『SPSS统计分析基础教程』（第4版）配套PPT使用说明

- 如果你是初上讲台的新手老师需要找感觉，可以直接上b站收看免费的入门篇课程，该课程就是用本套PPT作为课件在授课
  - 本PPT是以素材形式提供，已根据本科教学实际需求删除了部分过于复杂的内容，请各位老师根据情况选用
- 进阶参考资料
  - 系统学习：张文彤, 董伟主编. 《SPSS统计分析高级教程》（第3版）. 高等教育出版社, 2018.1
  - 实战演练：张文彤, 钟云飞, 王清华编著. 《IBM SPSS数据分析实战案例精粹》（第2版）. 清华大学出版社, 2020.5
- 教材配套数据下载及其它需求：wintone_zhang@qq.com

SPSS统计分析教程(2024版)



IBM® SPSS® Statistics

29版

@ 文彤老师

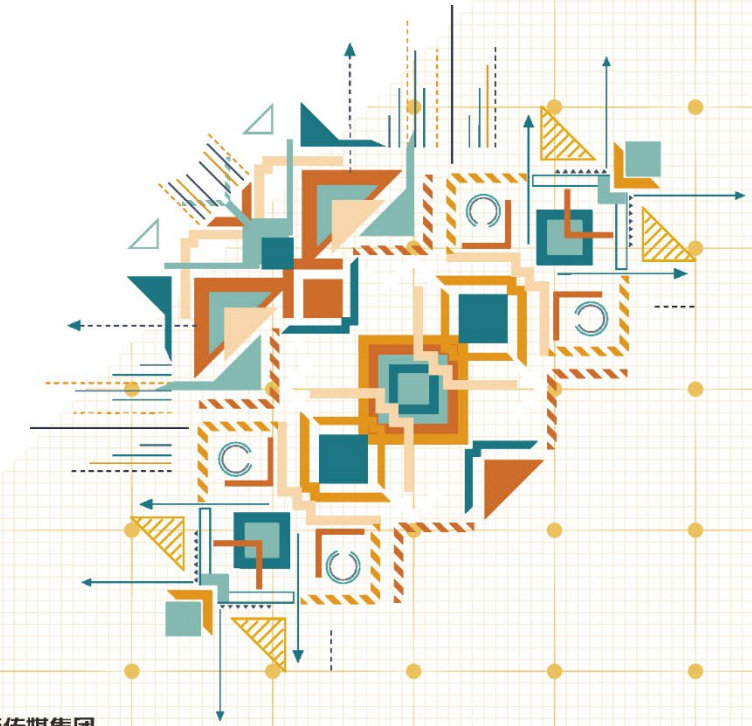
高等学校教材

SPSS统计分析基础教程

Basic Statistical Analysis Using SPSS

（第4版）

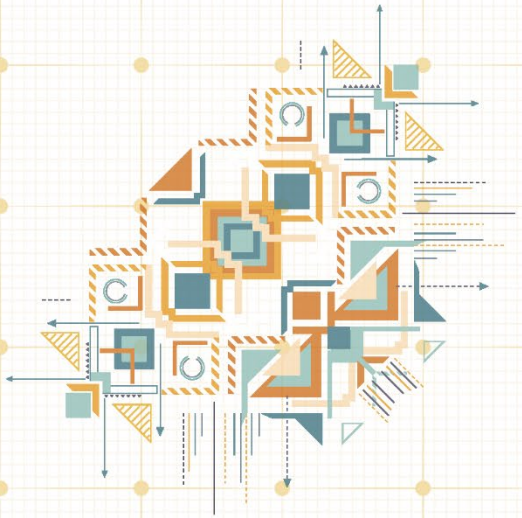
张文彤 编著



中国教育出版传媒集团
高等教育出版社

SPSS统计分析基础教程（第4版）

张文彤 编著



数字课程网站

网址：http://abook.hep.com.cn/1878413
http://abook.hep.edu.cn/1878413

定价 00.00 元

学习统计软件的意义是什么

- 统计分析软件是数据分析的主要工具
- 统计设计完成后，完整的数据分析过程包括
  - 数据的收集
  - 数据的整理
  - 数据的分析
  - 结果的报告
  - 结果的实施与应用（线上系统或线下应用）
- 统计学为数据分析过程提供一套完整的科学的方法论，统计软件为数据分析提供了实现手段

SPSS发展简史

- 软件名称
  - 最早：Statistical Package for Social Science
  - 现在：IBM SPSS Statistics
- 发展历程
  - 60年代：美国斯坦福大学三位研究生研制
  - 70年代：SPSS总部成立于芝加哥，推出SPSS中小型机版－SPSSX
  - 80年代：微机版（V1~4）SPSS/PC+
  - 90年代： Windows版（V6~10）
  - 本世纪：11~29版，中文版，被IBM收购
    - 基本成熟，不再有大的变化

SPSS的基本特点

- 优势
  - 功能强大
  - 兼容性好
  - 易用性强，是应用统计人员的首选统计分析工具
  - 扩展性高（以一种不同的方式）
- 劣势
  - 计算速度相对较慢
  - 在统计模型的纳入上速度较慢
    - SPSS够用吗？

SPSS的C/S结构与模块化结构

表 1.1　SPSS 模块与功能对应表

模块名称	功　　能
Statistic Base	提供最常用的数据管理和统计分析功能。
Advanced Statistics	一般线性模型、混合线性模型、对数线性模型、生存分析等。
Regression	Logistic回归，非线性回归、Probit回归等。
Categories	对应分析、感知图、PROXSCAL等。
Missing Value	缺失数据的报告与填补等。
Conjoint	正交设计、联合分析等，适用于市场研究。
Forecasting	Arima模型、指数平滑、自回归等。
Custom Tables	交互式创建各种表格（如堆积表、嵌套表、分层表等）。
Complex Samples	多阶段复杂抽样技术等。
Bootstrap	提供计算统计学中的Bootstrap方法用于参数估计。
Decision Trees	提供树结构模型分析方法。
Neural Network	提供BP神经网络和RBF神经网络方法。
Data Preparation	提供数据核查、自动清理等一系列数据准备工具。
Statistic Adapter	实际上属于SPSS和Deployment Family产品的接口，可以在企业应用程序、工具和解决方案环境中管理对象的生命周期。
Direct Market	提供了一组用于改善直销活动效果的工具，以针对特定目标群体最大限度地提高促销措施的相应率。

SPSS的四种操作窗口：数据编辑窗口

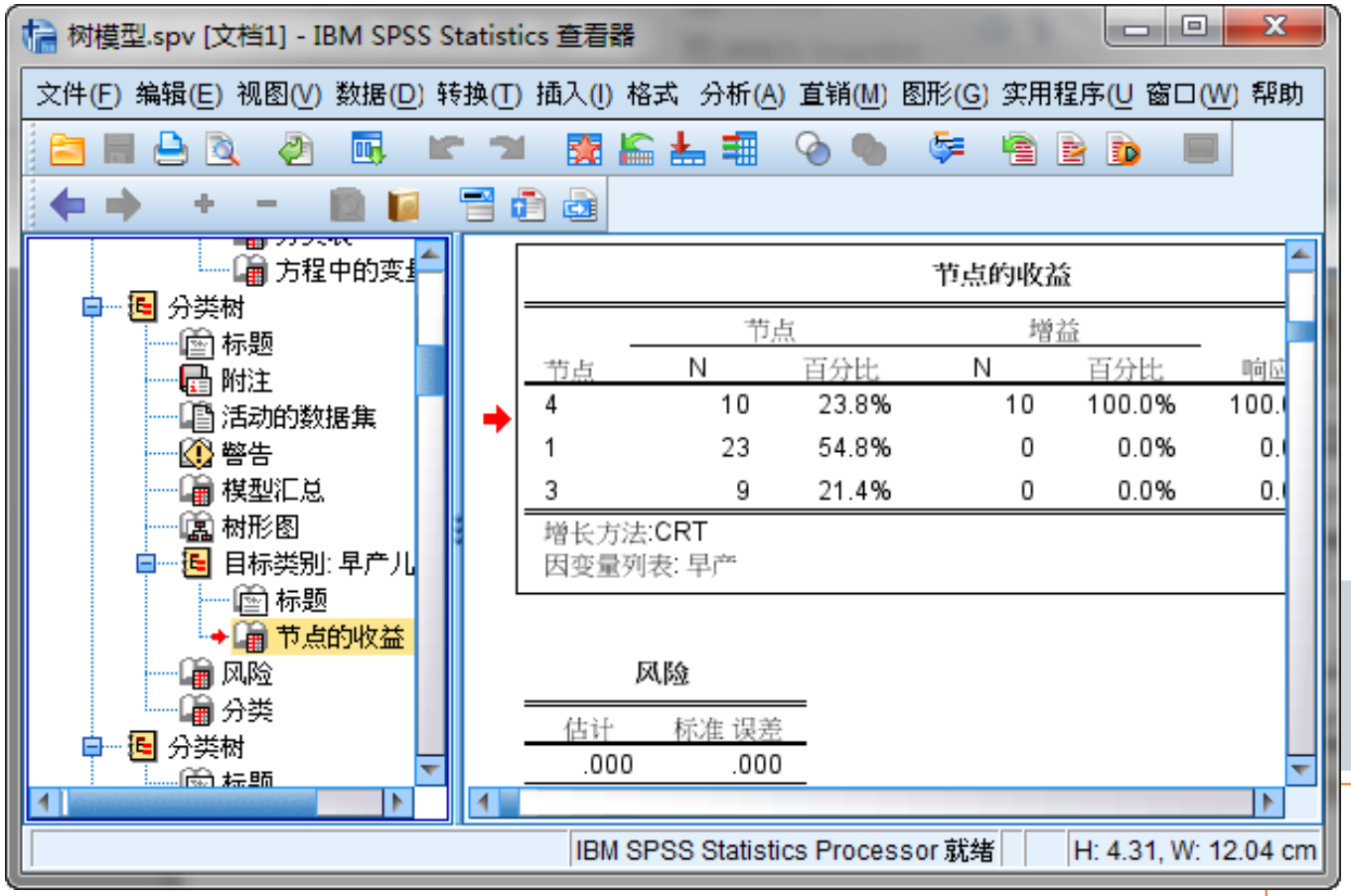
- 数据表格
  - 标准的电子表格
  - 数据视图：主操作界面
  - 变量视图：定义变量
- 单元格编辑器
  - 显示精确数值
  - 不能直接输入公式
- 菜单、快捷工具栏、状态栏
- 同时打开多个数据窗口



	中文变量名	low	age	lwt	race	smoke	ptl	ht	ui
1	85	正常	19	182	黑人	不吸烟	无	无	
2	86	正常	33	155	其他...	不吸烟	无	无	
3	87	正常	20	105	白人	吸烟	无	无	
4	88	正常	21	108	白人	吸烟	无	无	
5	89	正常	18	107	白人	吸烟	无	无	
6	91	正常	21	124	其他...	不吸烟	无	无	
7	92	正常	22	118	白人	不吸烟	无	无	
8	93	正常	17	103	其他...	不吸烟	无	无	
9	94	正常	29	123	白人	吸烟	无	无	
10	95	正常	26	113	白人	吸烟	无	无	
11	96	正常	19	95	其他...	不吸烟	无	无	
12	97	正常	19	150	其他...	不吸烟	无	无	
13	98	正常	22	95	其他...	不吸烟	无	是	

SPSS的四种操作窗口：结果浏览窗口

- 统计软件里最美观的结果输出
- 提供类似资源管理器的界面和操作方式
- 可对内容进行
  - 加密保存
  - 隐藏
  - 修改编辑
  - 在线动态展示



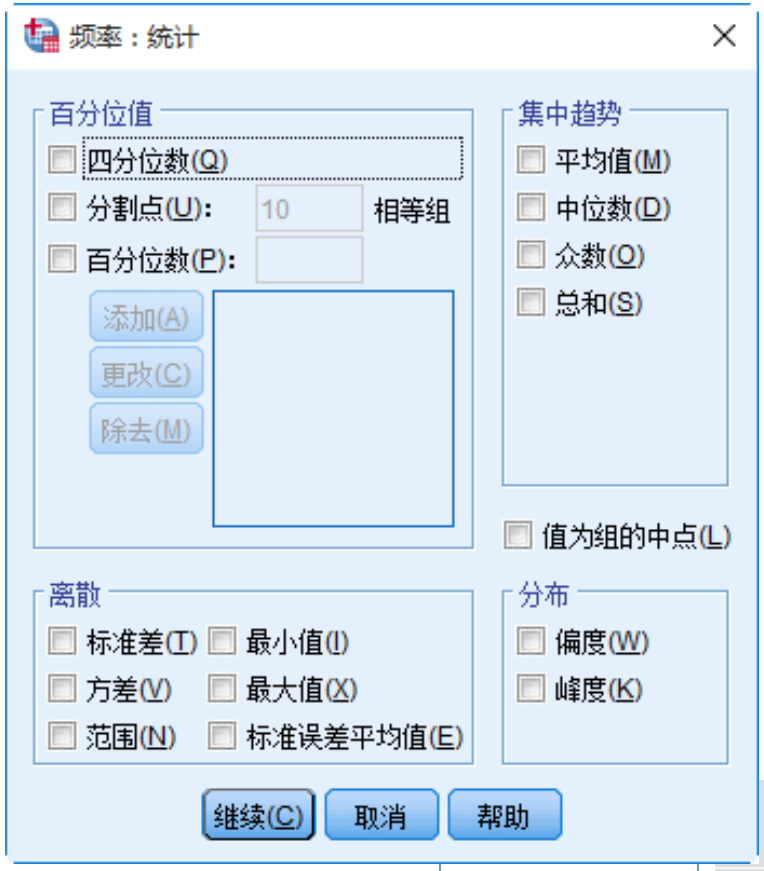
SPSS的四种操作窗口

- 语法窗口
  - 编写SPSS程序时使用
  - 功能集中体现在菜单“RUN”中
- 脚本窗口
  - 高级用户使用basic或者python语言进行SPSS二次平台开发时使用
  - Script菜单：提供了基本的开发功能
  - Debug菜单：提供程序调试的常用功能

统计分析软件的基本使用方式

- 人机交互对话方式：Stata为代表
- 批处理（程序）方式：SAS为代表
- 菜单方式和对话框方式：SPSS为代表





SPSS的四种运行方式

- 菜单对话框方式
  - 最常用，最好记，重用性差，效率最低
- 程序方式
  - 适合于批量数据分析（建议了解基本知识）
- Include命令方式
  - 适合于较大规模的分析项目
- SPSS Production Facility方式
  - 适合于分布式环境中在服务器上执行分析

SPSS的四种结果输出

- 枢轴表
  - 增强的多维立体表格，拥有动态展示功能
- 文本格式
  - 标准的RTF文本
- 统计图
  - 具有动态展示功能
  - 每几年图形引擎都会出现大的变化，幸好操作方式基本保持稳定
- 模型
  - 具有更强大的分析结果展示功能

SPSS的菜单项

- 文件
- 数据与转换
- 分析
- 图形
- 帮助

SPSS对话框操作基本规范

- 候选变量列表框
  - 测量尺度+标签+变量名方式
- 工作变量列表框
- 变量移动钮
  - 根据当前框的状态不断变化
  - 也可直接使用鼠标拖放操作



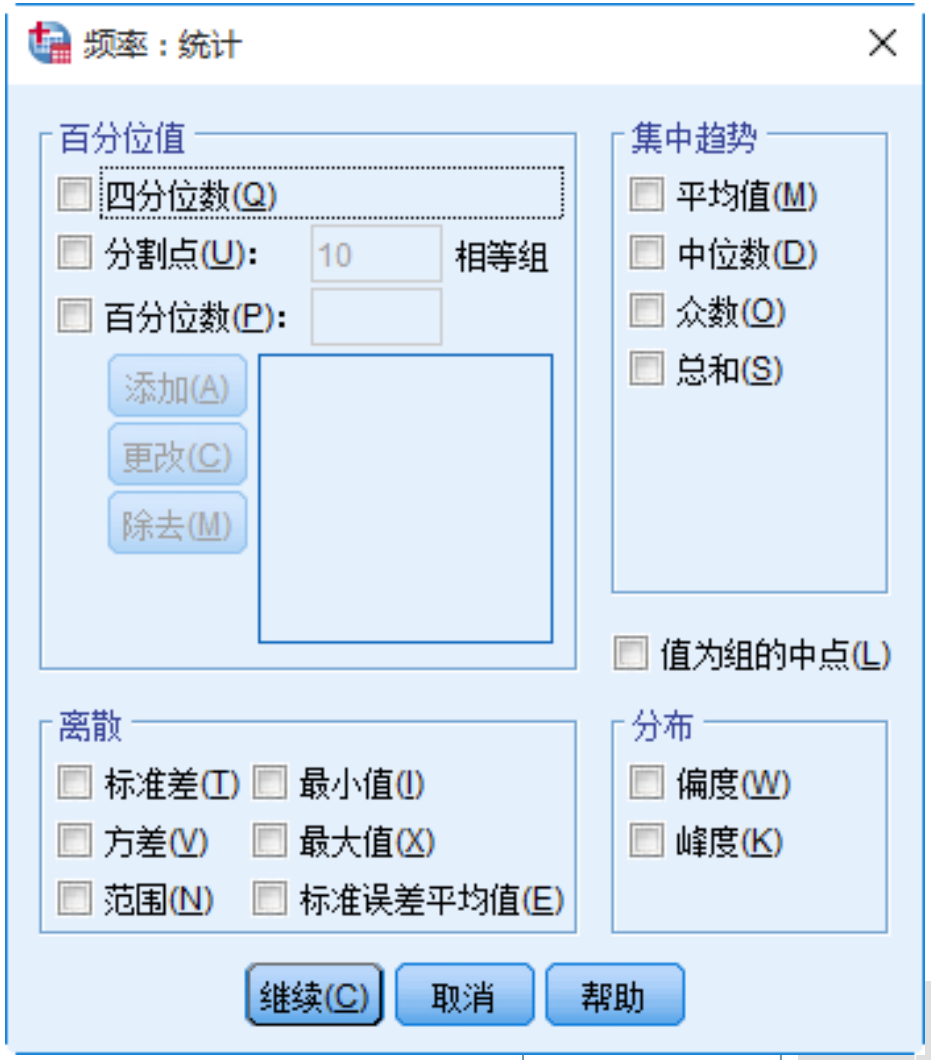
SPSS对话框操作基本规范

- 五个标准按钮
  - 确定、取消
  - 粘贴：用于自动生成SPSS程序
  - 重置：恢复对话框为初始状态
  - 帮助：方法简介、用法简介
- 其他按钮
  - 多数情况下会弹出二级对话框
  - 灰色表示不可用



SPSS对话框操作基本规范

- 单选框（组）
  - 一般均成组出现，多选一
- 复选框（组）
  - 可成组，也可单独出现，多选多
- 下拉列表框
  - 功能实际上和单选框组相同
- 文本框
  - 往往会和单选框、复选框结合起来使用
- 框组
  - 各种元素的集合，集中完成某项功能



SPSS结果窗口操作

- 结果文件的打开和保存
- 结果文件的导出操作
  - 注意中文兼容性问题，必要时需要对软件进行设置
- 结果内容与office软件的交互操作
  - 选择性粘贴
- 结果内容的移动和删除
  - 展开和折叠
  - 标准的拖放式操作
  - 在大纲视图中操作最方便

案例：消费者信心指数

- 消费者信心 (Consumer Confidence or Consumer Sentiment) 是指消费者根据国家或地区的经济发展形势，对就业、收入、物价、利率等问题综合判断后得出的一种看法和预期
  - 居民在决策家庭的开销，特别是购买诸如住宅、汽车等耐用商品时，是经过深思熟虑才决定的。不仅是以目前的经济条件为基础，同时结合了对未来收入、就业、物价、利率变动等因素的判断和预期
  - 理论体系最早由美国密歇根大学调查研究中心的乔治•卡通纳(George Katona) 在上世纪40年代后期提出
  - 六十多年的时间已经充分验证了上述理论在解释宏观经济现状时的有效性与合理性

对中国消费者信心指数的需求

- 中国经济的重要性逐渐增高
- 国外对中国官方统计数据的质疑持续存在
  - 地方数据和全国数据的较大差异
  - 过高和过低数据时的修正（这个全世界其实都一样）
- 官方统计信息对于商业分析需求的匹配度较差
- 第三方数据：一大堆烂番茄！

中国消费者信心调研的研究方法

发布频率	每月一次
数据采集方式	CATI (电脑辅助电话访问)，随机抽样，第一手调研
样本选择	18-64岁中国城市居民，家庭开支或投资决策者
	抽取中国经济最发达的30个城市
	代表中国1/7的人口与1/3的GDP
访谈时间长度	10~15 分钟
问卷问题数量	不少于30题
样本量	每月不少于1,000样本

中国消费者信心指数研究问卷

- S0 受访者所在城市：
  - 100 北京 200 上海 300 广州
- S1 请问您贵姓是？____
- S2 记录被访者性别：
  - 1 男性 2 女性
- S3 请问您的十足年龄是？____
- S4 请问您的学历是？
  - 1 初中/技校或以下 2 高中/中专 3 大专 4 本科 5 硕士或以上

中国消费者信心指数研究问卷

C0 请问您的家庭目前有下列还贷支出吗？

C0_1 房贷 1 有 2 无 99 拒答

C0_2 车贷 1 有 2 无 99 拒答

C0_3 其他一般消费还贷 1 有 2 无 99 拒答

A3 首先，请问与一年前相比，您的家庭现在的经济状况怎么样呢？是变好、基本不变还是变差？

1 明显好转 2 略有好转 3 基本不变
4 略有变差 5 明显变差 9 说不清/拒答

A3a 为什么您这样说呢？（最多限选两项）___

0 中性原因

10 改善：收入相关

20 改善：就业状况相关

30 改善：投资相关

40 改善：家庭开支相关

50 改善：政策/宏观经济

90 不知道/拒答

110 恶化：收入相关

120 恶化：就业状况相关

130 恶化：投资相关

140 恶化：家庭开支相关

150 恶化：政策/宏观经济相关

指数值的计算方法

- 指数的计算基于所对应题目的汇总进行
  - 1~5分别转换为0~200，3分转换为中值100
  - 拒答选项9也直接转换为100
- 在得到相应题目的得分之后，将用于指数计算的各题目得分直接相加，然后除以“基线”调查时的这一数值，即得到相应的指数值
  - Qa3 + Qa4 + Qa8 + Qa10 + Qa16
- 因此，所计算出的指数代表的是当期数值相对于“基线”调查数值的变动比例。如果乐观人群的比例高于“基线”，则指数大于100(100%)，反之，则小于100
  - 作为基线水平的是2007年4月的数值

如何在统计软件/数据管理软件中录入数据?

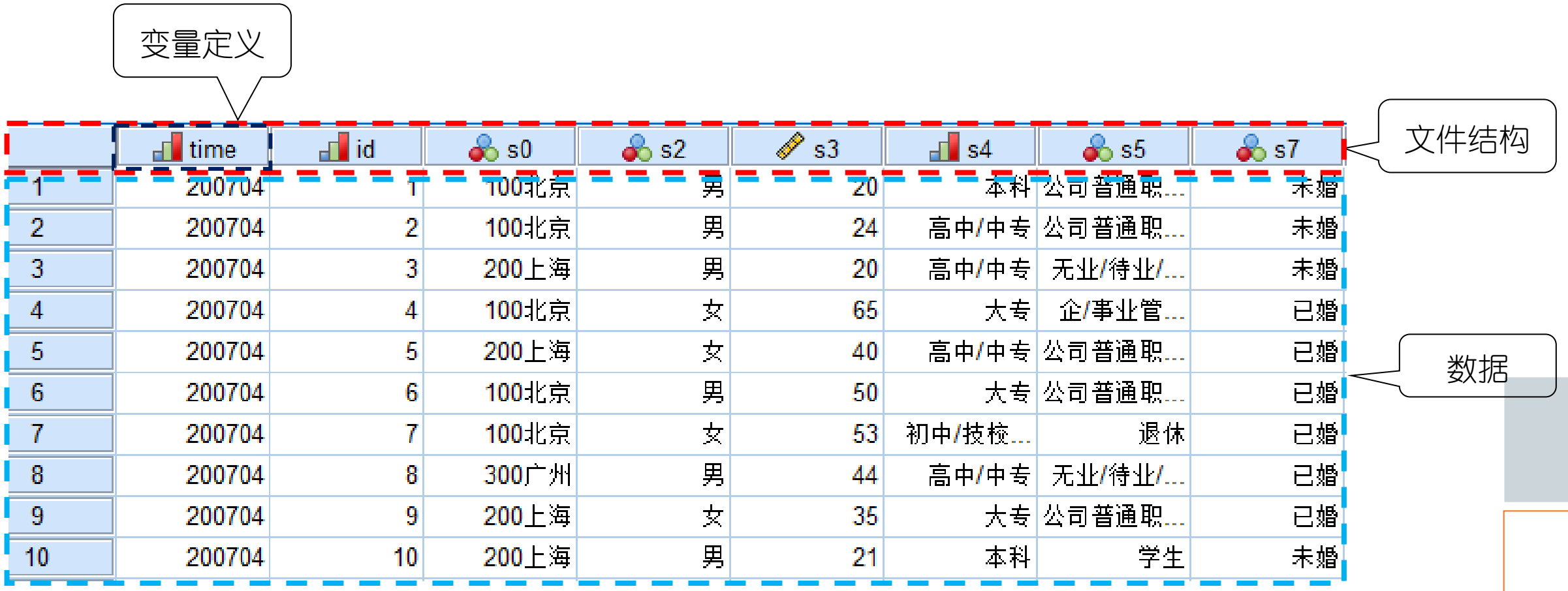
- 纸上谈兵，设计变量
- 实现蓝图，定义变量
- 脚踏实地，输入数据
- 防患于未然，存盘以免断电

统计分析方法，都是以关系型数据库的二维表作为基本数据结构的

所有的统计软件也都以二维表作为数据存储和分析的基础



统计分析所用的二维表数据结构



	time	id	s0	s2	s3	s4	s5	s7
1	200704	1	100北京	男	20	本科	公司普通职...	未婚
2	200704	2	100北京	男	24	高中/中专	公司普通职...	未婚
3	200704	3	200上海	男	20	高中/中专	无业/待业/...	未婚
4	200704	4	100北京	女	65	大专	企/事业管...	已婚
5	200704	5	200上海	女	40	高中/中专	公司普通职...	已婚
6	200704	6	100北京	男	50	大专	公司普通职...	已婚
7	200704	7	100北京	女	53	初中/技校...	退休	已婚
8	200704	8	300广州	男	44	高中/中专	无业/待业/...	已婚
9	200704	9	200上海	女	35	大专	公司普通职...	已婚
10	200704	10	200上海	男	21	本科	学生	未婚

数据的统计格式与日常数据记录格式

	组1测量值	组2测量值
1	12	45
2	24	34
3	16	26
4	23	37
5	.	25

	测量值	组别
1	12	1
2	24	1
3	16	1
4	23	1
5	45	2
6	34	2
7	26	2
8	37	2
9	25	2

数据文件结构设计的基本原则

- 不同观察对象的数据不能在同一条记录中出现
  - 同一个观察对象的数据应当占同一行
- 不同测量指标/影响因素的数据不能在同一列出现
  - 即同一个指标的测量数值都应当录入到同一个变量列中去
- 最终的数据集应当能够包含原始数据的所有信息
- 可以出现违反这些原则的例外情况
  - 重复测量数据出于分析模型的需求，会违反上述要求

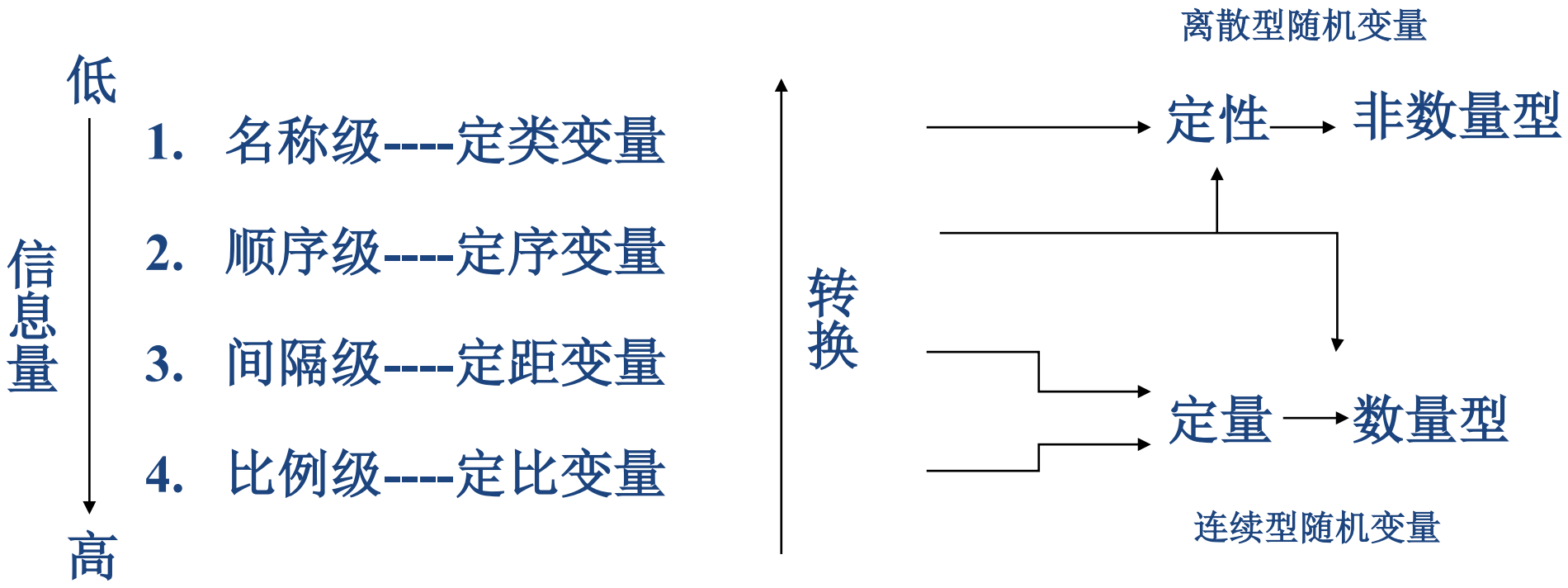
数据文件结构设计的具体思路

- 对于调查表数据的录入，往往一张调查表对应了一个受访者，大多数情况下也就对应了一条案例
- 非调查问卷数据，需要根据具体的情况设计变量结构
- 出于数据管理的便捷性考虑，很可能在出现重复测量时，将每次测量单独考虑，从而形成多个（变量列不一致的）数据文件；或者将每次测量看作单独的变量，形成重复结构的记录格式

变量的测量尺度

- 测量尺度
  - Level of Measurement，指用怎样的精确程度来测量所感兴趣的指标
- 名义尺度（Nominal Measurement）
  - 等价于无序多分类
- 顺序尺度（Ordinal Measurement）
  - 默认字母顺序表示大小顺序，等价于有序分类
- 标度尺度（Scale Measurement）
  - 定距尺度（Interval）：如温度
  - 定比尺度（Ratio）：如体重
- 问题：二分类是什么测量尺度？

变量的测量尺度


低
信息量
高
1. 名称级----定类变量
2. 顺序级----定序变量
3. 间隔级----定距变量
4. 比例级----定比变量
转换
离散型随机变量
定性
非数量型
定量
数量型
连续型随机变量

定义变量的核心属性

- 变量名
- 变量存储类型
  - 数值型：应用最为广泛
  - 字符型：由于分析、整理都较困难，建议尽量少用，改为编码录入
  - 日期型：实际上是特殊的数值型变量，尽量少用
- 变量测量尺度
  - 新对话框体系要求测量尺度设定完全正确
  - 模型类对话框、统计图、统计表对话框

定义变量的其他附加属性

- 变量名与变量值标签
  - 可用于对变量及变量值含义进行说明，使结果更易于阅读
- 缺失值
  - 除非问卷中有特殊编码，否则按默认情况处理
- 其它设置
  - 均按默认值处理，一般不需要更改

CCSS项目的数据录入：开放题和单选题

- 开放题的录入
  - ID：可考虑使用数值变量
  - 姓名：足够长度的字符串变量
    - Unicode问题
- 单选题的录入
  - 原始字符串方式
  - 字符代码 + 值标签方式
  - 数值代码 + 值标签方式
  - 带有半开放选项的单选题录入
    - 注意命名规律

CCSS项目的数据录入：多选题

- 多重二分法（C0）
  - 每个选项单独用一个二分类变量列来记录，可直接单独分析每个选项
- 多重分类法（A3a）
  - 更节约存储空间
  - 可以很明确的得知受访者的选项选择顺序
- 半开放多选题的处理方式

我对数据录入软件的看法

- 数据量较少时：EXCEL或SPSS
  - 注意：EXCEL对变量类型不能正确控制，容易出错
- 录入质量要求较高时： Epidata等专用软件
- 数据量极大时：利用Access等专用数据库软件平台做二次开发

新建数据文件与打开数据文件

- 软件打开时就默认在内存中新建一个空数据文件
- 数据文件的工作名称与存储名称
- 打开其他格式的数据文件
  - 直接读入常用格式的数据文件
  - 利用通用数据库ODBC接口读取数据文件
  - 使用文本导入向导读入文本文件

保存数据文件

- 首次保存与另存
  - 善于使用^S
- 加密保存
  - 注意有遗失密码的风险
- 筛选需要保存的变量
- 存为其他格式
  - 注意兼容性，很可能会丢失许多变量设定
  - XLS：是否保存标签而不是原始数值
  - SAS：将标签设定保存为程序文件

数据管理包括的内容

- 信息的录入和保存
- 数据的整理/清理
- 数据的排列格式变换
- 数据文件的拼接、拆分
- 数据信息的脱敏/匿名化
- 数据文件及相关信息的建档
- 数据的预处理

转换（Transform）菜单的功能设定

- 集中了针对变量级别的信息进行处理的功能
  - 对原始数据进行四则运算
  - 对数据重新编码，包括分段、编秩、匿名化、哑变量化等
  - 对时间序列数据做预处理，包括生成序列、差分运算、缺失值填补等
  - 针对建模需求对数据做自动准备
- 这些功能往往在统计分析的预处理中起着非常重要的作用

数据（Data）菜单的功能设定

- 对整个数据集做加工整理
- 对数据集进行操作
  - 简单命令：插入变量、插入个案、到达某条个案、复制数据集等
  - 常用的简单过程：排序、拆分文件、个案筛选和个案加权，其功能并不复杂，但使用的极为频繁
- 数据格式重组与合并
  - 数据重组向导：用于对重复测量数据进行长型、宽型记录格式转换
  - 文件合并向导：包括横向合并和纵向合并两种情况

数据（Data）菜单的功能设定

- 数据质量控制
  - 重复个案查找向导，异常个案查找向导，比较数据集
  - 数据字典相关向导界面
  - 数据验证模块
- 与统计模型密切相关的过程
  - 正交设计
  - 倾斜权重/得分匹配/个案控制匹配
  - 定义日期时间/定义多重响应集/数据汇总过程等

变量赋值（Compute）

- 用于计算新变量，或者给老变量赋值
- 有大量的SPSS函数可供使用
  - 注意利用函数说明文档
- 学会如何写相应的语句可能更方便
- 实例
  - CCSS项目中的受访者年龄在18-65岁，他们在分析中会被分为18-34、35-54、55-65三组。为便于使用，年龄变量S3重新赋值后将会被存储为新变量TS3，其取值1、2、3分别代表上述三种情况



变量赋值（Compute）操作的复杂用法

- 含有条件组合的赋值方式
  - & | () ~=
  - 例：标示出未婚男性案例
- 字符型变量的操作
  - STRING var1 (A8).
  - concat(string(var1,F1),string(var2,f1)).
- 缺失值的处理
  - If missing(varA) varB=1.
- 几个特殊赋值符号
  - $casenum $sysmis

已有变量值的重编码

- 重编码（Recode）
  - 可直接替换原变量，也可重编码为新变量
- 对连续变量进行分组
  - 有重叠界值时的处理
  - Else
  - Missing、Sysmis
- 分类变量类别的合并
  - 需要注意输出类型

其余的数值重编码操作

- 变量自动重新编码
  - 可用于将字符变量转换为从1开始的数值变量
- 指定数值的查找与计数
  - 可按照筛选条件快速将所有案例分为两类

数值变量的离散化分段（分箱）

- 可视分箱
  - 希望用简单的方法进行分段时使用
  - 在可视界面下将连续变量进行分段
  - 可以使用百分位数、标准差范围或者等间距方式进行分段操作
- 最优分箱
  - 对前述可视化分段的进一步自动化
  - 用于建模分析前对连续变量的最优分段方式进行探索
  - 根据某些作为“关键指示变量”的分类变量，将原有的一个或多个连续变量按照该分类变量类间差异最大化的优化原则离散化为分类变量

文件级别的数据管理-- Data菜单(一)



案例排序

- 排序的用途
  - 将案例按ID变量排序，有利于查找、修改
  - 将某个变量按升序或降序排列，可以非常容易的发现输入错误，因为他们往往就是最大/最小值
  - 缺失值在排序中会排在最小值的前面，可以通过排序的方法发现哪些记录为缺失
- 排序的实现
  - 数据视图，变量名处右键菜单
  - 数据➔个案排序

变量排序

- 将原有的变量顺序保存为新属性
- 菜单方式与变量视图右键操作方式

拆分文件

- 拆分文件时的可选操作
  - 不拆分文件
  - 按所选变量拆分文件，各组分析结果紧挨在一起便于相互比较
  - 按所选变量拆分文件，各组分析结果单独放置
- 拆分结果可以保存，因此使用后要注意取消
- 将拆分结果直接保存为多个文件

选择个案

- 按要求筛选出数据集的一部分
- 可以使用组合条件进行筛选
- 可以直接在数据集中抽样
- 较老的版本只能使用数值或日期变量进行筛选
  - 字符型变量可以先使用自动重编码过程进行转换

个案加权

- 分类数据的录入格式
  - 枚举格式

	name	sex
1	张三	男
2	李四	男
3	王二麻子	女

  - 频数格式

	sex	count
1	男	32
2	女	20

- 只有先进行加权设定，统计软件才能正确识别和处理数据

数据汇总

- 根据要求将原数据文件分类汇总，相应的信息合并入原数据文件，或者成为新的数据文件以便进行分析
  - 当需要生成汇总指标作为变量变换的中间步骤时，该功能非常有用
- 例：分Time月份和S0城市对CCSS案例数据中的变量index进行均数汇总，并将结果输出到新数据文件Sum_index1中
- 例2：分Time月份和S0城市计算出每个组的变异系数
- 例3：分Time月份和S0城市计算出每个受访者的信心值index1和分组均数的差值

文件级别的数据管理-- Data菜单(二)



数据文件的纵向合并

- 从外部数据文件中增加案例到当前数据文件
- 相对而言比较简单，操作上没有需要特别注意之处
- 对话框界面下只能两两合并，多个文件一次就完成合并需要写程序
- 实例：将文件a.sav和b.sav合并

	id	sex	age	height
1	2	1	16	158
2	4	1	34	164
3	6	2	56	170
4	8	1	68	172
5	10	2	25	178

	id	sex	age	h	w
1	1	2	19	165	53
2	3	1	30	175	70
3	5	2	28	162	48
4	7	1	44	169	68

数据文件的横向合并

- 从外部数据文件增加变量到当前数据文件
- 默认是按照记录号顺序对应起来合并，由于在许多情况下是需要按照某个ID变量取值相同的原则进行对应和合并，此时就存在是否正确对应的问题，需要加以注意
- 实例：将文件a.sav和c.sav合并

	id	sex	age	height
1	2	1	16	158
2	4	1	34	164
3	6	2	56	170
4	8	1	68	172
5	10	2	25	178

	id	shape	weight
1	2	169	46
2	6	182	92
3	8	169	51
4	12	175	70

标识重复个案

- 数据管理中的常见需求，基于基本原理，可以自行用程序实现
- 相应功能被整合到一个对话框中实现，使用上更方便
  - 通过简单的菜单操作，用户可以迅速地发现个别变量值重复，或者所有数值完全重复的记录
- 例：在CCSS_bad.sav中，按照time、id均相同的标准查找重复记录

标识异常个案

- 基本原理
  - 聚类：采用两步聚类法，将所有个案按照其相似性自动分为若干类
  - 评分：使用该聚类模型，对每一个案例进行其相对于所属类别的异常度评估，并计算出相应的异常索引，索引值最高的一部分将被识别为异常案例
  - 报告：对所识别出的每一个异常案例，列出基于偏差度指标筛选出的导致异常的具体变量情况
- 注意：默认只显示异常信息最明显的一个变量作为标识变量
- 例：根据index1，index1a，index1b三个变量的取值在CCSS数据中标识异常个案

什么是数据字典

- 在大型数据分析项目中，由数据管理员事先定义好的非常详细的数据格式，包括变量格式、变量标签、值标签、缺失值定义
- 数据字典是使用者定义和维护数据库文件的标准，在长期项目的维护中起着**至关重要**的作用
- 在SPSS中，数据字典其实就是一个数据文件
  - 只有结构没有数据的空文件
  - 有预实验数据存储在内的一个实际数据文件
- SPSS中与数据字典相关的功能
  - 定义变量属性、复制数据属性、新建定制属性、设置未知测量级别

已有定义好的数据字典

- 生成一个空数据文件，将全部的数据字典定义设定好
- 使用新建定制属性设定默认情况下软件未提供的字典设定属性
- 在批量录入数据的过程，使用复制文件属性向导套用该字典，并检查是否有不合规数据存在，必要时进行字典的更新
- 在数据全部录入完毕后再次使用复制文件属性向导套用该字典

没有定义好的数据字典

- 在录入一批数据后使用变量属性定义向导完成数据字典的设定工作
- 随着录入工作的进行经常扫描数据的情况，以及时更新字典
- 在录入工作完毕后使用复制文件属性向导应用字典的最终版本
- 只是小项目
  - 可以利用设置未知测量级别向导来快速设定数据字典
  - 也可以直接录入数据后在变量视图中修改属性

使用变量子集

- 大型研究项目的数据集中可能包括上千个变量，分析时直接在上千个变量中进行选择会非常麻烦，此时可以用变量子集简化操作
- 默认已经设定好的变量集
  - ALLVARIABLES：包括所有的变量
  - NEWVARIABLES：包括数据集打开后所有新建立的变量
- 定义变量集与使用变量集

个案控制匹配

- 基于所指定的匹配变量取值近似相等或精确相等的条件，在数据集中为病例匹配符合条件的对照
- 可以为案例同时查找出多个匹配对照
- 匹配容差：匹配时容许的数值误差
  - 字符串变量必须严格相等
  - 数值变量可以容许存在数值误差，具体范围由用户指定
- 可以要求优先进行精确匹配，然后进行模糊匹配

倾向得分匹配（PSM）

- 个案控制匹配可以做到精确控制，但是在需要匹配的变量较多时往往难以实现
- PSM是近些年来比较流行的病例对照二次匹配方法，通过将所有需要匹配的变量转换为一个综合分值进行匹配的方式，解决了多变量匹配的难题
  - 将病例/对照指示变量作为因变量，匹配用变量全部作为自变量来拟合Logistic回归模型
  - 根据该模型为每个案例计算出倾向得分（得分本质上是发病概率的估计值）
  - 根据倾向得分，从对照组中按照得分接近的原则为每个病例选择匹配项

都这年月了，怎么用SPSS还要学编程?

- 实现10%的高级分析功能
- 实现批量分析的自动化和重复化
- 快速实现所需的分析功能
- 让自己显得更像个砖家

那么。。。学点什么呢?

- 基本语法规则
- 宏程序
- OMS系统
- 程序自动化执行

基本语法规则：主命令格式

- 每条SPSS命令必须从新行开始
- 每个命令应该以句点为命令终止符
  - 也可以将空行解释为命令终止符
- COMPUTE NEWVAR = OLDVAR * 2 .
  - 命令动词、分隔符、表达式、终止符

基本语法规则：子命令格式

- 是对命令的进一步说明和设置
- 很多非关键的子命令都会有其默认设定
- FREQUENCIES VARIABLES = var1 var2
  /STATISTICS = MEAN
  /ORDER = ANALYSIS.
  - 子命令名、分隔符、命令顺序

用DATASET命令控制工作数据集

- DATASET ACTIVATE 数据集 .
  - 将指定数据集切换为当前/活跃数据集
- DATASET DECLARE 数据集 .
  - 新建指定名称的数据集
- DATASET NAME 数据集 .
  - 重命名当前活跃数据集
- DATASET CLOSE 数据集 .
  - 关闭指定工作名称的数据集

保留字、系统变量和特殊命令

- 关键字和保留字：var1 to var5
  - 注意这里是按变量顺序，而不是变量名规律来定位！
- $系统变量：$casenum, $sysmis
- 特殊命令：execute

判断语句： DO IF -- END IF

- DO IF (age <20) .
  - COMPUTE ageclass = 1 .
  - COMPUTE younger = 1 .
- ELSE IF (age < 30) .
  - COMPUTE ageclass = 2 .
- ELSE IF (age < 50) .
  - COMPUTE ageclass = 3 .
- ELSE .
  - COMPUTE ageclass = 4 .
- END IF .

循环语句：LOOP -- END LOOP

- 基本结构
  - LOOP 控制变量名=起始值 TO 终止值 [BY 步长]
    - 程序段
  - END LOOP
- 示例：
  - LOOP #LOP = 1 TO 5 .
    - COMPUTE X = X + 1 .
  - END LOOP .
  - EXECUTE .

SPSS程序的三种编写方式

- 直接敲代码。。。是大家都知道的，但是很傻的做法
- 对话框粘贴程序
  - 随写随用，最为方便
- 输出LOG粘贴程序
  - 对过去数小时的分析操作进行代码保存
- 日志文件编辑程序
  - 对过去曾经完成过的分析操作进行代码保存

统计描述中可用的工具

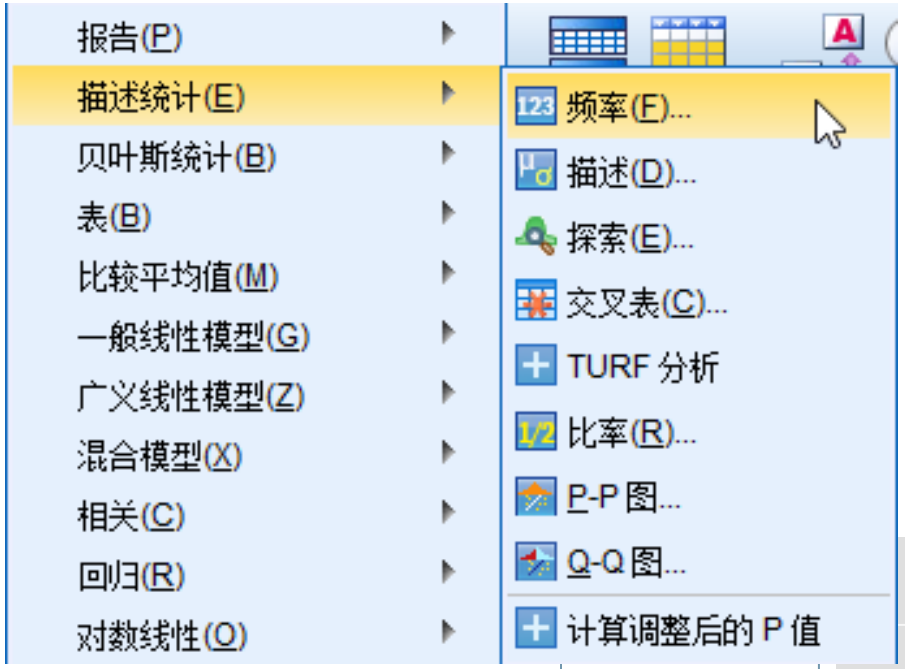
- 各种初步汇总描述方法
  - 分组汇总，最终汇总成相应的分组频数图/表，可以反应出数据的大致趋势
  - 多个百分位数组合起来，也能反应数据的分布特征
  - 对信息的利用仍然比较粗糙
- 各种统计描述指标
  - 实际上是更复杂的各种描述工具的基础，是针对数据的某种特征进行精确的数字呈现的一系列指标
  - 对于样本而言，这些统计描述指标也可被称为统计量
  - 均数、标准差、四分位间距，比、率等

统计描述中可用的工具

- 统计表
  - 当数据比较复杂，所计算的统计指标较多时，直接观察所计算出的数值比较麻烦，为此人们又会按照一定的排列方式将统计指标组织为一张表格，以方便使用
- 统计图
  - 统计表虽然能做到非常精确，但是不够直观，如果希望结果更为直观一些，则可以按照统计指标的大小将其绘制为一张图形，这就是所谓的统计图

SPSS中的统计描述相关功能

- SPSS中所有专业的制表过程都能够完成统计描述的任务
  - 报告（Reports）菜单项：主要输出文本化报表
  - 表（Tables）菜单项：输出专业标准的SPSS结果表格
  - 多重响应（Multiple Response）菜单：专门用于多选题数据描述
  - 比较平均值等其余菜单项
- 但描述统计（Descriptive）菜单项中提供的是最基本的统计描述过程



描述统计（Descriptive）菜单项中的功能

- 频率（Frequencies）过程
  - 计算各种常用的描述指标
  - 特色是产生频数表
  - 对分类变量和连续变量都适用
- 描述（Descriptive ）过程
  - 一般性的统计描述
  - 适用于服从正态分布的连续变量
- 探索（Explore） 过程
  - 用于对数据分布状况不清时的探索性分析，功能极为强大

描述统计（Descriptive）菜单项中的功能

- 交叉表（Crosstabs）过程
  - 完成有序/无序分类变量的统计描述
  - 完成分类变量各种各样“常规”的统计检验
- 比率（Ratio）过程
  - 用于两个连续变量相对比指标的描述，不太常用
- P-P图和Q-Q图
- TURF分析等其余插件

分类变量的统计描述



分类变量的常用描述指标

- 原始数据：
  - 频数列表
  - 百分比
  - 累计频数
  - 累计百分比
- 集中趋势
  - 众数（实用价值不大）

血型

		频率	百分比	有效百分比	累积百分比
有效	A	67	30.6	30.6	30.6
	AB	30	13.7	13.7	44.3
	B	37	16.9	16.9	61.2
	O	85	38.8	38.8	100.0
	合计	219	100.0	100.0	

分类变量的常用描述指标

- 比（ratio）：
  - 任意两个变量之比，这两个变量可以是性质相同的两个指标，如两个地区相同时期内某病新发病例数之比；也可以是性质不相同的两个指标之比
  - 性别比，货物/销售人员比
- 构成比（proportion）：
  - 部分占总体的比例，分观察对象为k个部分（A1、…、Ak），其中某一个部分观察对象数与观察对象总数之比为构成比
- 率（rate）：
  - 某个时期内某个事件发生的频率或强度，实际上有速度的概念在里面，可以和物理中速率的定义相对照

分类变量的联合描述：列联表

- 常见的是二维表，但三维及更高维列联表的使用原理完全相同
- 频数
- 行、列、总百分比
- 行、列、总合计

血型* 性别 交叉制表

			性别		
			男	女	合计
血型	A	计数	16	51	67
		血型 的 %	23.9%	76.1%	100.0%
		性别 的 %	22.2%	34.7%	30.6%
		总数的 %	7.3%	23.3%	30.6%
	AB	计数	8	22	30
		血型 的 %	26.7%	73.3%	100.0%
		性别 的 %	11.1%	15.0%	13.7%
		总数的 %	3.7%	10.0%	13.7%
	B	计数	17	20	37
		血型 的 %	45.9%	54.1%	100.0%
		性别 的 %	23.6%	13.6%	16.9%
		总数的 %	7.8%	9.1%	16.9%
	O	计数	31	54	85
		血型 的 %	36.5%	63.5%	100.0%
		性别 的 %	43.1%	36.7%	38.8%
		总数的 %	14.2%	24.7%	38.8%
合计		计数	72	147	219
		血型 的 %	32.9%	67.1%	100.0%
		性别 的 %	100.0%	100.0%	100.0%
		总数的 %	32.9%	67.1%	100.0%

多选题的描述指标

- 对多选题进行分析时所遇到的最大困难就是各选项间存在着关联，将它们单独进行分析并不恰当
  - 单独分析的结果仅反映选项自身情况，不能完全代表整个题目的情况
- 因此在描述时就需要采用基于频次的百分比等一些特殊指标

多选题的描述指标

- 应答人数与应答人次（Case & Response）
- 应答人数百分比（Percent of Cases）
  - 选择该项的人占总人数的比例
  - 用于反映选项自身被选择的情况
- 应答次数百分比（Percent of Responses）
  - 选择该项的次数占总次数（总反应数）的比例
  - 用于选项间受欢迎程度的比较

分类变量统计描述的SPSS实现

- Frequencies过程
  - 专门为产生频数表而设计，默认选项即可给出需要的结果
- Crosstabs过程
  - 生成二、三维交叉表
  - 默认只输出频数，计算百分比需要更改选项
- Multiple Response菜单
  - 专门为多选题数据的描述而设计
  - 输出格式较为简单，但使用方便
- Tables模块
  - 为制表提供了全面的支持，是对多选题的支持也很完善
  - 也可以顺带完成一些假设检验工作

连续变量的统计描述



频数表/图

- 是一种非常直观的方法，但比较粗糙，如果希望进行深入的分析，则必须要使用一些严谨的统计指标对各方面特征加以描述
- 操作步骤：确定组数、确定组距、确定各组段的上下限
  - 各步骤的操作都存在着一些主观性，但因为只是进行初步的观察，这并无大碍

在频数表中可以观察到的信息

- 集中趋势
- 离散趋势
- 分布形状
  - 是否对称，分布曲线的形状
- 分布特征
  - 单、双峰，有无极端值等

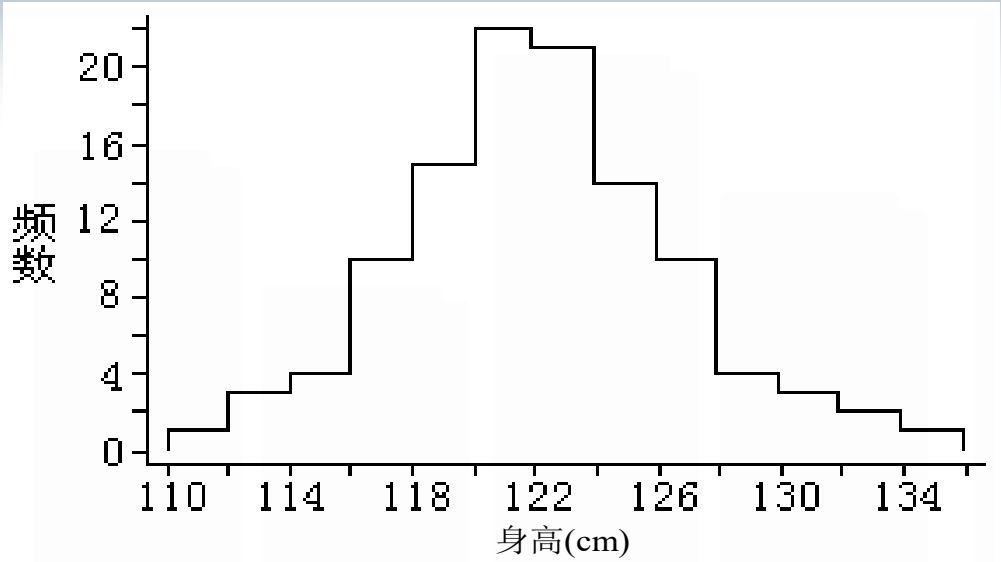


图 2.1　110 名 7 岁男童身高分布

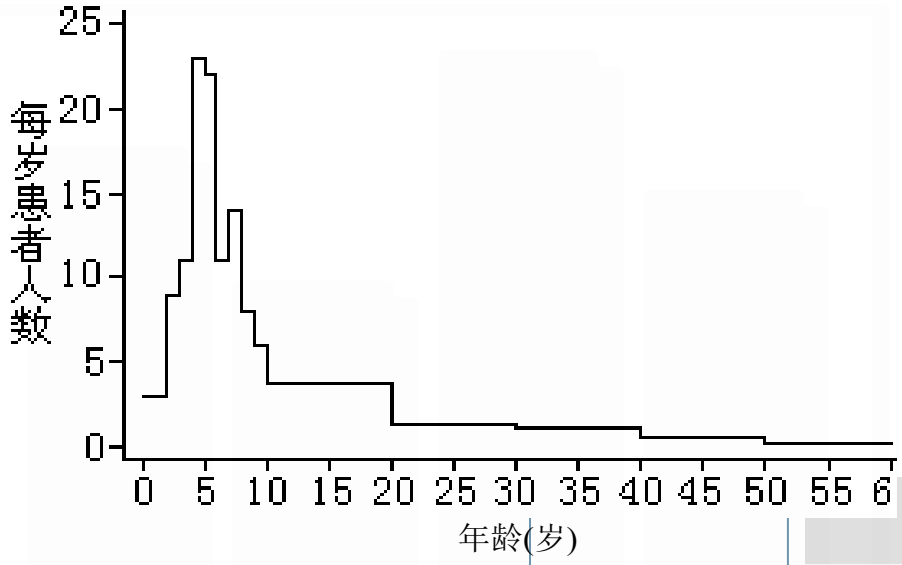


图 2.2　某市某年乙脑患者的年龄分布

集中趋势（Central tendency）

- 均数：mean
  - 描述一组数据在数量上的平均水平，总体均数和样本均数用不同符号表示
  - 适用范围：对称分布，特别是正态分布资料
- 几何均数：G
  - 针对正偏态资料集中趋势的描述而提出
  - 适用范围：对数正态分布资料，但往往被进一步扩大到等比资料

集中趋势（Central tendency）

- 中位数：median
  - 适用范围：万金油
  - 和均数相比较为迟钝，只有样本量较为充足时结果才稳定下来
- 众数：mode
  - 所有数值中出现次数最多的一个
  - 适用范围：国外

离散趋势（Dispersion tendency）

- 全距：Range
  - 适用范围最广，但是也最不稳定
- 方差：Variance
  - 离均差　　描述个体变异
  - 离均差和　　直接相加必然为0
  - 离均差绝对值之和　　数学推导困难
  - 离均差平方和

离散趋势（Dispersion tendency）

- 标准差：std.deviation
  - 解决了阅读时量纲的问题
- 变异系数：coefficient of variation
  - 解决了不同资料间变异程度对比的问题
    - 测量尺度相差太大：比较蚂蚁和大象的体重变异
    - 量纲不同：比较身高和体重的变异程度

百分位数

- 适用于各种分布
- 教科书上传统将其列为集中趋势的描述指标，但实际上不仅限于此，单独列为一类可能更为妥当
  - P2.5 P5 P10 P25 P50 P75 P90 P95 P97.5
- 使用条件：只有样本量较大时结果才比较稳定，对位于两端的百分位数而言更是如此
  - 100例时，P95右侧只有5例！

百分位数

- 四分位数
  - 实际上是P25、P50和P75分位数的总称，他们正好能够将全部总体单位按标志值的大小等分为四部分的三个数值
  - 上(P75)、下(P25)四分位数的差值被称为四分位间距

其他描述指标

- 分布形状描述指标： Distribution
  - 针对某种分布进行进一步的特征描述，主要是用于正态分布
  - 偏度系数Skewness 正态峰 正偏态 负偏态
  - 峰度系数Kurtosis 正态峰 平阔峰 尖峭峰
- 离群值与极端值列表

探索过程

- 可对变量进行更为深入详尽的描述性分析
- 主要用于对资料的性质、分布特点等完全不清楚时
- 常用描述性统计指标/图形更加详细、全面
- 可分组进行描述

探索过程：Statistics 子对话框

- Descriptives复选框
  - 输出均数、中位数、众数、5%修正均数、标准误、方差、标准差、最小值、最大值、全距、四分位全距、峰度系数及其标准误、偏度系数及其标准误，以及指定的均数可信区间。
- M-estimators复选框
  - 作中心趋势的最大似然估计，输出四个不同权重的最大似然估计值。
- Outliers复选框
  - 输出五个最大值与五个最小值。
- Percentiles复选框
  - 输出第5%、10%、25%、50%、75%、90%、95%位数。

探索过程：plot子对话框

- Boxplots单选框组：确定箱图的绘制方式
  - 按组别分组绘制(Factor levels together),
  - 不分组绘制(Dependents together)
  - 不绘制(None)
- Descriptive复选框组：
  - 茎叶图(Stem-and-leaf)和直方图(Histogram)。
- 案例：分月份time对总指数index1进行统计描述，以详细了解其分布情况

用SPSS绘图到底有啥优势?

切合数据分析的需求，基于数据对图形做动态编辑和分析。

对多选题图形分析的完善支持

基于数据的图形种类互转



SPSS为啥老是改绘图引擎?

因为要方便用户，与时俱进呀！

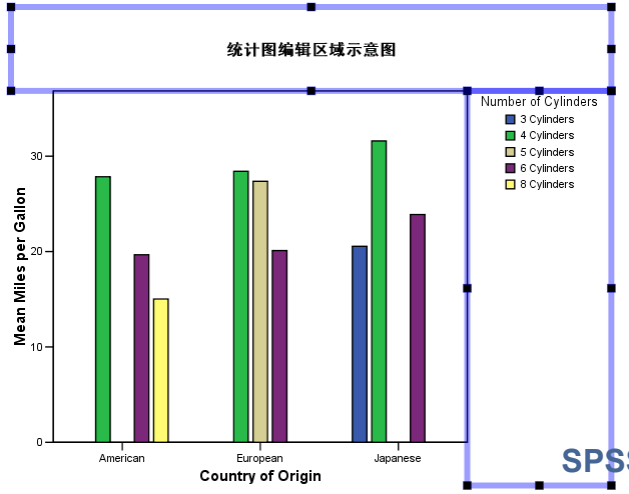


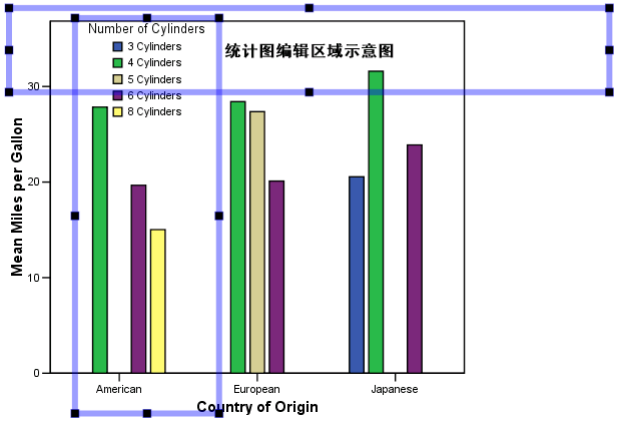
SPSS绘图的三种对话框操作方式

- 可视化的图表构建器：绝对主力
- 图形可视化模版选择器：
  - 是一个类似于绘图向导的可视化界面，很多操作细节非常类似与交互图
  - 实际来自于SPSS Modeler
- 继承自老版本的传统对话框：
  - 方便老用户使用
  - 注意有些特殊图形只能用该界面生成（主要整合在分析菜单中）
- 隐藏的第四种：程序方式
  - 普通用户可以忽略

自由灵活的图形框架

- 外框区和内框区
  - 选中后出现控制框，利用鼠标可进行框区大小的改变和位置的移动
  - 改变外框区大小时将对称的改变除文本外所有图形元素的大小
  - 在改变大小时图形长宽比例不再固定，完全做到了随心所欲
- 其余各区域
  - 区域的相对位置完全自由，可拖动到任意位置，区域内元素的各种设置保持不变



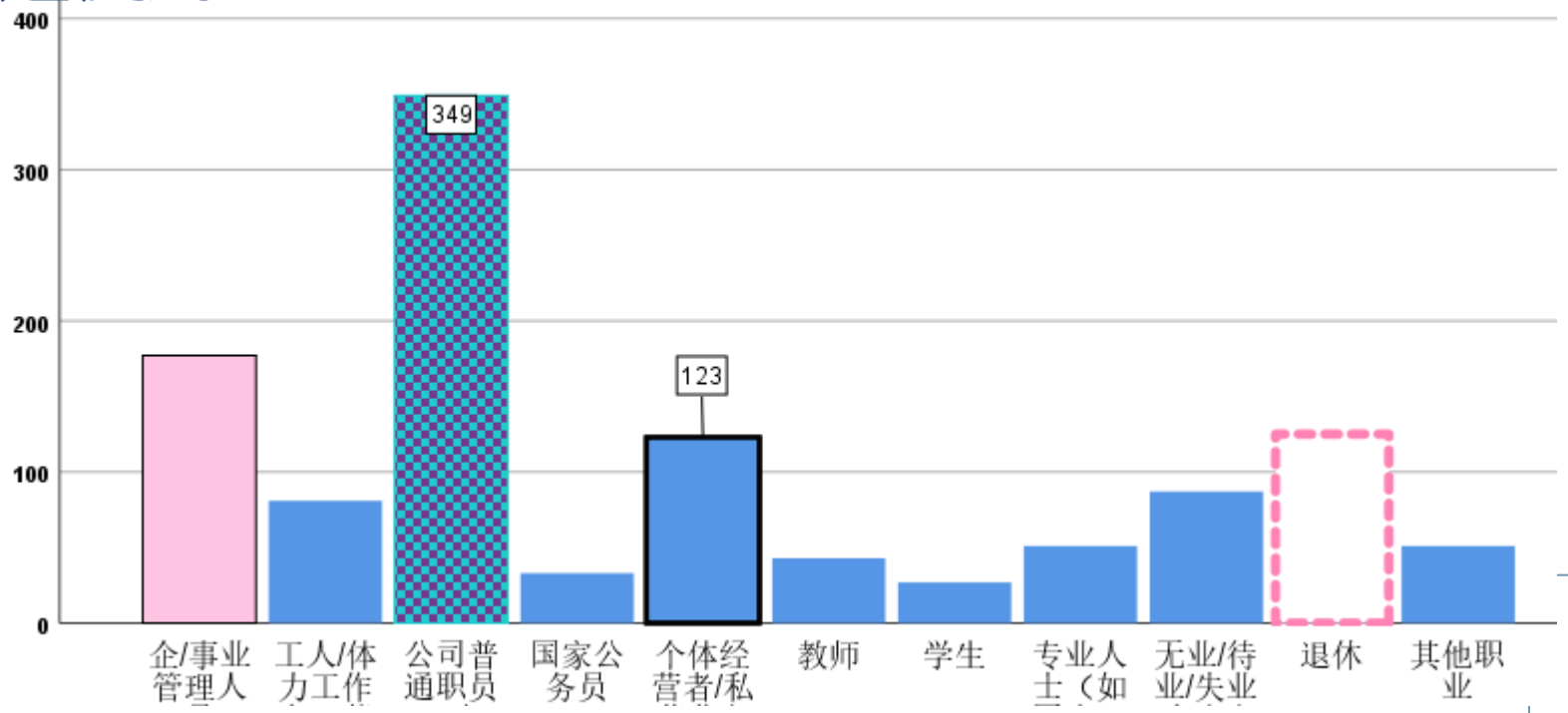


自由的元素选择方式

- 标准方法：
  - 单击图形元素，选中所有同类元素
  - 二次单击，则选中同组元素
  - 三次单击，则只选中该图形元素
- 快捷方法：
  - 有图例时，单击图例则选中所有相应的图形元素
  - 使用套索方式，一次选中多个对象

自由的元素选择方式：可选中即可编辑

- 可以单独设置某一个图形元素的格式
  - 包括单独标出具体的数值、ID号等
- 注意：并非所有操作都可以用于单独的图形元素，许多选项只能整个图形一起使用



直方图与茎叶图

- SPSS统计描述功能中的绘图选项
- 案例：绘制CCSS中年龄和信心指数的直方图
  - 增加参考线
  - 直方图组的绘制
  - 累积直方图的绘制
  - 分段直方图（Stacked Histogram）
  - 面积直方图（频数多边形，Frequency Polygon）
- 茎叶图

箱 图

- 案例：分月份比较信心指数的变化趋势
  - 四分位数将整个取值区间拆分为等样本量的四段
  - 离方框上/下界的距离超过四分位数间距1.5倍的为离群值，以“O”表示
  - 超过3倍的则为极值，用“*”表示
  - 散点旁边默认标出相应案例号备查
- 特色功能
  - 选择性显示某些标签
  - 从绘图本质上讲，每个箱图都是两个条图的组合

饼 图

- 案例：考察职业的分布情况
- 案例：分城市、月份考察样本性别比例
  - 可以考虑将分组设定为面板变量
  - 从使用习惯上讲，列面板变量更易于观察
- 特色功能
  - 旋转方向
  - 饼块的突出显示与合并

条图与误差图

- 案例：比较不同职业人群的消费者信心指数均值
  - 带误差线的条图与误差图
- 分段条图与百分条图案例：比较不同月份的A3a选项比例分布
- 特色功能
  - 条图与其他统计图形的相互转换
  - 复式条图和分段条图的相互转换
  - 分段条图和百分条图的相互转换

线图、面积图、摘要点图与垂线图

- 案例：分城市比较信心指数随时间的变化趋势
  - 北上广三地的信心指数变化规律
  - 三地信心的平均水平
  - 三地差异随时间的变化
- 特色功能
  - 更改数据点的显示方式
  - 更改数据点间的连接方式
  - 突出显示某一段连线
  - 半对数线图

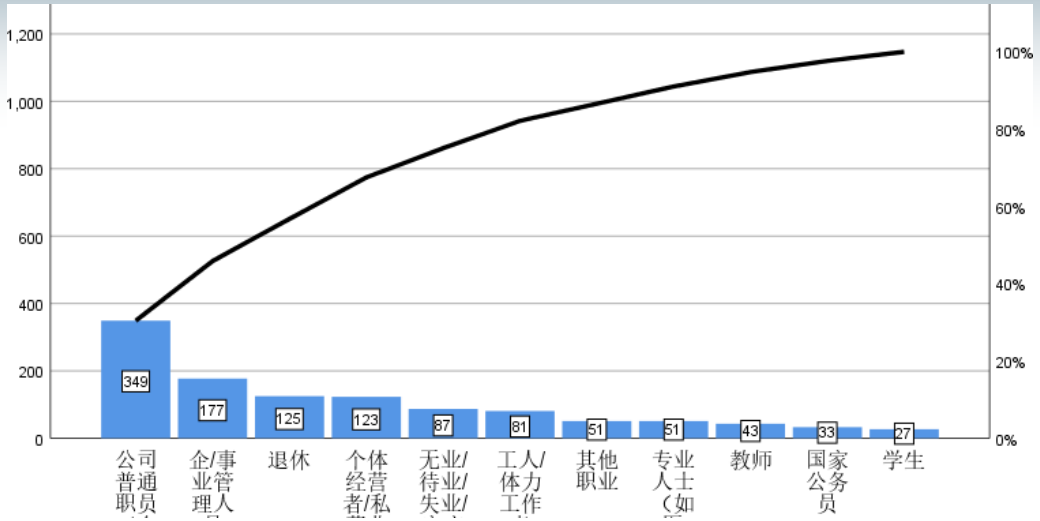
散点图

- 案例：考察年龄与信心指数之间的关系
- 案例：分性别考察年龄与信心指数之间的关系
  - 用拖放操作绘制多变量分组图
- 特色功能
  - 用套索模式选择离群散点
  - 改变过密散点图的显示方式
  - 添加回归趋势线和可信区间
  - 三维散点图的旋转观察

控制图

- 用于分析和判断生产工序是否处于稳定状态的一种统计图
  - 假定生产过程仅受随机因素的影响
  - 产品的质量特征的平均值和变异都基本保持稳定时，称之为处于受控状态
  - 当产品的误差波动超过警戒线时，就需要查明原因，不使失控状态延续而发展下去
- 控制图的分类
  - 计量控制图和计数控制图
  - 个体指标/汇总指标控制图
  - 基于全距/百分位数/标准差范围的控制图

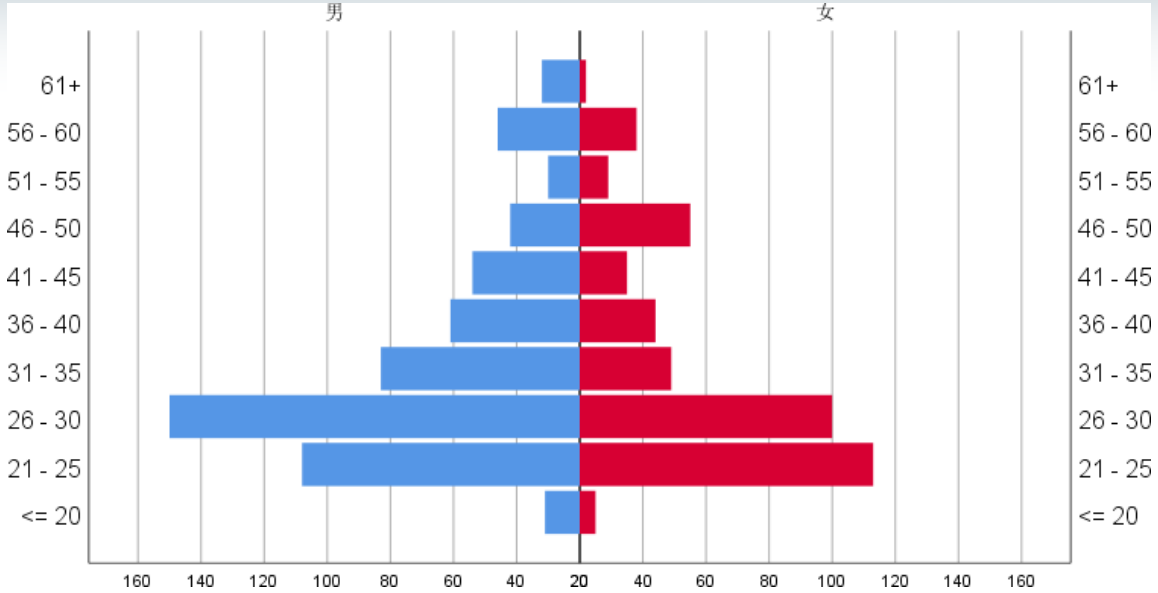
Pareto图



- 以意大利经济学家V.Pareto的名字命名
- Pareto法则：
  - 也称二八原理，或者80/20法则
  - 相对数量较少的原因往往造成绝大多数问题或缺陷
- 将出现的质量问题/质量改进项目按照重要程度依次排列，表示有多少结果是由已确认类型或范畴的原因所造成
- 协助用户区分“微不足道的大多数”和“至关重要的极少数”，便于分析者关注于重要的类别。是质量控制领域进行优化和改进的有效工具

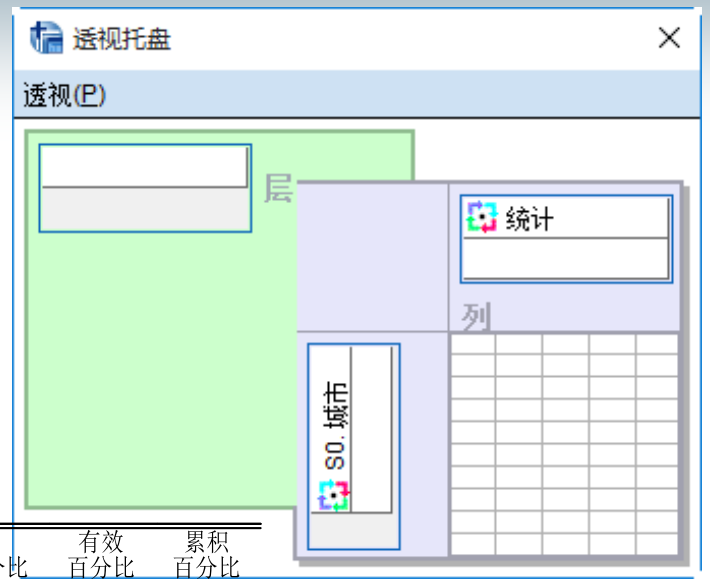
人口金字塔

- 人口统计中用于描述人口年龄和性别分布状况的图形，用于呈现人口现状及其未来发展趋势
- 本质上是横向绘制的男/女两个直方图，然后拼接而成
- 绘制要点在于需要手工生成分段汇总数据



表格的基本维度框架

- 行（Row）：形成表格横行的元素
- 列（Column）：形成表格列的元素
  - 行、列元素相交就会形成一个二维表，行、列元素不同取值的组合就确定了一个单元格
- 层（Layer）：表格中的第三个维度
  - 可以把表格想象成一个立方体，行、列、层就对应了立方体的长、宽和高
- 单元格：表格呈现数据信息的基本单位，可以是计数、百分比、均值等任何信息



S0. 城市

		频率	百分比	有效百分比	累积百分比
有效	100北京	378	33.0	33.0	33.0
	200上海	387	33.7	33.7	66.7
	300广州	382	33.3	33.3	100.0
	合计	1147	100.0	100.0	

几种基本表格类型：叠加表（Stacking）

	S0. 城市		S2. 性别	
100北京	200上海	300广州	男	女
计数	计数	计数	计数	计数
378	387	382	637	510

- 指在同一张表格中对两个变量进行描述，或者说表格中有一个维度的元素是由两个以上的变量构成
- 叠加表其实可以被简单的理解为为每个变量分别绘制两个简单的报表，然后将它们拼接到一起
- 也存在横向拼接的叠加表

几种基本表格类型：交叉表（Crosstabulation）

		S0. 城市			
		100北京	200上海	300广州	合计
S2. 性别	男	188	221	228	637
	女	190	166	154	510
合计		378	387	382	1147

- 是观察两个分类变量间联系时最常用的表格技术，它的两个维度都是由分类变量的各类别（及汇总）构成

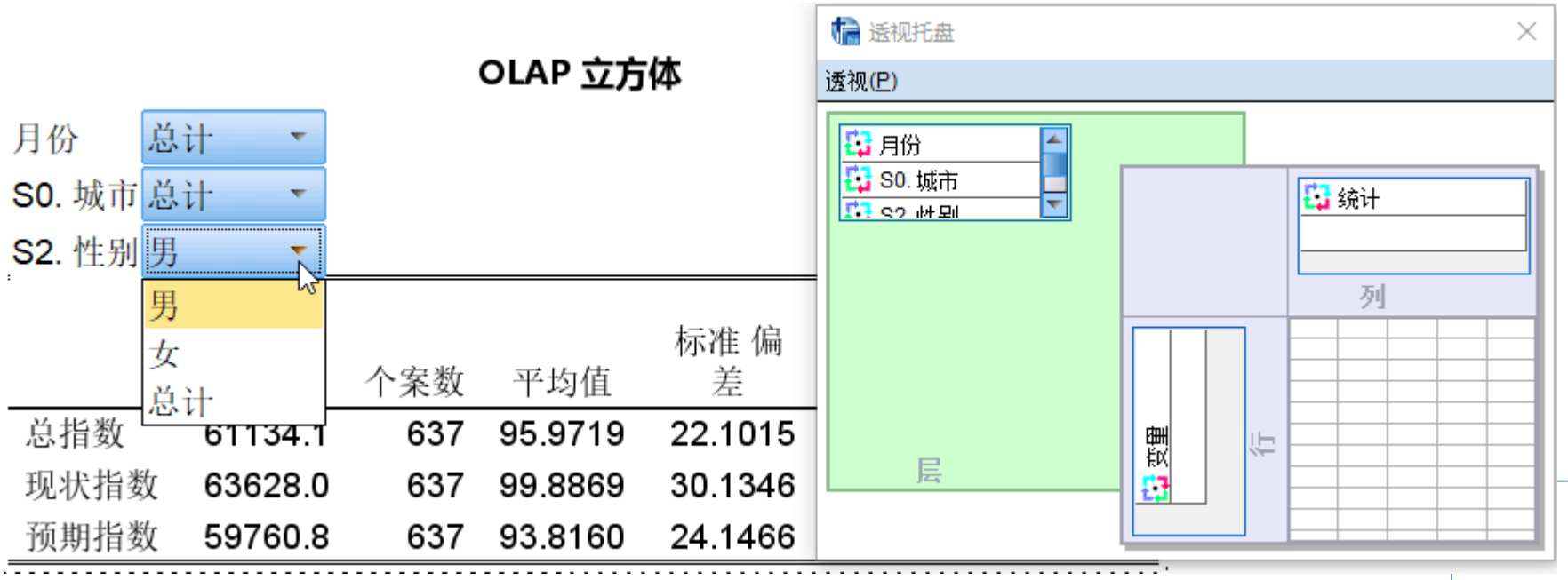
几种基本表格类型：嵌套表（Nesting）

- 也可以用于显示两个分类变量间的联系，但是在嵌套表中，这两个变量被放置在同一个表格维度中，即该维度是由两个变量的各种类别组合构成
- 嵌套表并不如交叉表直观。但是当每个单元格内需要呈现的统计指标非常多时，嵌套表更为美观和紧凑

				计数
S0. 城市	100北京	S2. 性别	男	188
			女	190
	200上海	S2. 性别	男	221
			女	166
	300广州	S2. 性别	男	228
			女	154

几种基本表格类型：多层表（Layers）

- 如果指定了层元素，则表格就由二维扩展到了三维，即多层表
- 事实上，多层表和嵌套表也非常的类似，只是现在只能每次观察到其中一层的数据而已
- 在数据仓库技术中，多层表也被称为数据立方体



几种基本表格类型：复合表格

- 在实际的工作中，上述表格类型还有可能互相组合，以更好的达到相应的分析目的
- 比如叠加－交叉表（一个维度是分类变量，另一个维度则是两个分类变量的叠加）、嵌套－交叉表（一个维度是分类变量，另一个维度则是两个分类变量的嵌套）等

表格绘制的基本步骤

- 确定所需绘制表格的基本结构，如行、列都由什么构成，是否在表格中会出现多个元素的嵌套，有多少种汇总，是否有嵌套汇总等
- 绘制表格的基本结构
  - 将注意力集中在是否已经得到了所需表格结构上，格式细节可暂时忽略
- 对细节进行完善，使单元格的基本输出格式符合要求
- 向表格中添加其余内容，使表格输出满足相应问题的需求
  - 添加其余变量、统计量、小计、总计等
- 对表格的附加文本和格式进行修饰
- 最后一次审核所绘制的表格，考虑有无需要改进之处

案例：题目A3的标准统计报表制作

- CCSS项目每月都会生成固定格式的统计表格，下图为对题目A3的固定表格格式，行标题首先为A3选项的占比，随后为题目感受值的均数，列标题则为受访月份。请用SPSS的制表模块实现该表格

	2009.9	2009.10	2009.11	2009.12
明显改善	12.3	10.3	11.7	12.2
略有改善	20.1	22.7	34.2	31.1
基本不变	46.3	53.1	41.3	50.6
略有恶化	8.8	7.6	6.9	3.7
明显恶化	11.9	4.4	3.4	0.2
不知道/无回答	0.5	1.8	2.5	2.2
感受值	**106.0**	**113.5**	**121.9**	**125.7**

操作流程

- 绘制表格基本框架
- 设置摘要统计量及格式
  - 对分类变量的摘要统计量设定
  - 对连续变量的摘要统计量设定
- 调整各种显示细节
  - 隐藏变量名标签
  - 使百分比和均数同列显示
  - 隐藏统计量标签“行N%”

复杂案例：题目A3a的标准统计报表制作

- 对多选题A3a按照下列汇总方式生成统计表

	2009.9	2009.10	2009.11	2009.12
导致家庭经济状况改善的原因	**25.6**	**20.4**	**30.7**	**25.0**
与收入相关的原因	17.3	15.8	24.6	22.8
与就业状况相关的原因	2.2	1.8	1.9	0.6
与投资相关的原因	1.9	0.5	1.2	0.0
与家庭开支相关的原因	4.2	2.3	2.2	1.9
与政策/宏观经济相关的原因	1.6	0.8	1.7	2.3
导致家庭经济状况恶化的原因	**38.4**	**20.9**	**24.2**	**12.3**
与收入相关的原因	11.7	7.0	5.6	6.2
与就业状况相关的原因	4.8	3.1	6.6	2.2
与投资相关的原因	1.5	0.5	0.2	0.0
与家庭开支相关的原因	24.1	12.0	14.4	4.1
与政策/宏观经济相关的原因	2.2	1.5	0.3	1.4

分析难点

- A3a是多选题，首先需要将变量A3a_1、A3a_2设定为多选题变量集才能进行制表
- 并非所有的选项都需要在表格中出现，中性原因、拒答/不知道这两个选项是需要被隐藏掉的
- 乐观选项、悲观选项分别需要进行小计，而不是全部选项进行合计，且小计位置在上方而不是常见的下方
- 小计项采用的是应答人数百分比（列N%），而选项采用应答人次百分比（列响应%），指标不同，需要分别设定

操作

- 设定多选题变量集
- 绘制表格基本框架
- 设定摘要统计量
- 设定分类变量小结和汇总项

假设检验原理入门



为什么要做检验

- 从统计描述结果中发现可能的数据规律
- 但如果是抽样研究的样本，此时获取的只是样本的信息
- 研究者关心的并不仅仅是样本，更希望了解相应的总体特征
  - 参数估计：推估样本所在的总体特征
  - 假设检验：对提出的一些总体假设进行分析判断，做出统计决策

假设检验的应用价值：药物筛选研究

- 制药业的研发周期长达10~15年，最终上市药物的背后，掩埋着被淘汰的数十万种候选药物，以及高达几十亿美元的开支
  - 本质上就是一个极端化的新品研发流程
  - 计算机分子模拟=>实验室探针筛选=>动物毒性实验=>人体毒性试验=>临床试验=>上市后副作用监测
  - 从实验室筛选开始，假设检验方法就在深入而广泛的加以应用，以尽量准确的甄别出符合需求的化合物
  - 一个候选药物的失败，可能会使得公司几年内都元气大伤
- 真实案例：倒霉的澳大利亚制药公司

假设检验步骤之前需要做的工作

- 运用统计知识根据研究设计和资料的性质正确选择分析过程
- 初步的统计描述（集中趋势、离散趋势）和统计分析
  - 集中趋势：均数、P50
  - 离散趋势：标准差/方差、四分位数间距
  - 分布特征
  - 异常值及其他

先来热热身

- 一家大型超市连锁店上个月接到几例消费者投诉说，某品牌的60克装薯片包装内的土豆片太少
- 店方猜想引起这些投诉的原因是运输过程中沉积在食品袋底部的土豆片碎屑过多，而不是整包的重量不足，但为了保险起见，店方仍然决定对这批存货的平均重量（克）进行检验
- 问题是：
  - 检验多少包才合适？少了没有说服力，多了成本高
  - 究竟差异要和60g大到什么程度才能认为确实有差异呢？

假设检验原理

- 现有的样本均数和已知总体均数不同，其差别可能有两个方面的原因造成
  - 样本来自已知总体（假定的那个总体），现有差别为抽样误差
  - 样本所来自的总体与已知总体不同，存在本质差异
- 为识别这两种可能，应当对其做假设检验

生活中隐含的假设检验

- 掷骰子，猜到点数为胜
  - 其实大家都明白如果筛子没问题，则六个点的出现概率应当相等（均为1/6，这就是一个事先假设），我们只是看每次具体的试验中谁的运气好
- 今天一共下了600次注，竟然一共只猜中了一次
  - 虽然平均应当赢约100次，但今天忘了查皇历，不宜搏彩，运气实在太差
  - 骰子有鬼，掷骰子的人可以人为控制结局，从而利用这种能力使自己得到了更多的收益

生活中隐含的假设检验

- 虽然第一种解释是可能的，但我们认为在筛子公平的前提假设下出现如此结果实在是太不可能了（概率小到不应当被我们一次就碰上），因此我们认为骰子实际上不均匀
- 结论：对方出老千！！！



假设检验原理

- 基础：小概率原理，即认为小概率事件在一次随机抽样中不会发生
  - 最经典的小概率事件：瞎猫碰到死耗子
- 基本思想：先建立一个关于样本所属总体的假设，考察在假设条件下随机样本的特征信息是否属小概率事件，若为小概率事件，则怀疑假设成立有悖于该样本所提供特征信息，因此拒绝假设
  - 事实上，小概率事件在随机抽样中还是可能发生的，只是发生的概率很小。若正好碰上了，则假设检验的结论就是错误的。当然，犯这种错误的概率很小，是我们为了做出统计决策而愿意付出的代价

假设检验的基本步骤：一、建立假设

- 根据统计推断的目的而提出的对总体特征的假设
- 统计学中的假设有两方面的内容：
  - 检验假设(hypothesis to be tested)，亦称原假设或无效假设(null hypothesis)，记为H_0
  - 备择假设(alternative hypothesis)，记为H_1。与H_0相对立，意义在于当H_0被拒绝时供采用。两者是互斥的，非此即彼
  - H_0：$\mu = \mu_0$，　H_1：$\mu \neq \mu_0$
  - H_0：$\mu = 60g$，　H_1：$\mu \neq 60g$

假设检验的基本步骤：二、确定检验水准

- 实际上就是确定拒绝H_0时的最大允许误差的概率
- 检验水准(size of test)，常用α表示，是指检验假设H_0本来成立，却根据样本信息拒绝H_0的可能性大小，换言之，α是拒绝了实际上成立的H_0的概率
- 常用的检验水准为$\alpha = 0.05$
  - 其意义是：在所设H_0的总体中随机抽得一个样本，其均数比现有样本均数更偏离总体均数的概率不超过5%
  - 类似于考试中习惯用60分作为及格线

假设检验的基本步骤：三、计算统计量和P值

- 实际上在此之前还有一步叫做进行试验，样本数据即从此得来
- 统计量只是工具，概率值才是目的，它可以客观衡量样本对假设总体偏离程度
  - 从H_0假设的总体中抽出现有样本（及更极端情况）的概率，即P值
- 例：600次赢100次是H_0假设的情况，只赢1次就是现有样本情况
  - 更极端的情况就是连1次也没有赢
  - 以及居然赢了400多次这类更罕见的情况

假设检验的基本步骤：三、计算统计量和P值

- 检验统计量的特点
  - 该统计量应当服从某种已知分布，从而可以计算出P值
  - 各种检验方法所利用的分布及计算原理不同，从而检验统计量也不同
- 初学者往往本末倒置，很认真地在学工具，却忘记了统计学的本质是思维方式

假设检验的基本步骤：四、得出推断结论

- 按照事先确定的检验水准α界定上面得到的P值，并按小概率原理认定对H_0的取舍，作出推断结论
- 若P ≤α
  - 基于H_0假设的总体情况出现了小概率事件
  - 则拒绝H_0，接受H_1，可以认为样本与总体的差别不仅仅是抽样误差造成的，可能存在本质上的差别，属“非偶然的(significant)”，因此，可以认为两者的差别有统计学意义
  - 进一步根据样本信息引申，得出实用性的结论

假设检验的基本步骤：四、得出推断结论

- 若$P > \alpha$
  - 基于H_0出现了很常见的事件
  - 则样本与总体间的差别尚不能排除纯粹由抽样误差造成，可能的确属“偶然的(non-significant)”，故尚不能拒绝H_0
  - 因此，认为两者的差别无统计学意义，但这并不意味着可以接受H_0
    - 尚不能认为有罪 ≠ 可确认无罪

关于掷筛子的假设检验

- 建立假设
  - H_0：筛子均匀，π=1/6　　H_1：筛子不均匀
- 确定检验水准
  - α=0.05
- 进行试验，计算检验统计量和P值
  - 相应的试验结果在H_0下对应的概率为1/600略多一点
- 得出推断结论
  - 基于H_0出现了小概率事件，结果有非常非常显著的统计学意义，你出老千！

I型错误和II型错误

	拒绝 H_0，接受 H_1	不拒绝 H_0
H_0 真实	I 型错误(α)	正确推断($1-\alpha$)
H_0 不真实	正确推断($1-\beta$)	II 型错误(β)

- 检验效能：H_1是真的，实际拒绝H_0的概率=1-β，称为Power，又称为检验效能
  - 由于两种可能的结论中，往往希望得到的是拒绝H_0的结论，所以实际问题在分析时检验效能不应当太低

如何控制两种错误?

- α可以事先人为设定
  - 希望拒绝结论更可信，则减小α
- β只能间接控制
  - 增大样本量以减小标准误
  - 放大α来减小β
- 实质：牺牲一个来保障另一个

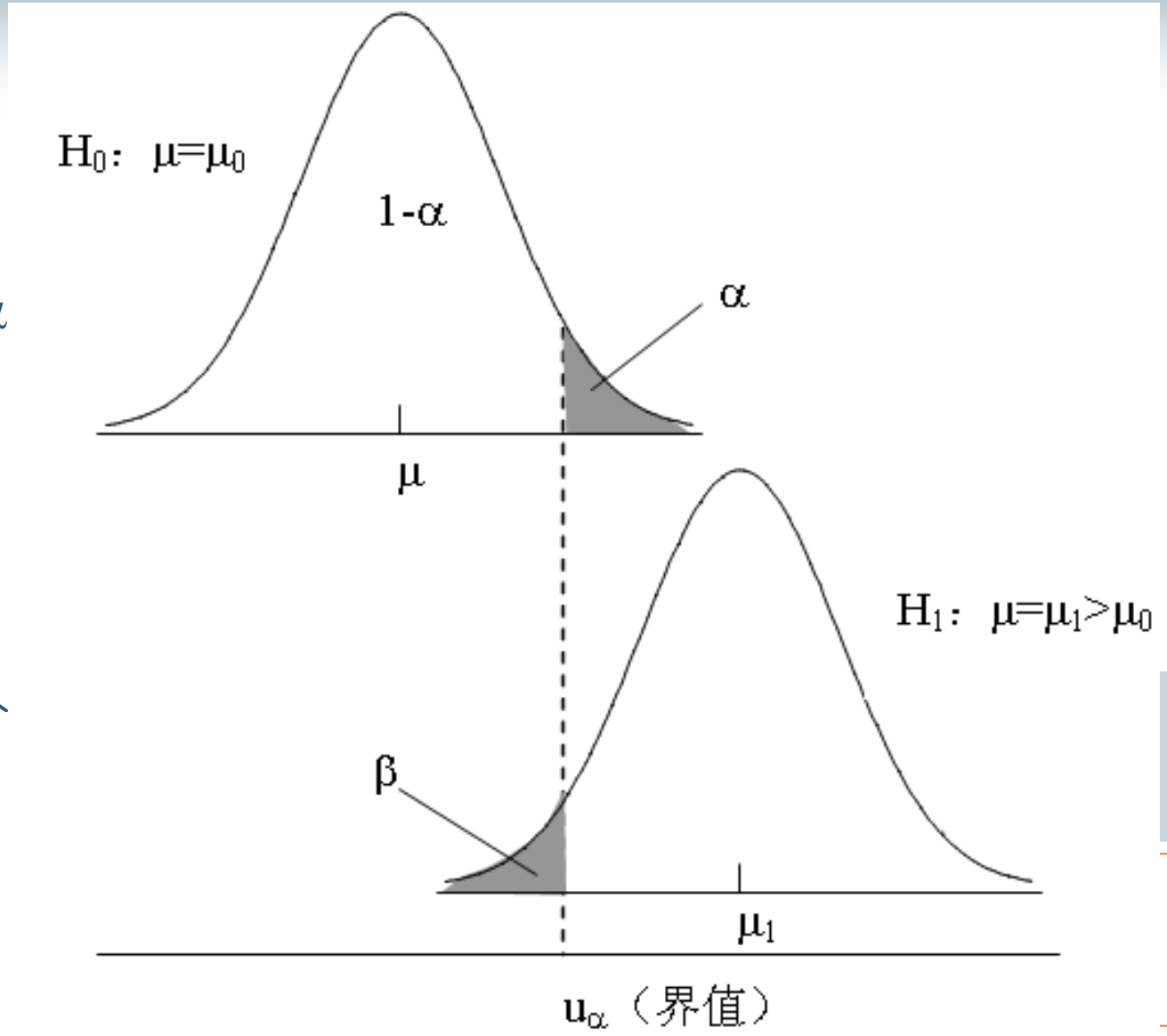


图 4.1　I、II 型错误示意（以单侧 t 检验为例）

假设检验的单侧和双侧问题

- 双侧检验
  - 不知道样本所在总体和假定总体的相应指标谁高谁低
  - 得到拒绝结论更困难，因此相应的结果也更稳妥
- 单侧检验
  - 在专业上可知所在总体的相应指标不可能更高/更低于假定总体值
  - 单侧检验更为敏感，但设定单侧检验需要有充分的专业知识来支持

统计方法应当注意其适用条件

- 独立性(independence)：各观察值间相互独立，不能互相影响
- 正态性(normality)：理论上要求样本取自正态总体
- 方差齐性(homogeneity)：两样本所对应的总体方差相等

假设检验的结论不能绝对化

- 本身就保留了犯错误的可能性
- 样本量导致的检验效能问题
  - 样本量太小，导致检验效能不足，从而无法检出可能存在的差异
  - 样本量太大，得出的有统计学意义的结论可能根本就没有实际意义

单样本t检验



统计理论复习

- CCSS项目基期的信息指数值被设定为100，但这是全部城市的平均水平，请考察基期时广州信心指数均值是否和基准值有差异
- 推断样本是否来自某已知总体，即要检验样本所在总体的均数是否等于已知的总体均数
- 为了回答该问题，统计学上采用了小概率反证法的原理：我们有如下两种假设：
  - H_0：样本均数与（假定的）总体均数的差异完全是抽样误差造成
  - H_1：样本均数与总体均数的差异除由抽样误差造成外，也反映了两个总体均数确实存在的差异

统计理论复习

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$$

- 先假设H_0成立，即一切都是抽样误差造成的。在这个前提下，现有的样本是从已知均数的大总体中抽出来的
- 显然，样本均数和假设总体均数之差就代表了偏离假设的程度
- 但此差异所对应的概率究竟是大还是小？仅看这一个数字很难做出判断。因为这还和数据的离散程度有关，为此需要找到某种方式对这一差值进行标准化

统计理论复习

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$$

- 标准化的基本方式就是将差值除以表示样本均数离散程度的指标
  - 在单样本的情况下，样本的均数服从t分布
- 这个被标化的差值，就是本次检验中所谓的统计量
  - 由于该统计量服从t分布，可利用该分布得到相应的概率值，故而此处的方法被称为为单样本t检验
- 最终求得的P值表示从假设总体中抽出当前样本均数（及更极端情况）的概率总和

统计理论复习

- 如果该P值太小，成为了我们所定义的小概率事件（小于等于α水准），则我们怀疑所做的假设不成立，从而拒绝H_0
  - 基本信念：小概率事件在一次实验中不可能发生
- 反之，我们就不能拒绝H_0，但一般也不太好说去接受他
  - 尚不能认为有罪 ≠ 可确认无罪

单样本t检验的适用条件

- 因为有中心极限定理，一般均数的抽样分布都不会有问题，真正会限制该方法使用的是均数是否能够代表相应数据的集中趋势
  - 强烈偏态分布时，均数已经无法正确代表数据的集中趋势
- 也就是说，只要数据分布不是强烈的偏态，一般而言单样本t检验都是适用的

完全随机的两样本t检验



分析目的与假设

- 目的：推断两个样本是否来自相同的总体，更具体地说，是要检验两样本所代表的总体均数是否相等
  - 当数据服从正态分布时，总体只需要两个参数就可以确定
- 检验假设
  - 无效假设 H_0：$\mu_1 = \mu_2$
  - 备择假设 H_1：$\mu_1 \neq \mu_2$
  - 检验水准：$\alpha = 0.05$

基本原理

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}}$$

- 其基本原理和单样本t检验相同
- 首先假设H_0：两样本来自同一总体
- 当该总体服从正态分布时，就可以采用两样本t检验来计算从该总体中抽得这样两个样本（及更加极端情况）的概率为多少，从而做出统计推断

适用条件

- 由于H_0假设的是两样本来自同一总体，分析目的只涉及到均值，因此两样本t检验在推导过程中除了要求总体服从正态分布外，还要求两样本各自所在总体方差相同
  - 独立性：对结果的影响较大，但一般没问题
  - 正态性：有一定的耐受能力，可以通过直方图等进行观察
    - 注意应当要分组考察
  - 方差齐性：相对而言对结论的影响较大，需要进行方差齐性检验

适用条件不被满足时的处理方式

- 情况较轻时可以采用校正t检验的结果
- 否则应使用变量变换使之满足条件
- 或采用
  - 非参数检验方法
  - 贝叶斯推断方法
  - 计算统计学方法（Bootstrap抽样等）

分析实例

- 现希望评价2007年4月第一次调查时中低收入和中高收入人群的消费者信心指数是否存在差异
- 分析：连续变量，成组设计，目的是两样本均数的比较
  - 正态性：可作直方图等
  - 方差齐性：在SPSS中会一并输出，python等工具中需要单独进行

配对设计的两种常见情况

- 对同一个受试对象处理前后的比较，或者同一个对象在不同时间段测量结果的比较
  - 个体的基本特征在各次测量中不变，导致同一个体的测量结果间存在关联
  - 结果中混杂了时间的影响，尽量少用
- 将受试对象按情况相近者配对（或者自身进行配对），分别给予几种处理，以观察这些处理的效果有无差别
  - 通过按照相应的因素配伍，不同个体的测量结果可能均受这些因素影响，因此它们之间也建立了关联

配对数据为什么不能用成组方法分析

1	2	1	11
2	3	2	10
3	4	3	9
4	5	4	8
5	6	5	7
6	7	6	6
7	8	7	5
8	9	8	4
9	10	9	3
10	11	10	2

配对数据为什么不能用成组方法分析

- 在配对设计得到的样本数据中，每对数据之间都有一定的相关，如果采用成组设计的分析方法就无法利用这种关系，浪费了大量统计信息
- 换言之，相应的统计分析方法一定要能够充分利用这些配对信息
  - 这也是设计方案高于分析方法的一个典型案例
- 解决方案：求出每对的差值，通过检验该差值总体均数是否为0，就可以得知两种处理有无差异

基本思路

$$t = \frac{\bar{D}}{S/\sqrt{n}}$$

- H0：两总体均值无显著差异，差值序列均值$\mu_0 = 0$
- 构造统计量：同单样本t检验
- 实质是先求出每对测量值的差值，然后检验差值序列的均值是否与0有显著差异
- 如果差值的均值与0有显著差异，则认为两总体均值存在显著差异
- 否则，认为两总体均值不存在显著差异
- 功能实际上和单样本t检验重复，但数据输入格式不同
- 方法适用条件也等同于单样本t检验

分析实例

- 为保证数据质量，接受过CCSS访问的受访家庭半年内不会再进行访问，但半年之后会进行抽样回访。在2007年12月，项目组对2007年4月的成功访问家庭进行了回访，共采集了88例有效样本，现希望比较这些样本的信心值是否发生变化，数据见CCSS_pair
- 这是一个典型的个体前后配对设计，应当采用配对t检验来进行分析
- 按照配对t检验对数据格式的要求，这里在输入数据时应当每个变量（一列）代表一个组，而每条记录（一行）代表一对数据

样本量太大时，假设检验的结果

其实没啥用！



什么是效应量（Effect Size）

- 用以衡量总体均值间差异大小（或者影响因素作用强度）的指标
  - 不受样本大小影响，从而可以在不同研究间进行比较
  - 这个东东的兴起实际上和Meta分析有关
- 差异/影响无统计学意义：收工走人
- 差异/影响有统计学意义：效应量很小
  - 多半是因为样本量过大，武德过于充沛所致，差异存在，但无专业价值
- 差异/影响有统计学意义：效应量足够大
  - 差异存在，且有专业价值
- 注意：仅适用于简单的分析方法

效应量的指标家族

- d-family（difference family）
  - 如Cohen's d、Hedges' g
- r-family（correlation family）
  - 如Pearson r、R^2、η^2、ω^2、f
- OR-family（categorical family）
  - 如odds ratio (OR)、risk ratio (RR)
- 注意：Meta分析中一般需要使用标化的效应量
  - 单纯衡量效应大小则使用未标化效应量即可

常见的效应量指标：t检验

- Cohen's d：表明两个均数标准化之后的差异大小
  - 独立样本t检验：$ES = (\mathrm{M}_1 - \mathrm{M}_2)/\mathrm{s_}pooled$，s_pooled为联合方差
  - 配对t：$ES = (\mathrm{M}_1 - \mathrm{M}_2)/\mathrm{s}$，两组的任一方差即可，需满足组间方差齐性
- Hedges g：在d值的基础上对联合方差做了进一步的校正
- Glass Delta：理论上能区分控制组和实验组时，只使用控制组方差
  - SPSS假定对话框中放在后面的一组为控制组
- 判断标准：可忽略<0.2<小效应<0.5<中等效应<0.8<高效应

常见的效应量指标：方差分析

- Cohen’s f
  - $ES = \sqrt{\mathrm{F/n}}$
  - 判断标准：可忽略<0.1<小\效应<0.25<中等效应<0.4<高效应
- 偏Eta^2：表示自变量X所导致的变异和误差相比的程度
  - 偏$\eta^2 = \mathrm{SS_A}/(\mathrm{SS_A} + \mathrm{SS_E})$
  - 判断标准：可忽略<0.01<小\效应<0.06<中等效应<0.14<高效应

d	r^*	η^2	Interpretation sensu Cohen (1988)	Interpretation sensu Hattie (2007)
< 0	< 0	-	Adverse Effect	
0.0	.00	.000	No Effect	Developmental effects
0.1	.05	.003		
0.2	.10	.010	Small Effect	Teacher effects
0.3	.15	.022		
0.4	.2	.039		Zone of desired effects
0.5	.24	.060	Intermediate Effect	
0.6	.29	.083		
0.7	.33	.110		
0.8	.37	.140	Large Effect	
0.9	.41	.168		
≥ 1.0	.45	.200		

效应量的计算方法

- d-family：SPSS 27版对话框已经内置
  - 包括t检验对话框和单因素ANOVA对话框
- r-family/OR-family：早都是老熟人了
  - 多年前SPSS已经在一般线性模型对话框选项中内置偏η^2
- 第三方工具：直接度娘“效应量计算器”

独立性

- 不同案例间的取值互相独立，不受到除研究中考虑到的分组因素、配对因素等研究因素之外任何其他因素的影响
  - 其他非研究因素的影响基本可以等同于随机误差
- 非独立的情形
  - 研究儿童生长发育，样本中存在兄弟、表兄弟等近亲关系
  - 流感等传染病的发病与否
- 数据非独立的影响较大，因为传统统计模型均按照数据独立假设进行推导，非独立会导致数据方差估计不准确，进而导致假设检验结果错误
  - 很严重的数据非独立也会使得参数估计值出现偏差

独立性

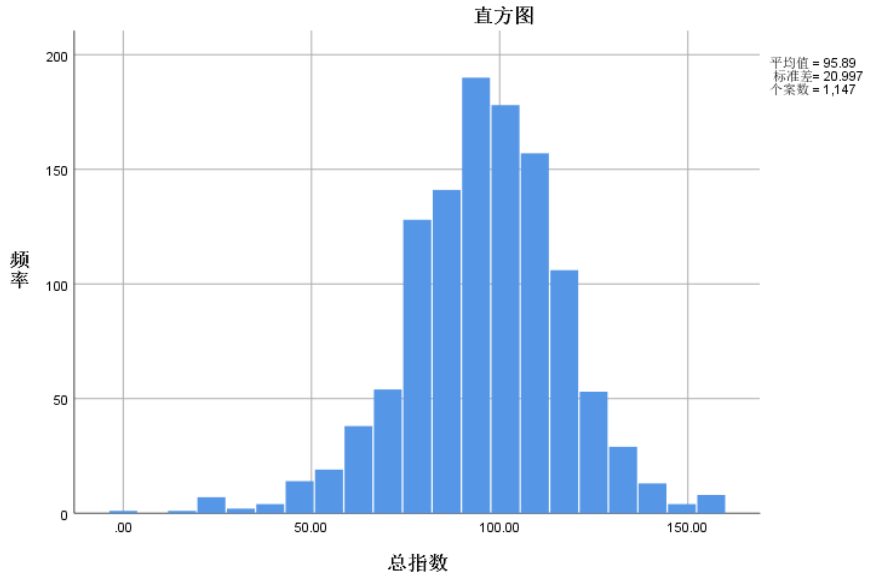
- 研究设计阶段
  - 尽量保证数据不违反独立性
  - 保证随机化真正得到实施
    - 随机抽样、随机分组
- 数据分析阶段
  - 游程检验、DW统计量、自相关分析等考察独立性，但实际上较少使用
- 应对策略
  - 采用专门的方差分析模型，对方差等进行校正
  - GEEs、混合效应模型等专门的非独立数据分析模型

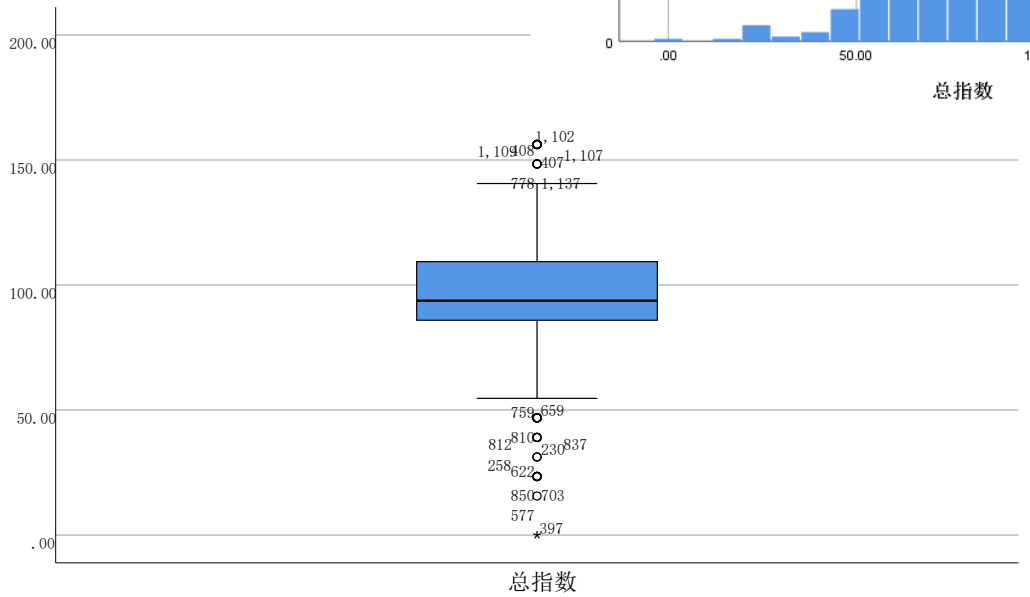
正态性

- 普通方法：如t检验，单因素方差分析等，事实上对正态性略微偏离都有比较好的耐受性，结果仍然是稳健的
  - 偏离程度过大时仍然需要处理，且此时均数的代表性也会有问题
- 简单统计模型：需要进行模型残差考察，光靠观察原始变量的分布并不能得到最终结论
  - 典型错误：多重线性回归，直接看Y的分布，而不去做残差分析
- 复杂统计模型：为避免结果受到干扰，原始变量都需要先做标准正态变换再建模
- 特例：临床试验中由于相关法规的要求，必须要进行正态性检验，并依据结果应用不同的分析手段

用图形方法考察正态性

- 直方图、茎叶图：基于初步汇总后的原始数据进行全面考察
  - 注意需要分组考察
  - 优点是简单直观，但缺点是结果不够精细
- 箱图：基于百分位数体系进行大致观察
  - 可同时考察和比较多组的数据分布
  - 精细度比直方图还要差



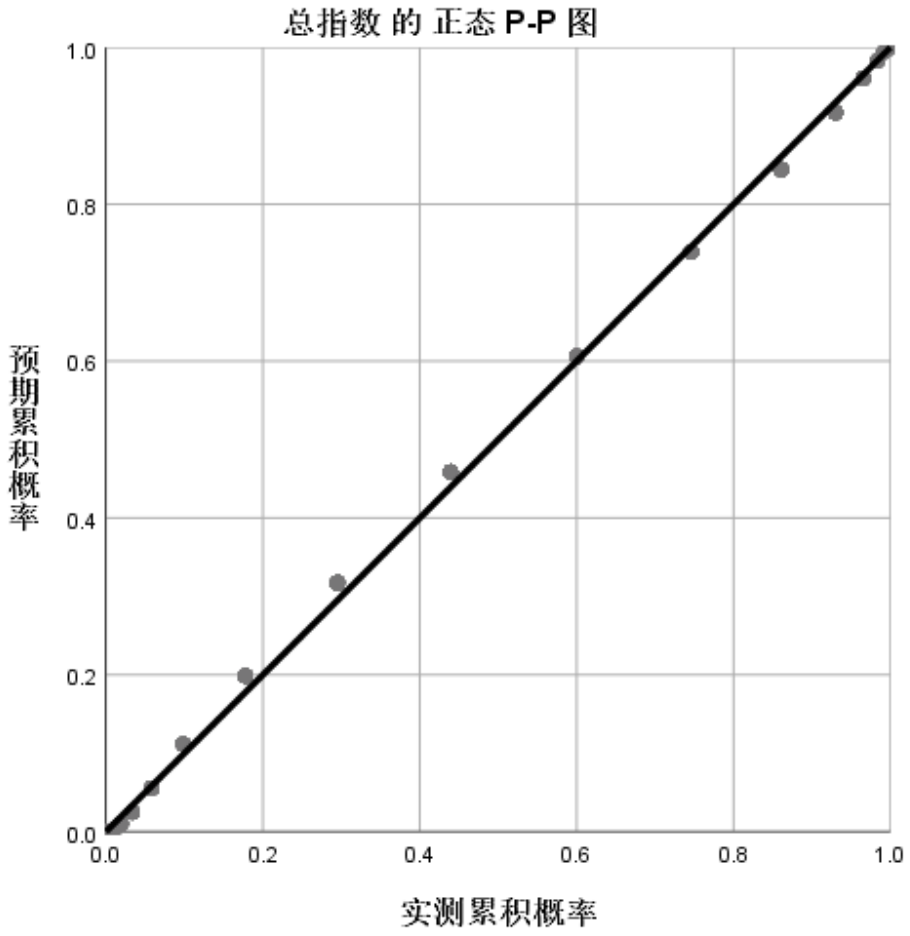


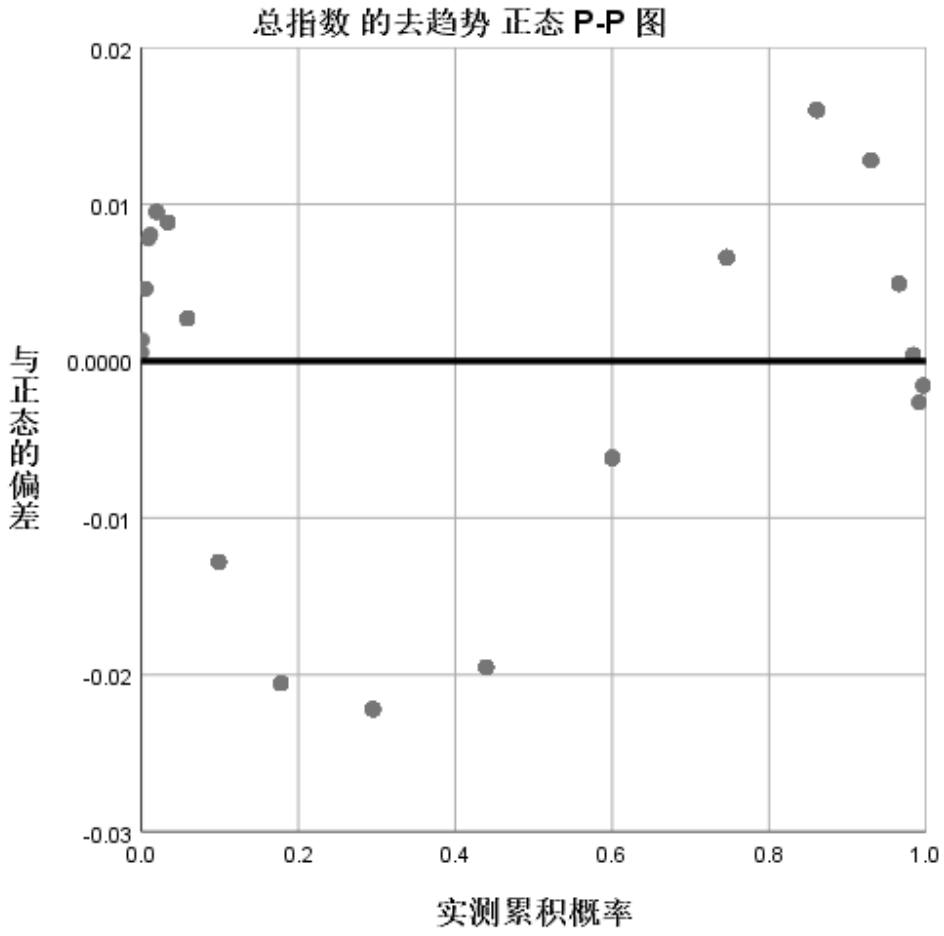
用图形方法考察正态性：PP图和QQ图

- 绘制出变量的实际累积概率与其假定理论分布累积概率的符合程度，以判断资料是否服从所考察的分布类型。如果变量服从理论分布，则实际累积概率与理论累积概率应该基本一致
  - 在图形上表现为实际数据的散点和理论上的直线完全重合
- P-P图比较的是两者的累计概率分布，而Q-Q图则是根据变量的实际百分位数与理论百分位数进行绘图，或者说得更通俗一点，相比之下Q-Q图的适用条件更宽松，结果也更稳健一些

用图形方法考察正态性：PP图和QQ图

- PP图的经验界值：上下偏离不超过0.05





用假设检验方法考察正态性

- Kolmogorov-Smirnov检验
  - 考察H_0理论分布（可以是任何分布）下累计相对频数和实际频数间的差异
  - 取频数差值绝对值的最大值为统计量
  - 结论比较严格，即使只有个别案例有偏离也会拒绝H_0
- Shapiro-Wilk检验
  - 基于数据的偏度系数和峰度系数进行正态性偏离程度的检验
  - 相对KS检验而言结果更稳妥一些
- 残差分析
  - 建模完毕后将残差存储为新变量，就可以进行图形观察甚至正态性检验

假设检验方式考察正态性的固有缺陷

- 样本量偏低时往往过于迟钝
- 样本较大时又会过敏，其结论失去使用价值
- 建议：除非临床试验这类法规强制性要求的领域，应当以图形分析为主

正态性检验的SPSS实现

- 描述统计=>探索：图
  - 提供KS检验和SW检验两种方法
  - 可直接给出分组后各组的检验结果
  - 会同步给出QQ图和箱图供考察
- 非参数检验=>单样本
  - 提供KS检验
  - 是可一次进行多个变量的检验
  - 新对话框和旧对话框两种操作方式

发现数据非正态时的应对策略：变量变换

- 评估严重程度
- 变量变换
  - 正偏态分布：对数变换或平方根变换
    - 有负值时，加常数再变换
  - 超过30%-60%区间的率：平方根反正弦变换
- 慎用过于复杂的变量变换方式
  - 传说中的Box-cox变换

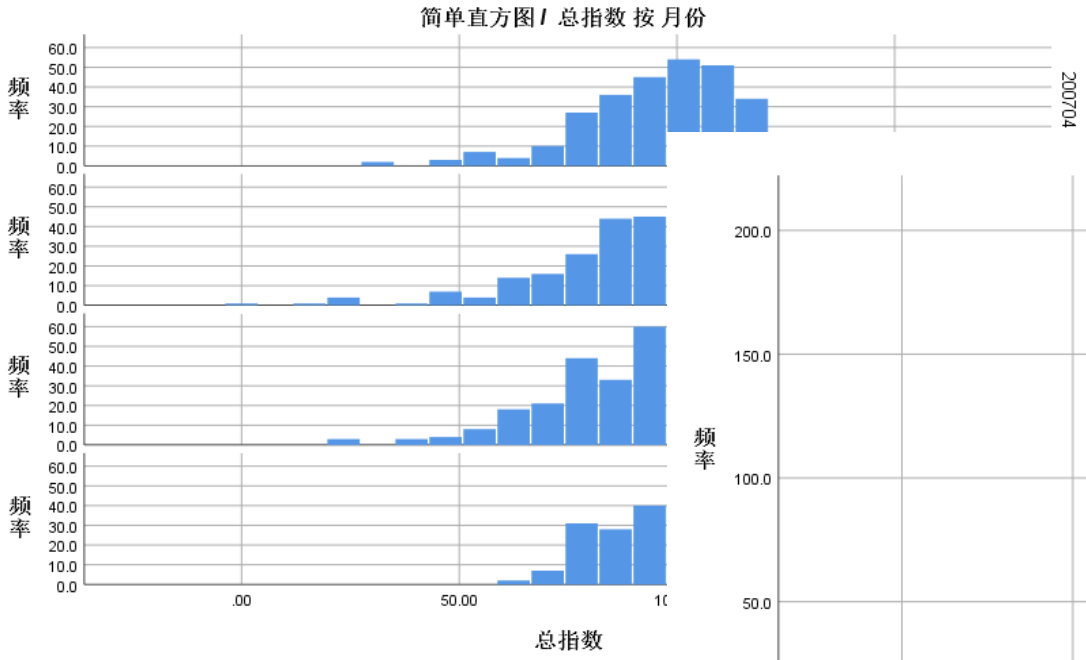
发现数据非正态时的应对策略：更换分析方法

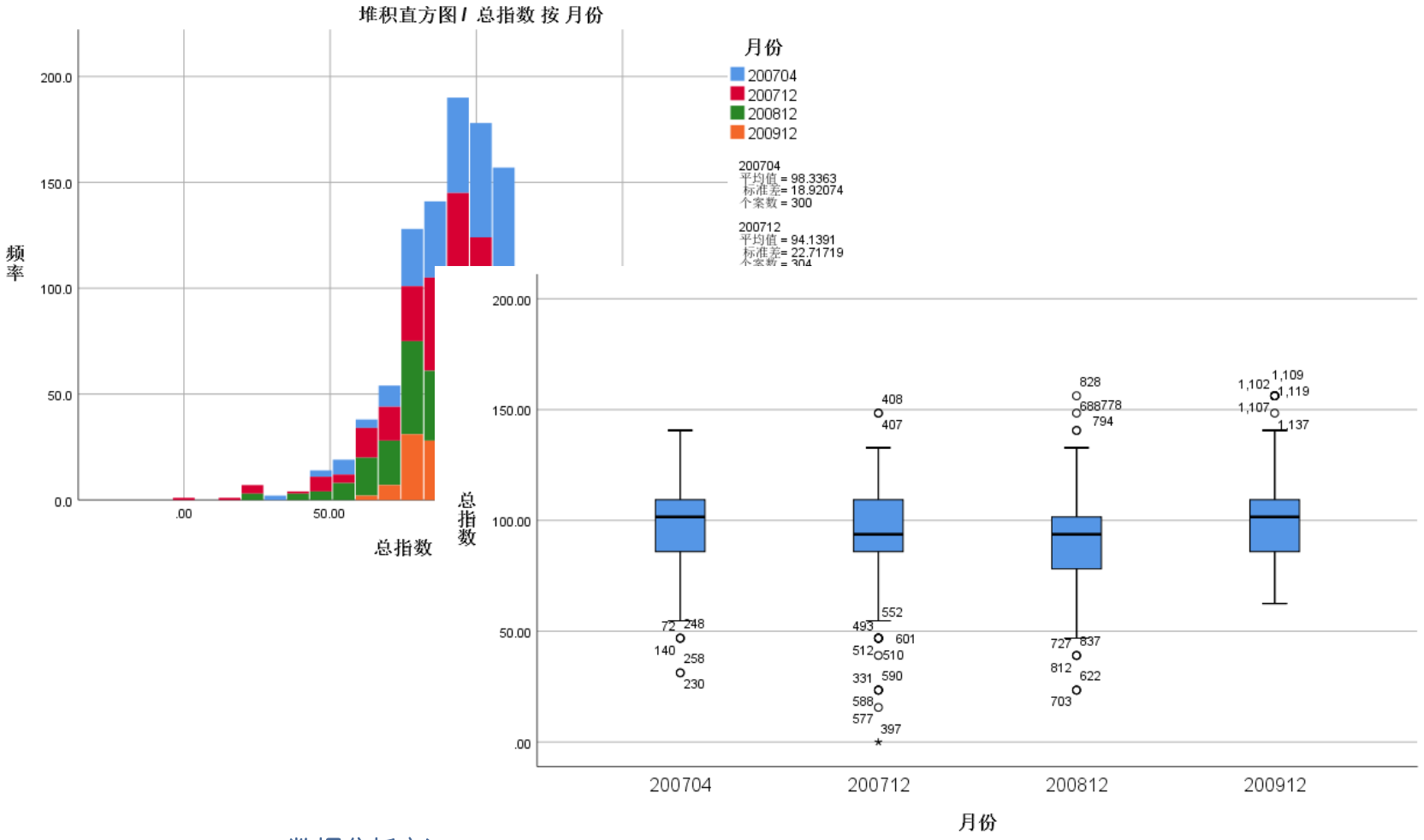
- 非参数方法
  - 秩和检验、秩变换分析
- 专用的分析方法
  - 校正的t检验、校正的方差分析等
- 一些特殊的计算统计学方法
  - Bootstrap技术
  - 贝叶斯推断方法

方差齐性

- 在各组间样本含量相差不太大时，方差轻微不齐仅会对方差分析的结论有少许影响
- 一般而言，只要最大/最小方差之比小于3，分析结果都是稳定的
- 然而，如果各组间样本含量相差过大时，方差不齐会对方差分析造成相当大的影响
  - 主要是基于H0假设下的标准误估计会严重失真
- 因此，统计软件往往会直接提供选项进行方差齐性检验的计算，甚至于进一步提供校正后的分析结果

用图形方法考察方差齐性





用假设检验方法考察方差齐性

- Bartlett's方差齐性检验
  - 基于数据服从正态分布的假设对其方差齐同性进行检验
  - 数据略微偏离正态分布时，检验结果就会出现较大偏差
- Levene方差齐性检验
  - 可用于正态或者非正态分布的资料（平均水平可以使用均数或者中位数等）
  - 在数据可能非正态的情况下，精度比Bartlett检验好
  - 可作为标准的方差齐性检验方法
- Fligner-Killeen检验
  - 一种非参数的方差齐性检验方法，不依赖于具体分布

方差齐性检验的SPSS实现

- t检验：默认直接给出结果
- 方差分析：提供相应选项
- 探索=>图：含Leven检验的分布-水平图：未转换
  - 分布水平图用于协助寻找使得各组方差齐的最佳转换方法，最好慎用

发现方差不齐时的应对策略

- 对原检验方法做校正
  - 校正的t检验
  - 校正的方差分析
  - 校正的事后两两比较方法
- 方差严重不齐时，最好考虑变量变换，或者改用非参数方法

单因素方差分析

@文彤老师



单因素方差分析

- 检验某一个影响因素的差异是否会给观察变量带来显著影响
- 实例：
  - 不同肥料对某农作物亩产量是否有显著差异
  - 不同的推销策略是否对推销额产生显著影响
  - CCSS案例中提供了2007年4月，以及2007、2008、2009年12月四个时间点的消费者信心监测数据，现希望考察这四个时间点的消费者信心指数平均水平是否存在差异

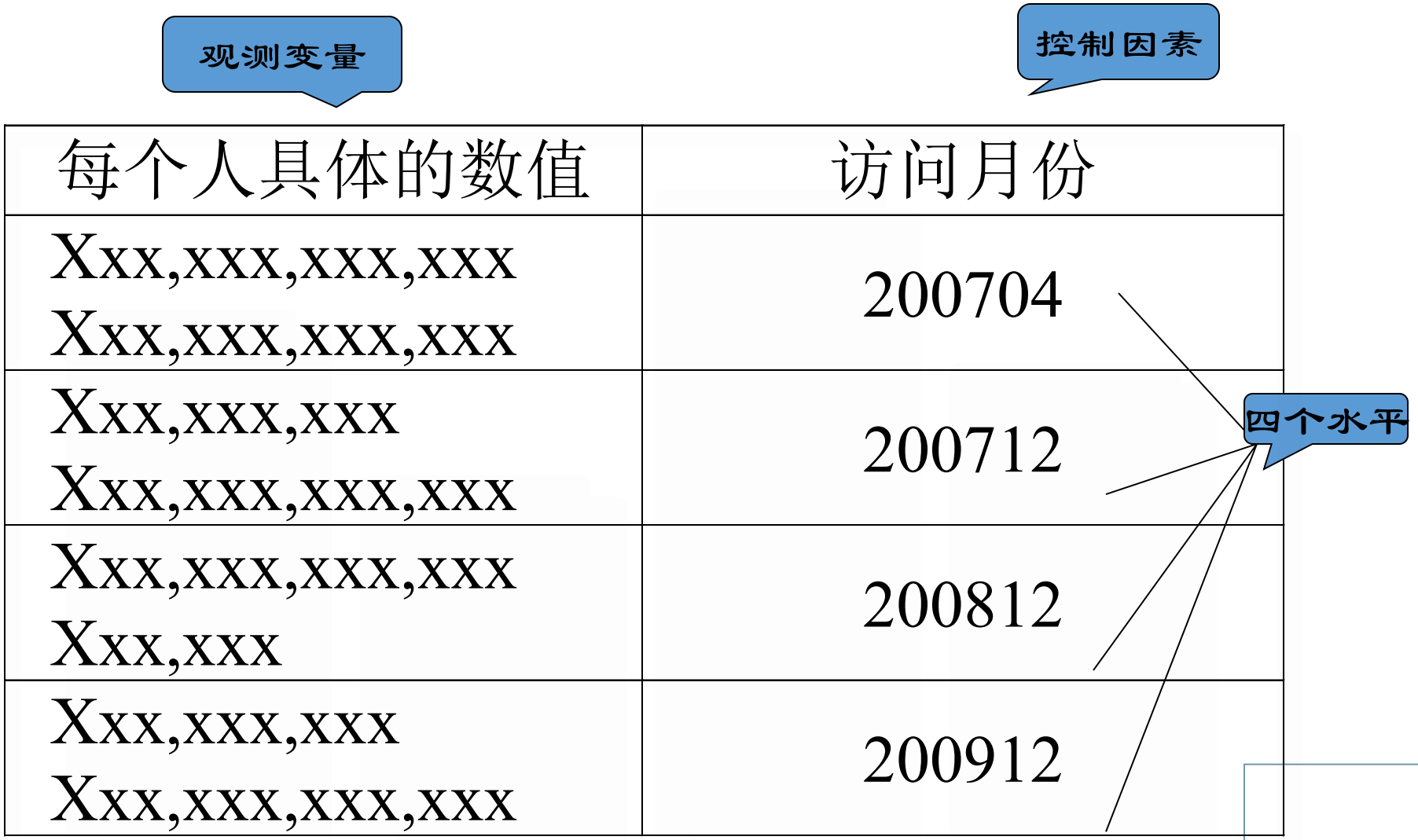
多组比较可以做两两检验吗

- 共有4组的均数需要比较。如果用t检验进行两两比较，共要进行6次t检验
- 如果每次t检验犯第一类错误的概率为0.05，则不犯第一类错误的概率为0.95，6次都不犯第一类错误的概率为0.95的6次方，即0.7351，因此在6次t检验中至少有一次犯第一类错误的概率为0.2649
- 由此可见用两两检验的方式进行多组间的比较会增大犯第一类错误的概率

基本原理

- 方差分析是基于变异分解的原理进行的，在单因素方差分析中，整个样本的变异可以看成由如下两个部份构成：
  - 总变异=随机变异 + 处理因素导致的变异
  - 信心指数总变异＝不同月份导致的变异＋不同受访者间的随机变异
- 处理因素导致的变异就是要研究的对象，我们希望能够证明他是否大于0

单因素方差分析的数据结构

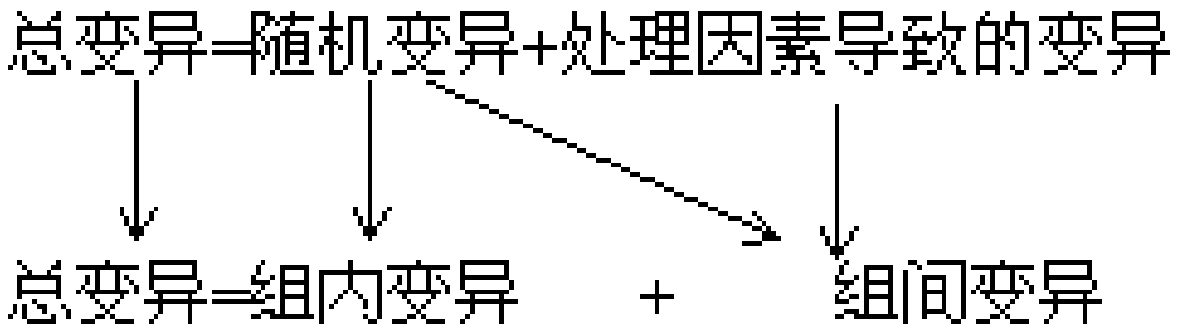


实际数据的变异分解

- 各组内部的变异（组内变异）
  - 反映个体差异（随机变异）的大小
- 各组均数的差异（组间变异）
  - 反映了个体差异（随机效应）的影响与可能存在的处理因素的影响之和
- 总变异 = 组内变异 + 组间变异

统计量的含义

- 采用合适的指标表示组内变异和组间变异的大小，将两者相比较，便可得知组内变异中是否真正包含了处理因素所导致的影响



统计量的含义

$$F = \frac{SSA/(k-1)}{SSE/(n-k)} = \frac{MSA}{MSE}$$

- F>>1：组间变异远大于组内变异，处理因素有影响
- F＝1：认为处理因素实际上无影响
- 究竟F值要多大才算非常大？采用P值来确定

事后检验与适用条件

- 方差分析本身是完美的，但在解决实际问题的时候，我们往往仍需要回答多个均数间究竟是哪些和哪些存在差异，这样问题又回到了两两比较上
- 和t检验时的情况类似，方差分析也要求各样本
  - 必须是相互独立的随机样本
  - 样本均来自正态分布的总体
  - 样本来自的总体方差相等，即方差齐性

两两比较所必须解决的问题

- 多组比较时，无论是方差分析还是卡方检验，在拒绝H_0之后，所接受的H_1都是各种情况的组合
  - 各组间两两均有差异
  - 各组间只有两组有差异
  - 各组间有若干组存在差异
  - …
- 为此，研究者还需要进一步回答各组间究竟是哪些和哪些存在差异，这样问题又回到了两两比较上
  - 最直接和原始的思路就是对总的一类错误大小进行校正
  - 关键：如何控制好总的一类错误大小

两两比较中会遇到的一类错误

- CER：比较误差，即每作一次比较犯一类错误的概率
- EERC：在完全无效假设下的试验误差率，即在H_0成立时做完全部比较所犯一类错误的概率
  - 事实上，很少会遇到H_0完全成立的情形
  - 方差分析检验/卡方检验本身控制的就是EERC
- MEER：最大试验误差率，即在任何完全或部分无效假设下，做完全部比较所犯一类错误的最大概率值
  - MEER的适用范围显然更广

直接校正P值：Sidak校正

- 当无效假设实际上成立，各组均数无差别时，完全两两比较犯第一类错误的概率为$1-0.95^{k(k-1)/2}$，此即EERC
- 因此，进行一类错误控制时最直接的想法就是将总的α水准控制到0.05，从而由上述公式反推得出每一个检验所使用的α_{ij}，这种校正方式被称为Sidak校正
- 注意：
  - 如果统计软件直接进行校正操作，则往往会倒过来将每个检验的P值放大（最大放大为1），而固定每个比较的α水准仍为0.05以便于阅读

直接校正P值：Bonferroni校正

- Sidak校正只是在无效假设成立的情况下才适用的校正方式
  - 多数实际问题中，都是有些组的均数相同，而有些组的均数不同，因此控制MEER更为合适
- Bonferroni不等式被广泛的用于此目的，它通过控制CER，使得MEER被控制在所设定的水准以内，其计算公式为：CER=α/c
  - c为需要进行的比较次数
- 只要CER，即每次比较使用的α_{ij}小于α/c，就可以保证MEER不会大于α

直接校正法的缺陷

- 将两两比较分别进行，不仅使用麻烦，也增加了误差的影响
  - 联合检验可解决此类问题
- 对一类错误的控制太严，结果往往偏于保守
- 但是，当研究者希望得到的阳性结论真实可信时，直接校正法过于严格的缺陷反而变得可资利用

联合检验：LSD-t检验

$$\text{LSD-}t = \frac{\overline{x}_{.i} - \overline{x}_{.j}}{S_{\overline{x}_{.i} - \overline{x}_{.j}}}$$

- 是least significant difference的缩写，即最小显著差异
- 显然，它只是t检验的一个变形，在标准误和自由度的计算上利用了全部样本信息，使得结果更为准确
  - t检验标准误和自由度的计算只利用了相应两组的信息
- 如果不对检验水准做任何校正，则可以认为LSD法是最灵敏，或者说最容易得出阳性结论的

两两比较的结果阅读和解释

- 常见的嵌套检验结果
  - ab，bc，但a组和c组有差异
- 统计结论本身就是概率性的，不存在递推关系
  - 也存在检验效能的问题
- 由统计结论进行的应用性外推，统计方法本身不负任何责任
- 如果结果可以直接外推
  - 头上一根头发都没有的人毫无疑问是秃子
  - 头上有一根头发的人和一根头发都没有的人之间看不出什么差别（差别无统计学意义），所以也是秃子
  - 依此类推，最后会得到一个满头黑发的人也是一个秃子的荒谬结论！

比较常用的联合比较方法

- S-N-K：使用Student 化的范围分布快速进行检验，将总样本分为多个同质子集
  - 以前常用，但现在发现组数很多时假阳性较高
- Tukey：原理类似于SNK
- Duncan：做了进一步的改进
- Scheffe：进行多组的联合检验而不是两两检验
- Dunnett：需要提供参照组，可进行单侧检验，SCI杂志中经常指定使用该方法

两两比较方法的选择

- 如两个均数间的比较是独立的，或者虽有多个样本均数，但事先已计划好要作某几对均数的比较，统计学者中多数赞同用一般t检验
- 总体检验如果没有拒绝H_0，不再考虑两两比较
- 如果存在明确的对照组，要进行的是验证性研究，即计划好的某两个或几个组间（都和对照组）的比较，宜用Bonferroni（LSD）法
  - 前期研究已有信息铺垫，因此常规性的控制一类错误即可

两两比较方法的选择

- 若需要进行的是多个均数间的两两比较（探索性研究），此时希望得到的阳性结论尽可能是真实的，因此要严格控制一类错误
  - 各组人数相等，适宜用Tukey法，该方法此时的检验效能较高
  - 比较较为复杂时，Scheffe法较为稳妥，但它相对比较保守
- 方差不齐时的两两比较方法：一般认为是Games-Howell法稍好一些，但最好直接使用非参数检验方法
- 注意：对于事先无严格设计的多重比较，各组间的差别只是一种提示，要确认这种差别最好重新设计实验

非参数检验



参数统计方法的局限

- 需要事先明确假定总体分布
  - t检验，方差分析
- 总体分布未知，或分布不符合要求时无法使用
- 结果为有序分类变量时无法使用
- 样本数据两端有不确定值时无法使用

非参数分析方法的特点

- 不依赖总体分布的具体形式
- 不是对分布参数进行估计或假设检验
  - 而是对总体的分布位置/形状进行估计或假设检验
- 适用范围广，几乎可用于任何情况
- 当资料符合参数检验方法的适用条件时，使用非参数方法的检验效能较低
  - 效能最高的秩和检验效能可达到参数方法的90%~95%

分布类型/形状的检验方法

- 亦称拟合优度检验方法
- 检验样本所在总体是否服从已知的理论分布
  - 多项分类变量分布的卡方检验
  - 二项分类变量分布的二项分布检验
  - 考察连续变量是否服从各种常用分布的单样本K-S检验
- 检验样本序列随机性的Runs过程

分布位置检验方法

- 检验样本所在总体的分布位置（集中趋势）是否相同
- 平时说的的非参数检验方法实际上指的就是这一类方法
- 两独立样本/多个独立样本中心位置的检验
  - 秩和检验
  - 中位数检验
  - 两样本K-S检验
- 两相关样本/多个相关样本中心位置的检验
  - 配对秩和检验
  - 配伍Friedman双向秩方差分析

非参数检验预备知识

- 首先应该明确该数据确实无法使用效率更高的参数方法！
- 可以考虑使用的信息量
  - 中位数：大于/小于基于H_0的中位数的样本个数分布是否均匀
  - 次序关系：数值和中位数的距离可以提供进一步的信息
- 顺序统计量：非参数检验的理论基础
  - 通过对数据从小到大排序（即排队），并用排序号（排队号）代替原始数据进行统计分析
- 秩（Rank）：排序号（排队号）在统计学上称为秩（秩次）
- 结（ties）：绝对值相等称为结，又称同秩，一般取平均秩次

独立样本比较的非参数方法

- Wilcoxon两样本秩和检验
  - H_0：两总体所在的中心位置相同
    - 中位数可以代表中心位置，但只使用中位数的话信息量太少
  - 基于H_0假定，混合编秩，分组求秩和
  - 考察各组秩和的大小是否明显偏离H_0
  - 在非参数检验方法中效能最高
- Mann-Whitney U
  - 基本等价于两样本秩和检验
  - 可近似理解为基于秩次进行了两样本t检验

独立样本比较的非参数方法

- Kolmogorov-Smirnov Z
  - 两样本K-S检验，检验效能不高
  - 考察的是整个分布是否相同，而不是只针对中心位置
- 二样本 Wald-Wolfowitz检验：以相同组成员一个游程的游程检验
- Moses极端反应检验：考察两组分布范围（秩次的全距）是否有明显差异，更看重的是两端极端值的分布
- Hodges-Lehman检验：为两组的中位数差异生成一个独立样本估计和置信区间

多个独立样本比较的非参数方法

- Kruskal-Wallis H检验
  - 本质上就是基于秩次的单因素方差分析，可用于两组或多组
- k样本Jonckheere-Terpstra检验：当分组变量为有序分类，即分组存在自然顺序（温度、剂量、病情组等）时，该方法比 Kruskal-Wallis方法的检验效能更高
  - 更准确的含义是按照该顺序，各组的平均水平只可能单调上升/下降
- 多组比较的事后两两比较
  - 本质上仍然需要考虑如何控制总的一类错误
  - 样本量较小时，也有统计学家倾向于不做检验水准的校正

两配对样本比较：Wilcoxon符号秩检验

- 求出各组配对差值
- 基于H_0假定成立，差值应当围绕0上下对称分布
- 按照差值绝对值计算秩次，然后分正、负组分别计算秩和
- 考察正负秩和的大小是否明显偏离H_0

配伍样本比较的非参方法：Friedman双向秩方差分析

- 配伍因素影响很大，不能忽视
- 在配伍因素必须考虑的前提下，基于H_0假设进行区组内编秩
- 按照比较组计算秩和，并比较其是否明显偏离H_0假设
- 这实际上是一个没什么用的方法
  - 配伍设计的资料很难会遇到不符合应用条件的情形
  - 该检验的检验效能非常低，与其采用秩和检验，不如进行变量变换

配伍样本比较的非参方法

- Kendall’s W
  - 主要用于计算Kendall和协系数，用于表示K个指标间的关联程度
- Cochran’s Q
  - 是McNemar检验针对多组的推广，只适用于两分类资料
  - Friedman双向秩方差分析只适用于连续变量

秩变换分析的基本原理

- 秩和检验方法都可以看作是基于H_0假设求出秩次，然后对秩次完成相应的参数检验
- 秩变换分析方法，就是利用这一原理，基于H_0假设成立的情况，先求出原变量的秩次，然后使用秩次代替原变量进行参数分析
- 当样本含量较大时，其分析结果和相应的非参数方法基本一致
- 但该方法可以进一步充分利用已知的参数方法，如多组样本的两两比较、多变量回归等，从而大大扩展了非参数分析方法的范围

秩变换分析的优缺点

- 优点
  - 适用范围广，样本量充足的情况下均可使用
  - 分析结果更为稳健，不易受极端值影响
- 缺点
  - 检验效能相对稍低，存在信息损失，不适用于中小样本
  - 而且其分析结果相对没有那么“定量”，毕竟其描述的是影响因素对因变量秩次的作用，而不是对因变量本身的作用

卡方检验

--两个分类变量之间的关系



χ^2检验的基本原理

- χ^2检验是用途很广的一种方法，主要用于分类数据的统计推断
  - 分类资料的分布是否符合假设
  - 两个率或两个构成比比较的卡方检验
  - 多个率或多个构成比比较的卡方检验
  - 分类资料的相关分析
  - 模型是否和样本数据完美拟合
- H_0：观察频数与期望频数没有差别
- 其原理为考察基于H_0的理论频数分布和实际频数分布间的差异大小，据此求出相应的P值

分析实例

- 在CCSS的分析报告中，所有受访家庭会按照家庭年收入被分为低收入家庭和中高收入家庭两类，现希望考察不同收入级别的家庭其轿车拥有率是否相同

家庭收入2级* O1. 是否拥有家用轿车 交叉制表

计数

		O1. 是否拥有家用轿车		
		有	没有	合计
家庭收入2级	Below 48,000	32	303	335
	Over 48,000	225	429	654
合计		257	732	989

理论轿车拥有概率

- 基于H_0成立，两样本所在总体无差别的前提下计算出各单元格的理论频数来
- 此时轿车拥有概率的最佳估计值就是样本合计的轿车拥有比例

家庭收入2级* O1. 是否拥有家用轿车 交叉制表

			O1. 是否拥有家用轿车		
			有	没有	合计
家庭收入2级	Below 48,000	计数	32	303	335
		家庭收入2级 中的 %	9.6%	90.4%	100.0%
	Over 48,000	计数	225	429	654
		家庭收入2级 中的 %	34.4%	65.6%	100.0%
合计		计数	257	732	989
		家庭收入2级 中的 %	26.0%	74.0%	100.0%

理论频数

$$T_{RC} = \frac{n_R n_C}{n}$$

- 基于轿车拥有概率的最佳估计值，就可以得出各单元格的理论频数

家庭收入2级* O1. 是否拥有家用轿车 交叉制表

			O1. 是否拥有家用轿车		
			有	没有	合计
家庭收入2级	Below 48,000	计数	32	303	335
		期望的计数	87.1	247.9	335.0
		家庭收入2级 中的 %	9.6%	90.4%	100.0%
	Over 48,000	计数	225	429	654
		期望的计数	169.9	484.1	654.0
		家庭收入2级 中的 %	34.4%	65.6%	100.0%
合计		计数	257	732	989
		期望的计数	257.0	732.0	989.0
		家庭收入2级 中的 %	26.0%	74.0%	100.0%

残差

- 设A代表某个类别的观察频数，E代表基于H_0计算出的期望频数，A与E之差被称为残差

家庭收入2级* O1. 是否拥有家用轿车 交叉制表

			O1. 是否拥有家用轿车		
			有	没有	合计
家庭收入2级	Below 48,000	计数	32	303	335
		期望的计数	87.1	247.9	335.0
		家庭收入2级 中的 %	9.6%	90.4%	100.0%
		残差	-55.1	55.1	
	Over 48,000	计数	225	429	654
		期望的计数	169.9	484.1	654.0
		家庭收入2级 中的 %	34.4%	65.6%	100.0%
		残差	55.1	-55.1	
合计		计数	257	732	989
		期望的计数	257.0	732.0	989.0
		家庭收入2级 中的 %	26.0%	74.0%	100.0%

残差平方和

- 残差可以表示某一个单元格中观察值和理论值的偏离程度
- 但残差有正有负，相加后会彼此抵消，总和仍然为0
- 为此可以将残差平方后求和，以表示样本总的偏离无效假设的程度

卡方统计量

- 另一方面，残差大小是一个相对的概念
  - 相对于期望频数为10时，20的残差非常大
  - 相对于期望频数为1000时20就很小
- 因此又将残差平方除以期望频数再求和，以标准化观察频数与期望频数的差别
  - 这就是卡方统计量，在1900年由英国统计学家Pearson首次提出，其公式为：

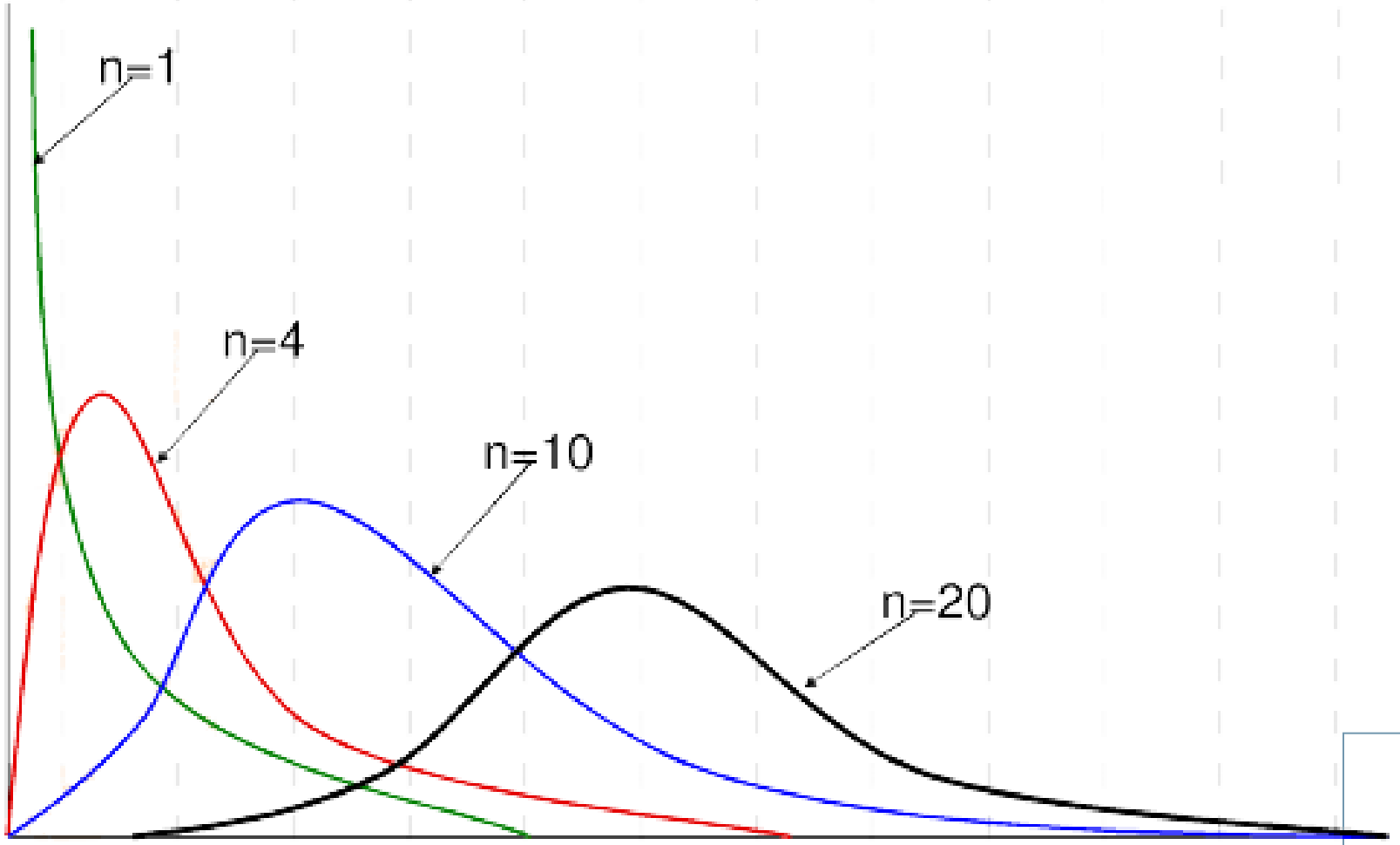
$$\chi^2 = \sum \frac{(A-E)^2}{E} = \sum_{i=1}^{k} \frac{(A_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(A_i - np_i)^2}{np_i} \qquad (\mathrm{i}=1,2,3,\cdots \mathrm{k})$$

卡方统计量

- 当观察频数与期望频数完全一致时，卡方值为0
- 观察频数与期望频数越接近，两者之间的差异越小，卡方值越小
- 反之，观察频数与期望频数差别越大，两者之间的差异越大，卡方值越大
- 当然，卡方值的大小也和自由度有关

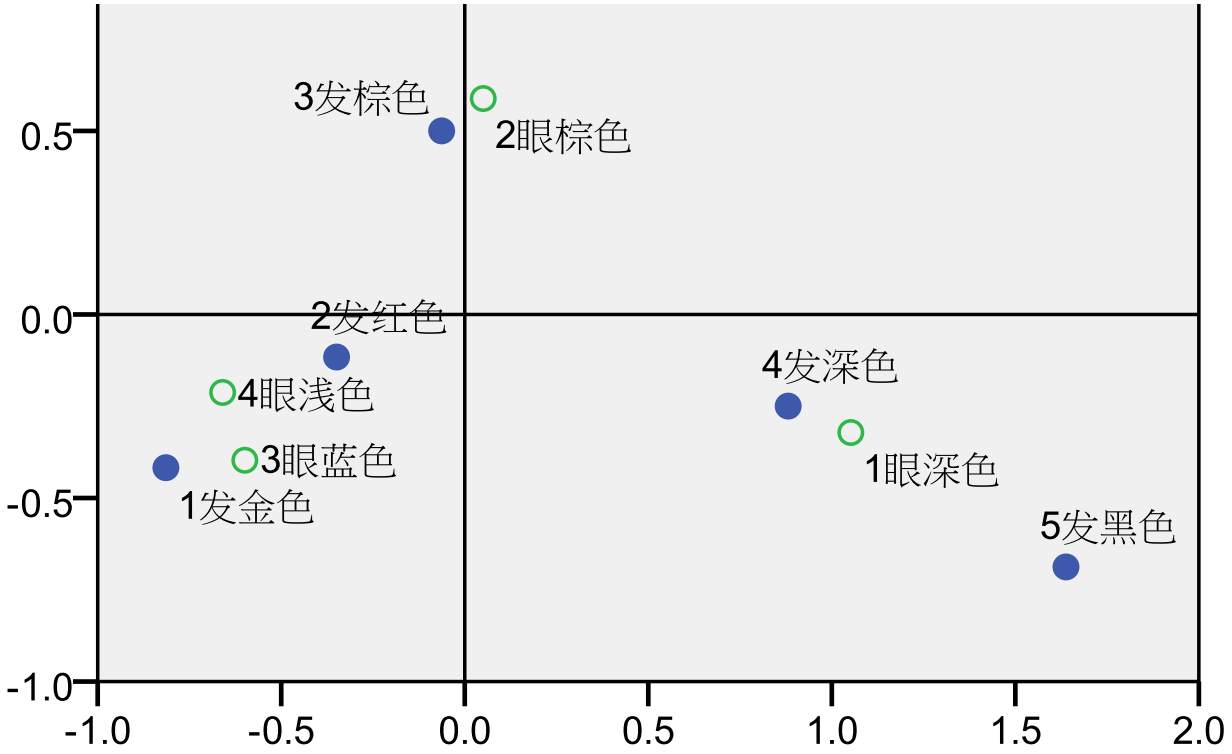
卡方分布

- 卡方值的大小不仅与A、E之差有关，还与单元格数（自由度）有关



当多分类变量检验出总体存在差异之后

- 仍然需要进行检验水准校正之下的两两比较
- 复杂建模方法，进行哑变量的精细比较
- 也可以换用图示分析的方法



χ^2值的校正

- 英国统计学家Yates认为，χ^2分布是一种连续分布，而四格表资料是分类资料，属离散分布，由此计算的χ^2值的抽样分布也应当是不连续的
- 当样本量较小时，两者间的差异不可忽略，应进行连续性校正
- Yates提出的办法是在每个单元格的残差中都减去0.5
  - 若n > 40 ，此时有 1< T <5时，需计算Yates连续性校正χ^2值
  - T <1，或n<40时，应改用Fisher确切概率法直接计算概率

如何进行列联表的检验

- 一般的看法
  - 当n≥40且所有T≥5时，用普通的卡方检验，若所得P约等于Alpha，改用确切概率法
  - 当n≥40但有1≤T<5时，用校正的卡方检验（如果有）
  - 当n<40或有T<1时，不能用卡方检验，改用确切概率法
- 建议的做法
  - 卡方检验可能有问题时，直接使用确切概率法
  - 计算量太大，无法得出确切概率时，使用蒙特卡洛抽样计算概率区间

蒙特卡罗估计

- 是“曼哈顿计划”计划的成员S.M.乌拉姆和J.冯·诺伊曼在核武器研究中产生的方法学副产品
- 因其计算原理仍然是和赌博有点关联，因此数学家冯·诺伊曼用驰名世界的赌城——摩纳哥的Monte Carlo城来命名这种方法
- 1777年，法国Buffon提出用投针实验的方法求圆周率，这被认为是蒙特卡洛方法的起源
- 计算机技术的快速发展，使得蒙特卡罗方法的大规模应用成为可能

蒙特卡罗估计的基本原理

配对卡方检验

- 用A、B两种方法检查已确诊的某种疾病患者140名，A法检出91名(65%)，B法检出77名(55%)，A、B两法一致的检出56名(40%)，问哪种方法阳性检出率更高？

A法	B法		合 口
	+	−	
+	56 (a)	35 (b)	91
−	21 (c)	28 (d)	49
合 口	77	63	140

配对卡方检验

- 对同一个体，分别有两次不同的测量，并最终构成了两组数据，因此研究框架是自身配对设计
- 基本思路：求出各对的差值，然后考察样本中差值的分布是否按照H_0假设的情况对称分布
- 显然，主对角线上的样本，两种检验方法的结论相同

A法	B法		合 口
	+	−	
+	56 (a)	35 (b)	91
−	21 (c)	28 (d)	49
合 口	77	63	140

配对卡方检验

- H_0：两法总体阳性检出率无差别，即B = C
- H_1：两法总体阳性检出率有差别，即B ≠ C
  - 非主对角线上的单元格才携带检验方法的差异信息
- 根据H_0得到b、c两格的理论数均为$T_b = T_c = (b+c)/2$，对应的配对检验统计量为

$$\chi^2 = \frac{(b-c)^2}{b+c}, \qquad \nu = 1$$

  - 一般在$b + c < 40$时，需用确切概率法进行检验，或者进行校正

分层卡方检验

- 卡方检验只能研究行、列两个变量间的关联
- 如何在控制第三个分类变量影响的前提下，得到更精确的分析结果？
- 例：
  - 收入级别间的拥有率不同
  - 各城市间轿车拥有率不同
  - 各城市的收入级别间轿车拥有率差距不同
  - 直接分析结论是否准确？
  - 关联强度在不同层间是否一致？

S0 城市			01. 是否拥有家用轿车 有	没有
100北京	收入	Below 48,000	8.8%	91.2%
		Over 48,000	38.2%	61.8%
	总计		28.8%	71.2%
200上海	收入	Below 48,000	3.7%	96.3%
		Over 48,000	30.4%	69.6%
	总计		22.0%	78.0%
300广州	收入	Below 48,000	15.1%	84.9%
		Over 48,000	34.8%	65.2%
	总计		27.3%	72.7%
总计	收入	Below 48,000	9.6%	90.4%
		Over 48,000	34.4%	65.6%
	总计		26.0%	74.0%

分层卡方检验

- 层间关联强度的齐性检验
  - H_0：各层间行、列变量的关联强度完全一致
- 控制分层因素影响之后的行列关联强度估计与检验
  - Mantel-Haenszel卡方检验
  - H_0：在控制了分层变量的影响之后，行、列变量的关联强度为0（或者为指定大小的数值）
  - 前提条件：各层关联强度相同，可以进行合并估计
- 无法合并估计时，考虑分层分析，或者使用logistic回归等方法建模，并加入交互项

Bernoulli试验序列

- 在重复实验中，如果对每一次实验，出现的结果只有两种情况，即Bernoulli试验
- 每次试验的条件不变，即每次试验中，结果A发生的概率不变（假设均为pi）
- 各次试验独立，即一次试验出现什么样的结果与前面已出现的结果无关
- 满足以上三个条件的n次Bernoulli试验构成的序列被称为Bernoulli试验序列

二项分布

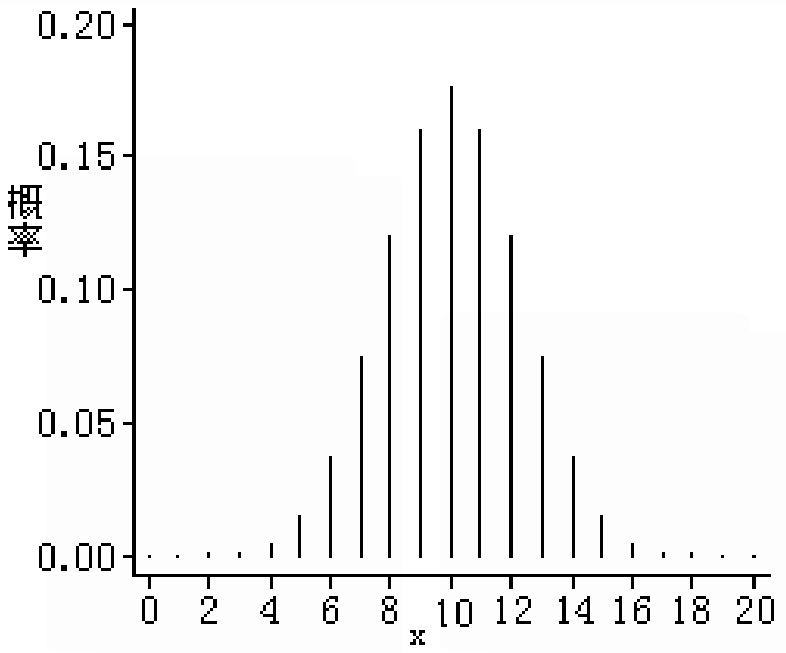
- 对于Bernoulli试验序列的n次试验，结局A出现的次数X的概率分布服从二项分布
- 二项分布指的是概率的分布
- 二项分布是一个离散型分布

X 的取值	0	1	…	k	…	n
取值概率	$\binom{n}{0}\pi^0(1-\pi)^{n-0}$	$\binom{n}{1}\pi^1(1-\pi)^{n-1}$	…	$\binom{n}{k}\pi^k(1-\pi)^{n-k}$	…	$\binom{n}{n}\pi^n(1-\pi)^{n-n}$

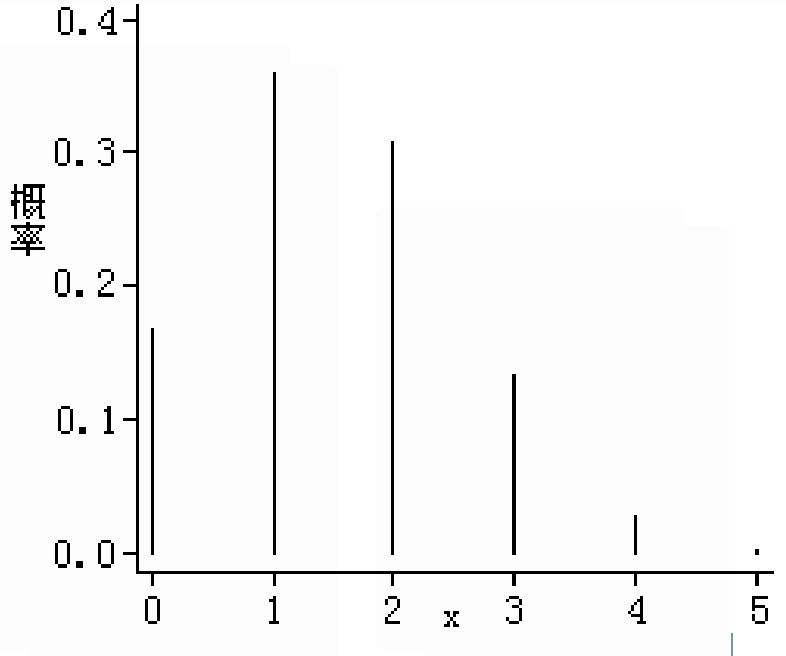
其相应取值概率为 $P(X=k)=\binom{n}{k}\pi^k(1-\pi)^{n-k}$

二项分布的两个参数

- 显然对于不同的n、不同的pi有不同的二项分布，它们是二项分布的两个参数
- 若*X*服从二项分布，则记*X*~B(n,pi)
  - 均数μ = n*pi，方差=n*pi*(1-pi)



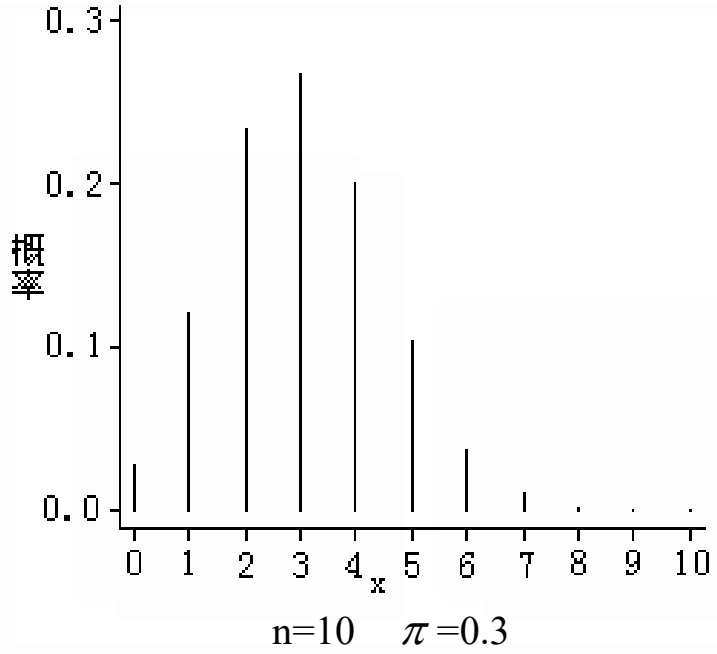
n=20， π=0.5



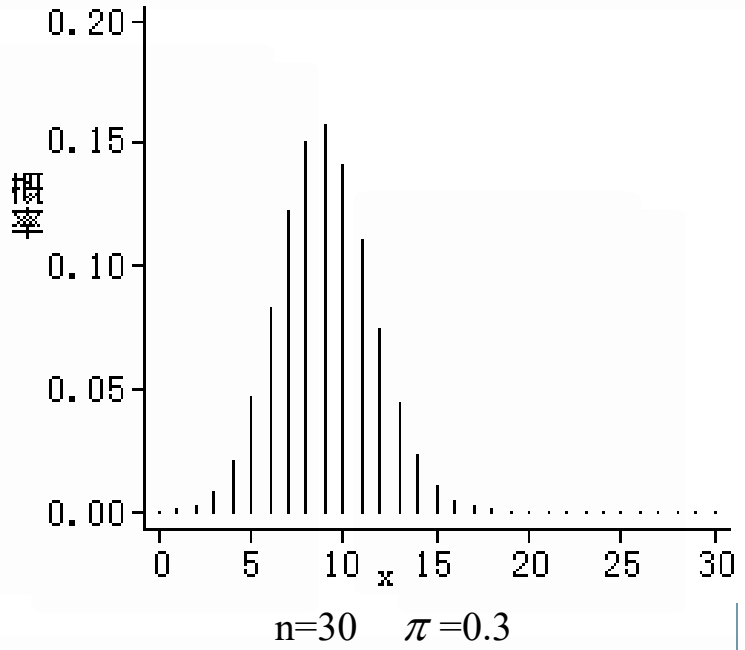
n＝5， π=0.3

二项分布的基本特征

- 当pi=0.5时，图形对称；当pi≠0.5时，图形呈偏态，但随n的增大，图形逐渐对称
- 因此，当n较大，pi不太极端时，可以采用正态近似方法计算概率分布规律（例如计算参考值范围）



n=10 π=0.3



n=30 π=0.3

样本率的抽样分布

- 对于大量重复随机抽样而言，样本率p围绕着总体率波动，样本量n的值越大，这种波动就越小
- 当n的值充分大时，p的分布就近似于均数为pi，标准差为sqrt(pi*(1-pi)/n)的正态分布
  - 一般的标准是n*pi和n*(1-pi)均大于5，且n>40
  - 当样本情况接近此标准时，往往会进行校正
- 注意：上文所说的样本率的标准差，实际上就是阳性数x的标准误

总体率的区间估计

- 直接计算概率
  - 在样本例数较小，且样本率接近1或0，即阳性事件发生率很高或很低时，应当按照率的抽样分布规律确定总体率的可信区间
- 正态近似
  - 当n较大，pi和1-pi均不太小时，样本率的抽样分布近似正态分布，因此可按正态近似法求总体率的1-α 可信区间

率的比较

- 如前所述，当n较大，pi和1-pi均不太小时，样本率的抽样分布近似正态分布，可利用正态分布的原理作假设检验
- 反之，则可使用二项分布自身的概率分布进行假设检验，这种方法被称为二项分布检验（确切概率法）

更多的二项分布置信区间计算方法

- Wald法：即正态近似法
  - N越大，计算得到的区间越准确？
  - P越居中，区间越准确？
- Wilson得分：震荡率和覆盖程度均优于Wald法

$$CI_W = \frac{X + \kappa^2/2}{n + \kappa^2} \pm \frac{\kappa n^{1/2}}{n + \kappa^2}(\hat{p}\hat{q} + \kappa^2/(4n))^{1/2}$$

- AC区间：覆盖率更高，但偏宽，小样本表现优于Wilson法

$$CI_{AC} = \tilde{p} \pm \kappa(\tilde{p}\tilde{q})^{\frac{1}{2}}\tilde{n}^{-\frac{1}{2}}$$

- Jeffreys贝叶斯区间：用Beta分布作为先验分布来推断二项分布参数
  - 小样本时覆盖率适中

更多的二项分布置信区间计算方法

- Clopper-Pearson法：也被称为确切概率法
  - 该名称实际上指的是他使用的分布是确切的概率分布，但并不意味着计算出的区间也是“确切”的
  - 计算中会用到Beta分布，因此也被称为Clopper-Pearson interval based on Beta distribution
- Logit法：利用logit变换进行区间求解
- Newcombe方法：用于率差置信区间的计算

变量的关联性分析



什么是变量的关联性分析（相关分析）

- 用于考察变量间数量关联密切程度的统计分析方法
  - 焦不离孟，孟不离焦
  - 身高越高，则体重一般也越大
- 几乎所有涉及到多个变量的假设检验方法，都可以被看作是这些变量间的关联性分析
  - t检验：分组变量与连续因变量间的关联性分析
  - 卡方检验：行、列分类变量间的关联性分析
  - 聚类分析：案例间的关联性分析
  - 多变量回归：因变量和一组自变量间的关联性分析

相关分析的分类

- 按照变量数量
  - 一个变量 vs. 另一个变量
  - 一个变量 vs. 一组变量
  - 一组变量 vs. 另一组变量
  - 多组变量之间的相关分析
- 按照变量种类
  - 连续变量
  - 有序分类变量
  - 无序分类变量
- 变量数量 × 变量种类：这下事情搞大了。。。

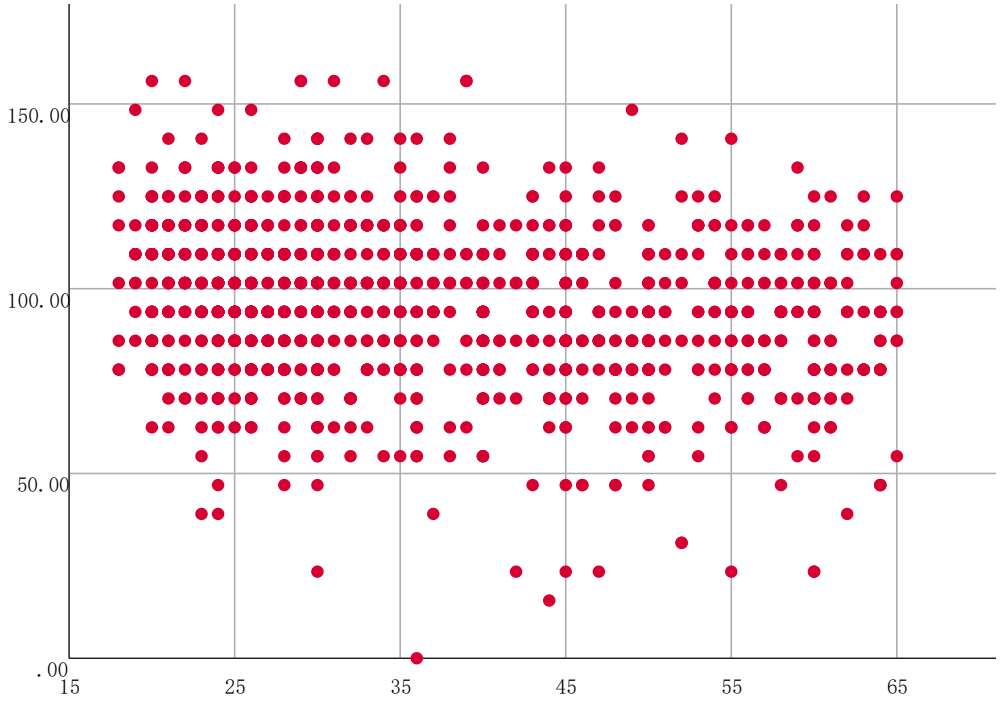
各种相关系数

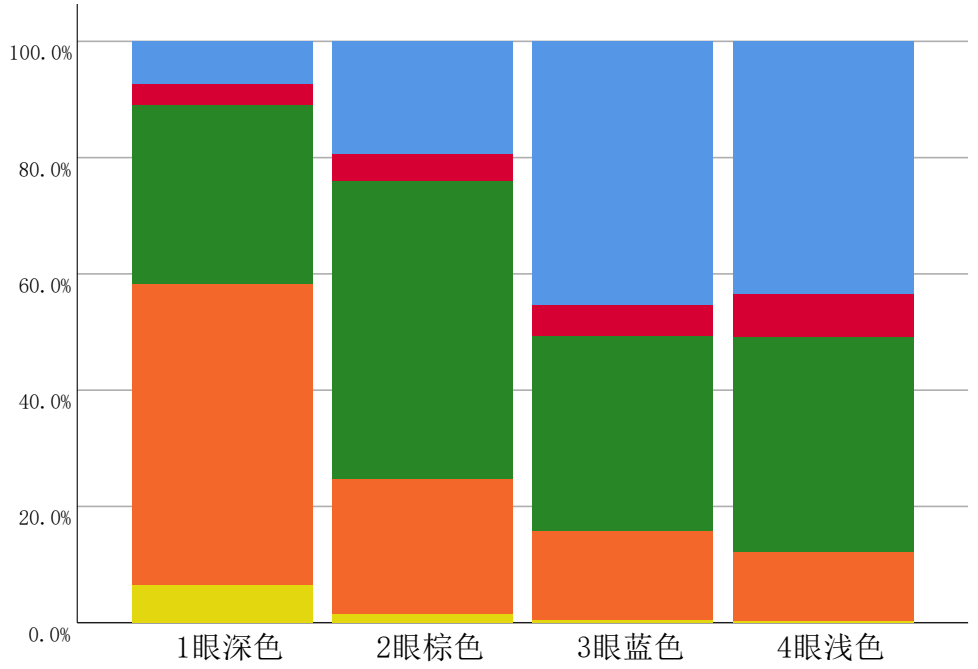
- 连续 vs. 连续：Pearson相关系数
  - Spearman秩相关系数
- 有序 vs. 有序：Gamma系数等
- 无序 vs. 无序：列联系数等
  - 基于卡方统计量进一步推导而来
  - 无方向，0~1
  - OR/RR：一类特殊的关联强度指标
- 连续 vs. 分类：Eta
  - 实质为方差解释度



统计图/统计表在相关分析中的重要性

- 连续变量：用散点图确认关联趋势是否为直线
- 分类变量：分组条图、马赛克图（分组百分条图）等工具





两连续变量的相关：常用术语

- 直线相关：两变量呈线性共同增大，或者呈线性一增一减的情况
- 曲线相关：两变量存在相关趋势，但是为各种可能的曲线趋势
  - 此时如果直接进行直线相关分析，有可能得出无相关性的结论
- 正相关与负相关：如果A变量增加时B变量也增加，则为正相关，如A变量增加时B变量减小，则为负相关
- 完全相关：不属于统计学的研究范畴
  - 可分为完全正相关和完全负相关两种
  - 零相关

两连续变量的相关：Pearson相关系数

- 在双变量正态分布的前提假设下，基于方差、协方差等指标而来
  - 方差：每个变量的离散程度（信息量大小）
  - 协方差：各变量共同携带的信息量大小
  - 协方差^2/(方差X *方差Y)：决定系数，两个变量总信息量中的共同部分占比
  - $\sqrt{决定系数}$：相关系数r
- 相关系数r是一个无单位的量值，且-1＜r＜1
  - 其正负反映了相关的方向
  - ｜r｜越接近于 1，说明相关性越好
  - ｜r｜越接近于 0，说明相关性越差

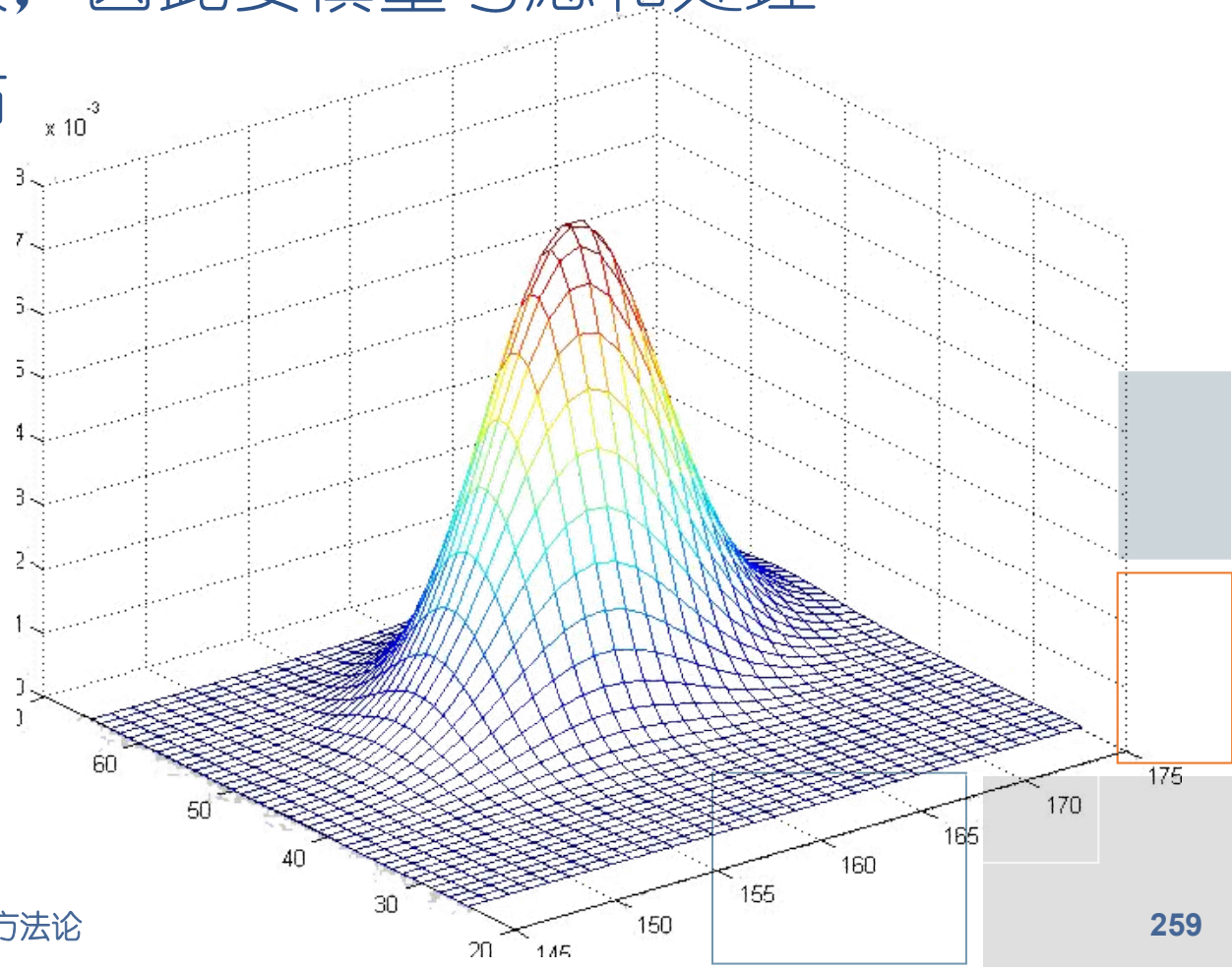
Pearson相关系数的检验

- H_0：两变量间无直线相关关系，ρ = 0
- H_1：两变量间有直线相关关系
- 检验的方法仍然是t检验

$$t = \frac{r - 0}{s_r}, \nu = n - 2 \qquad s_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

Pearson相关系数的适用条件

- 必须是线性相关的情形！！！
- 极端值对相关系数的计算影响极大，因此要慎重考虑和处理
- 要求相应的变量呈双变量正态分布
  - 其实基本做不到
  - 对此有一定的耐受性
  - 只要分布不是特别偏就成



Spearman秩相关系数

- 当相关分析的两变量X、Y并不符合适用条件时，r就不能正确地反应X、Y间的相关性
  - 方差/协方差能否正确代表数据的离散程度/信息量？
- Spearman提出首先对数据做秩变换，然后再计算两组秩间的直线相关系数
  - 显然，这实际上就是秩变换分析思想的一个具体应用
- 为了有别于r，该相关系数用r_S表示
- 该相关系数也是测量有序分类资料相关程度的指标之一

相对危险度（RR，Relative Risk）

$$RR = \frac{P_t}{P_c} = \frac{a/n_t}{c/n_c}$$

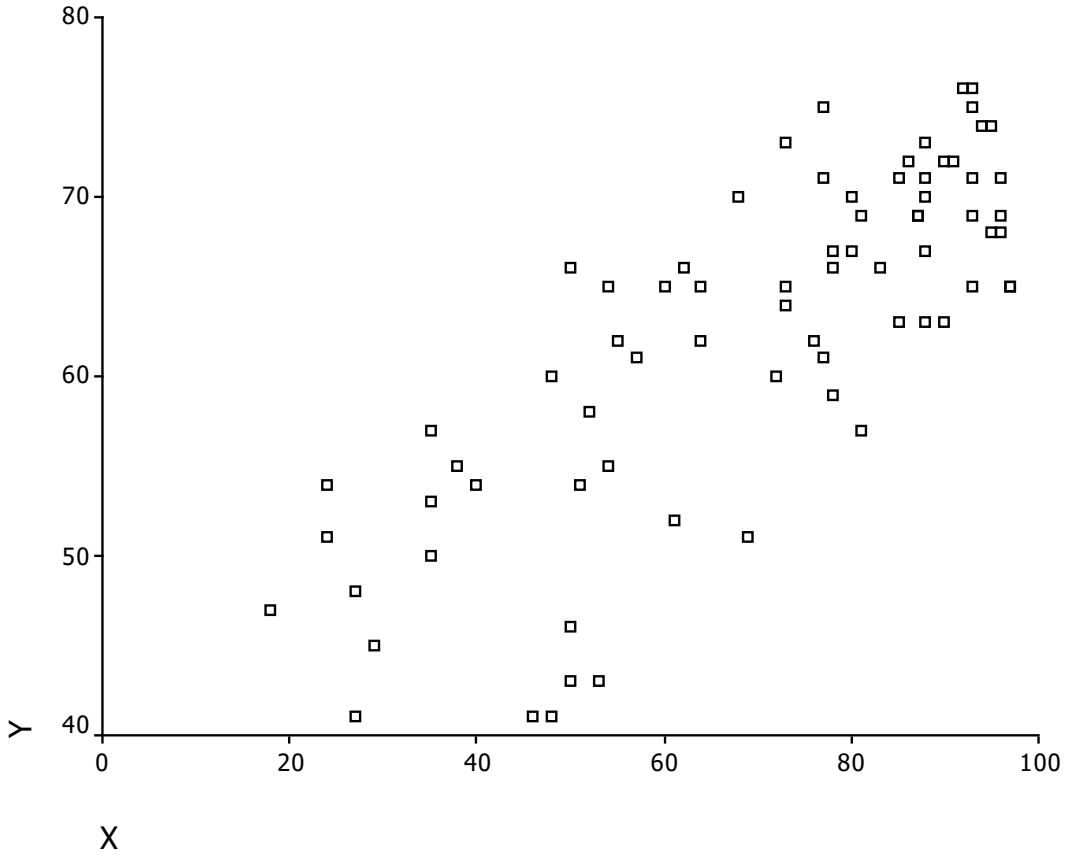
- 表示两种情况下发病密度或者说发病概率之比
  - P_t：实验组人群反应阳性概率
  - P_c：对照组人群反应阳性概率
- 如果RR>1，说明相应的自变量取值增加，会导致个体的发病/死亡风险增加若干倍
- 例：吸烟者的发病概率是非吸烟者的5倍
- RR在医学中得到了极为广泛的应用

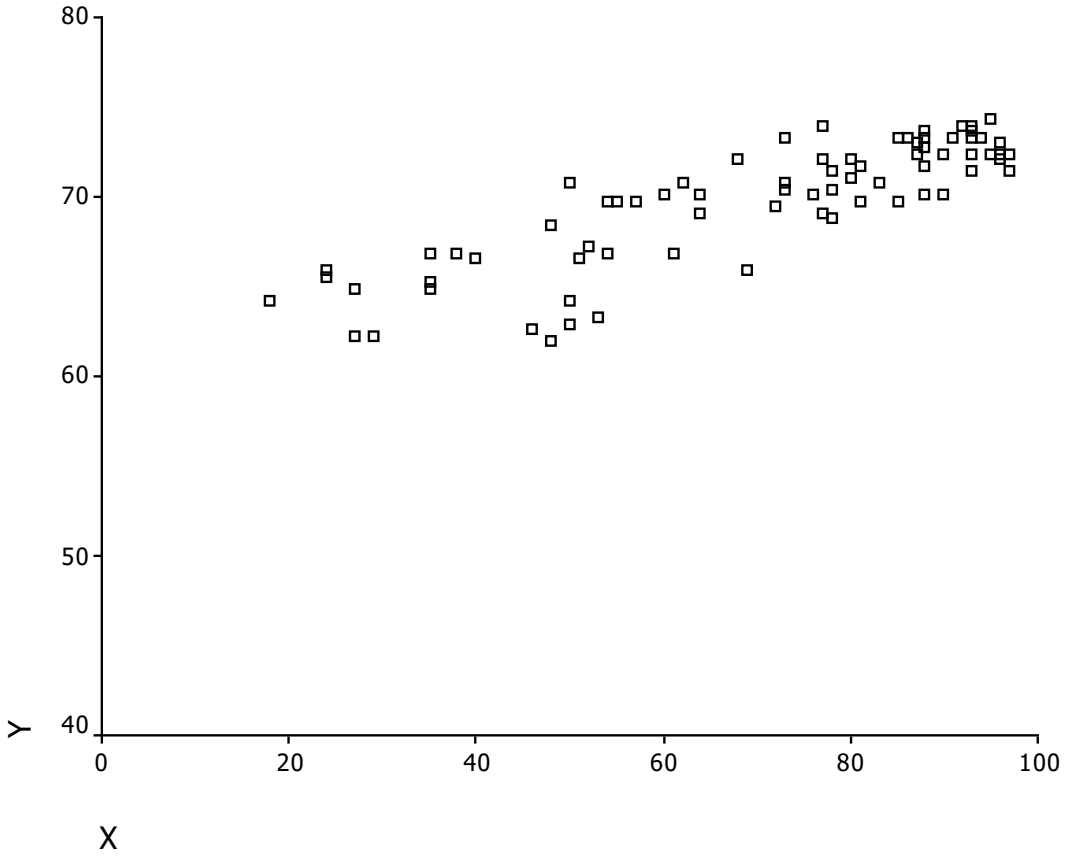
优势比（OR，Odds Ratio）

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}$$

- RR的计算要求得到各组的反应概率，这在回顾性研究中很难满足
  - 此时研究者往往使用OR值代替RR值
- OR：为下列两种比例之比
  - 反应阳性人群中实验因素有无的比例 a/b
  - 反应阴性人群中实验因素有无的比例之比 c/d
- 例：该疾病病例中吸烟/非吸烟者的比例是非病例中吸烟/非吸烟者比例的3倍
- OR可以间接反映关联强度，但是理解上比较困难
- 发病概率较低时，OR往往近似的在按照RR的含义进行解释和使用

相关与回归的区别和联系





- 反映关联紧密程度 vs. 反映自变量对因变量数量上的影响

回归分析概述

- 研究一个连续性变量（因变量）的取值随着其它变量（自变量）的数值变化而变化的趋势
- 通过回归方程解释两变量之间的关系显的更为精确，可以计算出自变量改变一个单位时因变量平均改变的单位数量，这是相关分析无法做到的
- 除了描述两变量的关系以外，通过回归方程还可以进行预测和控制，这在实际工作中尤为重要

线性回归模型的基本框架

- 线性回归假定自变量对因变量的影响强度始终保持不变

$$\hat{y} = \alpha + \beta x$$

- ŷ：给定自变量的取值时， y的估计值（所估计的平均水平）
  - 因变量的预测值可以被分解成两部分：常量+回归部分
- a/α：常量（constant），自变量x取值均为零时y的平均估计量
  - 可以被看成是一个基线水平
  - 多数情况下没有实际意义，研究者也不关心
- b/β：（偏）回归系数，自变量x 改变一个单位，y估计值的改变量
  - 因变量Y的变异中，可以由X直接估计的部分

线性回归模型的基本框架 $y_i = a + bx + e_i$ $e_i \sim N(0, \sigma^2)$

- 估计值和每一个实测值之间的差被称为残差。它刻画了因变量y除了自变量x以外的其它所有未进入该模型，或未知但可能与y有关的随机和非随机因素共同引起的变异，即不能由x直接估计的部分
- 为了方程可以得到估计，往往假定ε_i服从正态分布$N(0,\sigma^2)$
- 作为整体，模型可以对残差的离散程度进行估计

线性回归模型的适用条件

- 线性趋势：如果不是，则不能采用线性回归来分析
  - 可以通过散点图来加以判断
- 独立性：实际上就是要求残差间相互独立，不存在自相关，否则应当采用自回归模型来分析
  - 数据背景考察，DW检验，自相关/偏相关分析
- 正态性：就自变量的任何一个线性组合，因变量y均服从正态分布
  - 实际上是要求残差e_i服从正态分布
- 方差齐性：就自变量的任何一个线性组合，因变量y的方差均相同
  - 实质就是要求残差的方差齐

线性回归模型的适用条件：变量分布

- 因变量
  - 无法直接获得正态性、方差齐性等信息，但可以做初步观察
  - 一般而言，如果因变量分布极端偏态，则多半残差分布会有问题
- 自变量
  - 均为可被精确测量的确定数值（只能是连续变量）
- 自变量分布特征极端偏态时
  - 直接建模往往实用性不佳，因此可以考虑变量变换
  - 但是变量变换又会导致模型假定的线性关联发生改变，因此不能一刀切

线性回归模型的适用条件：样本量

- 可以对模型所需样本量做精确计算，但这种方法在筛选多个候选自变量的研究背景下基本没有实用价值
- 根据经验，案例数应当在希望分析的自变量数的20倍以上为宜
  - 实质上样本量和模型的决定系数有关，可通过迭代的方法进行计算
- 样本量不足并不意味着不能拟合模型，只是参数估计值可能不稳定，或者不能得到应当本来有统计学意义的检验结果

线性回归模型中的常用指标

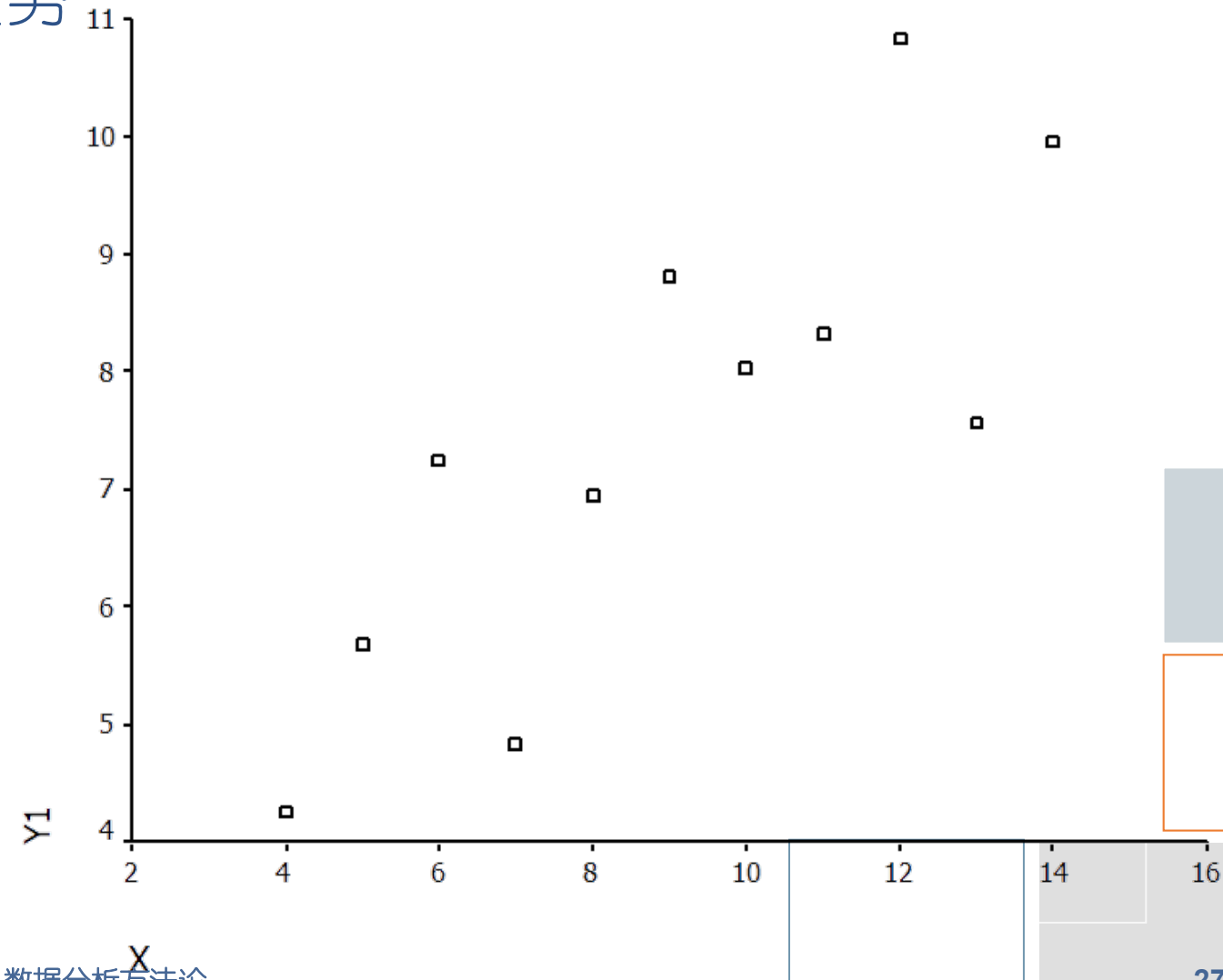
- 决定系数
  - 模型整体价值的衡量指标
  - 相应的相关系数的平方，用R^2表示
  - 反映因变量y的全部变异中能够通过回归关系被自变量解释的比例
- 偏回归系数
  - 反映某一个自变量在数量上对因变量的影响强度
  - 相应的自变量上升一个单位时，因变量取值的变动情况
- 标化偏回归系数：量纲问题
  - 用于自变量间重要性的比较

自变量在回归模型中的相对重要性

- 基于决定系数增量
  - 作为第一个输入变量时模型的 R^2
  - 作为最后一个输入变量时模型的增量 R^2
  - 所有模型的增量 R^2平均，这在输出中称为 LMG/Shapley值
- 基于标化系数（无法避免变量共线性的影响）
  - 标准化系数的平方
  - 标准化系数乘以相关系数
  - 边际相关系数

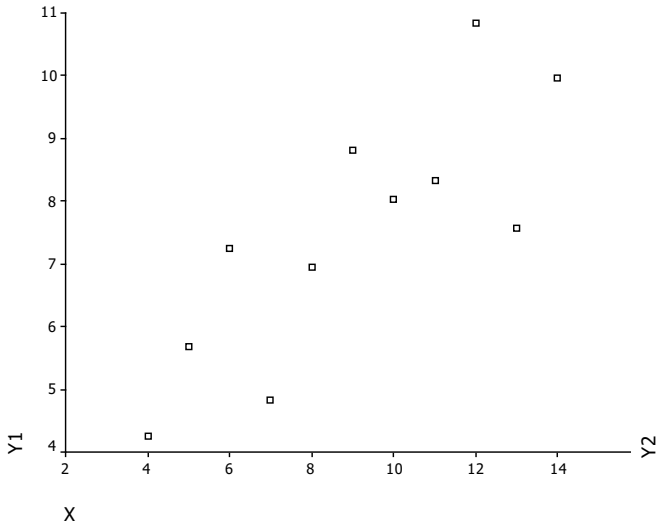
线性回归模型的标准建模步骤

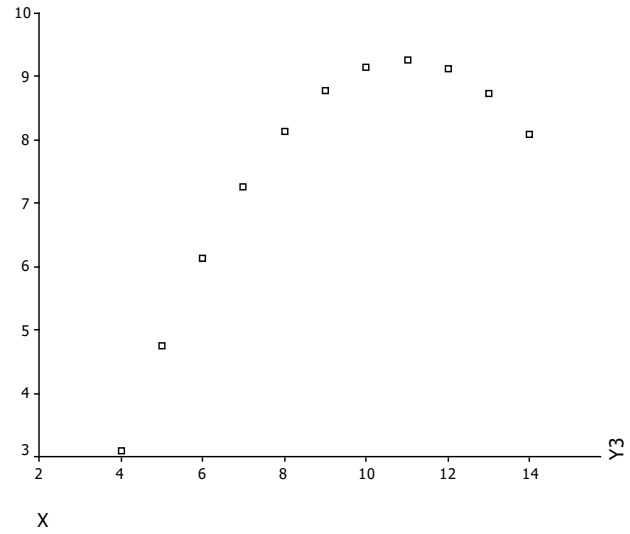
- 做出散点图，观察变量间的趋势
- 散点图考察的三个信息层次

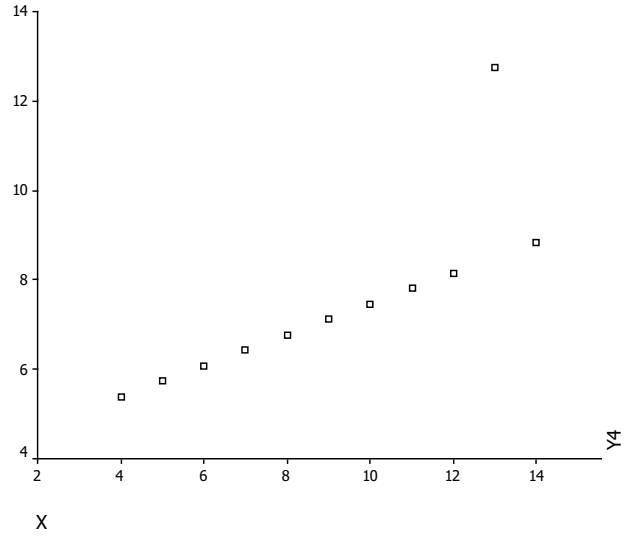


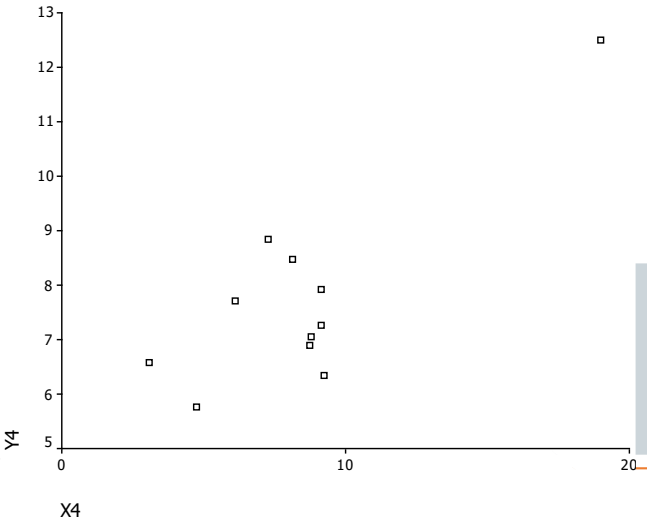
线性回归模型的标准建模步骤

- 只有确认为线性趋势的候选自变量才能直接纳入后续分析步骤





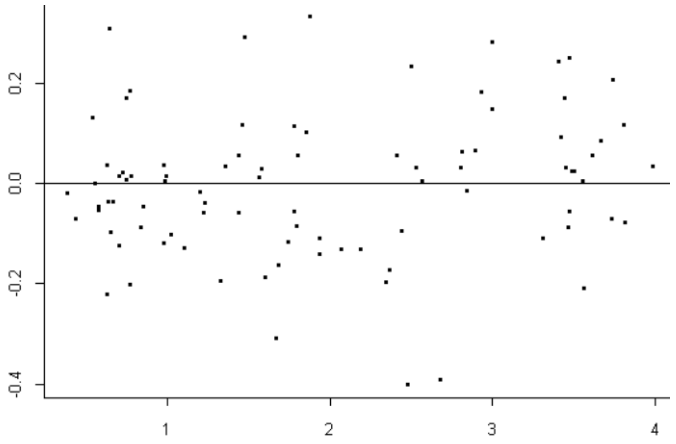


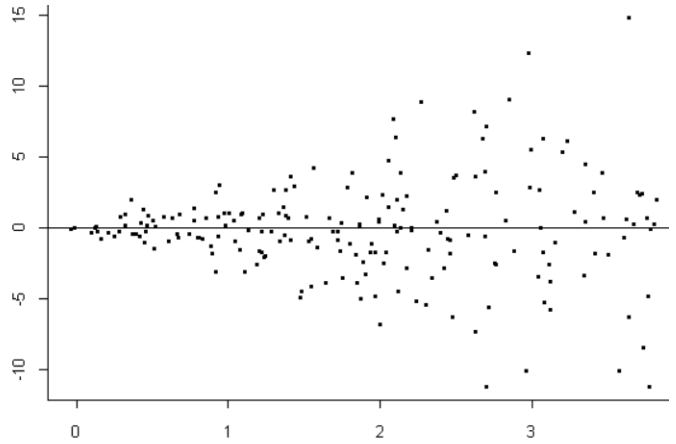


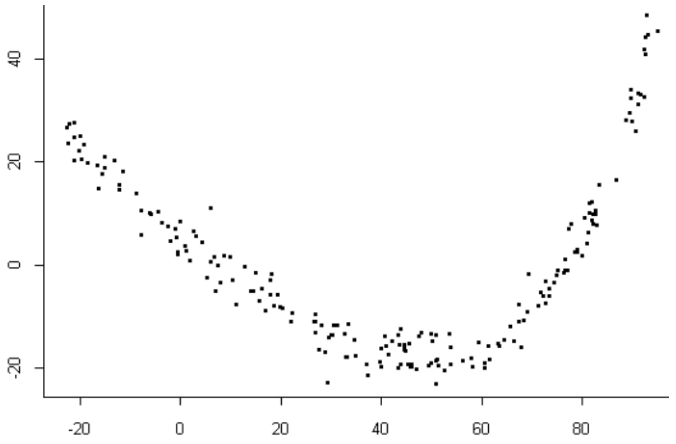
线性回归模型的标准建模步骤

- 考察数据的分布，进行必要的预处理
  - 观察自变量取值是否过于极端
  - 初步观察因变量的正态性
  - 初步观察可能的方差不齐迹象等问题
- 进行候选自变量筛选，初步建立直线回归模型
  - 科研领域的分析，尽量采用手工筛选
- 残差分析
  - 残差间是否独立（Durbin-Watson检验）
  - 残差分布是否为正态（图形或统计量）

线性回归模型的标准建模步骤







- 残差分析
- 强影响点的诊断
- 多重共线性问题的判断
- 这三个步骤往往混在一起，难以完全分出先后

多变量与逐步回归

- 多候选自变量的常见场景
  - 对因变量产生影响的因素不只一个
    - 研究中需要同时考虑年龄、性别等可能的混杂因素的作用
  - 分析哪些因素对因变量产生影响，或者排除了某些混杂因素的作用，分析所关心的因素对因变量的影响
- 一般说研究人员也会尽可能多地列出可能与因变量有关的自变量，生怕遗漏重要变量
  - 其中有一些变量对因变量实质上可能根本没有影响或影响很小
  - 错误地选进模型，会使模型应用的精度降低、代价升高。正是由于这些原因，在应用回归分析时，有必要对进入模型的自变量作精心选择

逐步回归：前进法

- forward-entry procedure
- 事先给定一个挑选自变量进入方程的标准
  - 一般给定一个P的界值，也有给定F值的
- 由无效模型开始，方程中除常数项外没有自变量
- 按自变量对y的贡献大小由大到小依次挑选进入方程
- 每选入一个变量进入方程，则重新计算方程外各自变量（在扣除了已选入变量的影响后）对y的贡献，直到方程外变量均达不到入选标准，没有自变量可被引入方程为止
- 该法只考虑选入变量，一旦某变量进入模型，就不再考虑剔除

逐步回归：后退法

- backward-elimination procedure
- 与前进法相反，后退法是事先给定一个剔除自变量的标准
- 以全模型开始，全部自变量都在方程之中
- 然后，按自变量对y的贡献大小由小到大依次剔除。每剔除一个变量，则重新计算未被剔除的各自变量对y的贡献
- 直到方程中所有变量均符合选入标准，没有自变量可被剔除为止
- 该法只考虑剔除，自变量一旦被剔除，则不再考虑进入模型

逐步回归

- 以上两种方法均有内在的缺陷，很自然的，人们希望能够将向前法和向后法相结合
- 逐步向前法 (forward stepwise）
  - 事先给定一个剔选变量的标准，按自变量对*y*的贡献大小由大到小依次挑选进入方程
  - 每选一个变量进入方程，则重新计算各自变量对*y*的贡献。并考察已在方程中的变量是否由于新变量的引入，其作用被新变量代替或部分代替了，抑制了它的作用并退化为不显著。如果有，则剔除。直到平衡为止
- 逐步向后法(backward stepwise)

变量自动筛选

绝对不是

万能的!



多变量建模的建议筛选策略

- 单自变量筛选，筛掉那些显然无关联的候选变量
  - 相关/回归等方法均可
  - 界值可以放大至0.2-0.5
- 尝试建立多自变量模型，可手动、也可利用自动筛选方法，但使用后者时要谨慎
- 多自变量和单自变量模型结果相矛盾时，以前者为准
- 结果不符合专业知识时，尽量寻找原因
  - 是否有交互项、有共线性、有强影响点、非直线关系

为什么要做样本量估计

- 太小：无法检验出预期存在的差异
  - 1985~1986年，15种临床医学杂志，450个两样本率比较，设差异为0.5倍，只有19%（85个）的Power在0.8以上
- 太大：造成无谓的浪费
  - 混杂因素可能更多
  - 管理实施上更加困难
  - 结果可能无专业意义
    - 只要样本量足够大，必然会有统计学差异

大数据时代，做样本量的估计还有意义吗？



样本量估计的应用场景

- 设计阶段
  - 按给定条件计算需要的样本量，以严格控制成本，提高产出效率
- 分析阶段
  - 对已进行的实验计算检验效能

计算样本量时所需的条件

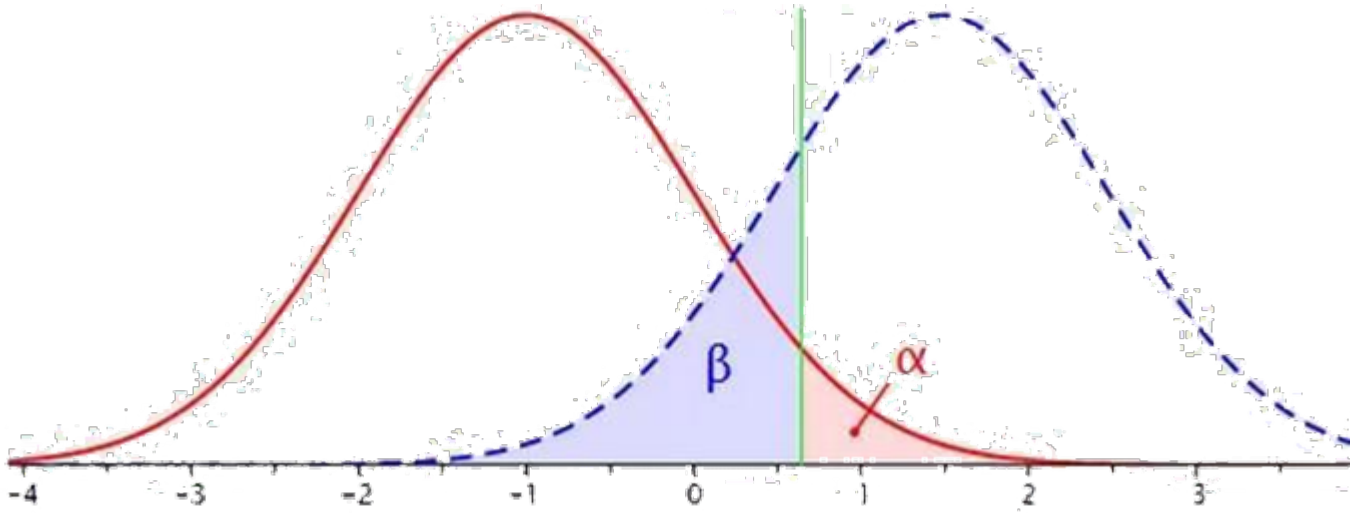
- 主要目的与主要指标
  - 当存在多个指标时，以最主要指标为准进行计算
- 设计/分析方法
  - 好的分析方法可以节约样本量
  - 但复杂的设计会给样本量计算带来困难
- 第一类错误的大小
- 单、双侧问题
- 基线水平
  - 一般不直接参与计算，但是有助于制定期望检出的差值

计算样本量时所需的条件

- 期望检出的差值
  - 一般为专业上认为有意义的差值
- 资料的离散程度
  - 预试验（I期临床试验，前期的小样本研究）
  - 查历史资料
  - 均数的1/2~1/4
  - 分类资料不用另行查找
- 检验效能的大小
  - 不应当低于0.75，一般都在0.8以上

样本量计算的基本原理

- 直接从假设检验的I、II类错误问题发展而来
- 原假设成立时，如果超过某个C点，则原假设被拒绝，即I类错误
- 在备择假设成立的情况下，如果低于某个C点，则原假设不能被拒绝，备择假设不能被接受，即II类错误
- 两个公式中的C点应当相同，联立求解，得到相应的N



单组/两组均值比较的样本量计算

- 单样本/配对样本 $N=\left[\frac{\left(u_\alpha+u_\beta\right)\sigma_d}{\delta}\right]^2$ $u_\beta=\delta\frac{\sqrt{N}}{\sigma_d}-u_\alpha$
- 成组两样本 $N=\left[\frac{2\left(u_\alpha+u_\beta\right)\sigma}{\delta}\right]^2$ $u_\beta=\frac{\delta\sqrt{N}}{2\sigma}-u_\alpha$

- 公式中的δ/σ也被称为effect size，即希望能够检验出的相对差异
- 实际使用中需要加上10%~20%的例数，以补充可能的损失

多组均值比较的样本量计算

- 实用：根据最重要的两样本组数据的特征进行样本总量外推
- 严格：基于组间变异进行计算
  - 难点在于effect size如何进行估计

$$f = \sqrt{\frac{\sum_{i=1}^{k} p_i * (\mu_i - \mu)^2}{\sigma^2}}$$

where $p_i = n_i / N$,
n_i = number of observations in group i
N = total number of observations
μ_i = mean of group i
μ = grand mean
σ^2 = error variance within groups

  - Cohen建议将f = 0.1, 0.25, 0.4分别作为小、中、大的效应值用于估计

率的比较的样本量计算

- 分类变量不需要另行估计变异，因为可以从“均数”计算得出
- 计算原理上和均数比较完全相同，但是难点在于如何给出effect size的合理估计值
- 方法1：effect size = 2*(arcsin(sqrt(p1))-arcsin(sqrt(p2)))
  - Cohen建议将h = 0.2, 0.5, 0.8分别作为小、中、大的效应值用于估计
- 方法2：正态近似估计

$$N = \frac{(u_\alpha + u_\beta)^2 4\pi_c(1-\pi_c)}{(\pi_1 - \pi_2)^2} \qquad u_\beta = \frac{\sqrt{N}|\pi_1 - \pi_2|}{2\sqrt{\pi_c(1-\bar{\pi})}} - u_\alpha$$

相关参考资料

- Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum
- https://www.statmethods.net/stats/power.html

分析实例

- 用新药降低高血脂患者的胆固醇，研究者规定试验组与对照组（安慰剂）相比，血清胆固醇平均降低0.5mmol/L以上，才有推广价值
- 引用文献中胆固醇的标准差为0.8mmol/L，规定单侧α = 0.05，power = 0.90，要求估计样本含量

$$N = \left[\frac{2(1.64 + 1.28)0.8}{0.5}\right]^2 = 87$$

分析案例

- 研究用新药降低高血脂患者胆固醇的效果，研究者规定试验组的血清胆固醇平均降低0.5mmol/L以上，才有进一步研究的价值
- 引用文献中胆固醇的标准差为0.8mmol/L，规定单侧α = 0.05，power =0.90，要求估计样本含量
- H = 0.5/0.8 = 0.625

分析案例

- 研究用新药降低高血脂患者胆固醇的效果，研究者规定试验组与对照组（安慰剂）相比，血清胆固醇平均降低0.5mmol/L以上，才有推广价值
- 引用文献中胆固醇的标准差为0.8mmol/L，规定单侧α = 0.05，power = 0.90，要求估计样本含量
- H = 0.5/0.8 = 0.625

分析案例

- 常规条件下某动物模型出现阳性结局的概率为40%，某研究人员考虑采用另一种方法进行试验，使用60个动物进行该研究，预期的成功概率为50%，请估计该检验的效能是否充足
  - 如果信息不足，宁可使用双侧检验进行估计
- h = 2 * (arcsin(sqrt(p1)) - arcsin(sqrt(p2))) = 0.2014
- 后续直接调用单样本t检验的样本量计算框架

分析案例

- 原方法下某动物模型出现阳性结局的概率为15%，现考虑采用改进的新方法进行比较，预期新方法阳性概率为30%，请估计该研究所需的动物样本量
  - 可以考虑使用单侧检验进行估计
- h = 2 * (arcsin(sqrt(p1)) - arcsin(sqrt(p2))) = 0.3639
- 后续直接调用两样本t检验的样本量计算框架

常用假设检验方法的正确选择



确定假设检验方法的两大要素

- 研究设计
  - 成组设计：比较男性和女性对某电影的评分
  - 配对设计/配伍设计：受试者在三种候选包装中选择最满意的一种
  - 更加复杂的试验设计：拉丁方设计、均匀设计等
- 变量类型
  - 连续变量：两种促销方式的销量比较
  - 有序变量：评价结果为优良中差的比较
  - 无序变量：不同学校的毕业生择业去向比较

对自身总体进行推断

- 连续变量：单样本t检验
  - 新样品评分是否达到基本要求
- 无序分类变量：二项分布、卡方分布检验
  - 筛子分布是否均匀
  - 性别比例是否为假定的数值
- 有序分类变量：秩和检验
  - 成绩优良中差的分布是否符合假定状况

因变量：连续

- 单个自变量：二分类　t检验
  - 成组t检验　　促销活动对销量有无作用
  - 配对t检验　　培训前后销售额的比较
- 单个自变量：多分类　单因素方差分析
  - 三种促销方式的销量比较
  - 后续有两两比较问题
- 单个自变量：连续
  - 相关分析：考察两变量在数量上的依存关系
  - 回归分析：考察一个变量对另一个变量影响程度的大小

因变量：连续

- 多个自变量：分类变量为主 方差分析模型
  - 便于两两比较
  - 有多种图形帮助分析
  - 自变量间可能有主次之分
- 多个自变量：连续变量为主 线性回归模型
  - 本质上和方差分析模型等价
  - 可利用散点图、回归线等图形帮助解释
  - 可用逐步法对自变量进行自动筛选

因变量：有序分类

- 当分类数较多时，可近似按照连续变量处理
- 单个自变量
  - 二分类　秩和检验
  - 多分类　秩和检验
    - 后续有两两比较问题
  - 连续　　Logistic回归
- 多个自变量：
  - Logistic回归
  - 判别分析
  - 最优尺度分析

因变量：无序分类（含二分类）

- 单个自变量
  - 二分类　卡方检验
  - 多分类　卡方检验
    - 后续有两两比较问题
  - 连续　　没有简单模型
- 多个自变量：分类变量为主
  - Logistic回归分析
  - 判别分析
  - 最优尺度分析

简单推断方法总结

- 对连续因变量的统计方法是最全面的
- 有序分类变量往往都近似的在当作连续性变量处理，或者退化成无序分类变量处理
- 无序分类因变量一般使用卡方检验，但只能得到粗糙的结果，精细分析困难
- 复杂的统计模型实际上都是让各种变量在拟合时尽量向连续性变量靠拢（变量变换、概率评分）