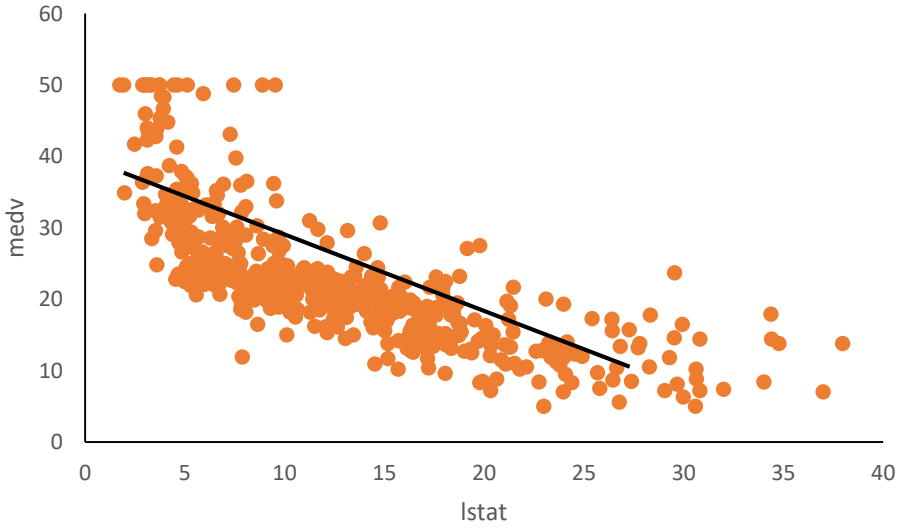
# 4.3.1 有监督学习-回归分析

- 波士顿房价预测
  - 首先我们先探索社区房价的中位数（medv）与社会经济地位低的家庭比例（lstat）之间的关系
  - 画出两个变量的散点图



# 4.3.1 有监督学习-回归分析

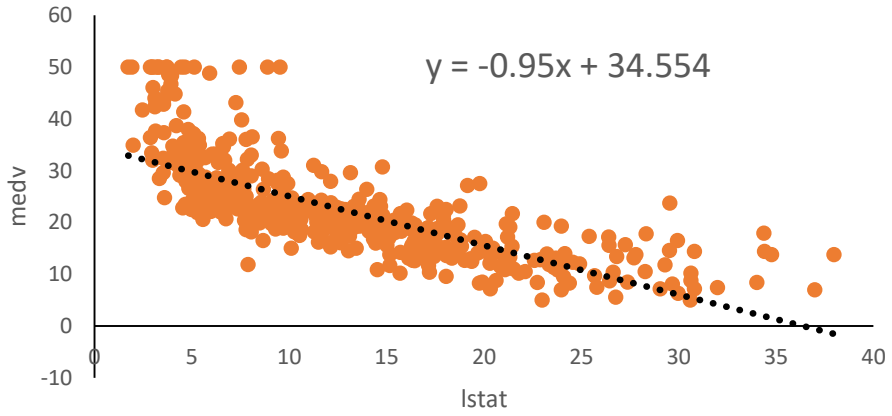
- 波士顿房价预测
  - 从散点图看到，社区房价的中位数（medv）与社会经济地位低的家庭比例（lstat）之间是大致存在线性反比的关系的，随着社会经济地位低的家庭比例（lstat）的增加，社区房价的中位数（medv）是趋向降低的。



# 4.3.1 有监督学习-回归分析

- 波士顿房价预测
  - 画出两个变量的趋势线，得到的线性方程为：

$$房价中位数 = 34.554 - 0.95 \times 经济地位低的家庭比例$$



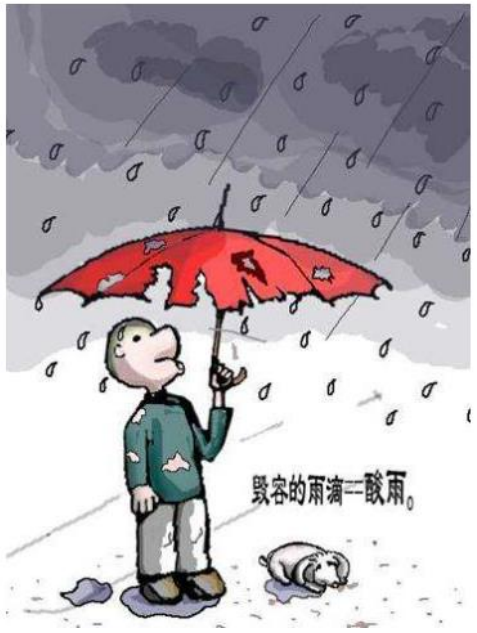

从上图可以看到房价的中位数与社会经济地位低的家庭比例是成反比的，社会经济地位低的家庭比例每增加10%，房价的中位数就减少9.5

- 问题：那么如果家庭比例是的值为30，得到房价的估计是多少呢？

# 4.3.1 有监督学习-回归分析

- 全国废气中主要污染物排放量与出现酸雨城市比例关系研究

当烧煤的烟囱排放出的二氧化硫酸性气体，或汽车排放出来的氮氧化物烟气上升到天上，这些酸性气体与天上的水蒸气相遇，就会形成硫酸和硝酸小滴，使雨水酸化，这时落到地面的雨水就成了酸雨。煤和石油的燃烧是造成酸雨的主要祸首。



# 4.3.1 有监督学习-回归分析

- 全国废气中主要污染物排放量与出现酸雨城市比例关系研究

  酸雨会对环境带来广泛的危害，造成巨大的经济损失：

  1）腐蚀建筑物和工业设备；

  2）破坏露天的文物古迹；

  3）损坏植物叶面，导致森林死亡；

  4）使湖泊中鱼虾死亡；

  5）破坏土壤成分，使农作物减产甚至死亡；

  6）饮用酸化造成的地下水，对人体有害。

# 4.3.1 有监督学习-回归分析

- 全国废气中主要污染物排放量与出现酸雨城市比例关系研究

  近几年来环境问题成为全社会极为关注的热点,空气污染是其中最热门的话题,同时也是也是最重要的民生问题。针对此情况，我们想要研究酸雨的形成与二氧化硫排放量，烟尘排放量，工业粉尘排放量之间的关系。

  酸雨数据集收集了2000年到2010年间二氧化硫排放量（万吨），烟尘排放量（万吨），工业粉尘排放量（万吨）和出现酸雨城市比例的数据。

# 4.3.1 有监督学习-回归分析

- 全国废气中主要污染物排放量与出现酸雨城市比例关系研究

  数据文件参见5_2.xls

| 全国废气中主要污染物排放量年际变化 | | | | |
|---|---|---|---|---|
| 年度 | 二氧化硫排放量（万吨） | 烟尘排放量（万吨） | 工业粉尘排放量（万吨） | 出现酸雨城市比例（%） |
| 2000 | 1995.1 | 1165.4 | 1092 | 49.8 |
| 2001 | 1947.2 | 1069.9 | 990.6 | 48 |
| 2002 | 1926.6 | 1012.7 | 941 | 48.7 |
| 2003 | 2158.5 | 1048.5 | 1021.3 | 54.4 |
| 2004 | 2254.9 | 1095 | 904.8 | 56.5 |
| 2005 | 2549.3 | 1182.5 | 911.2 | 56 |
| 2006 | 2588.8 | 1088.8 | 808.4 | 54 |
| 2007 | 2468.1 | 986.6 | 698.7 | 54.2 |
| 2008 | 2321.2 | 901.6 | 584.9 | 52.8 |
| 2009 | 2214.4 | 847.2 | 523.6 | 52.9 |
| 2010 | 2185.1 | 829.1 | 448.7 | 50.4 |

# 4.3.1 有监督学习-回归分析

- 回归分析结果

| 回归统计 | |
|---|---|
| Multiple R | 0.815626655 |
| R Square | 0.66524684 |
| Adjusted R Square | 0.5217812 |
| 标准误差 | 1.999912035 |
| 观测值 | 11 |

方差分析

| | df | SS | MS | F | gnificance F |
|---|---|---|---|---|---|
| 回归分析 | 3 | 55.63882659 | 18.54627553 | 4.636977 | 0.04342396 |
| 残差 | 7 | 27.99753704 | 3.999648149 | | |
| 总计 | 10 | 83.63636364 | | | |

| | Coefficients | 标准误差 | t Stat | P-value | Lower 95% | Upper 95% | 下限 95.0% | 上限 95.0% |
|---|---|---|---|---|---|---|---|---|
| Intercept | 27.88666706 | 7.893389165 | 3.53291425 | 0.00956 | 9.22176761 | 46.5515665 | 9.221768 | 46.55157 |
| X Variable 1 | 0.013634662 | 0.005197864 | 2.623127707 | 0.034253 | 0.00134367 | 0.02592566 | 0.001344 | 0.025926 |
| X Variable 2 | -0.014029171 | 0.019840597 | -0.707094202 | 0.502361 | -0.0609447 | 0.03288639 | -0.06094 | 0.032886 |
| X Variable 3 | 0.010410776 | 0.011555895 | 0.900906134 | 0.397573 | -0.0169146 | 0.03773612 | -0.01691 | 0.037736 |

从模型的$R$方来看，回归方程对观测数据的拟合度为0.67，说明回归方程的拟合效果一般，$F$统计量的$p$值低于0.05的水平说明模型是显著的。

# 4.3.1 有监督学习-回归分析

| 回归统计 | |
|---|---|
| Multiple R | 0.815626655 |
| R Square | 0.66524684 |
| Adjusted R Square | 0.5217812 |
| 标准误差 | 1.999912035 |
| 观测值 | 11 |

方差分析

| | df | SS | MS | F | gnificance F |
|---|---|---|---|---|---|
| 回归分析 | 3 | 55.63882659 | 18.54627553 | 4.636977 | 0.04342396 |
| 残差 | 7 | 27.99753704 | 3.999648149 | | |
| 总计 | 10 | 83.63636364 | | | |

| | Coefficients | 标准误差 | t Stat | P-value | Lower 95% | Upper 95% | 下限 95.0% | 上限 95.0% |
|---|---|---|---|---|---|---|---|---|
| Intercept | 27.88666706 | 7.893389165 | 3.53291425 | 0.00956 | 9.22176761 | 46.5515665 | 9.221768 | 46.55157 |
| X Variable 1 | 0.013634662 | 0.005197864 | 2.623127707 | 0.034253 | 0.00134367 | 0.02592566 | 0.001344 | 0.025926 |
| X Variable 2 | -0.014029171 | 0.019840597 | -0.707094202 | 0.502361 | -0.0609447 | 0.03288639 | -0.06094 | 0.032886 |
| X Variable 3 | 0.010410776 | 0.011555895 | 0.900906134 | 0.397573 | -0.0169146 | 0.03773612 | -0.01691 | 0.037736 |

出现酸雨城市比例 = 28.89 + 0.013二氧化硫排放量

−0.014烟尘排放量 + 0.01工业粉尘排放量

# 4.3.1 有监督学习-回归分析

| 回归统计 | |
|---|---|
| Multiple R | 0.815626655 |
| R Square | 0.66524684 |
| Adjusted R Square | 0.5217812 |
| 标准误差 | 1.999912035 |
| 观测值 | 11 |

方差分析

| | df | SS | MS | F | gnificance F |
|---|---|---|---|---|---|
| 回归分析 | 3 | 55.63882659 | 18.54627553 | 4.636977 | 0.04342396 |
| 残差 | 7 | 27.99753704 | 3.999648149 | | |
| 总计 | 10 | 83.63636364 | | | |

| | Coefficients | 标准误差 | t Sta | P-value | Lower 95% | Upper 95% | 下限 95.0% | 上限 95.0% |
|---|---|---|---|---|---|---|---|---|
| Intercept | 27.88666706 | 7.893389165 | 3.53291425 | 0.00956 | 9.22176761 | 46.5515665 | 9.221768 | 46.55157 |
| X Variable 1 | 0.013634662 | 0.005197864 | .623127707 | 0.034253 | 0.00134367 | 0.02592566 | 0.001344 | 0.025926 |
| X Variable 2 | -0.014029171 | 0.019840597 | -0.707094202 | 0.502361 | -0.0609447 | 0.03288639 | -0.06094 | 0.032886 |
| X Variable 3 | 0.010410776 | 0.011555895 | 0.900906134 | 0.397573 | -0.0169146 | 0.03773612 | -0.01691 | 0.037736 |

从各个变量的$p$值来看，如果选定0.05的显著性水平，二氧化硫排放量对酸雨城市比例是显著的，而烟尘排放量和工业粉尘排放量和出现酸雨城市比例之间的关系关系并不显著。

# 4.3.1 有监督学习-回归分析

- 下面是几个关于中国酸雨的常规结论(张新民等，2010)

1. 中国酸雨区是继欧洲和北美之后的世界三大酸雨区之一

2. 中国的酸雨主要由二氧化硫形成，但正在向硫酸-硝酸混合型转变，即氮氧化物的权重越来越大

3. 自上世纪80年代中国有相关记录开始，中国传统南方酸雨区并无太大变化，但有向北方拓展的趋势

- 从上面酸雨案例的分析，我们也佐证了酸雨是与二氧化硫有着重要关系的。

（张新民，柴发合，王淑兰，孙新章，and 韩梅：中国酸雨研究现状，环境科学研究，2010. 527-532，2010.）

# 4.3.2 有监督学习-逻辑回归

- 回归就相当于$y = f(x)$，表明自变量$x$与因变量$y$的关系。因变量$y$是一个连续变量，如果因变量$y$是一个分类变量呢？
- 如上所述，如果得到的响应变量$y$是离散的，例如只有“0”，“1”两类，应该如何处理？
- 例如，线性回归中利用年龄和教育程度对一个人的收入进行建模，假如我们并不关心一个人的实际收入，而是关心这个人的贫富状况，这种情况下，我们要建模的 $y$ 可以是“贫”和“富”这两个取值，可以用“0”，“1”替代。
- 处理上述问题的方法：**逻辑回归**（ Logistic Regression）

# 4.3.2 有监督学习-逻辑回归

- 逻辑回归适用场景：

- 医疗：判断特定手术或者治疗对病人的有效性。输入的变量可能是年龄、体重、血压、胆固醇水平等
- 金融：判断信贷是否违约；信用卡评级；风险控制；反欺诈等
- 营销：精准营销；判读客户是否流失等

需要说明的是通过逻辑回归的方法得到的不是一个确定的结果，而是事件的预测概率。举例来说，通过建立逻辑回归模型判断信贷是否违约，得到的只是一个预测的违约概率，是否将其判断为违约取决于设定的阈值。如果预测违约的概率为0.7，选择的阈值为0.5，那么可以判定这个客户是违约的。

# 4.3.2 有监督学习-逻辑回归

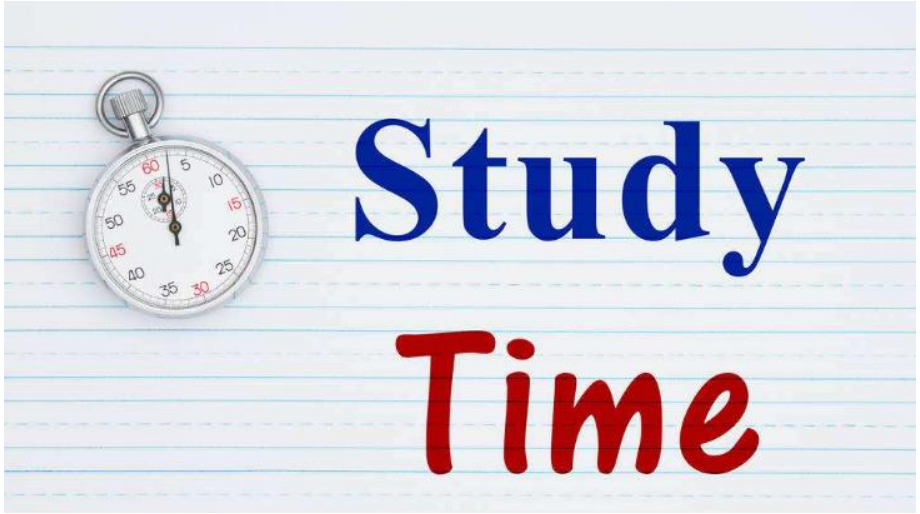
- 逻辑回归的因变量可以是**是否违约**这种二分类的，也可以是多分类的，但是二分类的更为常用，也更加容易解释。所以实际中最常用的就是二分类的逻辑回归。
- 一元线性回归形式：$y = \beta_0 + \beta_1 x + \varepsilon$
- 逻辑回归（非线性回归）：$y = \frac{\beta_0}{1+\exp(\beta_1+\beta_2 x)} + \varepsilon$

# 4.3.2 有监督学习-逻辑回归

天才就是百分之九十九的汗水加百分之一的灵感

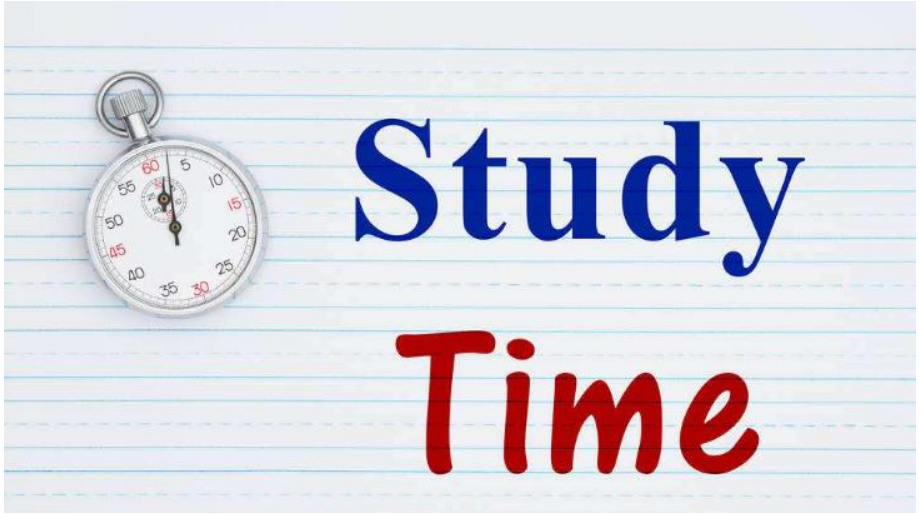
——爱迪生

- 那么这句话反映到学习上去，学习时间是否真的是会影响到考试成绩上去呢？



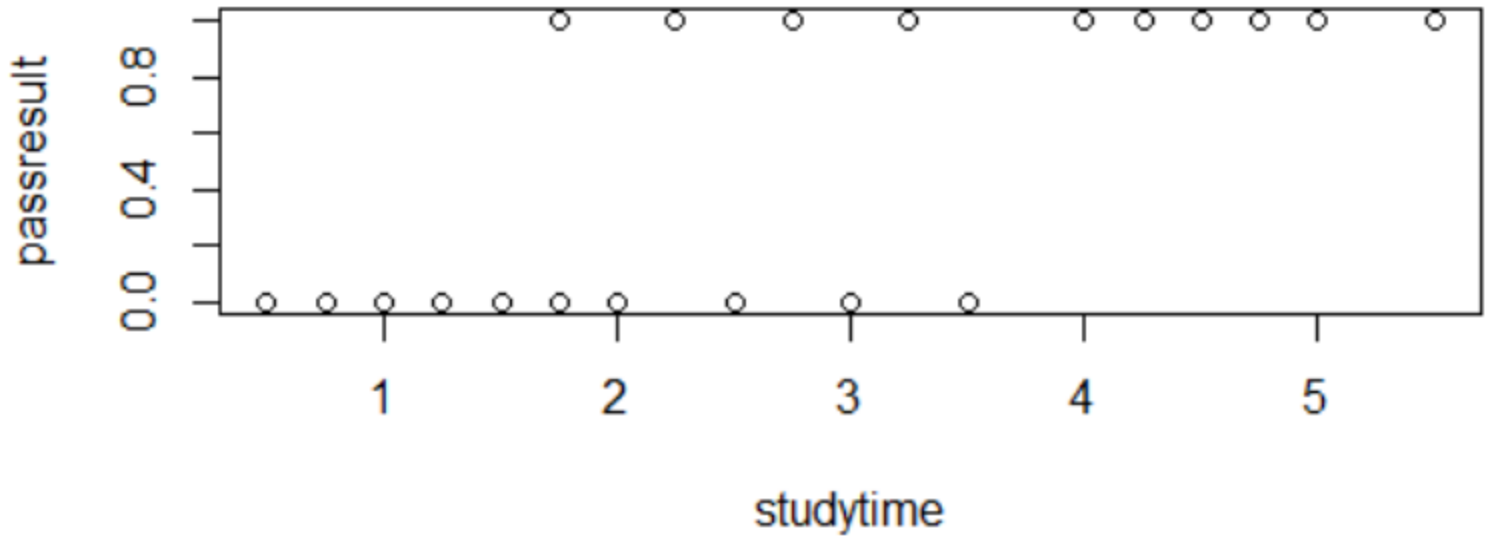
# 4.3.2 有监督学习-逻辑回归

- 我们收集了某校大一20个学生学习时间（studytime）与其是否通过某次考试（passresult）的数据。
- 我们感兴趣的问题是学生的学习时间是否与通过某次考试有关。
- 如果是有关系的，我们是否可以通过一位学生学习时间来预测他能否通过考试呢？



# 4.3.2 有监督学习-逻辑回归

- 画出两者的散点图

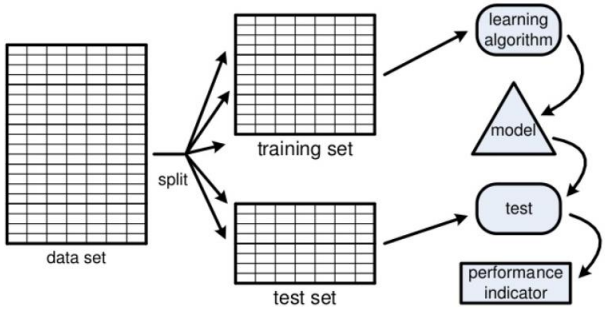



- Q：从上图我们能得到什么结论？

# 4.3.2 有监督学习-逻辑回归

- 这里我们用学习时间，来预测学生通过考试的概率有多大。
- 同时，我们也会用测试数据对模型进行检验，判断模型的预测效果

# 4.3.2 有监督学习-逻辑回归




- **训练集（train set）**
  - ✓ 作用：估计模型
  - ✓ 学习样本数据集，通过匹配一些参数来建立一个分类器。建立一种分类的方式，主要是用来训练模型的。
- **检验集（test set）**
  - ✓ 作用：检验最终选择最优的模型的性能如何
  - ✓ 主要是测试训练好的模型的分辨能力（识别率等）

# 4.3.2 有监督学习-逻辑回归

- 为什么要进行划分呢？

简而言之，为了防止**过度拟合**。如果我们把所有数据都用来训练模型的话，建立的模型自然是最契合这些数据的，测试表现也好。但换了其它数据集测试这个模型效果可能就没那么好了。

比如你给班上同学做校服，大家穿着都合适你就觉得按这样做就对了，那给别的班同学穿呢？不合适的概率会变高。总而言之训练集和测试集相同的话，模型评估结果可能比实际要好。

# 4.3.2 有监督学习-逻辑回归

- 建立逻辑回归模型

从验证数据结果如右表，准确率为

(1+4)/6=83.33%

| | predict | |
|---|---|---|
| real | 0 | 1 |
| 0 | 1 | 0 |
| 1 | 1 | 4 |

# 4.3.2 有监督学习-逻辑回归

- 标准普尔500指数2001年至2005年的每日收益率百分比数据（Smarket），数据集当中包括的变量有今天回报率（today），前一天到前五天的回报率（lag1-lag5），股票交易量（volume），某一天市场有正回报还是负回报（Direction），其中Direction包含两个分类：up/down
- 分析各个因素对市场回报方向的影响，并根据各个因素对市场回报方向进行预测。

# 4.3.2 有监督学习-逻辑回归

- 比较

| | Direction.2005 | |
|---|---|---|
| glm.pred | Down | Up |
| Down | 35 | 35 |
| Up | 76 | 106 |

从上面结果可以看到，预测的准确率为（35+106）/252=55.95%,

说明预测的效果并不理想。大家可以尝试使用Smarket数据集中的其他变量进行逻辑回归并展示效果。

# 4.3.2 有监督学习-逻辑回归

## ➢ 汽车贷款

- 20世纪90年代末，国内的一些银行开启了汽车信贷业务，新世纪初，经过初步尝试汽车信贷后，国内许多银行预感汽车信贷的获利机遇，纷纷开出了汽车信贷业务。在2002年，汽车信贷业务得到了前所未有的大发展，迎来了汽车信贷的春天。

# 4.3.2 有监督学习-逻辑回归

## ➢ 汽车贷款

- 汽车信贷这块在当时十分诱人的“蛋糕”，不仅遇到银行间的瓜分，且遭到保险公司的加入。随着购车人群的高峰迭起，保险公司亦涌入车贷市场，它们推出购车贷款的保证保险业务，在为银行车贷提供担保的同时，也为自己开辟了获利新径。
- 汽车信贷领域的另一位主角就是汽车经销商。他们借助银行、保险公司车贷业务的蓬勃兴起，推波助澜，大大地获取了客户购车的高额利润。

# 4.3.2 有监督学习-逻辑回归

## ➢ 汽车贷款风险

- 汽车信贷业务的超常规发展确实为企业带来了利润。在分享车贷“蛋糕”喜悦的同时，不断攀升的车贷违约率向银行业敲响了警钟。
- 对银行来说，车贷和房贷虽同属于个人贷款业务，但两者所发生风险的成因却大相径庭，首先，汽车相对房屋来说，具有固定与流动之分。住房有其固定的位置，而汽车则不然，它具有很大的流动性。汽车可以移动，可以在全中国移动。如果客户恶意逃债，他可以移动车辆逃走或隐藏；

# 4.3.2 有监督学习-逻辑回归

- 汽车贷款风险

- 其二，就当前市场而言，房产是升值的，而汽车则是降价的。它的信贷风险比房贷要大。其三，车贷相对于房贷而言，贷款期限短，还款数额多，贷款收回快，具有短平快的特点。因此车贷风险要比房贷来得快而高。
- 据有关车贷的统计，近年车贷的违约率大大升高，已达到0.5%至0.9%，有的已超过1%。更令人头疼的是，违约率还在不断升高。

# 4.3.2 有监督学习-逻辑回归

## ➢ 汽车违约贷款案例

- 某企业从事个人企业金融服务，向购车的个人提供信用贷款。该公司的风控部门根据贷款申请者的基本属性、信贷历史、历史信用情况、贷款标的物等信息构建贷款违约预测模型。
- 案例数据来自accepts.csv数据集，数据集中变量说明如下：
- 下面通过逻辑回归对数据集提供的汽车违约数据进行建模，预测违约与否。

| 名称 | 中文含义 |
| --- | --- |
| application_id | 申请者ID |
| account_number | 帐户号 |
| bad_ind | 是否违约 |
| vehicle_year | 汽车购买时间 |
| vehicle_make | 汽车制造商 |
| bankruptcy_ind | 曾经破产标识 |
| tot_derog | 五年内信用不良事件数量(比如手机欠费消号) |
| tot_tr | 全部帐户数量 |
| age_oldest_tr | 最久账号存续时间(月) |
| tot_open_tr | 在使用帐户数量 |
| tot_rev_tr | 在使用可循环贷款帐户数量(比如信用卡) |
| tot_rev_debt | 在使用可循环贷款帐户余额(比如信用卡欠款) |
| tot_rev_line | 可循环贷款帐户限额(信用卡授权额度) |
| rev_util | 可循环贷款帐户使用比例(余额/限额) |
| fico_score | FICO打分 |
| purch_price | 汽车购买金额(元) |
| msrp | 建议售价 |
| down_pyt | 分期付款的首次交款 |
| loan_term | 贷款期限(月) |
| loan_amt | 贷款金额 |
| ltv | 贷款金额/建议售价*100 |
| tot_income | 月均收入(元) |
| veh_mileage | 行使历程(Mile) |
| used_ind | 是否使用 |
| weight | 样本权重 |

数据集变量的简要说明

## 汽车违约贷款案例

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  9.3887384  0.6556662   14.32   <2e-16 ***
fico_score  -0.0159098  0.0009774  -16.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

从参数估计结果可以看到fico_score的系数为−0.0159，并且$p$值是显著的。回归方程可以写成下列形式（$P$是违约概率）：

$$\ln\left(\frac{P}{1-P}\right) = -0.0159 \times \text{"fico\_score"} + 9.389$$

## 汽车违约贷款案例

| real | predict 0 | 1 |
|---|---|---|
| 0 | 1152 | 26 |
| 1 | 250 | 23 |

得到的预测准确率为（1152+23）/1451=80.98%

这里选择的阈值为0.5，也就是说如果预测违约概率大于0.5那么可以认为是违约的，如果选择其他的值呢？

## 汽车违约贷款案例

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.864e+00  8.350e-01   7.023 2.17e-12 ***
fico_score  -1.531e-02  1.206e-03 -12.692  < 2e-16 ***
tot_derog    2.475e-02  1.630e-02   1.518    0.129
rev_util     5.589e-04  5.424e-04   1.030    0.303
ltv          3.410e-02  3.581e-03   9.524  < 2e-16 ***
veh_mileage  1.540e-06  1.438e-06   1.071    0.284
age_oldest_tr -3.512e-03  6.379e-04  -5.506 3.68e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

使用了多个变量的情况，从参数估计结果可以看到fico_score, ltv和age_oldest_tr的系数$p$值是显著的。

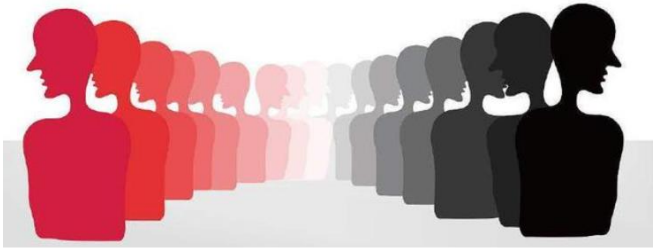
## 汽车违约贷款案例

| real | predict 0 | 1 |
|---|---|---|
| 0 | 1144 | 34 |
| 1 | 237 | 36 |

得到的预测准确率为
（1144+36）
/1451=81.23%

# 4.3.3 有监督学习-KNN

- 对于多分类问题，除了可以使用逻辑回归之外，**K-近邻算法**（KNN）也是很好的选择。
- 所谓K最近邻，就是$k$个最近的邻居的意思，说的是每个样本都可以用它最接近的$k$个邻居来代表。
- 比如在动物园当中，一只动物旁边离它最近的5只动物是猴子，那么我们也将这只动物划为猴子。
- KNN是通过测量不同特征值之间的距离进行分类。下面首先介绍一下距离的几种衡量方式。



# 4.3.3 有监督学习-KNN

- 在对样本进行分类时，样本之间的距离是如何度量的呢？
- 下面列举了几种常用的距离

设$x=(x_1,x_2,\ldots,x_p)'$和$y=(y_1,y_2,\cdots y_p)'$为两个样本

① **绝对值距离**：$d(x,y)=\sum_{i=1}^{p}|x_i-y_i|$

② **欧氏距离**：$d(x,y)=\left(\sum_{i=1}^{p}|x_i-y_i|^2\right)^{1/2}=\sqrt{(x-y)'(x-y)}$

这是聚类分析当中最常用的一种距离

# 4.3.3 有监督学习-KNN

③ **马氏距离：** $d(x,y)=\sqrt{(x-y)'S'(x-y)}$（其中S是协方差矩阵）。使用马氏距离的好处是考虑了各变量之间的相关性，并且与各变量的单位无关。但是马氏距离有一个很大的缺陷，聚类过程中的类一直变化着，这就使得样本协方差矩阵难以确定。因此，在聚类分析中，马氏距离一般不是理想距离。

# 4.3.3 有监督学习-KNN

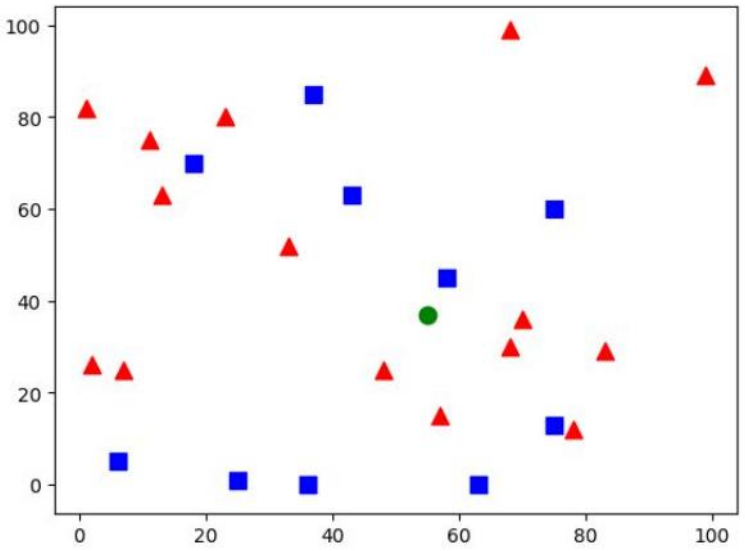
- 根据经验，今天和昨天的湿度差$x_1$及今天的压温差（气压与温度之差）$x_2$是预报明天是否下雨的两个重要因素，现得到两组样本（−1.9，3.2）和（0.2，6.2），试计算这两组样本之间的三种距离。

# 4.3.3 有监督学习-KNN

➢算法思路

如果一个样本在特征空间中的$k$个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别，其中$k$通常是**不大于20**的整数。

思考：绿色的点应该划为三角还是正方形？



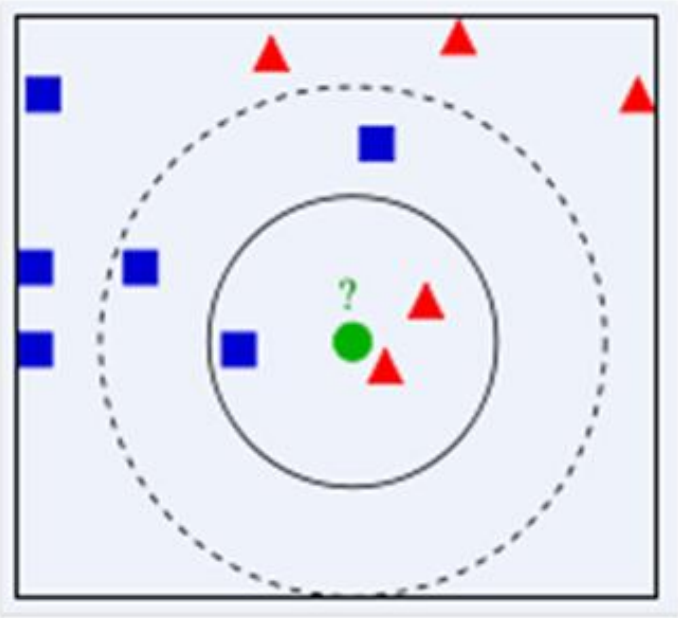
# 4.3.3 有监督学习-KNN

- $k$的选择

- 如果$k$值较小，就相当于用较小邻域中的训练实例进行预测，极端情况下$k=1$，测试实例只和最接近的一个样本有关，训练误差很小(0)，但是如果这个样本恰好是噪声，预测就会出错，测试误差很大。
- 如果$k$值较大，就相当于用很大邻域中的训练实例进行预测，极端情况是$k=n$，测试实例的结果是训练数据集中实例最多的类。

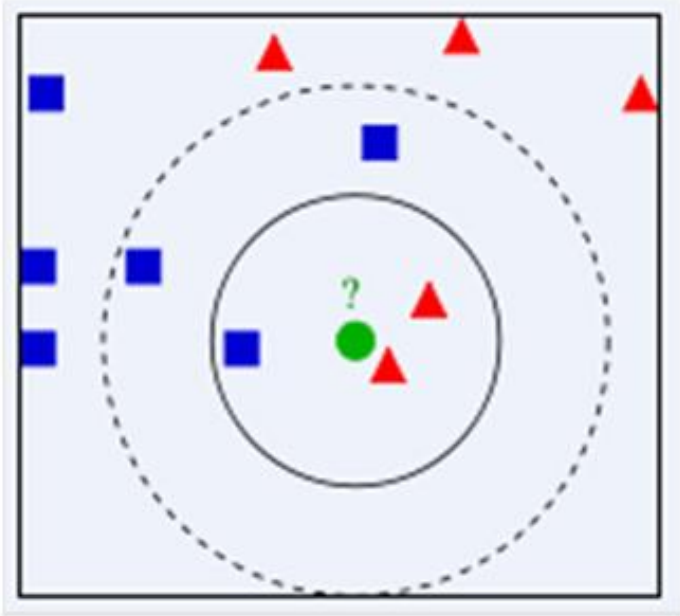
# 4.3.3 有监督学习-KNN

- 下面通过一个简单的例子说明一下：如下图，绿色圆要被决定赋予哪个类，是红色三角形还是蓝色四方形？如果$K=3$，由于红色三角形所占比例为2/3，绿色圆将被赋予红色三角形那个类。



# 4.3.3 有监督学习-KNN

- 如果$K=5$，由于蓝色四方形比例为3/5，因此绿色圆被赋予蓝色四方形类。
- 由此也说明了KNN算法的结果很大程度取决于$K$的选择。



# 4.3.3 有监督学习-KNN

- 算法描述：

① 计算测试数据与各个训练数据之间的距离

② 按照距离的递增关系进行排序

③ 选取距离最小的$K$个点

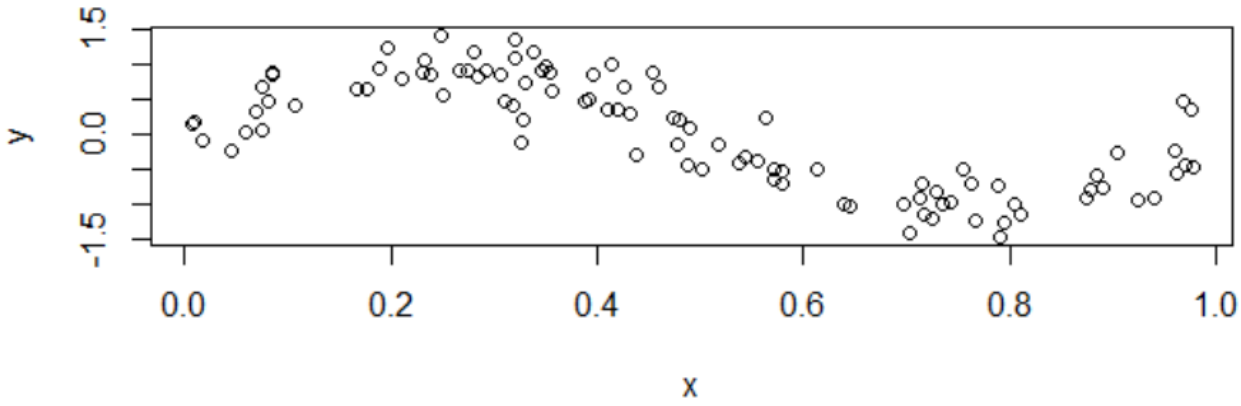
④ 确定前$K$个点所在类别的出现频率

⑤ 返回前$K$个点中出现频率最高的类别作为测试数据的预测分类

# 4.3.3 有监督学习-KNN

- 例子：如果我们假设响应变量是连续的情况，我们首先模拟产生来自三角函数的数据

得到的散点图如下：

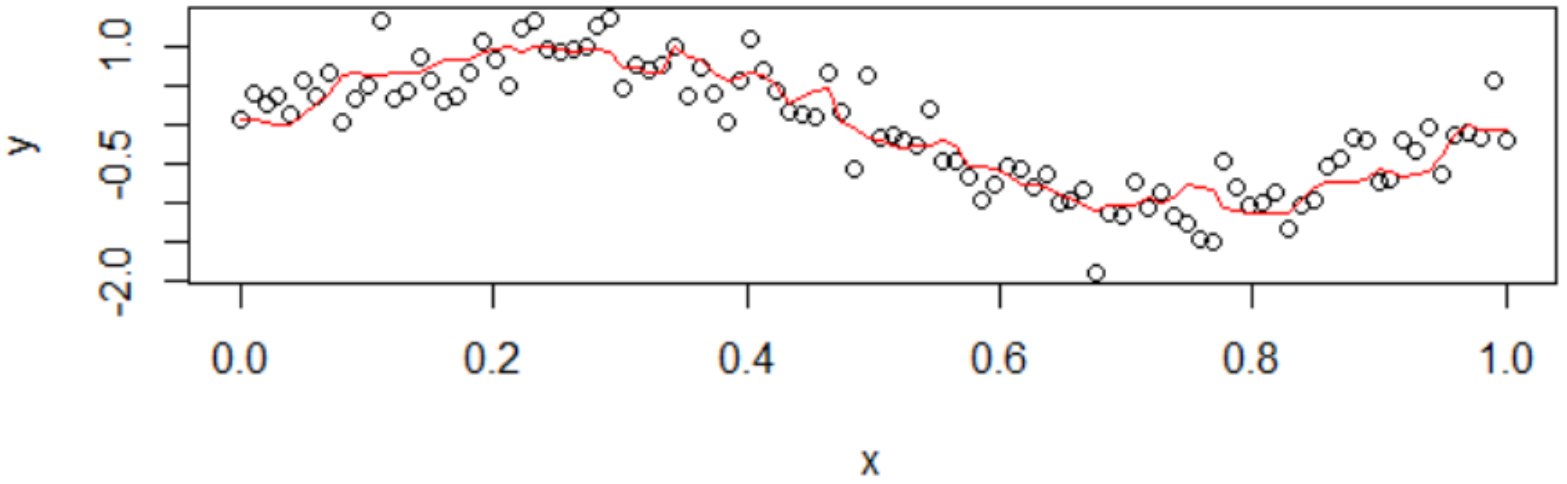



从上图可以看出三角函数的趋势。

# 4.3.3 有监督学习-KNN

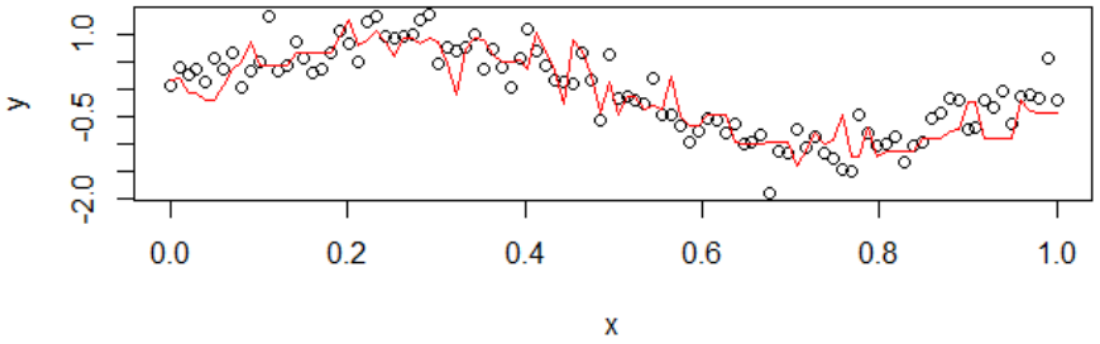
从下图可以看到通过KNN的方法对数据的拟合效果是很好的。

下面我们改变$K$的取值观察拟合的情况

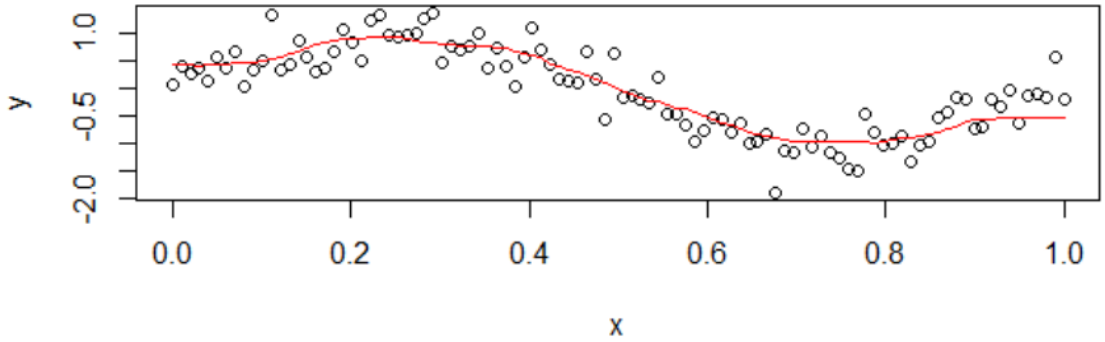


# 4.3.3 有监督学习-KNN

#$K = 1$时，拟合曲线




#$K = 20$时，拟合曲线



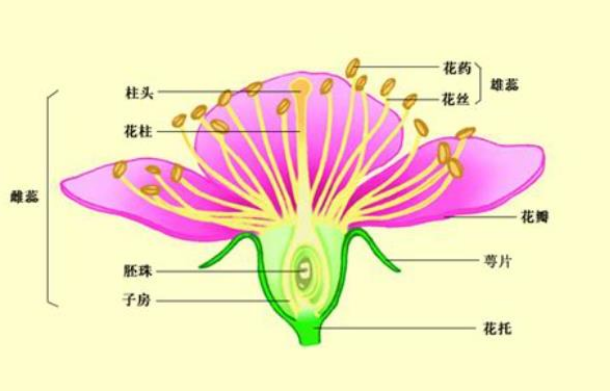

从上下两张图的比较中，可以看到当$k$的取值较小时，拟合曲线非常粗糙，而随着$k$的增大，曲线逐渐光滑。

# 4.3.3 有监督学习-KNN

- 下面我们讨论当响应变量是离散的情况下，利用KNN进行分类。

  IRIS数据集分别以厘米为单位测量了三种鸢尾中每种50朵花的萼片长度和宽度（ sepal length and width）以及花瓣长度和宽度（petal length and width）。品种有刚毛鸢尾、云彩鸢尾和维珍妮卡（Iris setosa, versicolor, and virginica）。

  能否通过花的萼片长度和宽度（ sepal length and width）以及花瓣长度和宽度（petal length and width）的信息区分出三种不同的花？



# 4.3.3 有监督学习-KNN

```
Call:
kknn(formula = Species ~ ., train = iris.learn, test = iris.valid,     distance
1, kernel = "triangular")

Response: "nominal"
          fit prob.setosa prob.versicolor prob.virginica
1   virginica           0       0.01531491      0.98468509
2  versicolor           0       1.00000000      0.00000000
3   virginica           0       0.00000000      1.00000000
4  versicolor           0       1.00000000      0.00000000
5  versicolor           0       1.00000000      0.00000000
6  versicolor           0       1.00000000      0.00000000
7      setosa           1       0.00000000      0.00000000
8  versicolor           0       0.96236501      0.03763499
9      setosa           1       0.00000000      0.00000000
10  virginica           0       0.00000000      1.00000000
11 versicolor           0       1.00000000      0.00000000
12 versicolor           0       0.74344918      0.25655082
13 versicolor           0       0.98242476      0.01757524
14  virginica           0       0.00000000      1.00000000
15 versicolor           0       0.90442536      0.09557464
```

输出的三列分别为划分三种花，选择概率最大的将样本划为此类

# 4.3.3 有监督学习-KNN

- 将检验集通过训练的结果输出和原来的结果进行比较

  程序文件参见5_7.R

| | fit | | |
|---|---|---|---|
| | setosa | versicolor | virginica |
| setosa | 14 | 0 | 0 |
| versicolor | 0 | 18 | 2 |
| virginica | 0 | 0 | 16 |

准确率为（14+18+16）/（14+18+2+16）=48/50=96%，说明使用KNN的方法进行这三类花的分类效果是很好的。

# 4.3.3 有监督学习-KNN

- 约会案例

某男士希望知道他登录婚恋网站之后，和喜欢的女性约会是否会成功，他收集了以往注册该网站的男士基本信息，以及约会是否成功的信息，由此希望知道自己注册之后能否约会到喜欢的女士。

这位男士可以使用KNN来完成这一分析。

# 4.3.3 有监督学习-KNN

- 约会案例

数据“date_data2.csv”记录了这些信息，该数据包含的变量，如下表所示：

| 变量名 | 含义 | 类型 |
| --- | --- | --- |
| income | 收入 | 连续 |
| attractive | 吸引力评分 | 连续 |
| assets | 资产 | 连续 |
| edueduclass | 教育程度 | 有序分类 |
| dated | 是否约会成功 | 无序分类（因变量） |

# 4.3.3 有监督学习-KNN

- 约会案例

```
        income attractive     assets edueduclass
[1,] -1.030402 -1.4404247 -0.9975692  -2.2120355
[2,] -1.030402 -1.2495250 -0.6078301   0.2367123
[3,] -1.030402 -1.5445518 -0.9975692  -2.2120355
[4,] -1.030402 -1.7180969 -0.9764679  -2.2120355
[5,] -0.944678 -1.2495250 -1.0131274  -1.3957863
[6,] -0.944678 -0.7809531 -0.9160641  -0.5795370
```

```
     income          attractive          assets          edueduclass
Min.   :-1.0304   Min.   :-1.71810   Min.   :-1.0131   Min.   :-2.2120
1st Qu.:-0.6875   1st Qu.:-0.78095   1st Qu.:-0.7064   1st Qu.:-0.5795
Median :-0.2589   Median : 0.01735   Median :-0.2773   Median : 0.2367
Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000
3rd Qu.: 0.4269   3rd Qu.: 0.63778   3rd Qu.: 0.3895   3rd Qu.: 0.2367
Max.   : 4.2845   Max.   : 1.70074   Max.   : 4.2852   Max.   : 1.8692
```

# 4.3.3 有监督学习-KNN

- 约会案例

将自变量和因变量分别提取出来，其中因变量是是否约会成功，收入、吸引力评分、资产和教育程度作为自变量，暂不考虑数据集中其他变量。在自变量中教育程度是不是连续变量，但是因为它是有序的，而且有较多的分类水平，在样本不是很大的情况时，可以作为连续变量来处理。

# 4.3.3 有监督学习-KNN

- 约会案例

```
Call:
kknn(formula = dated ~ ., train = date.learn, test = date.valid,     k = 3, distance = 2)

Response: "nominal"
   fit      prob.0     prob.1
1    0 1.00000000 0.00000000
2    1 0.00000000 1.00000000
3    1 0.00000000 1.00000000
4    1 0.00000000 1.00000000
5    1 0.08866211 0.91133789
6    0 1.00000000 0.00000000
```

划分为0或者1的概率

# 4.3.3 有监督学习-KNN

- 约会案例

```
  fit
    0  1
0 14  2
1  0 17
```

预测的准确率为（14+17）/（14+17+2+0）=31/33=94%，说明使用KNN的方法进行约会是否成功的预测效果是很好的。

如果尝试使用不同的$k$的取值呢?效果是否会有很大的差别，同学们可以尝试实验一下

# 4.3.4 逻辑回归 V.S. KNN

- 逻辑回归优缺点（KNN相比）

  优点：计算代价不高，速度很快，存储资源低

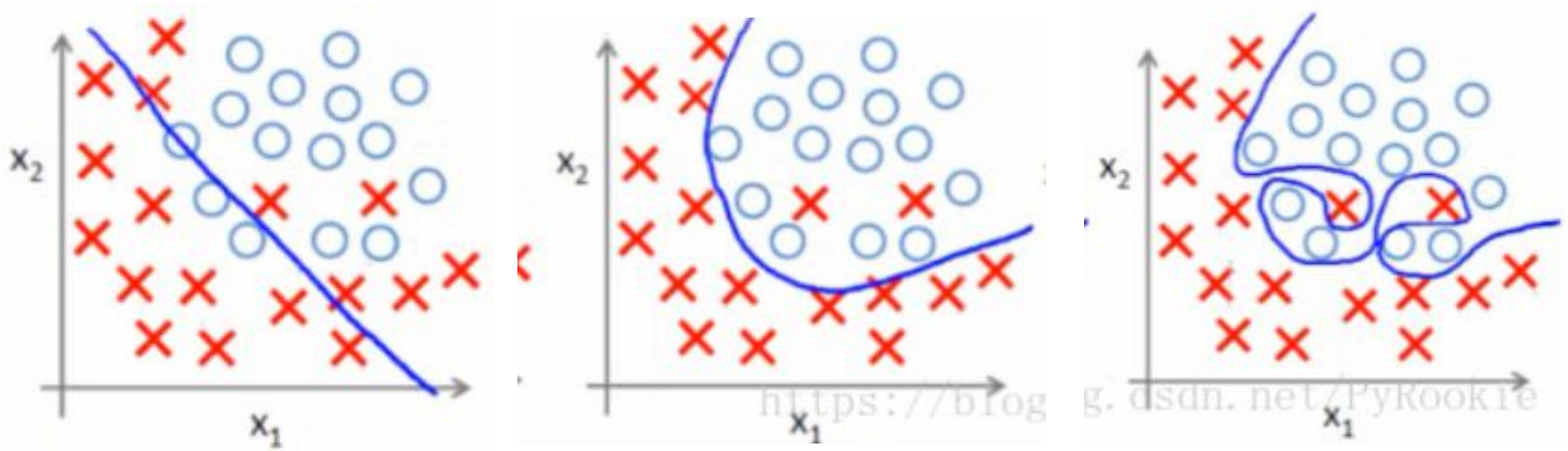
  缺点：容易欠拟合，分类精度可能不高

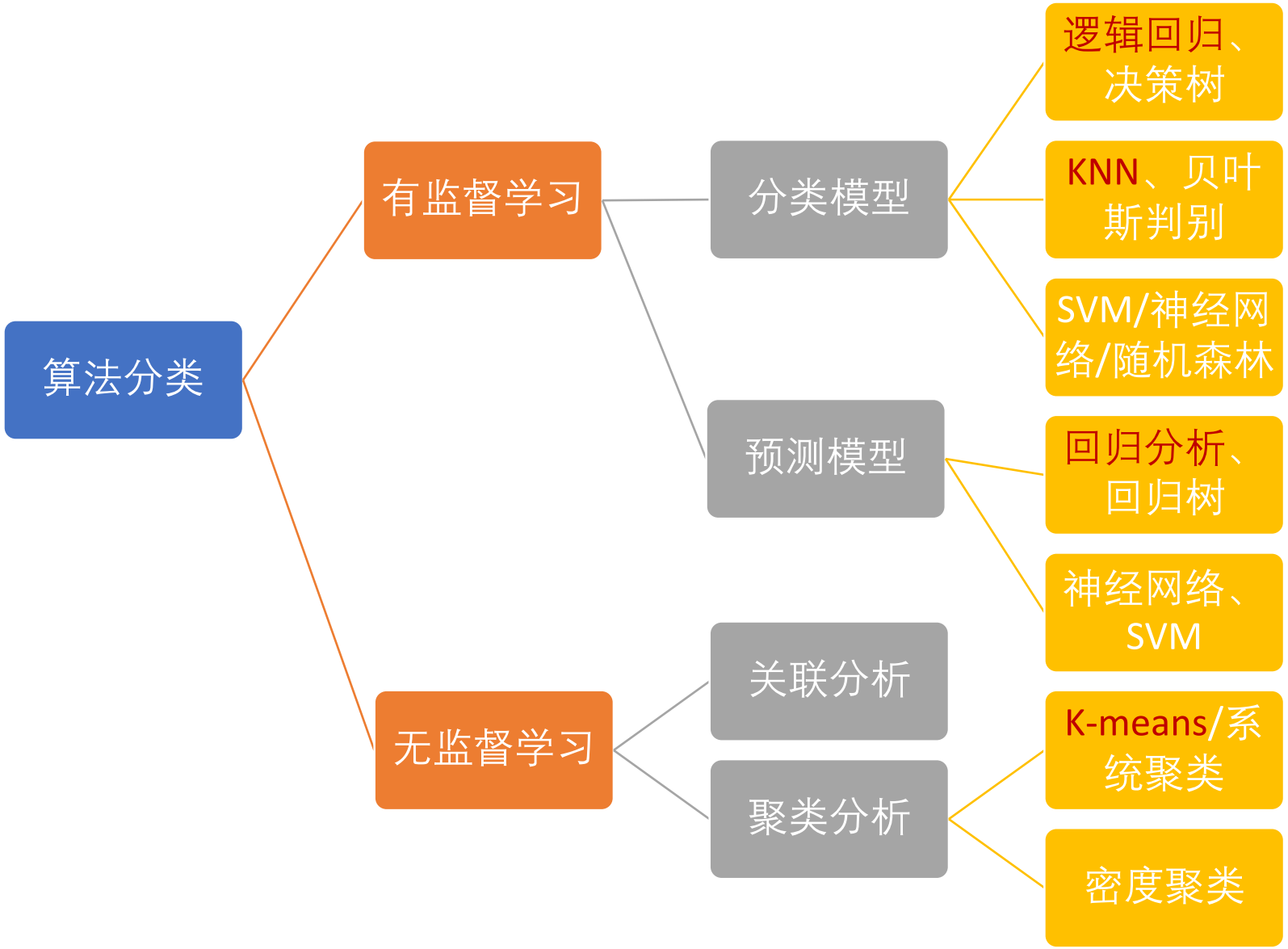
# 4.3.4 逻辑回归 V.S. KNN

- 拟合比较

KNN 相对过拟合例如图3

逻辑回归想对欠拟合例如图1




算法分类
有监督学习
无监督学习
分类模型
预测模型
关联分析
聚类分析
逻辑回归、决策树
KNN、贝叶斯判别
SVM/神经网络/随机森林
回归分析、回归树
神经网络、SVM
K-means/系统聚类
密度聚类

# 4.4、聚类分析

## ➢ 聚类分析

俗话说：物以类聚，人以群分，现实世界中存在着大量的**聚类**问题。

例如在商务上，市场分析人员希望将客户基本库中的客户分成不同的客户群，并且用购买模式来刻画不同客户群的特征进而可以达到一下三种目的：

- 有的放矢—精细化营销活动，生成可控的目标客户群
- 量体裁衣—发现各个细分的客户特性和需求，有针对性地设计营销计划
- 高瞻远瞩—发现战略焦点和业务发展方向

# 4.4、聚类分析

- 聚类分析

聚类分析是客户分群的常见实现方法之一。聚类分析力求使组内客户高度相似而不同组客户差异明显，从而达到分群的目的。

**客户分群案例：**

某世界500强的保险公司在过去几年间通过数据系统的升级，实现了客户数据的积累和打通。最核心的数据信息包括投保人的相关信息数据，详细的保单数据和第三方数据提供商提供的信用评分数据。由第三方专业数据咨询公司建立基于数据分析的聚类分群，并量化群体特征画像。大致的画像如下图所示：

# 4.4、聚类分析

**客户分群案例：**

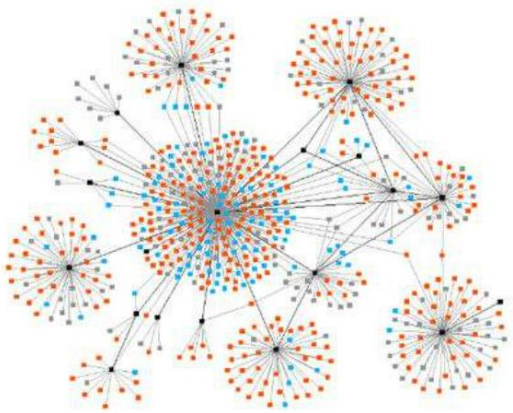
| 人群 | 青年精英 | 初为人母 | 财务自由 | 中年女士 | 居家老人 |
|---|---|---|---|---|---|
| 在客户总体中所占的比例 | 35%（其中50%对寿险产品感兴趣） | 25%（其中60%对寿险产品感兴趣） | 20%（其中40%对寿险产品感兴趣） | 10%（其中30%对寿险产品感兴趣） | 10%（其中25%对寿险产品感兴趣） |
| 群体特征 | 青年中高等收入男性居多受教育程度高 | 青年中等收入女性居多中等教育程度 | 中老年中高收入男性居多公司管理层 | 中年中等收入女性居多 | 中老年中低等收入男性/女性 |
| 消费取向 | 愿意尝试新鲜事物喜好科技产品风险爱好型价格敏感度高 | 使用社交媒体不易接受新科技产品对寿险相关产品了解甚少，需要专业指导价格敏感程度较高 | 对保险有深刻认识对财务状况充满信心价格敏感度最低对寿险产品有兴趣 | 不愿意接受新事物风险厌恶型非家庭的决策者对寿险产品兴趣低 | 使用社交媒体不信任金融机构对寿险产品没兴趣 |

# 4.4、聚类分析

**客户分群案例：**

相对应每一个群体，该保险公司和数据咨询公司紧密合作，并结合公司业务现状和发展方向，以及该公司在市场上现有产品的表现制定了一套公司层面的统一的营销知道策略。

事实证明，在基于多维度数据的科学客户分群基础上的营销策略比基于主观或者局部信息的营销策略能够更加准确地把握用户，从而达到更加有效的营销。



# 4.4、聚类分析

➢ 聚类分析

- 聚类分析的目的是把分类对象按照**一定的规则**分成若干类，这些类不是事先给定的，而是根据数据的特征确定的（无监督学习）。
- 聚类分析对于类的数目和类的结构不必做任何假设，在同一类的这些对象在某种意义上倾向于彼此相似，而在不同类里倾向于不相似。

动态聚类有许多方法，在此我们将讨论一种比较流行的动态聚类的方法—**$K$均值法**。

# 4.4、聚类分析

- **$K$均值聚类**（MacQueen, 1967）

① 选择$K$个样本作为初始凝聚点，或者将所有样本分成$K$个初始类，然后将这$K$个类的均值作为初始凝聚点

② 对除凝聚点之外的点逐个归类，将每个样本点归入凝聚点离它最近的那个类（通常采用欧氏聚类），给类的凝聚点更新为这一类目前的均值，直至所有样本都归了类

③ 重复步骤2，直至所有样本都不能再分配为止。

最终的聚类结果在一定程度上依赖于初始凝聚点或初始类的选择。经验表明，聚类过程绝大多数重要变化均发生在第一次再分配中。

# 4.4、聚类分析

- 设有五个样本，每个只测量了一个指标，分别是1，2，6，8，11，采用$K$均值聚类方法进行聚类。指定$K=2$，具体步骤如下：

① 随意将样本分成$G_1^0=\{1,6,8\}$和$G_2^0=\{2,11\}$两类，则这两个初始类的均值分别为5和6.5

② 计算1到两个类的欧式距离

$$d(1,G_1^0)=|1-5|=4$$

$$d(1,G_2^0)=|1-6.5|=5.5$$

由于1到$G_1^0$的距离小于到$G_2^0$的距离，因此1不用重新分配，计算6到两个类的距离

# 4.4、聚类分析

$$d(6, G_1^0)=|6-5|=1$$

$$d(6, G_2^0)=|6-6.5|=0.5$$

故6应该重新分配到$G_2^0$中，修正后的两个分类为$G_1^1=\{1,8\}$和$G_2^1=\{2,6,11\}$，新的类的均值分别为4.5和6.33，计算

$$d(8, G_1^1)=|8-4.5|=3.5$$

$$d(8, G_2^1)=|8-6.33|=1.67$$

结果8重新分配到$G_2^1$中，新的分类为$G_1^2=\{1\}$和$G_2^2=\{2,6,8,11\}$，其均值分别为1和6.75，再次计算

# 4.4、聚类分析

$$d(2, G_1^2) = |2-1| = 1$$

$$d(2, G_2^2) = |2-6.75| = 4.75$$

重新分配2到$G_1^2$中，两个新类为$G_1^3 = \{1,2\}$和$G_2^3 = \{6,8,11\}$，新类的均值为1.5和8.33

③ 再次计算每个样本到类均值的距离，可以发现每个样本都已经被分到了离类均值更近的类中，最终得到的两个类为$= \{1,2\}$和$\{6,8,11\}$

# 4.4、聚类分析

- USArrests数据集包含1973年美国50个州中，每10万居民中就有人因袭击、谋杀和强奸而被捕的比例以及生活在城市地区的人口百分比这些变量。
- 使用$K$均值将这50个州进行分类。

# 4.4、聚类分析

- 数据的前5行

| | Murder | Assault | UrbanPop | Rape |
|---|---|---|---|---|
| Alabama | 1.24256408 | 0.7828393 | -0.5209066 | -0.003416473 |
| Alaska | 0.50786248 | 1.1068225 | -1.2117642 | 2.484202941 |
| Arizona | 0.07163341 | 1.4788032 | 0.9989801 | 1.042878388 |
| Arkansas | 0.23234938 | 0.2308680 | -1.0735927 | -0.184916602 |
| California | 0.27826823 | 1.2628144 | 1.7589234 | 2.067820292 |

# 4.4、聚类分析

- 如果我们选定$K=4$，使用kmeans进行聚类

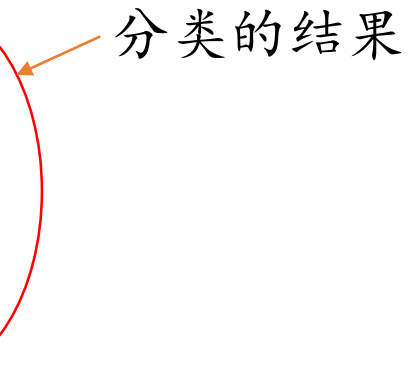
```
K-means clustering with 4 clusters of sizes 8, 13, 13, 16

Cluster means:
      Murder    Assault   UrbanPop        Rape
1  1.4118898  0.8743346 -0.8145211  0.01927104
2  0.6950701  1.0394414  0.7226370  1.27693964
3 -0.9615407 -1.1066010 -0.9301069 -0.96676331
4 -0.4894375 -0.3826001  0.5758298 -0.26165379
```

# 4.4、聚类分析

- 合并标签

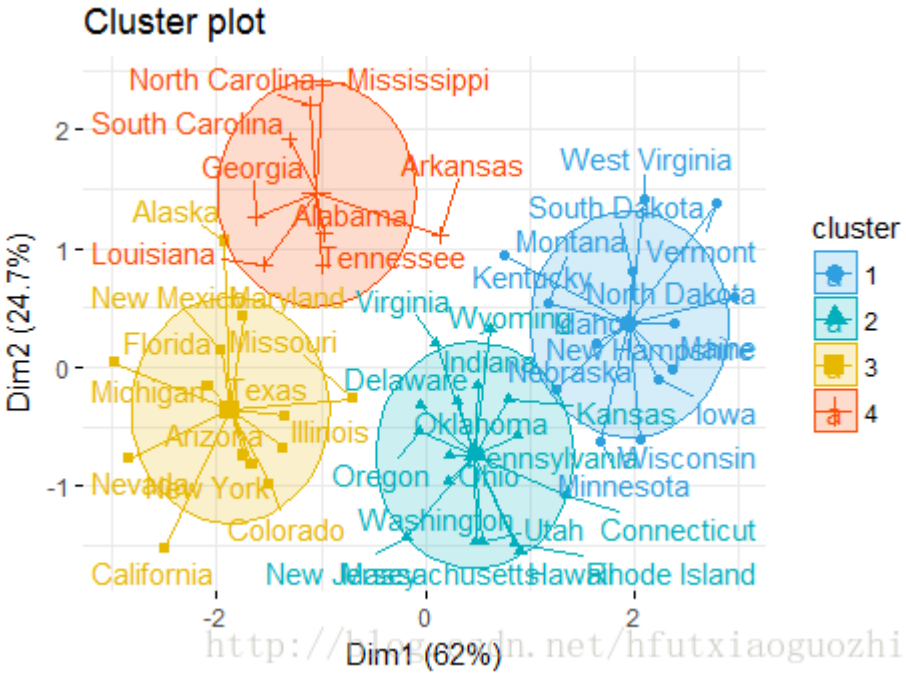
| | Murder | Assault | UrbanPop | Rape | cluster |
|---|---|---|---|---|---|
| Alabama | 13.2 | 236 | 58 | 21.2 | 1 |
| Alaska | 10.0 | 263 | 48 | 44.5 | 2 |
| Arizona | 8.1 | 294 | 80 | 31.0 | 2 |
| Arkansas | 8.8 | 190 | 50 | 19.5 | 1 |
| California | 9.0 | 276 | 91 | 40.6 | 2 |
| Colorado | 7.9 | 204 | 78 | 38.7 | 2 |



# 4.4、聚类分析

- 查看分类数量

```
 1  2  3  4
 8 13 13 16
```




从上面可以看到，第一类为8个州，第二类和第三类为13个，第四类为16个。右上的图展示了四个分类各个州的分布情况

# 4.4、聚类分析

## ➢ 电信运营商客户分群

一家电信运营商希望根据客户在几个主要业务中的消费情况，对客户分群分析，数据“profile_telecom.csv”记录了这些信息，该数据包含以下4个变量，如下表所示：

| 变量名 | 类型 | 解释 |
| --- | --- | --- |
| ID | 离散 | ID |
| cnt_call | 连续 | 打电话次数 |
| cnt_msg | 连续 | 发短信次数 |
| cnt_wei | 连续 | 发微信次数 |
| cnt_web | 连续 | 浏览网页次数 |

# 4.4、聚类分析

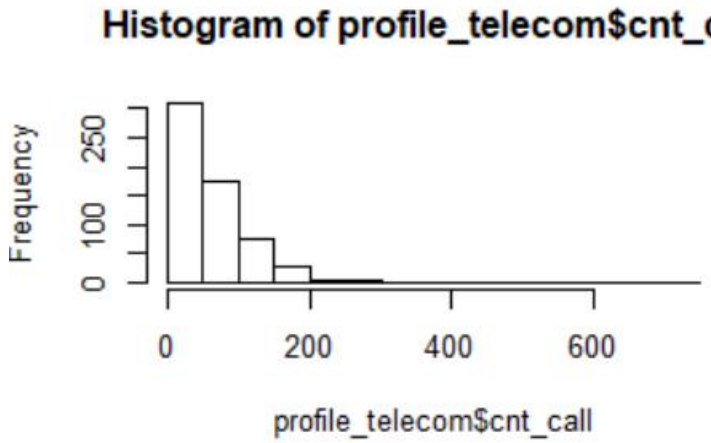
- 电信运营商客户分群

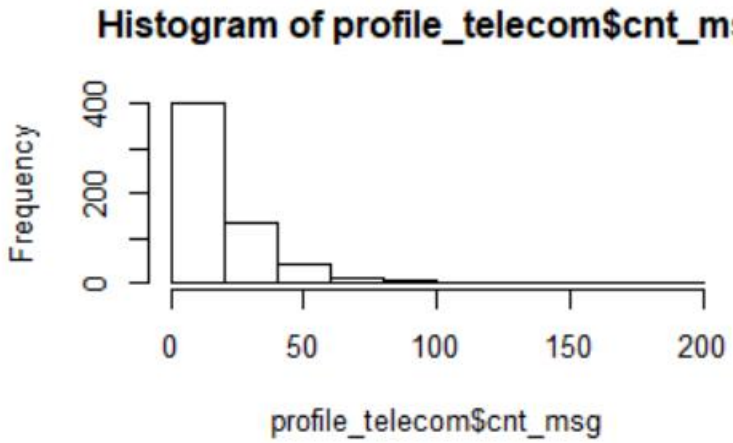
电信客户业务数据部分数据展示：

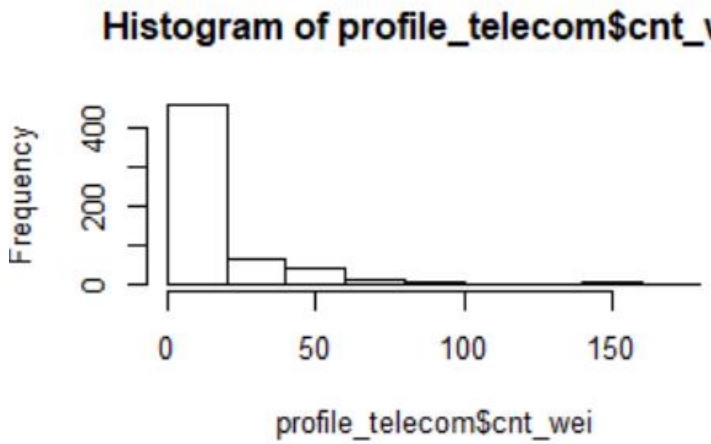
| | ID | cnt_call | cnt_msg | cnt_wei | cnt_web |
|---|---|---|---|---|---|
| 1 | 1964627 | 46 | 90 | 36 | 31 |
| 2 | 3107769 | 53 | 2 | 0 | 2 |
| 3 | 3686296 | 28 | 24 | 5 | 8 |
| 4 | 3961002 | 9 | 2 | 0 | 4 |
| 5 | 4174839 | 145 | 2 | 0 | 1 |
| 6 | 5068087 | 186 | 4 | 3 | 1 |

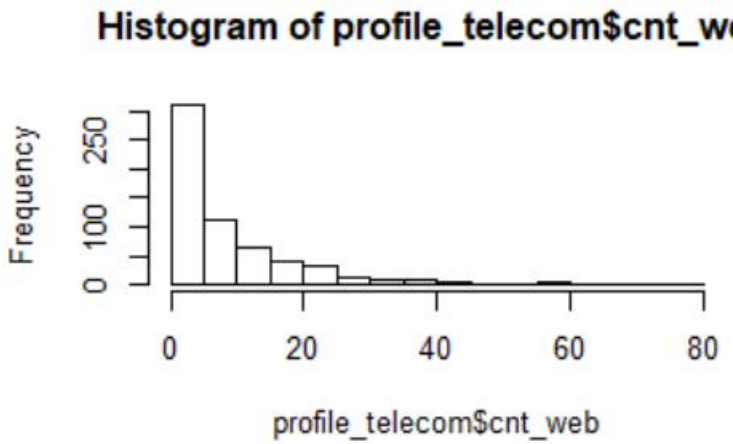
| ID | cnt_call | cnt_msg | cnt_wei | cnt_web |
|---|---|---|---|---|
| Min. : 1964627 | Min. : 2.00 | Min. : 2.00 | Min. : 0.00 | Min. : 0.00 |
| 1st Qu.:214527782 | 1st Qu.: 22.00 | 1st Qu.: 2.00 | 1st Qu.: 0.00 | 1st Qu.: 2.00 |
| Median :470609728 | Median : 49.00 | Median : 11.00 | Median : 4.00 | Median : 5.00 |
| Mean :448745381 | Mean : 65.59 | Mean : 17.76 | Mean : 14.81 | Mean : 8.88 |
| 3rd Qu.:679742515 | 3rd Qu.: 87.25 | 3rd Qu.: 26.00 | 3rd Qu.: 19.00 | 3rd Qu.:13.00 |
| Max. :873574458 | Max. :729.00 | Max. :186.00 | Max. :162.00 | Max. :77.00 |

# 4.4、聚类分析


Histogram of profile_telecom$cnt_call
Frequency
0
100
250
0
200
400
600
profile_telecom$cnt_call


Histogram of profile_telecom$cnt_msg
Frequency
0
200
400
0
50
100
150
200
profile_telecom$cnt_msg


Histogram of profile_telecom$cnt_wei
Frequency
0
200
400
0
50
100
150
profile_telecom$cnt_wei


Histogram of profile_telecom$cnt_web
Frequency
0
100
250
0
20
40
60
80
profile_telecom$cnt_web

# 4.4、聚类分析

- 电信运营商客户分群

从四个变量的直方图可以看到每个变量都是偏态的。一般来说，高消费的用户总是少数，多数消费都发生在中低档次。通信业务由于采用套餐制，所以很多用户都会在套餐当中的某项业务上面出现零消费。

# 4.4、聚类分析

- 电信运营商客户分群

聚类之后得到的类中心如下：

```
K-means clustering with 4 clusters of sizes 354, 62, 27, 157

Cluster means:
     cnt_call    cnt_msg    cnt_wei    cnt_web
1 -0.37888336 -0.4890277 -0.4601545 -0.5231555
2  1.94218693 -0.2939760 -0.4019650 -0.3951752
3  0.37289171  1.7696297  3.3646553  3.2292527
4  0.02319136  0.9144096  0.6176485  0.7803062
```

# 4.4、聚类分析

- 电信运营商客户分群

| | ID | cnt_call | cnt_msg | cnt_wei | cnt_web | cluster |
|---|---|---|---|---|---|---|
| 1 | 1964627 | 46 | 90 | 36 | 31 | 4 |
| 2 | 3107769 | 53 | 2 | 0 | 2 | 1 |
| 3 | 3686296 | 28 | 24 | 5 | 8 | 1 |
| 4 | 3961002 | 9 | 2 | 0 | 4 | 1 |
| 5 | 4174839 | 145 | 2 | 0 | 1 | 2 |
| 6 | 5068087 | 186 | 4 | 3 | 1 | 2 |

分类的结果

# 4.4、聚类分析

- 电信运营商客户分群

从上面的输出结果可以看到，绝大部分客户都被归为1类，而第4类的客户数量居中，而第2，3类客户数量很少。这种聚类结果是符合实际情况的，根据二八法则，价值低的客户人数最多，甚至占到90%也是常见的情况。

| 分类 | 频数 |
| --- | --- |
| 1 | 354 |
| 2 | 62 |
| 3 | 27 |
| 4 | 157 |

分类的结果

# 4.4、聚类分析

- 航空公司客户价值分析
- 航空行业竞争背景

- 随着中国加入WTO，民航市场政府管制的逐步开放，中国的民航市场竞争必然将愈来愈激烈。
- 中国巨大的市场机会吸引了众多的竞争对手加入。（中国目前的航空消费者仅仅只占全部人口的5%，而美国等成熟市场则达到95%以上，中国的发展趋势：将成为世界最大的民航市场之一）
- 民航的竞争除了三大航空公司之间的竞争之外，还将加入新崛起的各类小型航空公司、民营航空公司，甚至国外航空巨头。

# 4.4、聚类分析

| | |
|---|---|
| **国营航空垄断** | 上个世纪80年代初，<br>国内航空市场的绝大部分，都被三大国有航空公司占据，<br>其中包括中国航空、南方航空、东方航空 |
| **地方航空增加** | 上个世纪90年代初，<br>民航公司从原来民航总局直属的9家，<br>一下增至20多家，最多时曾达到34家。 |
| **民营航空涌现** | 2005年3月5日，民营奥凯航空有限公司，<br>首架航班启动，正式投入航线运营。<br>除此之外，尚有鹰联、春秋航空有限公司等。 |
| **外资航空进入** | 2005年、中美航空协议的正式实施，让美国各<br>航空公司，终于领到了“通行证”；中国航空<br>公司正式面对来自国际巨头的挑战。 |

# 4.4、聚类分析

## ➢ 航空行业的竞争形势

- 更新营销手段抢占市场 ：

各航空公司用直销这一营销新手段抢占航空客运市场，包括官网、呼叫中心、直属售票处、手机客户端、在竞争中收到了良好的成效。

- 价格战烽烟四起：

各航空公司纷纷采取优惠措施拉拢旅客，加强机票销售，市场竞争日趋激烈。

# 4.4、聚类分析

## ➢ 航空行业的竞争形势

- 混战航空枢纽：

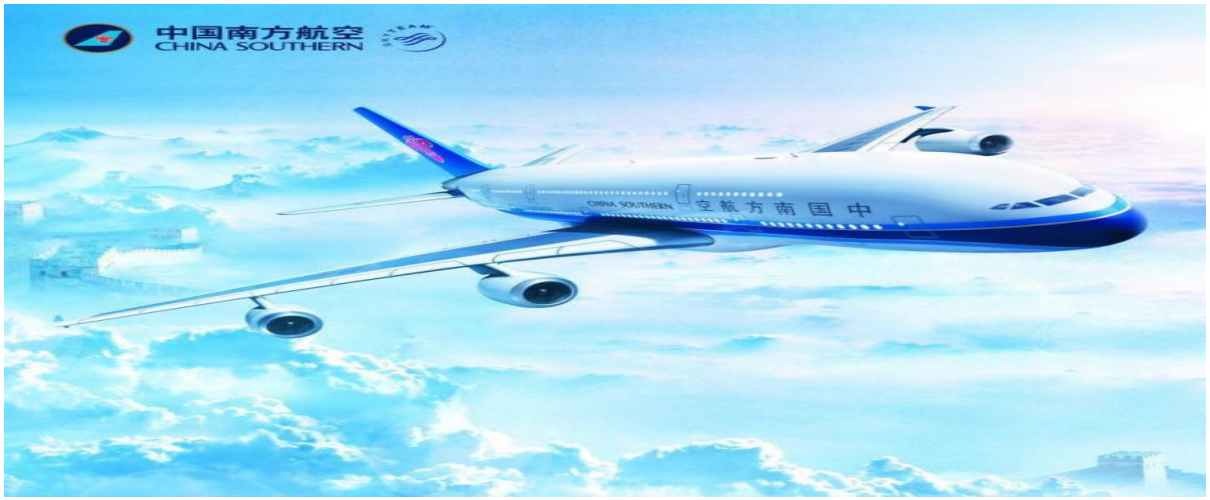
北京、上海、广州三大航空枢纽开放，国内航空市场进入战国时期，原有市场格局产生变化，重新开始洗牌。

- 航空公司扩大运力以保市场份额：

各航空公司为确保行业地位及份额，纷纷进行新一轮的招兵买马活动，扩大自身运力。

# 4.4、聚类分析

- 航空公司面临的现状
  - 客户流失
  - 资源没有充分利用
  - 竞争力下降



# 4.4、聚类分析

## ➢ 航空公司客户价值分析

- 随着市场经济的快速发展，我国的航空运输市场早已从卖方转向买方。
- 价格竞争已走到了尽头，躲在政府的保护伞下坐享其成已不再可能，通过先进的管理来获得竞争优势，成了各航空公司必然的选择。
- 由于航空产品生产过剩，产品同质化特征愈加明显，航空业务规则的主导已不再是产品价值，而是客户需求，于是航空公司从价格、服务间的竞争逐渐转向对客户的竞争，客户关系管理因此变得尤为重要。

# 4.4、聚类分析

## ➢ 关系营销

- 关系营销的目的是创造“忠诚的客户”。90年代以来，营销领域中越来越多的人转向美国北卡罗莱纳大学教授R.E劳特朗提出的4Cs理论，即消费者的需求和欲求（Consumer wants and needs）、成本（Cost）、方便（Convenience）、沟通（Communication）
- 它强调企业应该把追求顾客满意放在第一位，努力降低顾客的购买成本，注重顾客购买过程中的便利性，与客户保持有效的营销沟通。

# 4.4、聚类分析

## ➢ 关系营销

- 客户关系管理（Customer Relationship Management）的理念来源于关系营销。
- 客户的有效管理——客户细分
- 客户价值区间细分（例如大客户、重要客户、普通客户、小客户等）
- 根据“2/8 定律”的原理重点锁定高价值客户。客户价值区间的变量包括：客户响应力、客户消费频率、客户销售收入、客户利润贡献、客户忠诚度等

# 客户细分的目的

***客户获取***

- 谁是我们的客户？
- 我们的客户有何特征？
- 我们的客户需要什么？


***客户保有***

- 建立企业化的客户资源
- 持续的客户关系维护
- 提高客户满意度
- 延长客户生命周期

***客户价值提升***

- 保持VIP客户的价值贡献
- 推动客户向VIP转移

# 航空公司的常旅客计划


模式“常旅客”计划是国航特别设计的里程奖励活动。

载体：会员卡发放数量大。

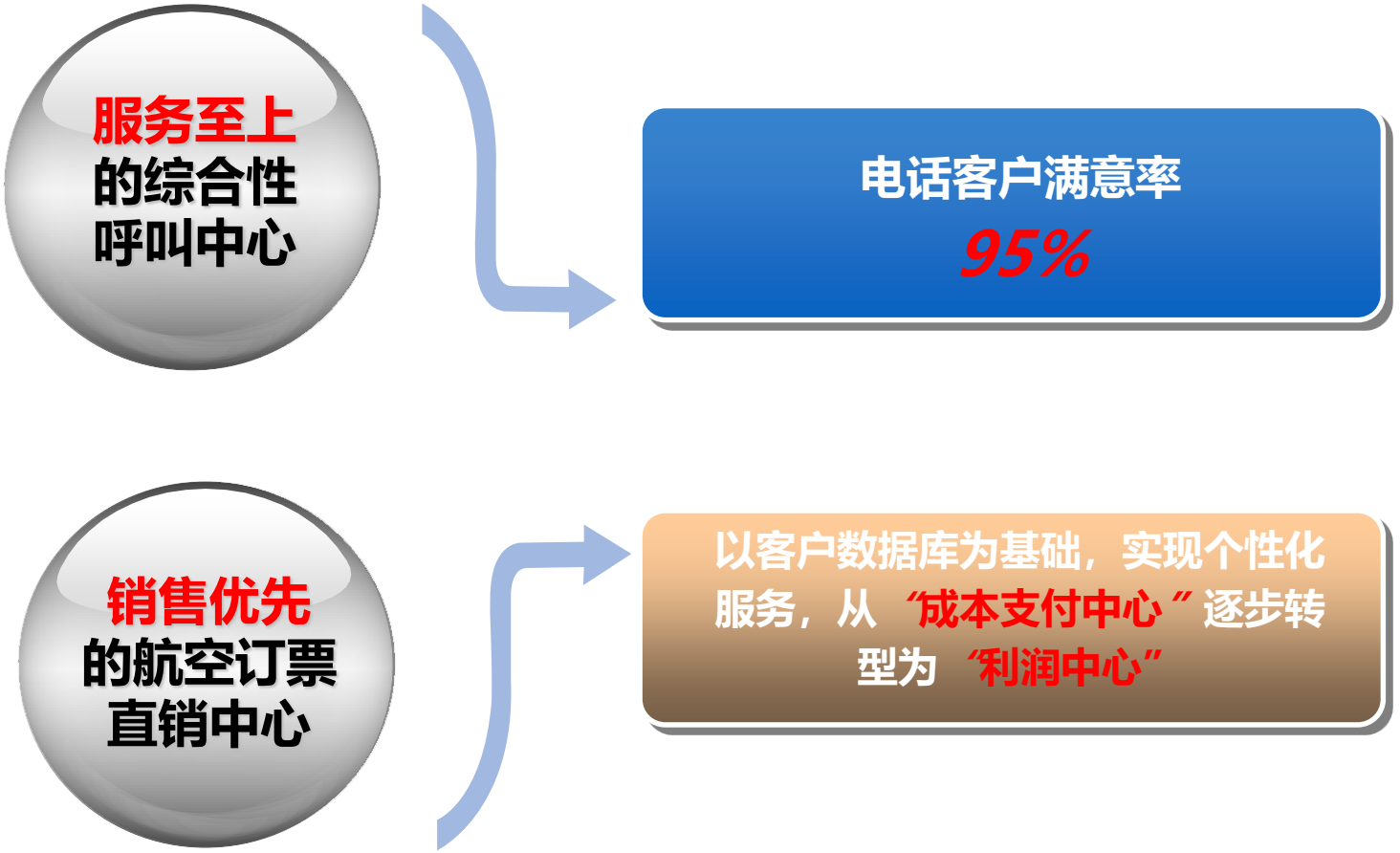
按标准将会员分为4级：普通卡会员、银卡会员、金卡会员及白金卡会员。

奖励：级别越高的会员获得的奖励也越多，如优先订座候补、优先机场候补、保留订座至航班起飞前48小时、额外免费行李额、优先行李处置、优先登机、 优先保证座位等。

# 高价值客户：“亲切关怀活动”

# 低价值客户：咨询服务呼叫中心

- 呼叫中心-从被动服务到主动营销



# 客户价值分析——RFM模型原理

## RFM模型主题、原理

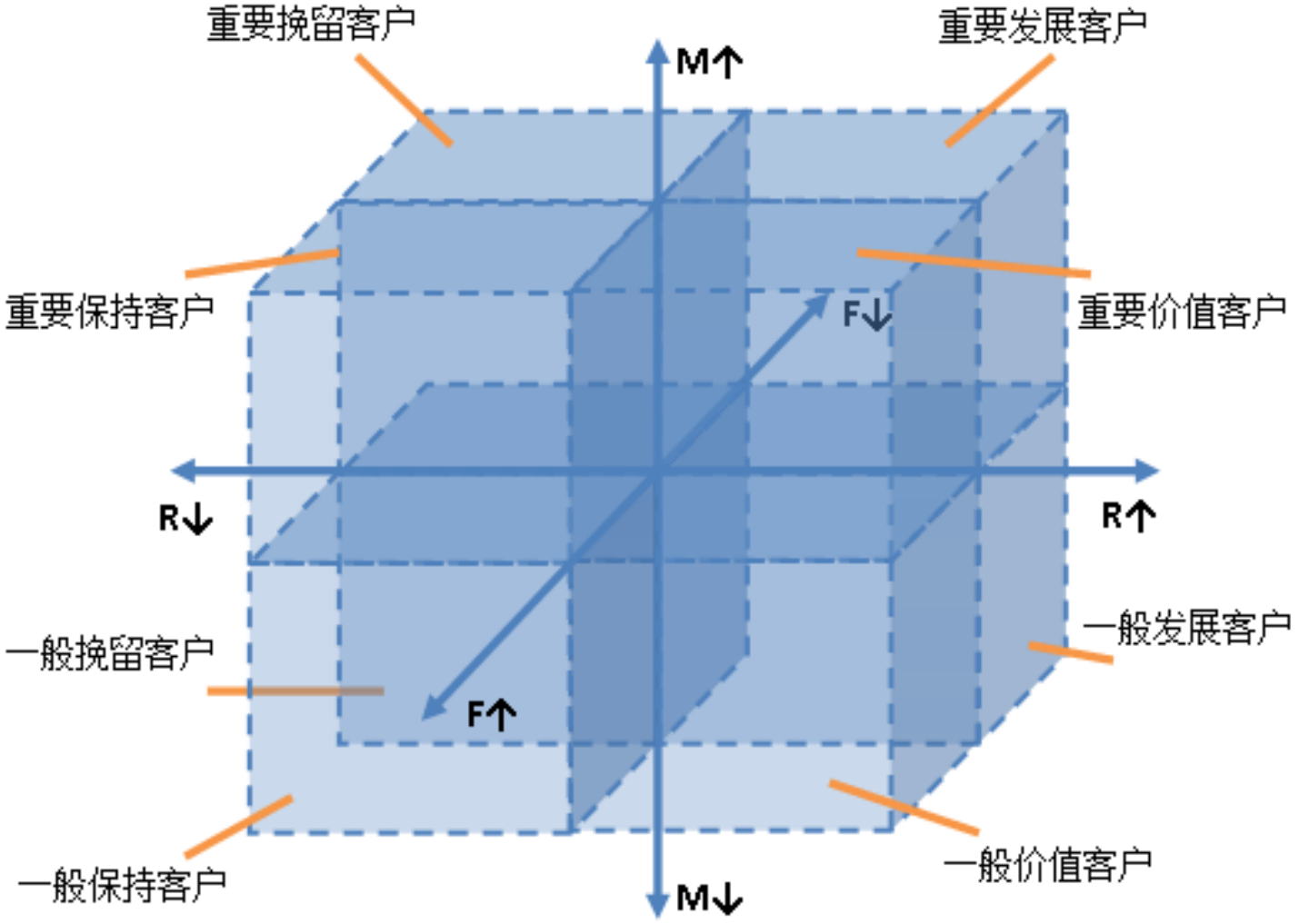
- 近度R(Recency-近度）：R代表客户最近的购买时间距离数据采集点的时间距离，R越大，表示客户越久未发生交易，R越小，表示客户越近有交易发生。R越大则客户越可能会“沉睡”，流失的可能性越大。在这部分客户中，可能有些优质客户，值得公司通过一定的营销手段进行激活。
- 额度M（Monetary-额度）：表示客户每次消费金额的多少，可以用消费总金额，也可以用过去的平均消费金额，根据分析的目的不同，可以有不同的标识方法。一般来讲，单次交易金额较大的客户，支付能力强，价格敏感度低，是较为优质的客户，而每次交易金额很小的客户，可能在支付能力和支付意愿上较低。

# RFM模型原理

- 频度F(Frequency-频度）：F代表客户过去某段时间内的活跃频率。F越大，则表示客户同本公司的交易越频繁，不仅仅给公司带来人气，也带来稳定的现金流，是非常忠诚的客户；F越小，则表示客户不够活跃，且可能是竞争对手的常客。针对F较小、且消费额较大的客户，需要推出一定的竞争策略，将这批客户从竞争对手中争取过来。

思考：
频度是不是越大越好？
频度和近度、金额有着怎样的关系？

# RFM模型思路


重要挽留客户
重要发展客户
M↑
重要保持客户
F↓
重要价值客户
R↓
R↑
一般挽留客户
一般发展客户
F↑
一般保持客户
一般价值客户
M↓

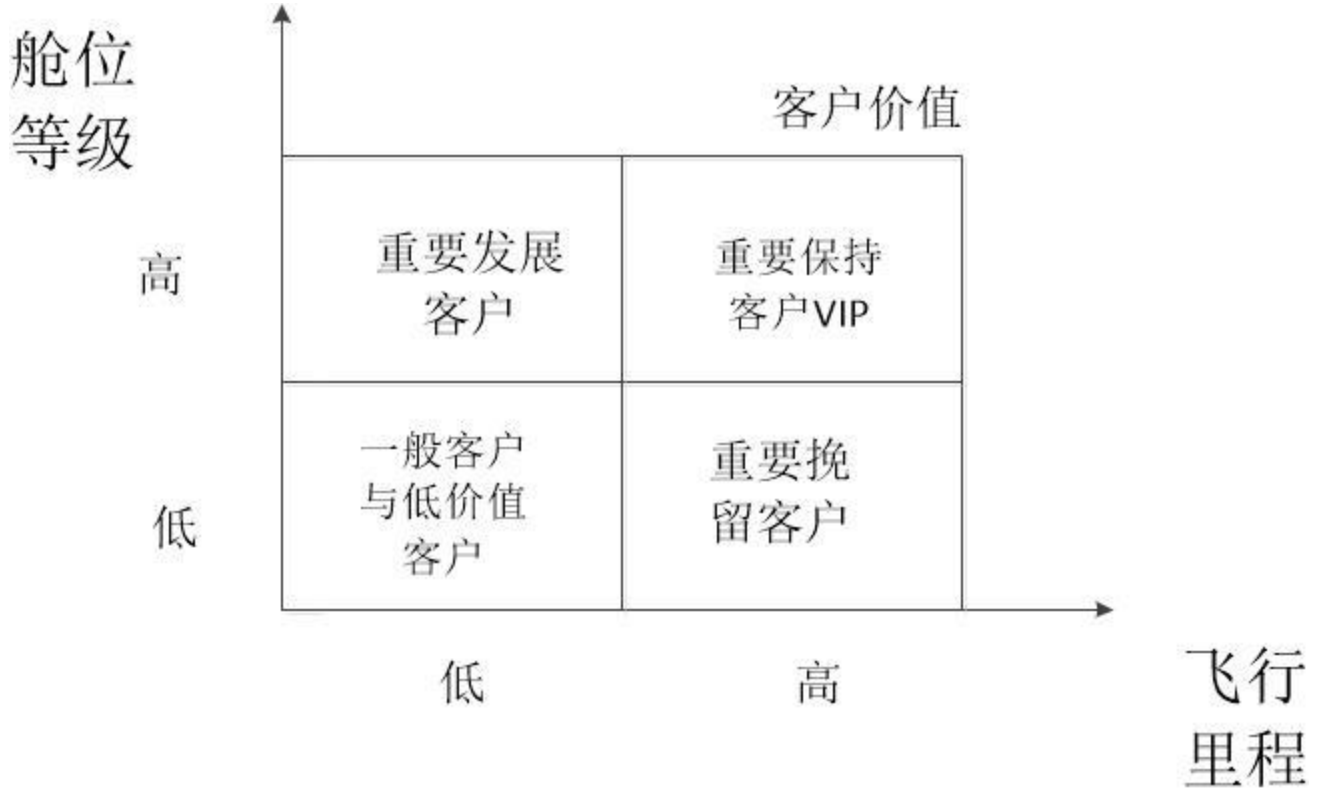
# RFM模型思路

| Recency | Frequency | Monetary | 客户类型 |
|---|---|---|---|
| ↑ | ↑ | ↑ | 重要价值客户 |
| ↑ | ↓ | ↑ | 重要发展客户 |
| ↓ | ↑ | ↑ | 重要保持客户 |
| ↓ | ↓ | ↑ | 重要挽留客户 |
| ↑ | ↑ | ↓ | 一般价值客户 |
| ↑ | ↓ | ↓ | 一般发展客户 |
| ↓ | ↑ | ↓ | 一般保持客户 |
| ↓ | ↓ | ↓ | 一般挽留客户 |

——“↑”表示大于均值，“↓”表示小于均值。

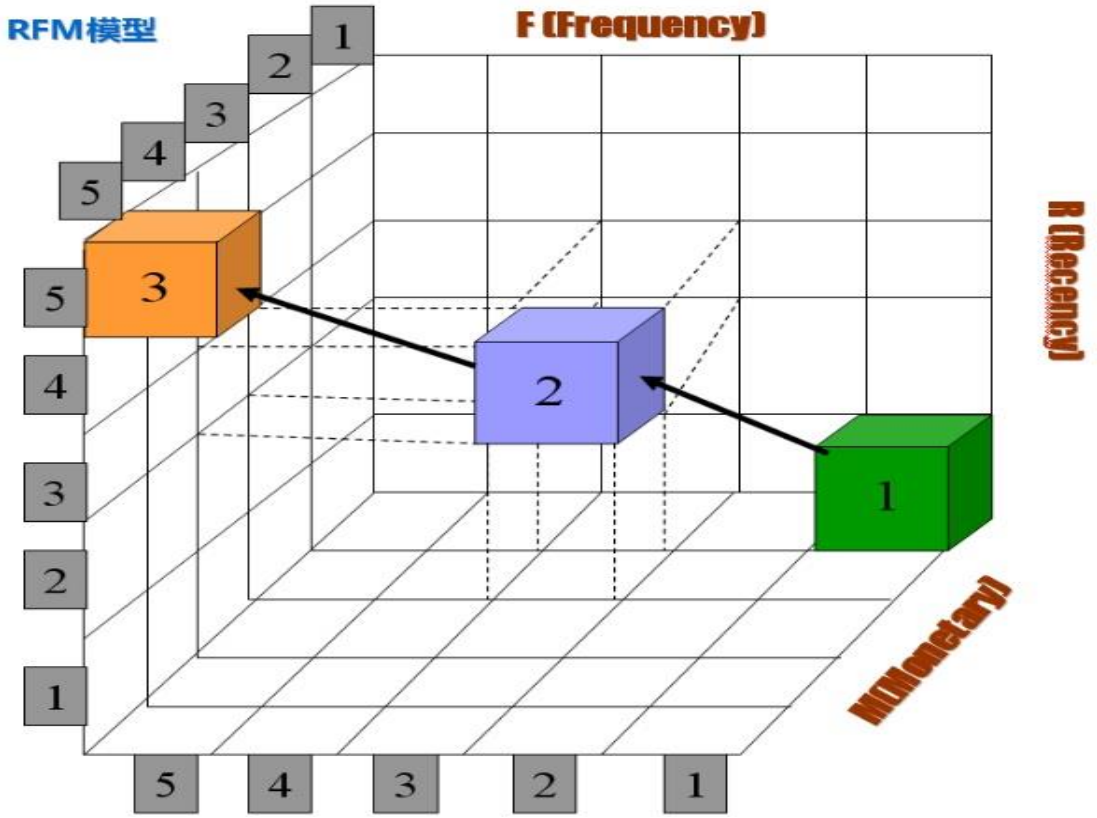
# 传统RFM模型金额不能代表客户价值

- 在模型中，消费金额表示在一段时间内，客业产品金额的总和。因航空票价受到**运输距离**、**舱位等级等**多种因素影响，同样消费金额的不同旅客对航空公司的价值是不同的。因此这这些指标不能直接用到航空公司的客户价值分析。



# 传统RFM模型细分客户太多

- 传统模型分析是利用属性分箱方法进行分析，但是此方法细分的客户群太多，需要一一识别客户特征和行为,提高了针对性营销的成本。



# 改进的RFM模型

| 模型 | L | R | F | M | C |
| --- | --- | --- | --- | --- | --- |
| 传统**RFM**模型 | | 客户最后一次购买距今时间长度 | 客户一段时间内购买该企业产品的次数 | 客户某一段时间内购买企业产品**金额** | |
| 航空公司**LRFMC**模型 | 会员入会时间距今月数 | 客户最近一次乘坐本公司飞机距今月数 | 客户一段时间内乘坐本公司的次数 | 客户一段时间内在航空公司累计的**飞行里程** | 客户在一段时间内所在仓位所对应的折扣系数平均值 |

# 原始数据情况

- 航空客户信息（数据见air_data.csv），其中已经包含会员档案信息和其乘坐航班记录等

| MEMBER_ | FFP_DATE | FIRST_FLIG | GENDE | FFP_TI | WORK_CIT | WORK_PROVIN | WORK_COUI | AGE | LOAD_TIME | FLIGHT | BP_SUM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 54993 | 2006/11/02 | 2008/12/24 | 男 | 6 | . | 北京 | CN | 31 | 2014/03/31 | 210 | 505308 |
| 28065 | 2007/02/19 | 2007/08/03 | 男 | 6 | | 北京 | CN | 42 | 2014/03/31 | 140 | 362480 |
| 55106 | 2007/02/01 | 2007/08/30 | 男 | 6 | . | 北京 | CN | 40 | 2014/03/31 | 135 | 351159 |
| 21189 | 2008/08/22 | 2008/08/23 | 男 | 5 | Los Angele | CA | US | 64 | 2014/03/31 | 23 | 337314 |
| 39546 | 2009/04/10 | 2009/04/15 | 男 | 6 | 贵阳 | 贵州 | CN | 48 | 2014/03/31 | 152 | 273844 |
| 56972 | 2008/02/10 | 2009/09/29 | 男 | 6 | 广州 | 广东 | CN | 64 | 2014/03/31 | 92 | 313338 |
| 44924 | 2006/03/22 | 2006/03/29 | 男 | 6 | 乌鲁木齐市 | 新疆 | CN | 46 | 2014/03/31 | 101 | 248864 |
| 22631 | 2010/04/09 | 2010/04/09 | 女 | 6 | 温州市 | 浙江 | CN | 50 | 2014/03/31 | 73 | 301864 |
| 32197 | 2011/06/07 | 2011/07/01 | 男 | 5 | DRANCY | | FR | 50 | 2014/03/31 | 56 | 262958 |
| 31645 | 2010/07/05 | 2010/07/05 | 女 | 6 | 温州 | 浙江 | CN | 43 | 2014/03/31 | 64 | 204855 |
| 58877 | 2010/11/18 | 2010/11/20 | 女 | 6 | PARIS | PARIS | FR | 34 | 2014/03/31 | 43 | 298321 |
| 37994 | 2004/11/13 | 2004/12/02 | 男 | 6 | 北京 | . | CN | 47 | 2014/03/31 | 145 | 256093 |
| 28012 | 2006/11/23 | 2007/11/18 | 男 | 5 | SAN MARI | CA | US | 58 | 2014/03/31 | 29 | 210269 |
| 54943 | 2006/10/25 | 2007/10/27 | 男 | 6 | 深圳 | 广东 | CN | 47 | 2014/03/31 | 118 | 241614 |
| 57881 | 2010/02/01 | 2010/02/01 | 女 | 6 | 广州 | 广东 | CN | 45 | 2014/03/31 | 50 | 289917 |
| 1254 | 2008/03/28 | 2008/04/05 | 男 | 4 | BOWLAND | CALIFORNIA | US | 63 | 2014/03/31 | 22 | 286164 |
| 8253 | 2010/07/15 | 2010/08/20 | 男 | 6 | 乌鲁木齐 | 新疆 | CN | 48 | 2014/03/31 | 101 | 219995 |
| 58899 | 2010/11/10 | 2011/02/23 | 女 | 6 | PARIS | | FR | 50 | 2014/03/31 | 40 | 249882 |
| 26955 | 2006/04/06 | 2007/02/22 | 男 | 6 | 乌鲁木齐市 | 新疆 | CN | 54 | 2014/03/31 | 64 | 215013 |
| 41616 | 2011/08/29 | 2011/10/22 | 男 | 6 | 东莞 | 广东 | CN | 41 | 2014/03/31 | 38 | 191038 |
| 21501 | 2008/07/30 | 2008/11/21 | 男 | 6 | . | 北京 | CN | 49 | 2014/03/31 | 106 | 220641 |
| 41281 | 2011/06/07 | 2011/06/09 | 男 | 6 | VECHEL | NORD BRABAN | AN | | 2014/03/31 | 23 | 255573 |
| 47229 | 2005/04/10 | 2005/04/10 | 男 | 6 | 广州 | 广东 | CN | 69 | 2014/03/31 | 94 | 193169 |
| 28474 | 2010/04/13 | 2010/04/13 | 男 | 6 | | CA | US | 41 | 2014/03/31 | 20 | 256337 |
| 58472 | 2010/02/14 | 2010/03/01 | 女 | 5 | | | FR | 48 | 2014/03/31 | 44 | 204801 |
| 13942 | 2010/10/14 | 2010/11/01 | 男 | 6 | PARIS | FRANCE | FR | 39 | 2014/03/31 | 62 | 241719 |
| 45075 | 2007/02/01 | 2007/03/23 | 男 | 6 | 湛江 | 广东 | CN | 46 | 2014/03/31 | 213 | 217809 |

# 原始数据情况

## 数据字段解释

| | |
|---|---|
| 会员卡号 | MEMBER_NO |
| 入会时间 | FFP_DATE |
| 第一次飞行日期 | FIRST_FLIGHT_DATE |
| 性别 | GENDER |
| 会员卡级别 | FFP_TIER |
| 工作地城市 | WORK_CITY |
| 工作地所在省份 | WORK_PROVINCE |
| 工作地所在国家 | WORK_COUNTRY |
| 年龄 | AGE |
| 观测窗口的结束时间 | LOAD_TIME |
| 飞行次数 | FLIGHT_COUNT |
| 观测窗口总基本积分 | BP_SUM |
| 第一年精英资格积分 | EP_SUM_YR_1 |
| 第二年精英资格积分 | EP_SUM_YR_2 |
| 第一年总票价 | SUM_YR_1 |
| 第二年总票价 | SUM_YR_2 |
| 观测窗口总飞行公里数 | SEG_KM_SUM |
| 观测窗口总加权飞行公里数（Σ | WEIGHTED_SEG_KM |
| 末次飞行日期 | LAST_FLIGHT_DATE |
| 观测窗口季度平均飞行次数 | AVG_FLIGHT_COUNT |
| 观测窗口季度平均基本积分累积 | AVG_BP_SUM |
| 观察窗口内第一次乘机时间至M | BEGIN_TO_FIRST |
| 最后一次乘机时间至观察窗口末 | LAST_TO_END |
| 平均乘机时间间隔 | AVG_INTERVAL |
| 观察窗口内最大乘机间隔 | MAX_INTERVAL |
| 观测窗口中第1年其他积分（ | ADD_POINTS_SUM_YR_1 |

**观测窗口**

以过去某个时间点为结束时间，某一时间长度作为宽度，得到历史时间范围内的一个时间段。

**折扣系数**

飞机一般分头等舱、公务舱和经济舱3三种。以经济舱票价为100，头等舱是150，商务舱是130。

**里程累计**

舱累计公里数的%150，商务舱%100，经济舱%50

# 属性构造思考


原始数据中包含40多个属性变量，这些属性变量能否满足模型需求，需不需要重新构造指标建模？若是需要重新构造指标，我们又该从哪些角度出发呢？

1、航空公司
2、客户价值分析

# 回顾目标

1. 借助航空公司客户数据，对**客户进行分类**；
2. 对不同的客户类别进行**特征分析**，比较不同类客户的**客户价值**；
3. 对不同价值的客户类别提供**个性化服务**，制定相应的**营销策略**。

# 分析方法与过程

## 第1步：数据抽取

- 以2014-03-31为结束时间，选取宽度为两年的时间段作为分析观测窗口，抽取观测窗口内有乘机记录的所有客户的详细数据形成历史数据。对于后续新增的客户详细信息，利用其数据中最大的某个时间点作为结束时间，采用上述同样的方法进行抽取，形成增量数据。
- 根据末次飞行日期，从航空公司系统内抽取2012-04-01至2014-03-31内所有乘客的详细数据，总共62988条记录。

# 分析方法与过程

## 第2步：探索分析

- 原始数据中存在票价为空值，票价为空值的数据可能是客户不存在乘机记录造成。
- 票价最小值为0、折扣率最小值为0、总飞行公里数大于0的数据。其可能是客户乘坐0折机票或者积分兑换造成。

| 属性名称 | SUM_YR_1 | SUM_YR_2 | ... | SEG_KM_SUM | AVG_DISCOUNT |
|---|---|---|---|---|---|
| 空值记录数 | 551 | 138 | ... | 0 | 0 |
| 最大值 | 239560 | 234188 | ... | 580717 | 1.5 |
| 最小值 | 0 | 0 | ... | 368 | 0 |

# 分析方法与过程

## 第3步：数据预处理

1. 数据清洗：从业务以及建模的相关需要方面考虑，筛选出需要的数据
    a) 丢弃票价为空的数据。
    b) 丢弃票价为0、平均折扣率不为0、总飞行公里数大于0的数据。

# 分析方法与过程

## 第3步：数据预处理

2. 属性规约：原始数据中属性太多，根据LRFMC模型，选择与其相关的六个属性，删除不相关、弱相关或冗余的属性。

| LOAD_TIME | FFP_DATE | LAST_TO_END | FLIGHT_COUNT | SEG_KM_SUM | AVG_DISCOUNT |
|---|---|---|---|---|---|
| 2014/4/1 | 2006/03/31 | 6.6 | 3 | 18770 | 0.66 |
| 2014/4/1 | 2006/03/31 | 3.8 | 24 | 35087 | 0.62 |
| 2014/4/1 | 2006/04/07 | 2.8 | 9 | 20660 | 0.52 |
| 2014/4/1 | 2006/08/10 | 1 | 12 | 23071 | 0.51 |
| 2014/4/1 | 2008/02/07 | 3.17 | 3 | 2897 | 0.95 |
| 2014/4/1 | 2010/09/16 | 1.57 | 3 | 4608 | 0.65 |
| 2014/4/1 | 2011/04/28 | 17.83 | 2 | 3390 | 0.48 |
| 2014/4/1 | 2012/03/09 | 4.13 | 8 | 11797 | 1.35 |
| 2014/4/1 | 2012/08/31 | 5.9 | 6 | 6355 | 0.75 |
| 2014/4/1 | 2005/09/29 | 0.4 | 54 | 62170 | 0.79 |
| 2014/4/1 | 2009/01/23 | 2.33 | 24 | 15894 | 0.6 |
| 2014/4/1 | 2009/01/30 | 0.07 | 13 | 19517 | 0.72 |
| 2014/4/1 | 2009/01/30 | 4.63 | 10 | 12686 | 0.55 |
| 2014/4/1 | 2009/02/13 | 14.93 | 13 | 10992 | 1.33 |
| 2014/4/1 | 2009/04/25 | 10.27 | 3 | 4137 | 0.67 |
| 2014/4/1 | 2009/06/12 | 0.33 | 19 | 37415 | 0.63 |
| 2014/4/1 | 2010/03/25 | 1.63 | 13 | 24156 | 0.79 |
| 2014/4/1 | 2010/04/15 | 22.23 | 3 | 1559 | 0.87 |
| 2014/4/1 | 2010/05/20 | 23.17 | 2 | 1870 | 0.6 |
| 2014/4/1 | 2010/07/08 | 2.6 | 30 | 46621 | 0.93 |
| 2014/4/1 | 2011/03/10 | 4.57 | 4 | 7999 | 0.58 |

# 分析方法与过程

## 第3步：数据预处理

### 3. 数据变换

a) **属性构造：**因原始数据中并没有直接给出LRFMC五个指标，需要构造这五个指标。（构造后数据见zscoredata.csv）

L = LOAD_TIME - FFP_DATE

会员入会时间距观测窗口结束的月数 = 观测窗口的结束时间 - 入会时间[单位：月

R = LAST_TO_END

客户最近一次乘坐公司飞机距观测窗口结束的月数 = 最后一次乘机时间至观察窗口末端时长[单位：月]

F = FLIGHT_COUNT

客户在观测窗口内乘坐公司飞机的次数 = 观测窗口的飞行次数[单位：次]

M = SEG_KM_SUM

客户在观测时间内在公司累计的飞行里程 = 观测窗口总飞行公里数[单位：公里]

C = AVG_DISCOUNT

客户在观测时间内乘坐舱位所对应的折扣系数的平均值 = 平均折扣率[单位：无]

# 分析方法与过程

## 第3步：数据预处理

### 3. 数据变换

a) **数据标准化：**因五个指标的取值范围数据差异较大，为了消除数量级数据带来的影响，需要对数据进行标准化处理。

| 属性名称 | L | R | F | M | C |
|---|---|---|---|---|---|
| 最小值 | 12.23 | 0.03 | 2 | 368 | 0.14 |
| 最大值 | 114.63 | 24.37 | 213 | 580717 | 1.5 |

# 分析方法与过程

## 第4步：构建模型

### 1. 客户K-Means聚类

采用K-Means聚类算法对客户数据进行分群，将其聚成五类（需要结合业务的理解与分析来确定客户的类别数量）。

表 7-9 客户聚类结果

| 聚类类别 | 聚类个数 | 聚类中心 | | | | |
|---|---|---|---|---|---|---|
| | | ZL | ZR | ZF | ZM | ZC |
| 客户群 1 | 5337 | 0.483 | -0.799 | 2.483 | 2.424 | 0.308 |
| 客户群 2 | 15735 | 1.160 | -0.377 | -0.087 | -0.095 | -0.158 |
| 客户群 3 | 12130 | -0.314 | 1.686 | -0.574 | -0.537 | -0.171 |
| 客户群 4 | 24644 | -0.701 | -0.415 | -0.161 | -0.165 | -0.255 |
| 客户群 5 | 4198 | 0.057 | -0.006 | -0.227 | -0.230 | 2.191 |

*由于 K-Means 聚类是随机选择类标号，因此上机实验得到结果中的类标号可能与此不同

# 分析方法与过程

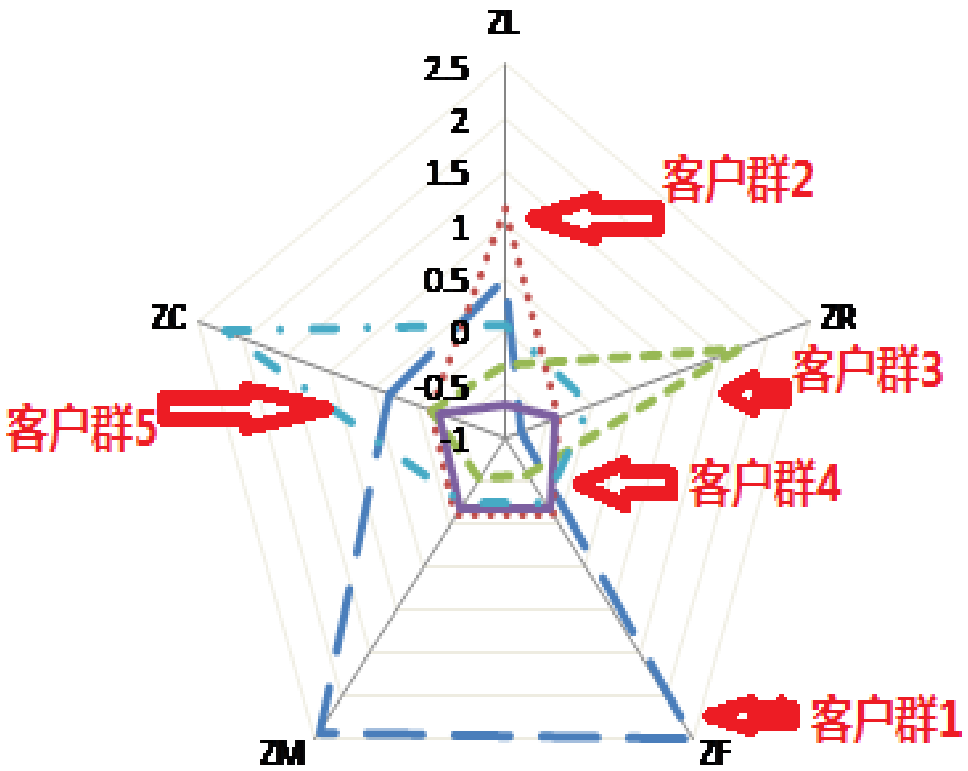
## 第4步：构建模型

### 2. 客户价值分析

对聚类结果进行特征分析，其中客户群1在MF、属性最大，在R属性最小；客户群2在L属性上最大；客户群3在R属性上最大，在F、M属性最小；客户群4在L、C属性上最小；客户群5在C属性上最大。

# 分析方法与过程

## 第4步：构建模型

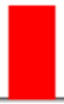
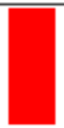


# 分析方法与过程

## 第4步：构建模型

### 2. 客户价值分析

根据业务定义五个等级的客户类别：重要保持客户、重要发展客户、重要挽留客户、一般客户、低价值客户。

| | 重要保持客户 | 重要发展客户 | 重要挽留客户 | 一般客户与低价值客户 |
|---|---|---|---|---|
| 平均折扣系数（C） |  |  |  |  |
| 最近乘机距今的时间长度（R） |  |  |  |  |
| 飞行次数（F） |  |  |  |  |
| 总飞行里程（M） |  |  |  |  |
| 会员入会时间（L） |  |  |  |  |

# 分析方法与过程

## 第4步：构建模型

- 表 7-9 客户群特征描述表

| 群类别 | 优势特征 | | | 弱势特征 | | |
|---|---|---|---|---|---|---|
| 客户群 1 | F | M | *R* | | | |
| 客户群 2 | L | **F** | **M** | | | |
| 客户群 3 | | | | *F* | *M* | R |
| 客户群 4 | | | | *L* | | *C* |
| 客户群 5 | | C | | **R** | F | M |

***注：正常字体表示最大值，加粗字体表示次大值，斜体字体表示最小值，带下划线的字体表示次小值**

- 表 7-10 客户群价值排名

| 客户群 | 排名 | 排名含义 |
|---|---|---|
| 客户群 1 | 1 | 重要保持客户 |
| 客户群 5 | 2 | 重要发展客户 |
| 客户群 2 | 3 | 重要挽留用户 |
| 客户群 4 | 4 | 一般客户 |
| 客户群 3 | 5 | 低价值客户 |

# 分析方法与过程

## 模型应用：根据各个客户群的特征，可采取一些营销手段和策略。

a) 会员的升级与保级。

航空公司的会员可以分为白金卡会员、金卡会员、银卡会员、普通卡会员，其中非普通卡会员可以统称为航空公司的精英会员。成为精英会员都是有一些要求的，比如在入会时间、飞行里程上的指标。但是由于很多客户不了解会员升级报级的要求，白白错失了升级机会，导致客户不满，产生流失的问题。

航空公司可以在对会员升级或保级进行评价的时间点之前，对那些接近但尚未达到要求的较高消费客户进行适当提醒甚至采取一些促销活动，刺激他们通过消费达到相应标准。这样既可以获得收益，同时也提高了客户的满意度，增加了公司的精英会员。

# 分析方法与过程

**模型应用：根据各个客户群的特征，可采取一些营销手段和策略。**

b) 首次兑换。

航空公司常旅客计划中最能够吸引客户的内容就是客户可以通过消费积累的里程来兑换免票或免费升舱等。可以采取的措施是从数据库中提取出接近但尚未达到首次兑换标准的会员，对他们进行提醒或促销，使他们通过消费达到标准。一旦实现了首次兑换，客户在本公司进行再次消费兑换就比在其他公司进行兑换要容易许多，在一定程度上等于提高了转移的成本。另外，在一些特殊的时间点（如里程折半的时间点）之前可以给客户一些提醒，这样可以增加客户的满意度。

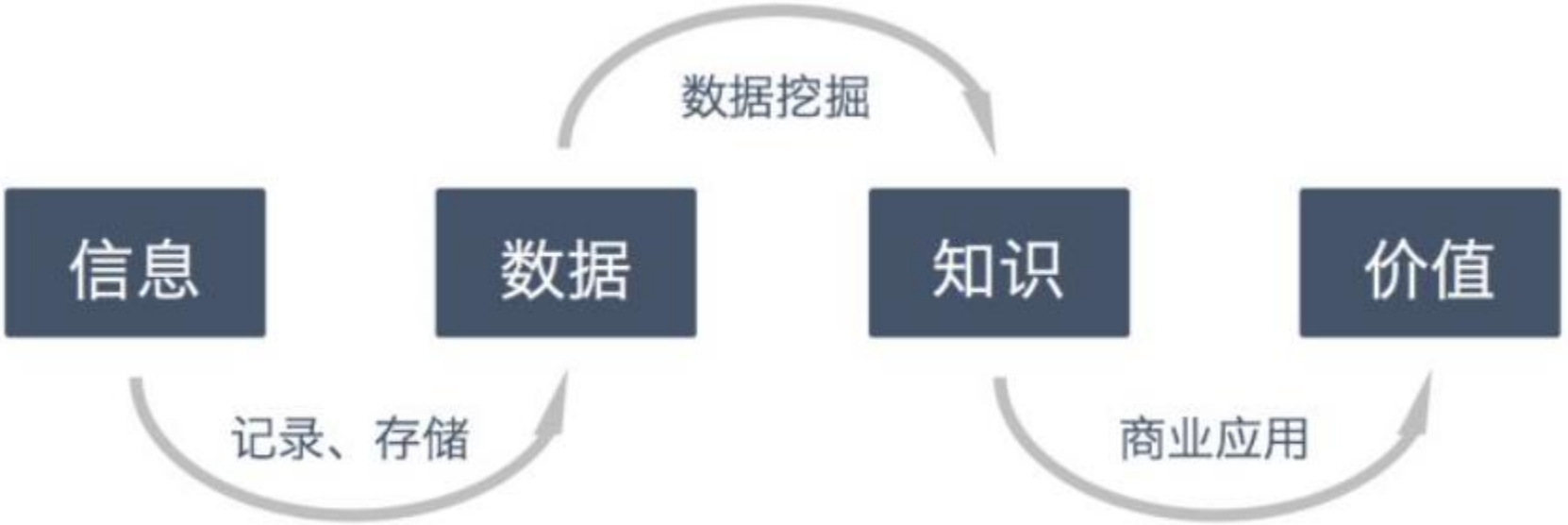
# 分析方法与过程

**模型应用：根据各个客户群的特征，可采取一些营销手段和策略。**

c) 交叉销售。

通过发行联名卡等与非航空类企业的合作，使客户在其他企业的消费过程中获得本公司的积分，增强与公司的联系，提高他们的忠诚度。如可以查看重要客户在非航空类合作伙伴处的里程积累情况，找出他们习惯的里程积累方式（是否经常在合作伙伴处消费、更喜欢消费哪些类型合作伙伴的产品），对他们进行相应促销。

- 通常我们把信息转化为价值，要经历信息、数据、知识、价值四个层面，数据挖掘是中间的重要环节，是从数据中发现知识的过程。



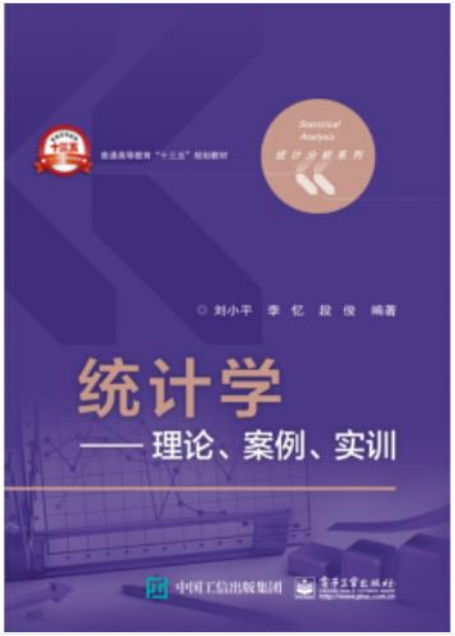
数据挖掘，不是简单的把数据呈现出来，而是要挖掘出数据之间隐藏着的不知道的关系、信息。

***数据是会说话的。***

# 推荐阅读


女士品茶
The Lady Tasting Tea
统计学如何变革了科学和生活
了解统计学的人，运气都不会太差


大数据挖掘
系统方法与实例分析
Using Big Data to Build Your Business


统计学
——理论、案例、实训

# 谢　谢！

## Thank You