

# 数据分析残卷

于淼

2018-12-31



# Contents

序	5
<b>1 导论</b>	<b>7</b>
1.1 数据科学 . . . . .	7
1.2 基本问题 . . . . .	7
1.3 工作流程 . . . . .	7
1.4 概率与分布 . . . . .	8
1.5 统计量 . . . . .	8
1.6 统计推断 . . . . .	8
1.7 统计模型 . . . . .	9
1.8 其他主题 . . . . .	9
1.9 应用 . . . . .	10
1.10 链接 . . . . .	10
<b>2 数据分析工具</b>	<b>11</b>
2.1 基础知识 . . . . .	11
2.2 命令行基础 . . . . .	11
2.3 版本控制 . . . . .	12
2.4 数据获取 . . . . .	12
2.5 高级命令 . . . . .	12
2.6 R . . . . .	13
2.7 Python . . . . .	31
2.8 Tex . . . . .	32
<b>3 可复算性研究</b>	<b>37</b>
3.1 Replication . . . . .	37
3.2 Reproducible . . . . .	37
3.3 研究流程 . . . . .	39
3.4 数据分析步骤 . . . . .	40
3.5 数据分析文件结构 . . . . .	41
3.6 文本化统计编程-Knitr . . . . .	41
3.7 结果通讯 . . . . .	41
3.8 检查列表 . . . . .	42
3.9 基于证据的数据分析 . . . . .	42
3.10 结果可解释 . . . . .	42
3.11 数据分析的理论 . . . . .	42
<b>4 探索性数据分析</b>	<b>43</b>
4.1 ACES 模型 . . . . .	43
4.2 探索绘图原则 . . . . .	43
4.3 探索性绘图 . . . . .	43
4.4 分层聚类 . . . . .	43

4.5 k-means 聚类 . . . . .	44
4.6 维度还原 . . . . .	44
4.7 可视化图形 . . . . .	44
<b>5 统计推断 . . . . .</b>	<b>45</b>
5.1 统计推断导论 . . . . .	45
5.2 概率 . . . . .	45
5.3 期望 . . . . .	46
5.4 方差 . . . . .	46
5.5 独立性 . . . . .	47
5.6 条件概率 . . . . .	48
5.7 贝叶斯定理 . . . . .	48
5.8 常见分布 . . . . .	49
5.9 渐进 . . . . .	50
5.10 置信区间 . . . . .	50
5.11 似然函数 . . . . .	51
5.12 贝叶斯推断 . . . . .	51
5.13 两独立样本 t 检验 . . . . .	52
5.14 假设检验 . . . . .	53
5.15 P 值 . . . . .	54
5.16 功效 . . . . .	54
5.17 多重比较 . . . . .	55
5.18 重采样推断 . . . . .	55
5.19 概念可视化 . . . . .	56
<b>6 回归模型 . . . . .</b>	<b>57</b>
6.1 回归模型导论 . . . . .	57
6.2 术语 . . . . .	57
6.3 回归线的最小二乘回归 . . . . .	58
6.4 统计线性回归模型 . . . . .	58
6.5 残差 . . . . .	58
6.6 回归推断 . . . . .	59
6.7 多元回归 . . . . .	60
6.8 模型诊断与选择 . . . . .	60
6.9 广义线性模型 . . . . .	61
6.10 二元响应 . . . . .	61
6.11 计数或速率响应 . . . . .	61
6.12 分段平滑 . . . . .	62
<b>7 最优化 . . . . .</b>	<b>63</b>
7.1 数学本质 . . . . .	63
7.2 简史 . . . . .	63
7.3 凸集 . . . . .	63
7.4 最小二乘法 . . . . .	63
7.5 线性规划 . . . . .	63
7.6 凸优化 . . . . .	64
<b>8 统计模型 . . . . .</b>	<b>65</b>
8.1 统计学习概论 . . . . .	65
8.2 统计学习简史 . . . . .	65
8.3 统计学习定义 . . . . .	65
8.4 预测 . . . . .	65
8.5 推断 . . . . .	65
8.6 估计模型 . . . . .	66
8.7 评价模型 . . . . .	66

8.8 研究设计 . . . . .	66
8.9 错误率 . . . . .	67
8.10 ROC 曲线 . . . . .	67
8.11 重采样技术 . . . . .	67
8.12 <b>caret</b> 包 . . . . .	68
8.13 数据分割 . . . . .	68
8.14 训练选项 . . . . .	68
8.15 预测变量作图 . . . . .	68
8.16 数据预处理 . . . . .	68
8.17 协变量生成 . . . . .	69
8.18 线性回归 & 多元线性回归 . . . . .	69
8.19 非线性 . . . . .	72
8.20 树 . . . . .	74
8.21 支持向量机 . . . . .	76
8.22 无监督学习 . . . . .	77
8.23 人工神经网络 . . . . .	78
8.24 模型联合 . . . . .	78
8.25 无监督预测 . . . . .	78
8.26 模型预测 . . . . .	79
8.27 模型可视化 . . . . .	79
<b>9 开发数据产品</b> . . . . .	<b>81</b>
9.1 shiny . . . . .	81
9.2 rCharts . . . . .	81
9.3 GoogleVis . . . . .	81
9.4 Slidify . . . . .	82
9.5 yhat . . . . .	82
9.6 swagger . . . . .	82
9.7 案例 . . . . .	82
<b>10 贝叶斯统计</b> . . . . .	<b>85</b>
10.1 贝塔分布 . . . . .	85
10.2 为什么击球的概率分布符合贝塔分布? . . . . .	86
10.3 先验与后验 . . . . .	86
10.4 经验贝叶斯 . . . . .	88
10.5 从整体到个人 . . . . .	91
10.6 可信区间与置信区间 . . . . .	92
10.7 后验错误率 . . . . .	95
10.8 错误发现率 . . . . .	98
10.9 q 值 . . . . .	99
10.10 贝叶斯视角的假设检验 . . . . .	100
10.11 比例检验 . . . . .	102
10.12 错误率控制 . . . . .	104
10.13 影响因子 . . . . .	105
10.14 混合概率模型 . . . . .	116
10.15 模拟验证结果 . . . . .	125
10.16 网络资源 . . . . .	141
<b>11 生存分析</b> . . . . .	<b>143</b>
11.1 Concepts . . . . .	143
11.2 Notation . . . . .	143
11.3 Cox proportional-hazards regression model . . . . .	143
11.4 Case: Recidivism . . . . .	143
11.5 further . . . . .	145

11.6 Time-Dependent Covariates . . . . .	146
11.7 Model Diagnostics . . . . .	147
11.8 Reference . . . . .	148
<b>12 生物信息 . . . . .</b>	<b>151</b>
12.1 数据结构 . . . . .	151
12.2 Pubmed 搜索 . . . . .	151
12.3 动态规划 . . . . .	152
12.4 得分矩阵 . . . . .	152
12.5 E 值 . . . . .	152
12.6 PSI-BLAST . . . . .	152
12.7 蛋白 . . . . .	152
12.8 蛋白结构预测 . . . . .	153
12.9 细菌基因组 . . . . .	153
12.10 病毒 . . . . .	153
12.11 单核苷酸多态性 (SNP) . . . . .	153
12.12 真核基因预测 . . . . .	153
12.13 DNA 指纹 . . . . .	154
12.14 Ensembl . . . . .	154
12.15 基因组学数据分析 . . . . .	154
12.16 链接 . . . . .	159
<b>13 流行病学 . . . . .</b>	<b>161</b>
13.1 声明 . . . . .	161
13.2 早期疾病的概念 . . . . .	161
13.3 近代流行病学关键人物 . . . . .	161
13.4 现代慢性病流行病学 . . . . .	162
13.5 流行病学基本概念 . . . . .	162
13.6 描述性流行病学 . . . . .	163
13.7 分析流行病学 . . . . .	165
13.8 疾病监控 . . . . .	166
13.9 疾病频率的测量 . . . . .	166
13.10 联系测量 . . . . .	168
13.11 随机误差 . . . . .	169
13.12 研究道德 . . . . .	169
13.13 临床实验 . . . . .	170
13.14 队列研究 . . . . .	170
13.15 病例对照研究 . . . . .	172
13.16 标准化 . . . . .	172
13.17 混杂 . . . . .	173
13.18 效应修饰 (EMM) . . . . .	173
13.19 多变量方法 . . . . .	173
13.20 篩选 . . . . .	173
13.21 因果推断 . . . . .	174
13.22 论文研读 . . . . .	175
<b>14 博弈论 . . . . .</b>	<b>179</b>
14.1 术语 . . . . .	179
14.2 支配策略 (dominate strategy) . . . . .	179
14.3 最佳回应 (Best response) . . . . .	179
14.4 纳什均衡 (Nash Equilibrium) . . . . .	179
14.5 帕累托最优 (Pareto Optimality) . . . . .	180
14.6 混合策略 (Mixed strategies) . . . . .	180
14.7 寻找纳什均衡 . . . . .	180

14.8 被支配策略 (dominated strategy) . . . . .	181
14.9 最大最小策略 (Maxmin strategies) . . . . .	181
14.10 扩展形式博弈 . . . . .	181
14.11 完美子博弈 . . . . .	181
14.12 信息不对称扩展形式博弈 . . . . .	181
14.13 混合与行为策略 . . . . .	182
14.14 重复博弈 . . . . .	182
14.15 随机博弈 (stochastic game) . . . . .	182
14.16 虚拟行动 (fictitious play) . . . . .	182
14.17 无悔学习 (No-regret learning) . . . . .	182
14.18 无限重复博弈的平衡 . . . . .	182
14.19 贝叶斯博弈 . . . . .	183
14.20 联盟博弈 . . . . .	183
14.21 夏普利值 (Shapley Value) . . . . .	183
14.22 核心 . . . . .	183
14.23 选举 . . . . .	184
14.24 机制设计 . . . . .	184
14.25 VCG 机制 . . . . .	184
14.26 拍卖 . . . . .	185
<b>15 量化投资</b> . . . . .	<b>187</b>
15.1 股票收益模型 . . . . .	187
15.2 风险 . . . . .	187
15.3 收益 . . . . .	187
15.4 一般性投资 . . . . .	188
15.5 风险管理原理 . . . . .	188
15.6 金融技术与发明 . . . . .	188
15.7 投资组合 . . . . .	188
15.8 保险 . . . . .	189
15.9 有效市场假说 . . . . .	189
15.10 行为经济学 . . . . .	189
15.11 金融监管 . . . . .	189
15.12 利率 . . . . .	190
15.13 银行 . . . . .	190
15.14 信用评级 . . . . .	190
15.15 股票 . . . . .	190
15.16 房产 . . . . .	191
<b>16 自然语言处理</b> . . . . .	<b>193</b>
<b>17 因果分析</b> . . . . .	<b>195</b>
17.1 Introduction . . . . .	195
17.2 Causal Information . . . . .	195
17.3 Theoretical Background . . . . .	196
17.4 Methods for Identification and Estimation . . . . .	196
17.5 链接 . . . . .	197



# 序

学习都是从模仿开始的，而模仿则意味着不加区别的接受，若是混口饭吃，自然也就够了，但模仿多了便能看到这知识表象下的东西。如果还能提炼一下，变成了自己的经验，各种经验互相联系影响，便有了理论。所以学习大都从知识点开始，而以形成一家之言为终，倘若这一家之言可以得到别人的认可，知识就开始传承了。且不论有多少东西会被反复发现发明出来，但是世界在某种程度上是可知的便是智慧生物生存的福利。

就数据分析而言，我自学过很多教材，可以说不同门派间手段差异非常大，但总又妄想一统江湖，所以便有了这份笔记来整合。不论最终的完成度如何，这都只会是一本残卷，因为这世界总有未知，也因此总有希望。

本书以 `bookdown` 写作，感谢相关工具开发者的努力，站在前人肩上看世界确实视角要开阔。



# 章 1

## 导论

### 1.1 数据科学

- 核心：数据处理
- 研究对象：实际问题（跨学科）
- 方法：统计学计算机科学专业领域
- 数据科学家：
  - 统计学水平高的程序员
  - 编程水平高的统计学家
  - 学术好奇心
  - 沟通交流能力
  - 产品经理
- 数据次于问题
- 大数据依赖科学而不是数据
- 实验设计重视可重复性随机与分组预测与推断不同不要选数据

### 1.2 基本问题

- 描述分析：对数据进行描述但不解释
- 探索分析：寻找未知的变量间关系（相关不代表因果）
- 推断分析：用小样本推断总体统计模型的目标强依赖采样过程
- 预测分析：用一组变量预测另一变量不一定有因果关系
- 因果分析：改变一个变量引发另一个变量变化的分析随机实验平均效果
- 机理分析：对个体改变一个变量所导致另一个变量的精确变化公式模拟与参数拟合

### 1.3 工作流程

- 数据收集
- 数据整理
- 数据探索
- 数据建模
- 模型评价
- 结果交流

## 1.4 概率与分布

概率与分布是统计的基本世界观，当我们用概率来理解世界时，所有事物便不仅仅是此时此刻的事，而是可能性中的一种。这种全局观好比从上帝视角开启有限平行宇宙，即使你知道每种状态及其概率，最后结果也无法预判。

- 从可能性到独立事件概率计算
- 从联合概率到条件概率到贝叶斯公式
- 事件的发生空间到分布
- 多事件发生概率比较到标准化分布-z 值
- 正态分布评价拟合
- 贝努利分布
- 二项分布，固定总数，成功概率，二项分布可用正态分布近似求值，也可用二项分布取精确值，求区间概率要扩大
- 负二项分布，固定成功次数概率
- 几何分布，最后一次成功概率
- 超几何分布，不放回抽样，成功概率
- 泊松分布，实验次数多，概率小，发生概率，泊松过程

## 1.5 统计量

统计量是对样本性质的一种描述或简化，用来提取设计者所关注的信号并尽可能排除掉噪音。

- 总体到样本
  - 多个事件的描述到众数中位数再到期望
  - 描述多个事件的变动到方差
  - 取样方法：随机，分层，分类
  - 样本独立性：简单随机取样，样本数少于 10% 的总体可认为独立样本
  - 估计的偏差为标准误
- 点估计到区间估计
  - 标准误只针对样本均值，理解为样本均值的估计标准差
  - 置信区间为对所有样本进行区间估计，95% 的区间包含真值，是对总体参数的估计，近似认为样本符合某分布
- 中心极限法则：样本均值的分布为正态分布

## 1.6 统计推断

统计推断基于构建的统计量来进行决策，这个决策过程涉及空假设、备择假设与 p 值。

- 假设检验
  - 不拒绝  $H_0$  不代表  $H_0$  是对的，拒绝  $H_0$  代表  $H_A$  可能正确，观察数值的区间重叠状况
  - 使用双重否定进行描述
  - type I 假阳性 type II 假阴性
  - 置信水平反映两种错误的可能性
  - p 值描述某数值在  $H_0$ （一般为等式）中出现的可能性，通常与置信水平对比，两边与单边
  - 构建符合某分布的统计量进行参数估计，通过标准误计算 p 值，进行假设检验过程
  - 功效表示  $H_A$  拒绝  $H_0$  的可能性，功效高，检验可靠
  - 统计差异显著不代表实际差异显著，甚至没有实际意义
- 均值比较（连续）

- 配对数据
- 均值比较
- t 分布与自由度及小样本均值的标准误估计
- 置信区间与 p 值
- 样本均值的 t 检验
- 多组数据均值的方差分析与 F 检验
- 多重比较的假阳性问题
- 样本数足够可用统计模拟的方法进行检验，数据存在层级结构则不可直接模拟
- 比例比较（计数）
  - 比例检验，计算基于  $H_0$  的标准误，计算 z 值，计算 p 值，可反推样品量
  - 比例差异检验， $H_0$  为比例相等，估计混合概率，计算标准误进行检验
  - 记分检验与 Wald 检验
- 优度拟合
  - 分布检验到卡方检验
- 独立性检验
- 精确检验

## 1.7 统计模型

统计模型是基于统计量的对事物的抽象，借助模型可以简化事物的复杂性或从某个角度更好理解事物。

- 变量关系到线性回归到线性诊断
- 参数估计到关系解释及误差分析
- 多元回归
- 模型选择
- 方差分析
- 非线性模型与平滑
- logistic 模型到广义线性模型
- 线性混合模型
- 主成分分析与因子分析

## 1.8 其他主题

- 非参数统计
- 贝叶斯统计
- 判别分析
- 岭回归与 lasso
- 广义加性模型
- 鲁棒模型
- 决策树到随机森林
- 人工神经网络
- 支持向量机
- 蒙特卡洛分析到统计模拟
- 图论
- 因果分析

## 1.9 应用

- 工具
- 实验设计
- 模式识别
- 流行病学
- 生物信息学
- 化学信息学
- 心理学
- 空间数据分析
- 时间序列分析与信号处理
- 量化投资

## 1.10 链接

- 统计问题
- R 问题
- R mailing list
- 数据分享
- 命令行数据科学
- 最流行的程序包
- 数据科学资料合集
- peerj 实用数据分析技巧特刊

## 章 2

# 数据分析工具

### 2.1 基础知识

- 层次：操作系统 - shell - 终端 - 命令行工具
- 分类：可执行文件、shell 内置命令、脚本、shell 函数、宏

### 2.2 命令行基础

- name of root is represented by a slash: /
- home directory is represented by a tilde: ~
- pwd print working directory
- recipe: command -flags arguments
- clear: clear out the commands in your current CLI window
- ls lists files and folders in the current directory
  - a lists hidden and unhidden files and folders
    - al lists details for hidden and unhidden files and folders
- cd stands for “change directory”
  - cd takes as an argument the directory you want to visit
  - cd with no argument takes you to your home directory
  - cd .. allows you to change directory to one level above your current directory
- mkdir stands for “make directory”
- touch creates an empty file
- cp stands for “copy”
  - cp takes as its first argument a file, and as its second argument the path to where you want the file to be copied
  - cp can also be used for copying the contents of directories, but you must use the -r flag
- rm stands for “remove”
  - use rm to delete entire directories and their contents by using the -r flag

- mv stands for “move”  
move files between directories  
use mv to rename files
- echo will print whatever arguments you provide
- date will print today's date

## 2.3 版本控制

```
$ git config --global user.name "Your Name Here" # 输入用户名
$ git config --global user.email "your_email@example.com" # 输入邮箱
$ git config --list # 检查
$ git init # 初始化目录
$ git add . # 添加新文件
$ git add -u # 更新改名或删除的文件
$ git add -A|git add --all # 添加所有改动
$ git commit -m "your message goes here" # 描述并缓存本地工作区改动到上一次commit
$ git log # 查看commit记录 用Q退出
$ git status # 查看状态
$ git remote add # 添加服务器端地址
$ git remote -v # 查看远端状态
$ git push # 将本地commit推送到github服务器端
$ git pull|fetch|merge|clone # 本地获取远端repo
$ exit # 退出
```

- Git = Local (on your computer); GitHub = Remote (on the web)

## 2.4 数据获取

- 复制: cp 或 scp (安全复制) > scp -i mykey.pem ~/Desktop/logs.csv ubuntu@ec2-184-73-72-150.compute-1.amazonaws.com:logs
- 解压: unpack > unpack logs.tar.gz
- 转化 excel 为 csv: in2csv、csvcut、csvlook > in2csv data/imdb-250.xlsx | head | csvcut -c Title,Year,Rating | csvlook
- 查询关系数据库:sql2csv > sql2csv -db 'sqlite:///data/iris.db' -query 'SELECT \* FROM iris' 'WHERE sepal\_length > 7.5'
- 互联网下载: curl -u 登录 -L 链接跳转 -I http 头文件 > curl -s http://www.gutenberg.org/cache/epub/76/pg76.txt | head -n 10 > curl -u username:password ftp://host/file > curl -L j.mp/locatbbar
- API: curlicue 来进行认证

## 2.5 高级命令

- !! 可重复上次命令
- chmod 增加权限
- #!/usr/bin/env bash 增加状况说明

- NUM\_WORDS="“\$1” 增加参数

## 2.6 R

### 2.6.1 语言导论

- R 语言是 S 语言的一种方言
- 1976 年 S 是 John Chambers 等在贝尔实验室作为 Fortran 的扩展库开发出来的
- 1988 年用 C 语言重写 S3 方法白皮书
- 1993 年 StatSci 从贝尔实验室获得 S 语言的独家开发售卖许可
- 1998 年 S4 方法绿皮书之后 S 语言稳定获得 Association for Computing Machinery's Software System Award
- 2004 年 Insightful (原 StatSci) 从 Lucent 收购了 S 语言
- 2006 年 Alcatel 收购了 Lucent 成立 Alcatel-Lucent
- 2008 年 TIBCO 收购 Insightful 之前 Insightful 开发并售卖 S-PLUS
- 1991 年 Ross Ihaka 与 Robert Gentleman 在 Zealand 开发了 R
- 1993 年发布 R 第一份许可
- 1995 年 R 作为自由软件发放 GUN 许可
- 1996 年 R 邮件列表创立
- 1997 年 R Core 成立控制 R 源码
- 2000 年 R version 1.0.0 放出
- 2013 年 R version 3.0.2 放出
- R 由 CRAN 掌控的 base 包与其他包组成
- 其余参考 R 主页
- 出色的 R 包
- 过时的 R 包

### 2.6.2 获得帮助

```
help()
?command
# 提问给出以下信息
version
str(.Platform)
```

### 2.6.3 数据类型及基本运算

- 所有数据都是对象所有对象都有类型
- 基本类型包括：字符 “” 数字整数 L 复数 (Re 实部 Im 虚部) 逻辑
- 向量储存同一类型数据
- list 存储不同类型数据 [[\*]] 引用相应向量 unlist 可用做紧凑输出
- 对象可以有属性 attributes
- 对象赋值符号为 <- 赋值同时展示加括号或直接输入对象名可累加赋值 a <- b <- c
- # 表示注释不执行
- : 用来产生整数序列也可以用 seq 生成
- 向量用 c 产生
- 空向量用 vector() 函数建立
- 向量中类型不同的对象元素会被强制转换为同一类型字符优先级最高其次数字其次逻辑 (0 or 1) 也可以用来串联字符
- 可使用 as.\* 来强制转化数据类型
- 对象可以用 names 命名

- 变量名开头不能是数字和. 大小写敏感下划线不要出现在名字里分割用. 变量名中不能有空格
- 保留字符

```
FALSE Inf NA NaN NULL TRUE break else for function if in next repeat while
```

- 清空 `rm(list = ls())`
- 矩阵
  - 带有 `dimension` 属性的向量为矩阵矩阵的生成次序为 upper-left
  - `matrix(1:6,nrow=2,ncol=3)` 表示建一个 2 行 3 列矩阵从 1 到 6 先列后行赋值可用 `byrow = T` 来更改
  - 可用 `c` 给 `dim` 赋值行和列数这样可把一个向量转为一个矩阵 `m<-1:6;dim(m)<-c(2,3)`
  - 矩阵可以用 `rbind` 或 `cbind` 生成
  - `t` 对矩阵转置
- 因子变量表示分类数据用标签名区分用 `level` 来命名排序默认是字母排序有些函数对顺序敏感可用 `levels = c()` 来命名 (例如低中高的排序) 数字表示 `drop = T` 表示显示截取数据的水平 `nlevels` 给出个数
- `NaN` 表未定义或缺失值 `NA` 表示无意义转换或缺失值 `NaN` 可以是 `NA` 反之不可以 `NA` 有数据类型 `is.NaN` 与 `is.NA` 可用来检验
- 数据框
  - 特殊 `list` 每个元素长度相等
  - 每一列类型相同矩阵所有数据类型相同
  - 特殊属性 `row.names`
  - 转为矩阵 `data.matrix`
  - 变量名自动转化可以不同
  - 因子变量保持为字符可以用 `I` `data.frame(x,y,I(c))`
- 数组
  - 表示更高维度的数据
  - `dim() = c(x,y,z)` 三维数组表示一组数
  - `dimnames` 给数组命名
  - 数组调用如果只有一行需要 `drop = F` 否则不会按照数组分类
- `ts` 产生时间序列对象
- `.Last.value` 引用前一个数值
- 取整数用 `round(x,n)` `n` 表示保留几位小数
- 截取整数 `trunc`
- 开平方 `sqrt`
- 绝对值 `abs`
- 指数函数 `exp`
- 自然对数函数 `log`
- 以 10 为底的对数函数 `log10`
- 三角函数 `sin cos tan asin acos atan`
- 常用的逻辑运算符有: 大于 `>` 小于 `<` 等于 `==` 小于或等于 `<=` 大于或等于 `>=` 与 `&` 非 `!` 或 `|`
- 判断向量 `x` 中是否与 `y` 中元素相等 `x %in% y` 结果返回逻辑值
- `sum` 求和 `prod` 求连乘
- `range` 给极值范围
- `duplicated` 给出有重复的值
- `unique` 给出无重复的值
- 向量操作 `union` 并集 `intersect` 交集 `setdiff` 除了交集的部分
- `rep` 用向量循环生成向量

```
x <- 1:4 # puts c(1,2,3,4) into x
i <- rep(2, 4) # puts c(2,2,2,2) into i
y <- rep(x, 2) # puts c(1,2,3,4,1,2,3,4) into y
z <- rep(x, i) # puts c(1,1,2,2,3,3,4,4) into z
w <- rep(x, x) # puts c(1,2,2,3,3,3,4,4,4,4) into w
```

- 整型变量后面加上 L x<-10L
- Inf 代表 1/0 同样 1/Inf 运算结果为 0

#### 2.6.4 环境／文件操作

- `getwd()` `setwd()` 设置工作目录
- `ls()` 列举环境中 bianliang
- `list.files()` 或 `dir()` 列举当前目录下文件
- `args()` 列举函数默认变量
- `dir.create()` 创建文件目录加上 `recursive=T` 可创建多级目录
- `file.create()` 创建文件
- `file.exists()` 检查文件是否存在
- `file.info()` 检查文件信息
- `file.rename()` 文件重命名
- `file.copy()` 文件复制
- `file.path()` 文件路径多个文件组成多级路径
- `unlink()` 删除文件

#### 2.6.5 下载

- 设定工作目录与数据存储目录

```
if (!file.exists("data")) {
  dir.create("data")
}
```

- url 下载与时间记录

```
fileUrl <- "yoururl"
download.file(fileUrl, destfile = "./data/XXX.csv", method = "curl")
list.files("./data")
dateDownloaded <- date()
```

#### 2.6.6 截取数据

- 可以用 `[x,y]` 提取特定数值
- `[-1,-2]` 可剔除第一行第二列
- `[]` 用来从 list 或者 frame 里提取元素类型固定可提取序列 `x[[1]][[3]]` 可部分匹配 `exact=FALSE`
- `$` 用名字提取元素可部分匹配
- 提取矩阵时默认只能提取向量但可以提取 `1*1` 矩阵 `x[1,2,drop=FALSE]`
- 先用 `is.NA()` 提取用! 排除缺失值可用 `is.element(x,y)` 来处理很多表示 NA 值的数字返回 `x %in% y` 的逻辑值
- 用 `complete.cases()` 提取有效数据用 `[]` 提取可用数据
- `head(x,n)` n 表示从头截取多少行
- `tail(x,n)` n 表示从尾截取多少行
- `subset(x,f)` x 表示数据 f 表示表达式
- 条件筛选中获得一个变量多个数值的数据使用 `[is.element(x,c(' ',NA,''))]` 或者 `[x%in%c(' ',' ',' ',' ')]` 使用 `x == c(' ',NA,'')` 会报错循环查找三个变量
- `x != 't'` 可能会把空白值输入应该使用 `is.element(x,'t')`
- `ifelse(con, yes, no)` 利用条件筛选返回 yes 或者 no 的值
- 支持正则表达式
- 增加行直接 `$`
- `seq` 产生序列
- 通过 [按行列或条件截取

- `which` 返回行号
- 排序向量用 `sort`
- 排序数据框 (多向量) 用 `order`
- `plyr` 包排序

```
library(plyr)
arrange(X, var1)
arrange(X, desc(var1))
```

### 2.6.7 读取数据

- `read.table` `read.csv` 读取表格反之 `write.table`
- `readLines` 读取文本行反之 `writeLines`
- `source` 读取 R 代码反之 `dump`
- `dget` 读取多个 R 代码反之 `dput`
- `load` 读取保存的工作区反之 `save`
- `unserialize` 读取二进制 R 对象反之  `serialize`
- `?read.table`
- 大数据读取提速
  - 计算内存
  - `comment.char = ""` 不扫描注释
  - 设定 `nrows`
  - 设定 `colClasses`

```
initial <- read.table("datatable.txt", nrows = 100)
classes <- sapply(initial, class)
tabAll <- read.table("datatable.txt",
                      colClasses = classes)
```

- 使用 `connections` 与 `file` 等保存外部文件指向

#### 2.6.7.1 读取本地文件

- `read.table`
- `read.csv` 默认 `sep=","`, `header=TRUE`
- `quote` 设定引用
- `na.strings` 设定缺失值字符
- `nrows` 设定读取字段
- `skip` 跳过开始行数

#### 2.6.7.2 读取 excle 文件

- `xlsx` 包

```
library(xlsx)
cameraData <- read.xlsx("./data/cameras.xlsx", sheetIndex=1, header=TRUE)
head(cameraData)
# read.xlsx2 更快不过运行读取时会不稳定
# 支持底层读取 如字体等
```

- `XLConnect` 包

```
library(XLConnect)
wb <- loadWorkbook("XLConnectExample1.xlsx", create = TRUE)
createSheet(wb, name = "chickSheet")
```

```
writeWorksheet(wb, ChickWeight, sheet = "chickSheet", startRow = 3, startCol = 4)
saveWorkbook(wb)
# 支持区域操作 生成报告 图片等
```

### 2.6.7.3 读取 XML 文件

- 网页常用格式
- 形式与内容分开
- 形式包括标签元素属性等
- XML 包

```
library(XML)
fileUrl <- "http://www.w3schools.com/xml/simple.xml"
# 读取 xml 结构
doc <- xmlTreeParse(fileUrl,useInternal=TRUE)
# 提取节点
rootNode <- xmlRoot(doc)
# 提取根节点名
xmlName(rootNode)
# 提取子节点名
names(rootNode)
# 提取节点数值
xmlSApply(rootNode,xmlValue)
```

- XPath XML 的一种查询语法
  - /node 顶级节点
  - //node 所有子节点
  - node(?) 带属性名的节点
  - node(?, ='bob') 属性名为 bob 的节点

```
# 提取节点下属性名为 name 的数值
xpathSApply(rootNode,"//name",xmlValue)
```

### 2.6.7.4 读取 json 文件

- js 对象符号结构化常作为 API 输出格式
- jsonlite 包

```
library(jsonlite)
# 读取 json 文件
jsonData <- fromJSON("https://api.github.com/users/jtleek/repos")
# 列出文件名
names(jsonData)
# 可嵌套截取
jsonData$owner$login
# 可将 R 对象写成 json 文件
myjson <- toJSON(iris, pretty=TRUE)
```

### 2.6.7.5 读取 MySQL 数据库

- 网络应用常见数据库软件
- 一行一记录

- 数据库表间有 index 向量
- 常见命令
- 指南
- RMySQL 包

```
library(RMySQL)
# 读取数据库
ucscDb <- dbConnect(MySQL(), user="genome",
                      host="genome-mysql.cse.ucsc.edu")
result <- dbGetQuery(ucscDb, "show databases;");
# 断开连接
dbDisconnect(ucscDb);
# 读取指定数据库
hg19 <- dbConnect(MySQL(), user="genome", db="hg19",
                  host="genome-mysql.cse.ucsc.edu")
allTables <- dbListTables(hg19)
length(allTables)
# mysql 语句查询
dbGetQuery(hg19, "select count(*) from affyU133Plus2")
# 选择子集
query <- dbSendQuery(hg19, "select * from affyU133Plus2 where misMatches between 1 and 3")
affyMis <- fetch(query); quantile(affyMis$misMatches)
```

#### 2.6.7.6 读取 HDF5 数据

- 分层分组读取大量数据的格式
- rhdf5 包

```
library(rhdf5)
created = h5createFile("example.h5")
created = h5createGroup("example.h5","foo")
created = h5createGroup("example.h5","baa")
created = h5createGroup("example.h5","foo/foobaa")
h5ls("example.h5")
A = matrix(1:10,nr=5,nc=2)
h5write(A, "example.h5","foo/A")
B = array(seq(0.1,2.0,by=0.1),dim=c(5,2,2))
attr(B, "scale") <- "liter"
h5write(B, "example.h5","foo/foobaa/B")
h5ls("example.h5")
df = data.frame(1L:5L,seq(0,1,length.out=5),
                c("ab","cde","fghi","a","s"), stringsAsFactors=FALSE)
h5write(df, "example.h5","df")
h5ls("example.h5")
readA = h5read("example.h5","foo/A")
readB = h5read("example.h5","foo/foobaa/B")
readdf= h5read("example.h5","df")
```

#### 2.6.7.7 读取网页数据

- 网页抓取 HTML 数据
- 读完了一定关链接
- httr 包

```

con = url("http://scholar.google.com/citations?user=HI-I6COAAAAJ&hl=en")
htmlCode = readLines(con)
close(con)
htmlCode
library(XML)
url <- "http://scholar.google.com/citations?user=HI-I6COAAAAJ&hl=en"
html <- htmlTreeParse(url, useInternalNodes=T)
xpathSApply(html, "//title", xmlValue)
library(httr)
html2 = GET(url)
content2 = content(html2, as="text")
parsedHtml = htmlParse(content2, asText=TRUE)
xpathSApply(parsedHtml, "//title", xmlValue)
GET("http://httpbin.org/basic-auth/user/passwd")
GET("http://httpbin.org/basic-auth/user/passwd",
    authenticate("user", "passwd"))
google = handle("http://google.com")
pg1 = GET(handle=google, path="/")
pg2 = GET(handle=google, path="search")

```

### 2.6.7.8 读取 API

- 通过接口授权后调用数据
- httr 包

```

myapp = oauth_app("twitter",
                  key="yourConsumerKeyHere", secret="yourConsumerSecretHere")
sig = sign_oauth1.0(myapp,
                     token = "yourTokenHere",
                     token_secret = "yourTokenSecretHere")
homeTL = GET("https://api.twitter.com/1.1/statuses/home_timeline.json", sig)
json1 = content(homeTL)
json2 = jsonlite::fromJSON(toJSON(json1))

```

### 2.6.7.9 读取其他资源

- 图片
  - jpeg
  - readbitmap
  - png
  - EBImage (Bioconductor)
- GIS
  - rgdal
  - rgeos
  - raster
- 声音
  - tuneR
  - seewave

### 2.6.8 数据总结

- head tail 查看数据

- `summary str` 总结数据
- `quantile` 按分位数总结向量
- `table` 按向量元素频数总结
- `sum(is.na(data)) any(is.na(data)) all(data$x > 0)` 异常值总结
- `colSums(is.na(data))` 行列求和
- `table(data$x %in% c("21212"))` 特定数值计数总结
- `xtabs ftable` 创建列联表
- `print(object.size(fakeData), units="Mb")` 现实数据大小
- `cut` 通过设置 `breaks` 产生分类变量
- Hmisc 包

```
library(Hmisc)
data$zipGroups = cut2(data$zipCode,g=4)
table(data$zipGroups)
library(plyr)
# mutate 进行数据替换或生成
data2 = mutate(data,zipGroups=cut2(zipCode,g=4))
table(data2$zipGroups)
```

## 2.6.9 数据整理

Raw data -> Processing script -> tidy data

- 前期需求
  - 原始数据
  - 干净数据
  - code book
  - 详尽的处理步骤记录
- 原始数据要求
  - 未经处理
  - 未经修改
  - 未经去除异常值
  - 未经总结
- 干净数据
  - 每个变量一列
  - 同一变量不同样本不在一行
  - 一种变量一个表
  - 多张表要有一列可以相互链接
  - 有表头
  - 变量名要有意义
  - 一个文件一张表
- code book
  - 变量信息
  - 总结方式
  - 实验设计
  - 文本文件
  - 包含研究设计与变量信息的章节
- 处理步骤记录
  - 脚本文件
  - 输入为原始数据
  - 输出为处理过数据
  - 脚本中无特定参数
- 每一列一个变量
- 每一行一个样本

- 每个文件存储一类样本
- `melt` 进行数据融合
- `reshape2` 包
- `dcast` 分组汇总数据框
- `acast` 分组汇总向量数组
- `arrange` 指定变量名排序
- `merge` 按照指定向量合并数据
- `plyr` 包的 `join` 函数也可实现合并

### 2.6.10 数据操作 `data.table` 包

- 基本兼容 `data.frame`
- 速度更快
- 通过 `key` 可指定因子变量并快速提取分组的行
- 可在第二个参数是 R 表达式

```
DT[,list(mean(x),sum(z))]
DT[,table(y)]
```

- 可用：生成新变量进行简单计算

```
DT[,w:=z^2]
DT[,m:= {tmp <- (x+z); log2(tmp+5)}]
```

- 进行数据条件截取

```
DT[,a:=x>0]
DT[,b:= mean(x+w),by=a]
```

- 进行计数

```
DT <- data.table(x=sample(letters[1:3], 1E5, TRUE))
DT[, .N, by=x]
```

### 2.6.11 文本处理

- 处理大小写 `tolower toupper`
- 处理变量名 `strsplit`

```
firstElement <- function(x){x[1]}
sapply(splitNames,firstElement)
```

- 字符替换 `sub gsub`
- 寻找变量 `grep`(返回行号) `grepl`(返回逻辑值)
- `stringr` 包 `stringr`
- `paste0` 不带空格
- `str_trim` 去除空格
- 命名原则
  - 变量名小写
  - 描述性
  - 无重复
  - 变量名不要符号分割
  - Names of variables should be
- 正则表达式
  - 文字处理格式
  - ^ 匹配开头

- \$ 匹配结尾
- [] 匹配大小写 ^ 在开头表示非
- . 匹配任意字符
- | 匹配或
- () 匹配与
- ? 匹配可选择
- \* 匹配任意
- + 匹配至少一个
- {} 匹配其中最小最大一个值表示精确匹配 m, 表示至少 m 次匹配
- \1 匹配前面指代

### 2.6.12 控制结构

- if else 条件

```
if(<condition>) {
    ## do something
} else {
    ## do something else
}
if(<condition1>) {
    ## do something
} else if(<condition2>) {
    ## do something different
} else {
    ## do something different
}
```

- for 执行固定次数的循环嵌套不超过 2 层

```
for(i in 1:10) {
    print(i)
}
```

- while 条件为真执行循环条件从左到右执行

```
count <- 0
while(count < 10) {
    print(count)
    count <- count + 1
}
```

- repeat 执行无限循环配合 break 中断并跳出循环
- next 跳出当前循环继续执行

```
for(i in 1:100) {
    if(i <= 20) {
        ## Skip the first 20 iterations
        next
    }
    ## Do something here
}
```

- return 退出函数
- 避免使用无限循环可用 apply 替代

### 2.6.13 函数

```
f <- function(<arguments>) {
  ## Do something interesting
}
```

- 函数中参数默认值可用 `formals()` 显示
- 参数匹配
  - 先检查命名参数
  - 然后检查部分匹配
  - 最后检查位置匹配
- 定义函数时可以定义默认值或者设为 `NULL`
- 懒惰执行：只执行需要执行的语句
- ... 向其他函数传参之后参数不可部分匹配

### 2.6.14 编程标准

- 使用文本文档与文本编辑器
- 使用缩进
- 限制代码行宽 80 为宜
- 限制单个函数长度

### 2.6.15 范围规则

- 自由变量采用静态搜索
- 环境是由数值符号对组成每个环境都有母环境
- 函数与环境组成环境闭包
- 首先从函数环境中寻找变量
- 之后搜索母环境
- 最高层为工作区
- 之后按搜寻列表从扩展包中寻找变量
- 最后为空环境之后报错
- 可以函数内定义函数
- S 都存在工作区函数定义一致 R 存在内存可根据需要调用函数环境

### 2.6.16 向量化操作

- 向量操作针对元素
- 矩阵操作也针对元素 `%*%` 表示矩阵操作

### 2.6.17 绘图系统

#### 2.6.17.1 基础绘图

- 艺术家绘画模式
- `graphics` 包括基础包的绘图函数如 `plot`, `hist`, `boxplot`
- `grDevices` 包括执行调用绘图设备函数如 X11, PDF, PostScript, PNG
- 叠加函数高度自由度
- 初始化新图然后标注
- 以下命令熟记
  - `pch`: the plotting symbol (default is open circle)
  - `lty`: the line type (default is solid line), can be dashed, dotted, etc.

- lwd: the line width, specified as an integer multiple
- col: the plotting color, specified as a number, string, or hex code; the colors() function gives you a vector of colors by name
- xlab: character string for the x-axis label
- ylab: character string for the y-axis label
- par(): 查找做图的画布参数具体如下
- las: the orientation of the axis labels on the plot
- bg: the background color
- mar: the margin size
- oma: the outer margin size (default is 0 for all sides)
- mfrow: number of plots per row, column (plots are filled row-wise)
- mfcol: number of plots per row, column (plots are filled column-wise)
- plot: make a scatterplot, or other type of plot depending on the class of the object being plotted
- lines: add lines to a plot, given a vector x values and a corresponding vector of y values (or a 2-column matrix); this function just connects the dots
- points: add points to a plot
- text: add text labels to a plot using specified x, y coordinates
- title: add annotations to x, y axis labels, title, subtitle, outer margin
- mtext: add arbitrary text to the margins (inner or outer) of the plot
- axis: adding axis ticks/labels
- 图形设备
  - 图像一定要有设备
  - 屏幕设备 Mac quartz() windows windows() Unix/linux x11()
  - 先调用后用 dev.off() 关闭设备
  - 矢量图设备保真放大元素过多体积庞大 pdf() svg() winmetafile() postscript()
  - 位图设备放大失真基于像素 png() jpeg() tiff() bmp()
  - 当前设备 dev.cur()
  - 设置设备 dev.set(<integer>)
  - 设备转移 dev.copy dev.copy2pdf

### 2.6.17.2 lattice

- 一站式解决
- lattice 包括框架图函数如 xyplot, bwplot, levelplot
- grid 包括独立于基础绘图系统的网格绘图系统
- 一个函数解决问题默认自定义空间少
- 返回 trellis 类型对象可单独存储
- 界面调整使用 panel 选项
- 以下为常见函数
  - xyplot: this is the main function for creating scatterplots
  - bwplot: box-and-whiskers plots (“boxplots”)
  - histogram: histograms
  - stripplot: like a boxplot but with actual points
  - dotplot: plot dots on “violin strings”
  - splom: scatterplot matrix; like pairs in base plotting system
  - levelplot, contourplot: for plotting “image” data
- 基本格式
  - `xyplot(y ~ x | f * g, data)`
  - 可同时展示分组信息及交互作用

### 2.6.17.3 ggplot2

- 基于图形语法理念

- 图形属性映射数据问题
- 自动处理界面允许后期添加结合 base 与 lattice
- 默认友好
- 基础绘图 `qplot()`
- `ggplot()` 通过叠加元素出图
- 细节调整 `xlab()`, `ylab()`, `labs()`, `ggtitle()`
- 主题调整 `theme()`
- 做图需求
  - 数据框 `data.frame`
  - 属性映射 `asesthetic mapping`
  - 几何对象 `geoms`
  - 条件 `facets`
  - 统计转换 `stats`
  - 范围量表 `scales`
  - 坐标轴系统 `coordinate system`

#### 2.6.17.4 数学绘图

- Tex 语法
- 使用 `expression()`
- `?plotmath`

#### 2.6.17.5 色彩管理

- `colorRamp` 返回 01 间数值表示颜色过度
- `colorRampPalette` 返回 8 位颜色代码调色盘
- `colors` 返回可用颜色
- RColorBrewer 包含有预先配色信息序列无序两级
- `rgb` 产生三原色颜色 `alpha` 控制透明度
- 绘图时用 `col` 调用调色盘颜色

```

pal <- colorRamp(c("red", "blue"))
pal(0)

##      [,1] [,2] [,3]
## [1,] 255    0    0
pal(1)

##      [,1] [,2] [,3]
## [1,]    0    0 255
pal(0.5)

##      [,1] [,2] [,3]
## [1,] 128    0 128
#####
pal <- colorRampPalette(c("red", "yellow"))
pal(2)

## [1] "#FF0000" "#FFFF00"
pal(10)

## [1] "#FF0000" "#FF1C00" "#FF3800" "#FF5500" "#FF7100" "#FF8D00" "#FFAA00"
## [8] "#FFC600" "#FFE200" "#FFFF00"

```

```
#####
library(RColorBrewer)
cols <- brewer.pal(3, "BuGn")
```

### 2.6.18 日期与时间

- 日期以 `data` 类型存储
- 时间以 `POSIXct` 或 `POSIXlt` 类型存储
- 数字上是从 1970-01-01 以来的天数或秒数
- `POSIXct` 以整数存储时间
- `POSIXlt` 以年月日时分秒等信息存储时间
- `strptime as.Date as.POSIXlt as.POSIXct` 用来更改字符为时间
- `format` 处理日期格式
  - `%d` 日
  - `%a` 周缩写
  - `%A` 周
  - `%m` 月
  - `%b` 月缩写
  - `%B` 月全名
  - `%y` 2 位年
  - `%Y` 4 位年
- `weekdays` 显示星期
- `months` 显示月份
- `julian` 显示 70 年以来的日期
- `lubridate` 包
  - `ymd`
  - `mdy`
  - `dmy`
  - `ymd_hms`
  - `Sys.timezone`

### 2.6.19 循环

#### 2.6.19.1 `lapply`

- 对列表对象元素应用函数
- 可配合匿名函数使用

```
x <- list(a = 1:5, b = rnorm(10))
lapply(x, mean)
```

```
## $a
## [1] 3
##
## $b
## [1] 0.0694
x <- 1:4
lapply(x, runif, min = 0, max = 10)

## [[1]]
## [1] 0.848
##
```

```
## [[2]]
## [1] 8.84 1.76
##
## [[3]]
## [1] 6.596 0.243 5.383
##
## [[4]]
## [1] 6.76 7.54 2.68 7.29
x <- list(a = matrix(1:4, 2, 2), b = matrix(1:6, 3, 2))
lapply(x, function(elt) elt[,1])
```

```
## $a
## [1] 1 2
##
## $b
## [1] 1 2 3
```

### 2.6.19.2 sapply

- `lapply` 的精简版
- 如果结果是单元素列表转化为向量
- 如果结果是等长向量转化为矩阵
- 否则输出依旧为列表

```
x <- list(a = 1:4, b = rnorm(10), c = rnorm(20, 1), d = rnorm(100, 5))
sapply(x, mean)
```

```
##      a      b      c      d
## 2.500 -0.224  0.841  4.906
```

### 2.6.19.3 vapply

- 类似 `lapply` 可用更复杂函数返回矩阵

### 2.6.19.4 replicate

- 用于将函数循环使用如返回随机矩阵

### 2.6.19.5 rapply

- 用 `how` 来调整输出方法如选取某列表中类型数据进行迭代

### 2.6.19.6 apply

- 数组边际函数常用于矩阵的行列处理
- 行为 1, 列为 2
- 可用 `rowSums` `rowMeans` `colSums` `colMeans` 来替代大数据量更快

```
x <- matrix(rnorm(50), 10, 5)
apply(x, 1, quantile, probs = c(0.25, 0.75))
```

```

##      [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9]
## 25% -0.474 -0.211 -0.490 -0.6426 -0.671 -0.966 -2.176 -0.404 -0.1987
## 75%  0.723  0.238  0.639  0.0973  0.904  1.085 -0.173  0.376 -0.0603
##      [,10]
## 25% -1.5974
## 75%  0.0797

a <- array(rnorm(2 * 2 * 10), c(2, 2, 10))
apply(a, c(1, 2), mean)

##      [,1]   [,2]
## [1,] -0.36172 -0.0338
## [2,] -0.00725  0.2848

```

### 2.6.19.7 tapply

- 对数据子集（因子变量区分）向量应用函数

```

x <- c(rnorm(10), runif(10), rnorm(10, 1))
f <- gl(3, 10)
tapply(x, f, mean)

##      1       2       3
## 0.446 0.438 1.518

```

### 2.6.19.8 by

- 对数据按照因子变量应用函数类似 tapply
- 按照某个分类变量 a 分类求均值 by(x[,-a], a, mean)

### 2.6.19.9 split

- 将数据按因子分割为列表常配合 lapply 使用
- 类似 tapply
- 可用来生成分组用 drop 来删除空分组

```

x <- c(rnorm(10), runif(10), rnorm(10, 1))
f <- gl(3, 10)
lapply(split(x, f), mean)

## $`1`
## [1] 0.0961
##
## $`2`
## [1] 0.62
##
## $`3`
## [1] 0.601

x <- rnorm(10)
f1 <- gl(2, 5)
f2 <- gl(5, 2)
str(split(x, list(f1, f2), drop = TRUE))

```

```
## List of 6
## $ 1.1: num [1:2] 0.742 -0.368
## $ 1.2: num [1:2] 0.461 -0.191
## $ 1.3: num 1.11
## $ 2.3: num -1.05
## $ 2.4: num [1:2] 0.101 -1.076
## $ 2.5: num [1:2] -0.202 0.341
```

### 2.6.19.10 mapply

- 多变量版 apply 从多个参数范围取值并用函数得到结果

```
noise <- function(n, mean, sd) {
  rnorm(n, mean, sd)
}

mapply(noise, 1:5, 1:5, 2)

## [[1]]
## [1] 0.567
##
## [[2]]
## [1] -0.559 1.782
##
## [[3]]
## [1] 0.787 4.063 -1.543
##
## [[4]]
## [1] 8.737 2.906 0.173 3.438
##
## [[5]]
## [1] 4.78 1.59 6.22 7.34 5.02

# 等同于如下循环

#list(noise(1, 1, 2), noise(2, 2, 2),
#     noise(3, 3, 2), noise(4, 4, 2),
#     noise(5, 5, 2))
```

### 2.6.19.11 eapply

- 对环境变量应用函数用于包

## 2.6.20 模拟

- 在某分布下产生随机数
  - d 分布概率密度
  - r 分布随机数
  - p 分布累计概率
  - q 分布分位数

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

- `set.seed` 保证重现性
- `sample` 对数据采样

### 2.6.21 调试

- 三种提示 `message warning error` 只有 `error` 致命
- 关注重现性
- 调试工具 `traceback debug browser trace recover`
- 三思而行

### 2.6.22 分析代码

- 先设计后优化
- `system.time` 计算代码运行时间返回对象类型 `proc_time`
  - `user time` 执行代码用时
  - `system time` CPU 时间
  - `elapsed time` 实际用时
  - 在多核或并行条件下实际用时可以短于执行代码用时
  - 明确知道耗时较长的函数时使用
- `Rprof` R 代码要支持分析函数
  - `summaryRprof` 可使结果易读
  - 不要与 `system.time` 混用
  - 0.02s 记录一次执行函数
  - `by.total` 记录单个函数用时
  - `by.self` 记录函数执行时被调用函数用时

### 2.6.23 包开发

- `DESCRIPTION` 指明包内容
  - `Package` 包名字
  - `Title` 全名
  - `Description` 一句话描述
  - `Version` 版本号
  - `Author` 作者
  - `Maintainer` 维护者
  - `License` 许可协议
  - `Depends` 依赖
  - `Suggests` 建议
  - `Date` 发布日期 YYYY-MM-DD 格式
  - `URL` 项目主页
- R 源码
- `Documentation` 文档 `Rd` 文件
- `NAMESPACE` 关键词输入输出的函数及类型
- R CMD `build/check newpackage` 构建检查包
- `roxygen2` 源文件注释文档

### 2.6.24 方法与类型

- R 面向对象编程
- 对象用 `setClass` 指定类型用 `setMethod` 指定处理类型的方法
- 对象一般指新的数据类型
- S3 函数对象不算严格 `generic` 处理对象开放没有指定类型就用通用方法

- S4 函数对象定义严格只处理指定类型对象不可直接调用方法针对性强
- `stats4` 有很多针对性的极大似然估计的对象定义与方法

### 2.6.25 并行计算

- 任务切分后多线程/多核/多机同时执行，然后汇总，需要调用配置管理
- 并行计算的优势在于利用独立计算单元同时计算汇总
- 单机可以多核或多进程，例如 OpenMP
- 也可以 GPU 加速，例如 CUDA
- 集群可在应用层定义后交给后端做分发例如 `snow`
- 有些函数已经进行了并行化优化可直接调用，有些需要声明用法才能调用
- 多机器临时集群可以跨主机分布或进行云计算，需要指定名称，可通过传统 socket 或符合 MPI 标准的方式来组建
- `BiocParallel` 包封装了常见并行函数方便编程
  - `bplapply` 对每个 x 进行函数计算，同 `lapply`
  - `bpmaply` 对多个函数参数并行运行函数，同 `mapply`
  - `bpiterate` 对迭代出得数据反复运行函数
  - `bpvec` 向量化运算，这样切分更快
  - `bpaggregate` 聚合运算

### 2.6.26 分布式计算

- Sparkly

### 2.6.27 异常值监测

- twitter 的断点检测

### 2.6.28 图片处理

- imager

## 2.7 Python

- 基础数据类型 NULL
- 数值类型
  - int
  - float
  - bool(逻辑运算)
- 列表
  - 从 0 开始
  - 元素可变
  - () 赋值为 Tuples 类型元素不可变
- 字符串
  - 文本处理
  - python 专长
- 字典
  - {} 包含
  - : 指定属性值
- python 中对象均有类型可自定义

### 2.7.1 工具包

- Numpy 数值计算包
- Pandas 数据清洗缺失值切分
- Matplotlib 数据可视化
- sklearn 机器学习包

## 2.8 Tex

### 2.8.1 语言基础

- 作者 Donald Knuth
- **tex** 排版引擎圆周率
- **metafont** 处理字体自然对数的底数
- 控制序列钩子为\
- 宏包对控制序列打包钩子为\
  - Lamport
  - **latex** 宏包分部分处理文档打包了大量命令
  - **latex 2e** 后基本停止
  - Hans 对 **latex** 不满认为可定制性不够遂进行二次开发有了 **context**
- 引擎处理控制序列进行排版
  - **pdftex** 可解决文档直接输出为 PDF 的问题避免产生 dvi
  - 早期不支持 unicode 对多国语言只能通过调用宏包来实现字符与图形对应 **cjk ctt ctex** 等都是此类宏包需要安装字体
  - **xetex** 可原生支持 unicode 的引擎并调用系统字体支持 plain tex xelatex 可支持 latex 宏包
  - **luatex** 合并 **metapost** 可直接绘图可直接调用字体可脱离宏包调用程序现与 **context** 结合紧密
- **tex** 格式 Knuth 为原始 300 个控制序列写的宏包有 600 命令这 900 个合称 **plain tex**
- 将引擎宏包格式辅助程序等打包即为发行版
  - **miktex texlive mactex**
  - **context minimals** 只有自己的引擎与宏包
- 字体最早是栅格后来是矢量
  - type I 是最早的矢量
  - truetype 是 type I 的竞争对手
  - opentype 是基于 truetype 的进化版
  - 最早格式为 DVI 为字体准备了字形盒子可通过上面编码调用字库显示之后出现了 PS 与 PDF
  - 原来要编译多次现在只需要用 **xetex** 或 **luatex** 引擎就可以了他们内置了库来实现字形盒子与字体的联系这个库有 cache 功能
- 字体分类
  - 衬线体起笔落笔有差异横竖粗细各不同易于识别宋体
  - 非衬线体笔画粗细一致无装饰醒目黑体

- 等宽体每个字宽窄相同汉字编程

### 2.8.2 关于 xetex

- xeCJK 使用 `xelatex` 引擎的中文宏包纠正了 `xelatex` 一些缩进等的不美观
- ctex 包含早期 CTT CJK 及 xeCJK 可用`\setCJKmainfont{SimSun}` 来调用系统字体下面是底层调用中英文混排

#### 2.8.2.1 实例讲解

```
\documentclass[12pt,a4paper]{article}
\usepackage{xltxtra,fontspec,xunicode}
\usepackage[slantfont,boldfont]{xeCJK} % 允许斜体和粗体
\setCJKmainfont{FZJingLeiS-R-GB} % 设置缺省中文字体
\setCJKmonofont{SimSun} % 设置等宽字体
\setmainfont{TeX Gyre Pagella} % 英文衬线字体
\setmonofont{Monaco} % 英文等宽字体
\setsansfont{Trebuchet MS} % 英文无衬线字体
```

### 2.8.3 常见问题

- 空白 tab 与多个空白认为是一个空白空行表示段落结束
- 保留字符 # \$ % ^ & \_ { } ~ \ 可使用`\# \$ \% \^{} \& \_ \{ \} \sim \` 来表示 `\backslash` 表示断行  
`\backslash$` 生成反斜杠
- latex 命令 `\tex{}` 后面加空格防止命令延长 {} 中为命令参数
- % 表示注释掉一行也可使用`\usepackage{verbatim}` 中的 `comment` 环境
- 源文件结构
  - `\documentclass[]{...}` 声明文档类型 [] 中为选项包括字体纸张公式对齐等文档格式
  - `\usepackage[]{...}` 加入需要的宏包 [] 中为触发功能的关键词
  - 以上为导言区
  - `\begin{document}` 开始正文
  - `\end{document}` 结束文档
- 页面样式`\pagestyle{style}` 不同页眉页脚样式
- `\include{lename}` 用来包含文档多用于大型文档在新页包含连续可用`\input{lename}`
- `\includeonly{lename, lename, . . .}` 导言区包含文档在所有`\include` 文档中只有`\includeonly` 中的会被处理
- 语法检查`\usepackage{syntonly} \syntaxonly`
- `\hyphenation{word list}` 给出断字列表完整的不允许断有-的表示允许的唯一断字点在文档中-表示唯一允许断字的地方
- `mbox fbox` 不允许断字的地方后者给出一个方框 `mbox` 可用来分割连字
- 特殊字符
  - ‘输入两个表示双引号

- - 输入 1 个连字号 2 个短破折 3 个长破折网址中波浪号用 `$\sim$` 而不是`\~` 表示
- 摄氏度用 `$-30^\circ\mathrm{C}$` 表示
- `\ldots` 表示省略号 bable 宏包可处理多种非中文语言
- ~ 用来强制取消大写字母后空格多出的一点 `\@` 用来表示大写字母作为最后一个词后句号的处理一般 `latex` 不会处理大写字母后的句号（加入多一点空格）认为是缩写
- `\frontmatter` 应接着命令 `\begin{document}` 使用它把页码更换为罗马数字
- 正文前的内容普遍使用带星的命令（例如，`\chapter*{Preface}`）以阻止 `latex` 对它们排序
- `\mainmatter` 应出现在书的第一章紧前面它打开阿拉伯页码计数器并对页码从新计数
- `\appendix` 标志书中附录材料的开始该命令后的各章序号改用字母标记
- `\backmatter` 应该插入与书中最后一部分内容的紧前面如参考文献和索引在标准文档类型中它对页面没有什么效果
- 交叉引用 `\label{marker}` 引用点 `\ref{marker}` 引用 `\pageref{marker}` 引用点页码交叉引用
- 产生脚注 `\footnote{footnote text}`
- 强调 `\underline{text}` 下划线 `\emph{text}` 斜体强调中强调会切换字体
- 环境
  - `itemize` 环境用于简单的列表 `enumerate` 环境用于带序号的列表 `description` 环境用于带描述的列表
  - `flushleft` 和 `flushright` 环境分别产生靠左排列和靠右排列的段落
  - `center` 环境产生居中的文本如果你不输入命令 `\\\` 指定断行点 `latex` 将自行决定
  - `quote` 环境对重要断语和例子的引用很重要
  - `quotation` 环境用于超过几段的较长引用，因为它对段落进行缩进
  - `verse` 环境用于诗歌，在诗歌中断行很重要。在一行的末尾用 `\\\` 断行，在每一段后留一空行
  - `verbatim` 环境直接输出其中内容可用断字表示可表示空格较短的用`\verb*|like this :-(|`
  - `\begin{tabular}{table spec}` 用来生成表格
  - `\begin{figure}[placement specifier]` or `\begin{table}[placement specifier]` 表示浮动体
  - `\caption{caption text}` 给浮动体加标签
  - `\listoffigures` 与 `\listoftables` 生成图表目录
- 数学公式
  - 段落中放于 `\(` 和 `\)` \$ 和 \$ 或者 `\begin{math}` 和 `\end{math}`
  - 单独一行可放于 `\[` 和 `\]` 或 `\begin{displaymath}` 和 `\end{displaymath}`
  - 带编号可放于 `equation` 数学环境中
  - 空格和分行都将被忽略所有的空格或是由数学表达式逻辑的衍生或是由特殊的命令如 `\quad` 或 `\quad` 来得到
  - 不允许有空行每个公式中只能有一个段落
  - 每个字符都将被看作是一个变量名并以此来排版如果你希望在公式中出现普通的文本（使用正体字并可以有空格），那么你必须使用命令 `\text{...}` 来输入这些文本
- `\newtheorem{name}[counter]{text}[section]` 定理环境 `name` 是短关键字，用于标识“定理”。`text` 定义“定理”的真实名称，会在最终文件中打印出来。

- 建立新命令
  - `\newcommand{name}[num]{definition}`
  - 第一个参数 `name` 是你想要建立的命令的名称
  - 第二个参数 `definition` 是命令的定义
  - 第三个参数 `num` 是可选的用于指定命令所需的参数数目（命令最多可以有 9 个参数）如果不给出这个参数那么新建的命令将不接受任何参数
  - `num` 可用来传参，`\renewcommand` 可用来建立与原命令名称相同的命令
- 建立新环境 `\newenvironment{name}[num]{before}{after}`
- 建立新宏包 `\ProvidesPackage{package name}` 命令环境打包起名字保存为 `sty` 可直接调用其实就是打包导言区
- 行距`\linespread{factor}`
- 首行缩进与段落间距 `\setlength{\parindent}{0pt} \setlength{\parskip}{1ex plus 0.5ex minus 0.2ex}`
- 水平距离`\hspace{length}` 橡皮擦 `\stretch{n}`  $x\hspace{\stretch{3}}x$
- 垂直距离`\vspace{length}`
- `\sum\limits_{k=1}^n k^2` 使求和符号上下标真正出现在上下位



# 章 3

## 可复算性研究

### 3.1 Replication

- 科学研究的的终极标准是研究证据可独立发现与验证
- 并非所有结果都可以重复

### 3.2 Reproducible

- 可重复的数据分析过程与代码
- 数据维度增高
- 现有数据可被整合入更大的数据集
- 计算机条件允许



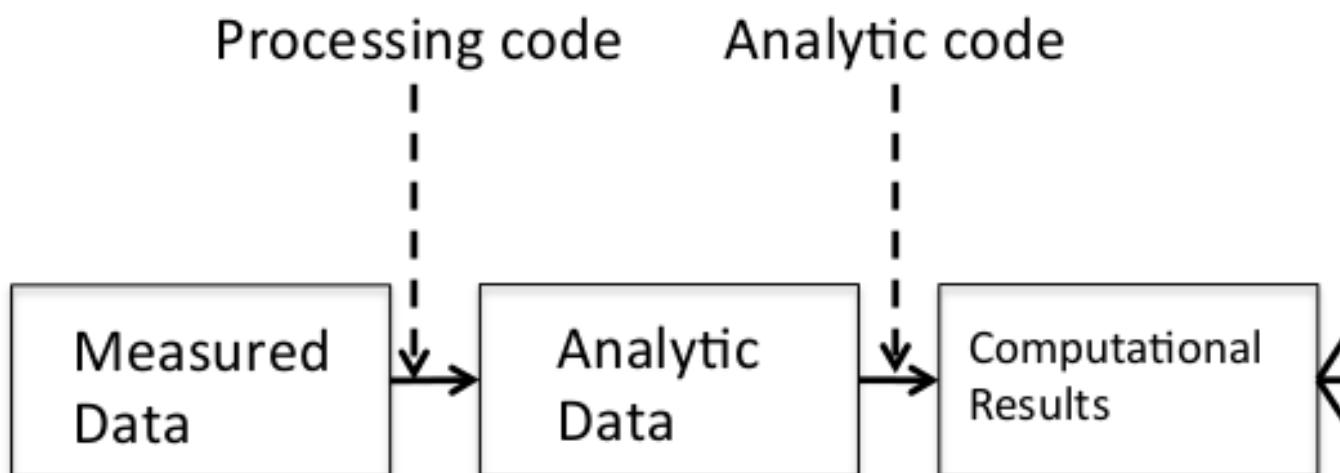
### 3.3 研究流程

# Research P

Author

---

Presentation



## 3.4 数据分析步骤

- 定义问题
  - 背后要有科学假设或问题
  - 从大到小具体定义
- 定义理想数据
  - 描述性的 <- 总体数据
  - 探索性的 <- 有属性测量的样本数据
  - 推断性的 <- 合适的总体随机采样
  - 预测性的 <- 来自同一总体有训练集与测试集的样本
  - 因果性的 <- 随机性研究
  - 机械性的 <- 系统中所有组成部分的数据
- 决定可获取数据
  - 网络免费数据
  - 购买数据
  - 注意使用条款
  - 数据不存在自己创造 <- 实验
- 获取数据
  - 原始数据
  - 引用来源
  - 网络数据注明数据来源 URL 与获取时间
- 整理数据
  - 原始数据需要整理
  - 如果事先处理过要搞清楚如何处理的
  - 了解数据来源
  - 需要重新格式化采样 <- 记录步骤
  - 判断数据是否合适不合适重新获取
- 探索性数据分析
  - 描述性总结数据
  - 检查缺失值
  - 绘制探索性图
  - 尝试探索性分析例如聚类
- 统计预测/建模
  - 基于探索性分析
  - 根据问题确定方法
  - 数据转换要解释
  - 测定的不确定性要考虑
- 解释结果
  - 描述
  - 相关
  - 推断
  - 预测
- 质疑结果
  - 问题
  - 数据源

- 处理过程
- 分析
- 结论
- 整合写出结果
  - 从问题角度出发
  - 形成一个故事
  - 不要包含分析过程除非用来说明问题消除质疑
  - 以故事而不是时间顺序描述
  - 图片要漂亮
- 写出可重复的 R 代码
  - Rmarkdown 文件

## 3.5 数据分析文件结构

- Data
  - Raw data 来自网络在 Readme 里注明 url 描述日期
  - Processed data 命名体现处理过程 Readme 里注明处理过程
- Figures
  - Exploratory figures 不必考虑装饰
  - Final figures 只考虑装饰
- R code
  - Raw scripts 不必过分注释版本控制不一定用得上
  - Final scripts 注释清晰包括处理细节只包括文章需要费分析
  - R Markdown files (optional)
- Text
  - Readme files 按步骤记录清晰
  - Text of analysis 包括前言方法结果结论讲故事有引用

## 3.6 文本化统计编程-Knitr

- markdown 是轻量化结构语言
- R markdown 是轻量化统计结构语言
- 文本 + 代码块逻辑清晰
- 文本语言可用 latex markdown
- 代码块可用 R
- 不用保存输出
- 可缓存结果 cache 包

## 3.7 结果通讯

- 研究论文的信息层级
  - 题目/作者名单
  - 摘要

- 主体/结果
  - 支持材料/细节
  - 代码/数据
- 邮件汇报的信息层级
    - 题目最好一行一句
    - 描述问题如何实验总结发现
    - 简明扼要
    - 如果有问题写成 yes/no 形式
    - 附件齐全严谨

### 3.8 检查列表

- 数据选取得当
- 问题简单专一
- 队友靠谱
- 兴趣驱动
- 不要手动处理数据全部交给计算机
- 少用交互界面用命令行界面并记录历史
- 使用版本控制处理降速而冷静
- 记录软件操作环境 `sessionInfo()`
- 不保存结果保证数据可重复
- 使用随机数要说明种子
- 原始数据-处理数据-分析-报告
- 考虑从哪一步开始数据重复性变差

### 3.9 基于证据的数据分析

- 可重复性研究不保证结果是对的
- 发表后研究存在动因应关注数据生成前的过程
- 设定基于证据研究的路线图
- 减少研究人员的自由度
- 提出区域研究范式

### 3.10 结果可解释

- 结果可解释模型

### 3.11 数据分析的理论

- 数据分析的核心应该是可重复性

# 章 4

## 探索性数据分析

### 4.1 ACES 模型

Letter	Step	Notes
A	Acquire the data and Assemble the data frame	Find data and import
C	Clean the data frame	Identify and limit columns, rows, indices, dates, etc.
E	Explore global properties	Visualize! Basic plots and stats appropriate to the data set
S	Subset comparisons	Look at (visualize!) initial emergenet variable relationships and subsets

### 4.2 探索绘图原则

- 表示可比的对比
- 表示因果解释机制系统结构
- 表示多元变量（超过 2）
- 证据整合目的驱动非工具驱动
- 证据描述要标注限定恰当
- 内容为王

### 4.3 探索性绘图

- 个人理解用
- 不用过分关注细节
- 基于问题或假设出发

### 4.4 分层聚类

- 找到最近的聚到一起找下个最近的
- 给出距离范围与距离计算方法

- 欧氏距离多维空间点距开平方
- manhattan 距离出租车距离绝对值
- 给出变量间或样本间的关系
- 图形可能不稳定多少样本多少类
- 结果是确定的
- 选定 cut 点并不明显
- 应该首先用来探索

## 4.5 k-means 聚类

- 固定聚类数给出聚类中心寻找最近的点循环
- 需要聚类数与聚类距离范围
- 需要大量聚类通过眼睛交叉检验
- $k$  的经验值  $\sqrt{n}/2$  或者根据解释的变量变化多少来选取
- 结果不确定根据聚类数与迭代次数而变化

## 4.6 维度还原

- 找到最不相关的数来解释整体方差（统计）在这些数中选取个数最少的来解释原始数据（压缩）
- 不一定是真实向量的叠加
- SVD 是 PCA 的一种解法 UDV 三个向量其中  $U$  表示行变化模式  $D$  表示方差  $V$  表示列变换模式这样有助于解释主成分变化
- 标准化与否影响结果
- 计算量大
- 类似探索分析还有因子分析独立成分分析潜在语义分析
- impute 包可补充缺失值

## 4.7 可视化图形

- 动态可视化
- 弦图
- 示意地图
- 变形地图绘制
- 重复模式可视化
- 不确定性可视化

# 章 5

## 统计推断

### 5.1 统计推断导论

- 定义用需要考虑不确定度的含噪音的统计学数据推断事实
- 工具随机化随机采样采样模型假设检验置信区间概率模型实验设计 bootstrapping 排列交换随机类型
  - 频率派使用概率的频率解释来控制错误率
  - 贝叶斯派给定概率与数据概率哪个靠谱

### 5.2 概率

- 术语
  - 样本空间  $\Omega$
  - 事件样本空间子集  $E$
  - 单独事件
  - 空事件
  - $E$  发生  $E$  发生
  - $E$  发生  $E$  不发生
  - $EF$   $E$  发生则  $F$  发生
  - $EF$   $EF$  一起发生
  - $EF$   $EF$  中至少一个发生
  - $EF = EF$  互斥
  - $E^c$  或  $\bar{E}$   $E$  不发生
- 概率
  1. 对事件  $E \subset \Omega$ ,  $0 \leq P(E) \leq 1$
  2.  $P(\Omega) = 1$
  3. 如果  $E_1$  与  $E_2$  互斥有  $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ .
  4. 概率无限可加性  $P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$
  5.  $P(\emptyset) = 0$
  6.  $P(E) = 1 - P(E^c)$
  7.  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  8. 如果  $A \subset B$  则  $P(A) \leq P(B)$
  9.  $P(A \cup B) = 1 - P(A^c \cap B^c)$
  10.  $P(A \cap B^c) = P(A) - P(A \cap B)$
  11.  $P(\cup_{i=1}^n E_i) \leq \sum_{i=1}^n P(E_i)$
  12.  $P(\cup_{i=1}^n E_i) \geq \max_i P(E_i)$
- 随机变量

- 实验的数值输出
- 离散随机变量取可数的概率  $P(X = k)$
- 连续随机变量取连续区间子集概率  $P(X \in A)$
- 概率质量函数 (PMF) <- 离散随机变量
  1. 对于所有  $x$   $p(x) \geq 0$
  2.  $\sum_x p(x) = 1$
- 概率密度函数 (PDF) <- 连续随机变量
  1. 对于所有  $x$   $f(x) \geq 0$
  2.  $f(x)$  下面积为 1
- 累计概率函数 (CDF)
  - 定义  $F(x) = P(X \leq x)$
  - 生存函数  $S(x) = P(X > x)$   $S(x) = 1 - F(x)$
  - 对于连续函数 CDF 是 PDF 的积分
- 分位数  $\alpha^{th}$ 
  - $F(x_\alpha) = \alpha$
  - $50^{th}$  分位数是中位数

### 5.3 期望

- 离散随机变量均值  $E[X] = \sum_x xp(x)$
- $E[X]$  代表质量与位置的中心  $\{x, p(x)\}$
- 连续随机变量均值  $E[X] =$  the area under the function  $t f(t)$
- 期望值是线性可加的
- 如果  $a$  与  $b$  不随机  $X$  与  $Y$  是随机变量
  - $E[aX + b] = aE[X] + b$
  - $E[X + Y] = E[X] + E[Y]$
- 样本均值是总体均值  $\mu$  的无偏估计的证明

$$\begin{aligned}
 E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] &= \frac{1}{n} E\left[\sum_{i=1}^n X_i\right] \\
 &= \frac{1}{n} \sum_{i=1}^n E[X_i] \\
 &= \frac{1}{n} \sum_{i=1}^n \mu = \mu.
 \end{aligned}$$

### 5.4 方差

- 描述随机变量的离散情况
- 如果  $X$  是均值  $\mu$  的随机变量其方差为  $Var(X) = E[(X - \mu)^2]$
- 离开均值距离期望的平方
- 计算公式  $Var(X) = E[X^2] - E[X]^2$
- 如果  $a$  是常数有  $Var(aX) = a^2 Var(X)$
- 方差的开方是标准差单位与  $X$  一致
- 车比雪夫不等式 (Chebyshev's inequality) 边界极为保守

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

## 5.5 独立性

- 独立事件

- 两事件  $A$  与  $B$  在  $P(A \cap B) = P(A)P(B)$  下独立
- 在  $P([X \in A] \cap [Y \in B]) = P(X \in A)P(Y \in B)$  下两随机变量  $X$  与  $Y$  独立
- 对于一组随机独立变量  $X_1, X_2, \dots, X_n$  有  $f(x_1, \dots, x_n) = \prod_{i=1}^n f_i(x_i)$
- iid 随机变量 (independent and identically distributed) 来自同一分布相互独立的随机变量

- 协方差 (covariance)

- $Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)] = E[XY] - E[X]E[Y]$
- $Cov(X, Y) = Cov(Y, X)$
- $Cov(X, Y)$  可以有正负
- $|Cov(X, Y)| \leq \sqrt{Var(X)Var(y)}$

- 相关性 (correlation)

- $X$  与  $Y$  的相关性  $Cor(X, Y) = Cov(X, Y)/\sqrt{Var(X)Var(y)}$
- $-1 \leq Cor(X, Y) \leq 1$
- 只有对常数  $a$  与  $b$  满足  $X = a + bY$  时  $Cor(X, Y) = \pm 1$
- $Cor(X, Y)$  无单位
- $Cor(X, Y) = 0$  时  $X$  与  $Y$  不相关
- $Cor(X, Y)$  越接近 1  $X$  与  $Y$  越正相关反之接近-1 负相关
- $\{X_i\}_{i=1}^n$  是一组随机变量当  $\{X_i\}$  不相关时  $Var(\sum_{i=1}^n a_i X_i + b) = \sum_{i=1}^n a_i^2 Var(X_i)$
- 如果一组随机变量  $\{X_i\}$  不相关方差的和等于和的方差非标准差

- 样本均值方差的推导

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} Var\left(\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n Var(X_i) \\ &= \frac{1}{n^2} \times n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

- 当  $X_i$  独立且方差为  $Var(X_i) = \frac{\sigma^2}{n}$
- $\sigma/\sqrt{n}$  为样本均值的标准误
- 样本均值的标准误就是样本均值分布的标准差
- $\sigma$  是一次观察分布的标准差
- 样本均值要比一次观察变化小因此除以  $\sqrt{n}$

- 样本方差

- $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$
- 总体方差  $\sigma^2$  的估计
- 计算  $\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2$
- 均值偏差平方的均值
- 样本方差是总体方差的无偏估计

$$\begin{aligned}
E \left[ \sum_{i=1}^n (X_i - \bar{X})^2 \right] &= \sum_{i=1}^n E[X_i^2] - nE[\bar{X}^2] \\
&= \sum_{i=1}^n \{Var(X_i) + \mu^2\} - n\{Var(\bar{X}) + \mu^2\} \\
&= \sum_{i=1}^n \{\sigma^2 + \mu^2\} - n\{\sigma^2/n + \mu^2\} \\
&= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 \\
&= (n-1)\sigma^2
\end{aligned}$$

- 澄清

- 假定  $X_i$  是 iid 均值  $\mu$  方差  $\sigma^2$
- $S^2$  估计  $\sigma^2$
- $S^2$  的计算涉及除  $n-1$
- $S/\sqrt{n}$  估计  $\sigma/\sqrt{n}$  是均值的标准误

## 5.6 条件概率

- $B$  为一个事件有  $P(B) > 0$
- $B$  出现条件下  $A$  的条件概率为  $P(A | B) = \frac{P(A \cap B)}{P(B)}$
- 如果  $A$  与  $B$  独立有  $P(A | B) = \frac{P(A)P(B)}{P(B)} = P(A)$

## 5.7 贝叶斯定理

$$P(B | A) = \frac{P(A | B)P(B)}{P(A | B)P(B) + P(A | B^c)P(B^c)}.$$

- 2\*2 列联表 - 诊断测试

		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	True Positive	False Positive (Type I error)	Positive predict $\frac{\sum \text{True Pos}}{\sum \text{Test Outcome}}$
	Test Outcome Negative	False Negative (Type II error)	True Negative	Negative predict $\frac{\sum \text{True Neg}}{\sum \text{Test Outcome}}$
		<b>Sensitivity =</b> $\frac{\sum \text{True Positive}}{\sum \text{Condition Positive}}$	<b>Specificity =</b> $\frac{\sum \text{True Negative}}{\sum \text{Condition Negative}}$	

## 5.8 常见分布

- 贝努力分布
  - 二元输出变量
  - 数值为 0 或 1 概率  $p$  与  $1-p$
  - $X$  的 PMF 是  $P(X=x) = p^x(1-p)^{1-x}$
  - 均值  $p$  方差  $p(1-p)$
  - 如果有 iid 的贝努力观察  $x_1, \dots, x_n$  似然函数  $\prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$
  - 似然函数依赖  $x_i$  的和  $\sum_i x_i/n$  包含了所有  $p$  的可能性
  - 最大化似然函数可以得到  $p$  的估计
- 二项分布
  - PMF
 
$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$$

对于  $x = 0, \dots, n$
- 正态分布
  - PDF  $(2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}$
  - $X$  为均值  $E[X] = \mu$  方差  $Var(X) = \sigma^2$  的 iid 随机变量
  - 写作  $X \sim N(\mu, \sigma^2)$
  - 均值  $\mu = 0$  方差  $\sigma = 1$  是标准正态分布
  - 标准正态函数写作  $\phi$
  - 标准正态随机变量用  $Z$  表示
  - 如果  $X \sim N(\mu, \sigma^2)$  并且  $Z = \frac{X-\mu}{\sigma}$  是标准正态函数
  - 如果  $Z$  是标准正态函数  $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$

- 非标准正态密度函数  $\phi\{(x - \mu)/\sigma\}/\sigma$
- 正态似然函数对方差的估计是有偏的
- 正态的和是正态样本均值正态
- 正态的平方是卡方
- 正态分布
- 泊松分布
  - PMF  $P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$
  - 均值方差均为  $\lambda$
  - 可看做很短时间间隔中发生事件的概率模拟速率其中  $\lambda * h$  小于 1 则各时间段独立
  - $X \sim Poisson(\lambda t)$   $\lambda = E[X/t]$  是速率  $t$  是总时间
  - $n$  大  $p$  小是对二项分布的模拟
  - $X \sim Binomial(n, p)$ ,  $\lambda = np$

## 5.9 演进

- 样本接近无穷大时统计量的行为
- 频率派的基石
- 大数理论 (LLN) 样本数量越多均值接近期望
- 中心极限理论 (CLT) iid 变量均值的分布标准化后随样本数增加接近标准正态分布

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\text{Estimate} - \text{Mean of estimate}}{\text{Std. Err. of estimate}}$$

- 可根据变量分布来知道均值方差计算出样本均值标准误就可以根据 CLT 计算逼近的统计量
- 置信区间
  - 根据 CLT 随机区间  $\bar{X}_n \pm z_{1-\alpha/2}\sigma/\sqrt{n}$  包括  $\mu$  的概率逼近于  $100(1 - \alpha)\%$   $z_{1-\alpha/2}$  为标准正态分布  $1 - \alpha/2$  的分位数  $100(1 - \alpha)\%$  为置信区间  $\sigma$  可用样本估计  $s$  来近似
  - 估计是基于分布假设的如果分布有解析解则置信区间可以更准确的得到估计
  - 先生成不依赖参数的统计量
  - 根据统计量的概率分布计算参数的边界

## 5.10 置信区间

- 卡方分布
  - 假定  $S^2$  是来自  $n$  个 iid  $N(\mu, \sigma^2)$  数据样本的方差有  $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$  符合自由度  $n-1$  的卡方分布
  - 不对称分布
  - 均值是自由度方差是两倍的自由度
  - 方差的置信区间

$$\begin{aligned} 1 - \alpha &= P\left(\chi^2_{n-1, \alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1, 1-\alpha/2}\right) \\ &= P\left(\frac{\chi^2_{n-1, 1-\alpha/2}}{(n-1)S^2} \leq \frac{\sigma^2}{(n-1)S^2} \leq \frac{\chi^2_{n-1, \alpha/2}}{(n-1)S^2}\right) \end{aligned}$$

- $\left[\frac{\chi^2_{n-1, 1-\alpha/2}}{(n-1)S^2}, \frac{\chi^2_{n-1, \alpha/2}}{(n-1)S^2}\right]$  是  $\sigma^2$  的  $100(1 - \alpha)\%$  置信区间
- 依赖正态性假设开方后得到  $\sigma$  的置信区间
- Gosset 的 t 分布
  - 比正态分布尾厚
  - 考虑自由度自由度大时接近正态分布

- $\frac{Z}{\sqrt{\frac{X^2}{df}}}$
- 假定  $(X_1, \dots, X_n)$  是 iid  $N(\mu, \sigma^2)$  有  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  是标准正态分布  $\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} = S/\sigma$  是卡方除以自由度的开方
- 有

$$\frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{S/\sigma} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

- 服从自由度  $n - 1$  的  $t$  分布
- 均值的置信区间

$$\begin{aligned} & 1 - \alpha \\ &= P\left(-t_{n-1, 1-\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1, 1-\alpha/2}\right) \\ &= P(\bar{X} - t_{n-1, 1-\alpha/2}S/\sqrt{n} \leq \mu \leq \bar{X} + t_{n-1, 1-\alpha/2}S/\sqrt{n}) \end{aligned}$$

$t_{df, \alpha}$  是  $t$  分布的  $\alpha^{th}$  分位数自由度  $df$   
 -  $t$  检验不适合有偏分布置信区间中心也不在均值上

## 5.11 似然函数

- 一组数据的似然函数是数据固定下参数的联合概率密度函数
- 似然函数可用来估计参数是参数的函数
- 似然函数比估计两个可能参数值的可能性
- 给定模型与数据似然函数包含所有参数可能性
- 样本独立时参数的似然函数是各独立样本似然函数的乘积
- 参数使似然函数概率取最大值时真实的可能性更大更支持这组数据这个估计是最大似然估计 (MLE)

## 5.12 贝叶斯推断

- Posterior  $\propto$  Likelihood  $\times$  Prior
- 先验 beta 分布
  - 01 之间
  - 依赖  $\alpha, \beta$  的概率密度函数

$$\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} \quad \text{for } 0 \leq p \leq 1$$

- 均值  $\alpha/(\alpha + \beta)$
- 方差  $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- $\alpha = \beta = 1$  为均匀分布
- 后验 beta 分布
  - 参数  $\tilde{\alpha} = x + \alpha$   $\tilde{\beta} = n - x + \beta$  的 beta 分布

$$\text{Posterior} \propto p^x (1-p)^{n-x} \times p^{\alpha-1} (1-p)^{\beta-1} \tag{5.1}$$

$$= p^{x+\alpha-1} (1-p)^{n-x+\beta-1} \tag{5.2}$$

- 后验均值

$$E[p \mid X] = \frac{\tilde{\alpha}}{\tilde{\alpha} + \tilde{\beta}} \quad (5.3)$$

(5.4)

$$= \frac{x + \alpha}{x + \alpha + n - x + \beta} \quad (5.5)$$

(5.6)

$$= \frac{x + \alpha}{n + \alpha + \beta} \quad (5.7)$$

(5.8)

$$= \frac{x}{n} \times \frac{n}{n + \alpha + \beta} + \frac{\alpha}{\alpha + \beta} \times \frac{\alpha + \beta}{n + \alpha + \beta} \quad (5.9)$$

(5.10)

$$= \text{MLE} \times \pi + \text{Prior Mean} \times (1 - \pi) \quad (5.11)$$

- 后验均值是先验均值与最大似然估计的混合
- 当  $n$  变大  $\pi$  接近 1 先验作用小
- 当  $n$  很小先验作用大
- 当数据量够大时先验概率作用就很小了
- 当先验概率足够稳定数据就作用不大了
- 信任区间
  - 95% 信任区间  $[a, b]$  会满足  $P(p \in [a, b] \mid x) = .95$
  - 最高后验密度 (HPD) 区间

### 5.13 两独立样本 t 检验

- $X_1, \dots, X_{n_x}$  为 iid  $N(\mu_x, \sigma^2)$
- $Y_1, \dots, Y_{n_y}$  为 iid  $N(\mu_y, \sigma^2)$
- $\bar{X}, \bar{Y}, S_x, S_y$  为均值与标准差
- 根据均值与方差的线性组合有  $\bar{Y} - \bar{X}$  也是正态均值  $\mu_y - \mu_x$  方差  $\sigma^2(\frac{1}{n_x} + \frac{1}{n_y})$
- 混合方差为  $S_p^2 = \{(n_x - 1)S_x^2 + (n_y - 1)S_y^2\}/(n_x + n_y - 2)$  为  $\sigma^2$  的良好估计
- 该估计为无偏估计

$$\begin{aligned} E[S_p^2] &= \frac{(n_x - 1)E[S_x^2] + (n_y - 1)E[S_y^2]}{n_x + n_y - 2} \\ &= \frac{(n_x - 1)\sigma^2 + (n_y - 1)\sigma^2}{n_x + n_y - 2} \end{aligned}$$

- 该估计独立于  $\bar{Y} - \bar{X}$  因为方差独立于均值
- 两个独立的卡方变量之和是自由度之和的卡方值

$$\begin{aligned}
(n_x + n_y - 2)S_p^2/\sigma^2 &= (n_x - 1)S_x^2/\sigma^2 + (n_y - 1)S_y^2/\sigma^2 \\
&= \chi_{n_x-1}^2 + \chi_{n_y-1}^2 \\
&= \chi_{n_x+n_y-2}^2
\end{aligned}$$

- 构建统计量

$$\frac{\bar{Y} - \bar{X} - (\mu_y - \mu_x)}{\sqrt{\frac{(n_x + n_y - 2)S_p^2}{(n_x + n_y - 2)\sigma^2}}} = \frac{\bar{Y} - \bar{X} - (\mu_y - \mu_x)}{S_p \left( \frac{1}{n_x} + \frac{1}{n_y} \right)^{1/2}}$$

- 该统计量为符合自由度  $n_x + n_y - 2$  的  $t$  分布
- 置信区间

$$\bar{Y} - \bar{X} \pm t_{n_x+n_y-2, 1-\alpha/2} S_p \left( \frac{1}{n_x} + \frac{1}{n_y} \right)^{1/2}$$

- 方差不等

$$\bar{Y} - \bar{X} \sim N \left( \mu_y - \mu_x, \frac{s_x^2}{n_x} + \frac{s_y^2}{n_y} \right)$$

- 统计量

$$\frac{\bar{Y} - \bar{X} - (\mu_y - \mu_x)}{\left( \frac{s_x^2}{n_x} + \frac{s_y^2}{n_y} \right)^{1/2}}$$

近似于自由度

$$\frac{(S_x^2/n_x + S_y^2/n_y)^2}{\left( \frac{S_x^2}{n_x} \right)^2/(n_x - 1) + \left( \frac{S_y^2}{n_y} \right)^2/(n_y - 1)}$$

的  $t$  分布

## 5.14 假设检验

- 使用数据做决定
- 空假设  $H_0$  无变化
- 备择假设  $H_a$  或大或小或不等
- 真值表

Truth	Decide	Result
$H_0$	$H_0$	Correctly accept null
$H_0$	$H_a$	Type I error
$H_a$	$H_a$	Correctly reject null
$H_a$	$H_0$	Type II error

- Z 检验
  - Z 检验  $H_0 : \mu = \mu_0$  与
    - \*  $H_1 : \mu < \mu_0$

- \*  $H_2 : \mu \neq \mu_0$
- \*  $H_3 : \mu > \mu_0$
- 检验统计量  $TS = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$
- 拒绝空假设条件
  - \*  $TS \leq -Z_{1-\alpha}$
  - \*  $|TS| \geq Z_{1-\alpha/2}$
  - \*  $TS \geq Z_{1-\alpha}$
- 样本数要足够否则选  $t$  检验
- 通过  $\alpha$  控制了 Type I error 但没控制  $\beta$  Type II error 所以结论为没有拒绝  $H_0$  而不是接受  $H_0$
- 拒绝  $H_0$  的值域为拒绝域
- 二项分布不易做正态假设可精确计算拒绝域

## 5.15 P 值

- 假定没有事发生出现状况的可能性
- 先定义分布然后计算相关统计量对比常见阈值看数值是否够极端
- 阈值为达到显著性水平与 p 值有区别
- p 值可设定任意显著性水平小于就可以拒绝
- 两尾检验单尾概率翻倍
- 独立于假设检验但常常一起使用

## 5.16 功效

- 错误拒绝空假设的概率为功效 (power)
- Power =  $1 - \beta$  对 Type II error 的控制
- 正态分布假设下的推导

$$1 - \beta = P\left(\frac{\bar{X} - 30}{\sigma/\sqrt{n}} > z_{1-\alpha} \mid \mu = \mu_a\right) \quad (5.12)$$

$$= P\left(\frac{\bar{X} - \mu_a + \mu_a - 30}{\sigma/\sqrt{n}} > z_{1-\alpha} \mid \mu = \mu_a\right) \quad (5.13)$$

(5.14)

$$= P\left(\frac{\bar{X} - \mu_a}{\sigma/\sqrt{n}} > z_{1-\alpha} - \frac{\mu_a - 30}{\sigma/\sqrt{n}} \mid \mu = \mu_a\right) \quad (5.15)$$

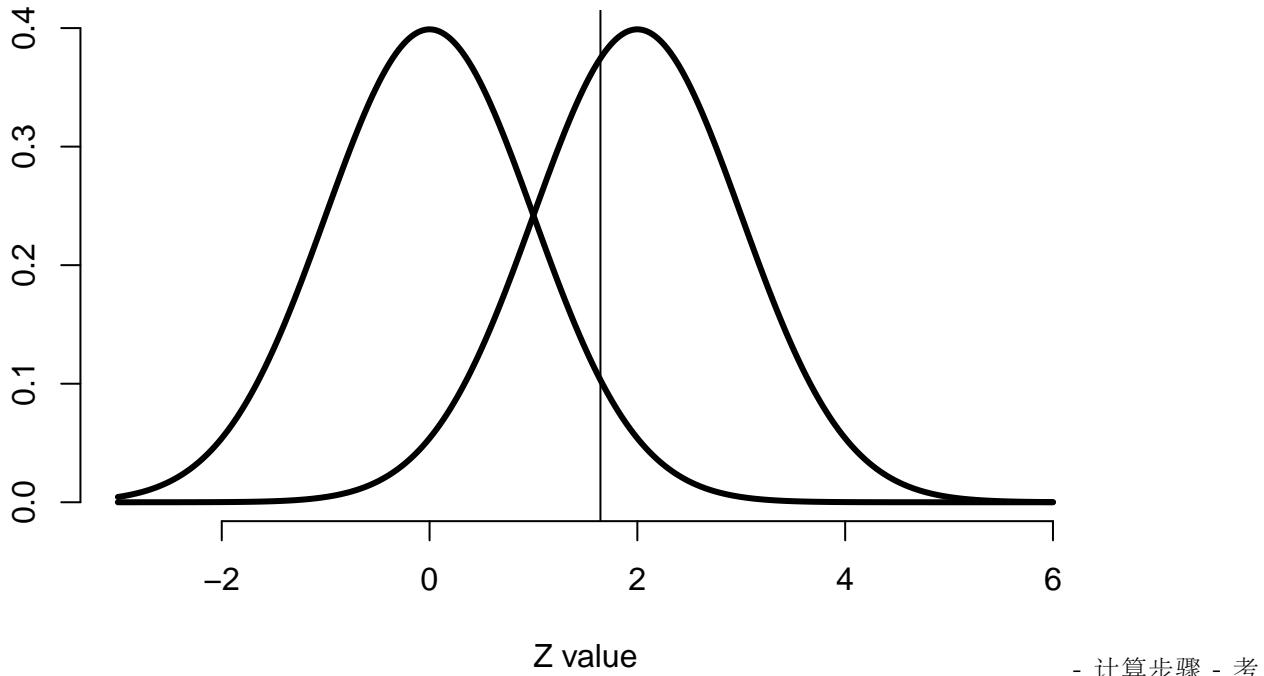
(5.16)

$$= P\left(Z > z_{1-\alpha} - \frac{\mu_a - 30}{\sigma/\sqrt{n}} \mid \mu = \mu_a\right) \quad (5.17)$$

(5.18)

(5.19)

```
sigma <- 10; mu_0 = 0; mu_a = 2; n <- 100; alpha = .05
plot(c(-3, 6), c(0, dnorm(0)), type = "n", frame = F, xlab = "Z value", ylab = "")
xvals <- seq(-3, 6, length = 1000)
lines(xvals, dnorm(xvals), type = "l", lwd = 3)
lines(xvals, dnorm(xvals, mean = sqrt(n) * (mu_a - mu_0) / sigma), lwd = 3)
abline(v = qnorm(1 - alpha))
```



考虑  $H_0 : \mu = \mu_0$  与  $H_a : \mu > \mu_0$  且在  $H_a$  下  $\mu = \mu_a$  - 在  $H_0$  下统计量  $Z = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}$  符合  $N(0, 1)$  - 在  $H_a$  下  $Z$  是  $N\left(\frac{\sqrt{n}(\mu_a - \mu_0)}{\sigma}, 1\right)$  - 如果  $Z > z_{1-\alpha}$  拒绝空假设也就是给定条件下功效不够 - 当检验  $H_a : \mu > \mu_0$ , 如果功效为  $1 - \beta$  那么  $1 - \beta = P\left(Z > z_{1-\alpha} - \frac{\mu_a - \mu_0}{\sigma/\sqrt{n}} \mid \mu = \mu_a\right) = P(Z > z_\beta)$  也就是  $z_{1-\alpha} - \frac{\sqrt{n}(\mu_a - \mu_0)}{\sigma} = z_\beta - \mu_a, \sigma, n, \beta, \mu_0, \alpha$  给定五个可解出剩余的 - 两尾检验考虑  $\alpha/2$  - 功效在  $\alpha$  提高单尾检验功效高于两尾  $\mu_1$  距离  $\mu_0$  远功效大样本数提高功效高 - 计算功效不需要特定样本只需要指定  $\frac{\mu_a - \mu_0}{\sigma}$  也就是有效样本大小无单位 - R 中使用 `power.t.test` 来计算 t 检验功效相关参数指定多数求一个

## 5.17 多重比较

- 多次进行比较会导致错误率与校正出现问题
- False positive rate 错误结果是显著的比率  $\alpha$  样本数增大错误增加
- Family wise error rate (FWER) 所有比较中至少一个假阳性比率
  - Bonferroni correction
  - 假设你进行  $m$  次测试控制  $\alpha$  在某水平计算所有测试的  $p$  值将  $\alpha$  设为  $\frac{\alpha}{m}$  所有测试都在这个置信度下进行
  - 容易计算过于保守
- False discovery rate (FDR) 声称显著是错误的概率
  - $m$  次测试水平  $\alpha$  计算  $p$  值
  - 排序  $P_{(i)} \leq \alpha \times \frac{i}{m}$  为显著
  - 相对容易计算不保守允许一定的假阳性
- 调节  $p$  值
  - $P_i^{fwer} = \max(m \times P_i, 1)$  类似 FWER 处理  $\alpha$  的方式处理  $p$  按照正常  $\alpha$  检测
- 一般情况对  $p$  值用 bonferroni/BH 纠正就够了
- 对比间依赖强烈考虑 method="BY"
- 多重比较从原理到应用 从实用角度分类适合常见科研实验结果处理

## 5.18 重采样推断

- jackknife

- 用来无偏估计偏差与标准误
- 每次估计删掉一个数据  $\bar{\theta} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_i$
- 偏差  $(n - 1)(\bar{\theta} - \theta)$
- 标准误  $\left[ \frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_i - \bar{\theta})^2 \right]^{1/2}$
- 可用来估计分位数是 bootstrap 的线性逼近但性质不好
- 假观察量角度理解 jackknife Pseudo Obs =  $n\hat{\theta} - (n - 1)\hat{\theta}_i$  生成原数据集
- bootstrap

## 5.19 概念可视化

- 统计概念可视化
  - 构建置信区间与求标准误
  - 假定采样分布是总体分布重采样估计统计量
  - 有放回的重采样  $B$  次  $N$  个样本得到估计统计量的一个分布直接计算置信区间
  - 非参方法偏差小 进阶指南
- 置换检验
  - 分组对比时取消原分组随机分组
  - 重复进行记录分组差异
  - 对比原参数与置换后参数差异进行推断

# 章 6

## 回归模型

### 6.1 回归模型导论

- Francis Galton 1885 年用父母身高预测子女身高的案例
- 考虑单变量的数据代表：最小二乘值
  - 最小二乘值物理意义为质心
  - 最小二乘统计学意义是平均值
  - 可用不等式解也可用求导方法解

$$\sum_{i=1}^n (Y_i - \mu)^2 = \sum_{i=1}^n (Y_i - \bar{Y} + \bar{Y} - \mu)^2 \quad (6.1)$$

$$= \sum_{i=1}^n (Y_i - \bar{Y})^2 + 2 \sum_{i=1}^n (Y_i - \bar{Y})(\bar{Y} - \mu) + \sum_{i=1}^n (\bar{Y} - \mu)^2 \quad (6.2)$$

$$= \sum_{i=1}^n (Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu) \sum_{i=1}^n (Y_i - \bar{Y}) + \sum_{i=1}^n (\bar{Y} - \mu)^2 \quad (6.3)$$

$$= \sum_{i=1}^n (Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu)(\sum_{i=1}^n Y_i - n\bar{Y}) + \sum_{i=1}^n (\bar{Y} - \mu)^2 \quad (6.4)$$

$$= \sum_{i=1}^n (Y_i - \bar{Y})^2 + \sum_{i=1}^n (\bar{Y} - \mu)^2 \quad (6.5)$$

$$\geq \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (6.6)$$

- 通过原点的回归
  - 最小化  $\sum_{i=1}^n (Y_i - X_i \beta)^2$
  - 两变量关系用回归线解释
- 回归分析种类大全

### 6.2 术语

- $X_1, X_2, \dots, X_n$  表示  $n$  个数据点
- $Y_1, \dots, Y_n$  表示另外  $n$  个数据点
- 用希腊字母表示不知道的东西如  $\mu$
- 大写字母表示概念值小写字母表示真实值如  $P(X_i > x)$
- $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  表示均值数据的中心趋向

- $\tilde{X}_i = X_i - \bar{X}$  表示对数据中心化均值为 0
- 均值为数据的最小二乘估计
- $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} (\sum_{i=1}^n X_i^2 - n\bar{X}^2)$  表示方差
- $S$  为标准差数据的离散程度
- $X_i/s$  表示数据缩放方差为 1
- $Z_i = \frac{X_i - \bar{X}}{s}$  表示数据的标准化先中心化再标准化
- $Cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1} (\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y})$  表示协方差
- $Cor(X, Y) = \frac{Cov(X, Y)}{S_x S_y}$  表示相关性
  - $Cor(X, Y) = Cor(Y, X)$
  - $-1 \leq Cor(X, Y) \leq 1$
  - $Cor(X, Y)$  度量线性关系强度
  - $Cor(X, Y) = 0$  表示无线性关系

### 6.3 回归线的最小二乘回归

- 用最小二乘法寻找回归线最小化  $\sum_{i=1}^n \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$
- 如果定义  $\mu_i = \beta_0 + \hat{\beta}_0 = \bar{Y}$  不考虑其他变量  $Y$  的均值就是最小二乘估计
- 如果定义  $\mu_i = X_i \beta_1 + \hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2}$  如果考虑过原点线的回归斜率如上
- 如果考虑  $\mu_i = \beta_0 + \beta_1 X_i$

$$\sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(\hat{\beta}_0 + \hat{\beta}_1 X_i - \beta_0 - \beta_1 X_i) \quad (6.7)$$

$$= (\hat{\beta}_0 - \beta_0) \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) + (\beta_1 - \hat{\beta}_1) \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i \quad (6.8)$$

(6.9)

- 解为  $\hat{\beta}_1 = Cor(Y, X) \frac{Sd(Y)}{Sd(X)}$   $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
- 如果标准化数据  $\{\frac{X_i - \bar{X}}{Sd(X)}, \frac{Y_i - \bar{Y}}{Sd(Y)}\}$  解为  $Cor(Y, X)$
- 回归是因变量向自己均值回归与向自变量相关回归的平衡

### 6.4 统计线性回归模型

- 最小二乘是一种估计方法，做推断需要模型
- 建立线性回归的概率模型  $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- $\epsilon_i$  为 iid  $N(0, \sigma^2)$
- $E[Y_i | X_i = x_i] = \mu_i = \beta_0 + \beta_1 x_i$
- $Var(Y_i | X_i = x_i) = \sigma^2$
- 对  $N(\mu_i, \sigma^2)$  独立变量  $Y$  进行极大似然估计
- $\mathcal{L}(\beta, \sigma) = \prod_{i=1}^n \{(2\pi\sigma^2)^{-1/2} \exp(-\frac{1}{2\sigma^2}(y_i - \mu_i)^2)\}$
- 取对数有  $-2 \log\{\mathcal{L}(\beta, \sigma)\} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu_i)^2 + n \log(\sigma^2)$
- 最小二乘估计就是极大似然估计
- $\hat{\beta}_1 = Cor(Y, X) \frac{Sd(Y)}{Sd(X)}$   $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
- 截距是自变量为 0 时  $Y$  的期望斜率是自变量变化一个单位对  $Y$  的影响

### 6.5 残差

- 模型  $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

- 预测值  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- $e_i = Y_i - \hat{Y}_i$  观察数据与回归线的垂直距离
- 最小二乘估计最小化残差  $\sum_{i=1}^n e_i^2$
- 残差  $e_i$  可看作  $\epsilon_i$  的估计
- 可证  $E[e_i] = 0$  模型中考虑截距  $\sum_{i=1}^n e_i = 0$  考虑自变量  $\sum_{i=1}^n e_i X_i = 0$
- 残差可用来评价模型效果
- 残差波动不同于模型波动
- 残差波动  $\sigma^2$  的极大似然估计为  $\frac{1}{n} \sum_{i=1}^n e_i^2$
- $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$  为无偏估计

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \quad (6.10)$$

$$= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \quad (6.11)$$

$$(6.12)$$

- 其中  $(Y_i - \hat{Y}_i) = \{Y_i - (\bar{Y} - \hat{\beta}_1 \bar{X}) - \hat{\beta}_1 X_i\} = (Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})$
- $(\hat{Y}_i - \bar{Y}) = (\bar{Y} - \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i - \bar{Y}) = \hat{\beta}_1(X_i - \bar{X})$
- 有  $\sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^n \{(Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})\} \{\hat{\beta}_1(X_i - \bar{X})\} = \hat{\beta}_1 \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) - \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 = \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 - \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 = 0$
- 综上  $\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$
- 有 Total Variation = Residual Variation + Regression Variation
- 模型解释部分  $R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$
- 已知  $(\hat{Y}_i - \bar{Y}) = \hat{\beta}_1(X_i - \bar{X})$   $\hat{\beta}_1 = Cor(Y, X) \frac{Sd(Y)}{Sd(X)}$  有  $R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \hat{\beta}_1^2 \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = Cor(Y, X)^2$
- $R^2$  实际上是相关性  $r$  的平方  $<$  线性模型的可解释性
- $R^2$  会伴随样本数增加而增加会因删除异常值而增加
- `data(anscombe); example(anscombe)`
- 小恐龙变换

## 6.6 回归推断

- $\frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}}$  总符合正态分布或  $t$  分布
- 假设检验  $H_0 : \theta = \theta_0$  与  $H_a : \theta >, <, \neq \theta_0$
- 置信区间  $\theta$  通过  $\hat{\theta} \pm Q_{1-\alpha/2} \hat{\sigma}_{\hat{\theta}}$  构建

$$Var(\hat{\beta}_1) = Var\left(\frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) \quad (6.13)$$

$$= \frac{Var(\sum_{i=1}^n Y_i(X_i - \bar{X}))}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} \quad (6.14)$$

$$= \frac{\sum_{i=1}^n \sigma^2 (X_i - \bar{X})^2}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} \quad (6.15)$$

$$= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (6.16)$$

$$(6.17)$$

- $\sigma_{\hat{\beta}_1}^2 = Var(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^n (X_i - \bar{X})^2$

- $\sigma_{\hat{\beta}_0}^2 = \text{Var}(\hat{\beta}_0) = \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \sigma^2$
- 这样  $\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}}$  遵守自由度为  $n - 2$  的  $t$  分布或正态分布
- 在  $x_0$  回归线的标准误  $\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$
- 在  $x_0$  预测值的标准误  $\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$
- CI 代表回归线在特定  $x$  处的变动 PI 代表预测值在此处的变动前者在回归线固定时不变后者还要考虑预测值围绕回归线的变动

The prediction interval is the range in which future observation can be thought most likely to occur, whereas the confidence interval is where the mean of future observation is most likely to reside. From here

## 6.7 多元回归

- 线性模型  $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} + \epsilon_i = \sum_{k=1}^p X_{ik} \beta_k + \epsilon_i$
- 最小化  $\sum_{i=1}^n (Y_i - \sum_{k=1}^p X_{ki} \beta_k)^2$  最小二乘估计也是误差正态化的极大似然估计
- 最小二乘估计等价于  $\sum_{i=1}^n (Y_i - X_{1i} \hat{\beta}_1 - \dots - X_{ip} \hat{\beta}_p) X_k = 0$  本质上使其他参数固定解出一个然后逐级代入最后全部解出参数值参考线性代数
- 参数代表固定其他参数后变动一个单位引发的变化
- 方差估计  $\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n e_i^2$
- 参数标准误  $\hat{\sigma}_{\hat{\beta}_k} \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}}$  符合自由度  $n - p$  的  $T$  分布
- 多元模型中加入变量会导致原有变量的参数估计发生变化甚至方向相反一般是由于加入变量与原有变量存在共相关导致两者参数估计都不准

```
n <- 100; x2 <- 1 : n; x1 <- .01 * x2 + runif(n, -.1, .1); y = -x1 + x2 + rnorm(n, sd = .01)
summary(lm(y ~ x1))$coef
summary(lm(y ~ x1 + x2))$coef
```

- R 会自动检测并消除变量生成的变量如上面 x2 中需要加入 `runif(n, -.1, .1)` 才能得到结果
- 多元模型中包括分类变量考虑加入虚拟变量  $Y_i = \beta_0 + X_{i1} \beta_1 + \epsilon_i$  属于该分类时  $E[Y_i] = \beta_0 + \beta_1$  否则为  $E[Y_i] = \beta_0$
- 分类变量截距有意义代表其中一个分类等同于其他分类与该分类进行 t 检验如果模型中去掉截距等同于所有分类与零进行 t 检验参数系数为均值差可用 `relevel(data, 'name')` 来指定比对对象
- 两变量均值差的标准误通过  $\text{Var}(\hat{\beta}_B - \hat{\beta}_C) = \text{Var}(\hat{\beta}_B) + \text{Var}(\hat{\beta}_C) - 2\text{Cov}(\hat{\beta}_B, \hat{\beta}_C)$  来计算进行推断
- 交互作用  $E[Y_i | X_{1i} = x_1, X_{2i} = x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$  中交互作用参数实际表示  $E[Y_i | X_{1i} = x_1 + 1, X_{2i} = x_2 + 1] - E[Y_i | X_{1i} = x_1, X_{2i} = x_2 + 1] - E[Y_i | X_{1i} = x_1 + 1, X_{2i} = x_2] - E[Y_i | X_{1i} = x_1, X_{2i} = x_2] = \beta_3$  各交互参数变化一单位响应变化
- 多元回归的参数解释需要考虑清楚变量类型与交互作用
- 多元回归中变量与响应变量与变量间的相关性要全盘考虑通过模拟观察决定

## 6.8 模型诊断与选择

- 通过残差诊断最小二乘决定均值为零方差通过  $\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-p}$  进行无偏估计
- 异常值判断对回归关系包括系数与其标准误的影响残差的分布检验等 `?influence.measures`

There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know. Donald Rumsfeld

- 随机化有助于平衡未知变量
- 杠杆点加入前后与回归线距离差的比值

- 参数方差膨胀共相关或随机相关 `vif` 来检验协变量在欠拟合下有偏
- 协变量的选择需要专业知识与经验

## 6.9 广义线性模型

- Nelder 与 Wedderburn 1972 年提出
- 响应是指数家族模型模型组成部分是线性的线性预测变量与响应通过连接函数联系
- 线性模型
  - $Y_i \sim N(\mu_i, \sigma^2)$
  - $\eta_i = \sum_{k=1}^p X_{ik} \beta_k$
  - $g(\mu) = \eta$
  - 似然模型为  $Y_i = \sum_{k=1}^p X_{ik} \beta_k + \epsilon_i$   $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$
- logistic 模型
  - $Y_i \sim Bernoulli(\mu_i)$
  - $\eta_i = \sum_{k=1}^p X_{ik} \beta_k$
  - $g(\mu) = \eta = \log\left(\frac{\mu}{1-\mu}\right)$   $g$  为 logit 函数
  - 似然函数为  $\prod_{i=1}^n \mu_i^{y_i} (1 - \mu_i)^{1-y_i} = \exp\left(\sum_{i=1}^n y_i \eta_i\right) \prod_{i=1}^n (1 + \eta_i)^{-1}$
- 泊松模型
  - $Y_i \sim Poisson(\mu_i)$
  - $\eta_i = \sum_{k=1}^p X_{ik} \beta_k$
  - $g(\mu) = \eta = \log(\mu)$
  - 似然函数为  $\prod_{i=1}^n (y_i!)^{-1} \mu_i^{y_i} e^{-\mu_i} \propto \exp\left(\sum_{i=1}^n y_i \eta_i - \sum_{i=1}^n \mu_i\right)$
- 似然函数与数据的联系  $\sum_{i=1}^n y_i \eta_i = \sum_{i=1}^n y_i \sum_{k=1}^p X_{ik} \beta_k = \sum_{k=1}^p \beta_k \sum_{i=1}^n X_{ik} y_i$  只有  $\sum_{i=1}^n X_{ik} y_i$
- 极大似然估计的解  $0 = \sum_{i=1}^n \frac{(Y_i - \mu_i)}{Var(Y_i)} W_i$   $W_i$  是连接函数的反函数的微分
- 响应的方差中线性模型  $Var(Y_i) = \sigma^2$  是常数 logistic 模型  $Var(Y_i) = \mu_i(1 - \mu_i)$  泊松模型  $Var(Y_i) = \mu_i$
- 可通过对模型方差增加调谐参数  $\phi$  使模型更灵活 quasi-likelihood
- 模型求解为  $\hat{\beta}_k$  及可能的  $\hat{\phi}$
- 线性预测变量关系  $\hat{\eta} = \sum_{k=1}^p X_k \hat{\beta}_k$
- 平均响应  $\hat{\mu} = g^{-1}(\hat{\eta})$
- 系数解释  $g(E[Y|X_k = x_k + 1, X_{\sim k} = x_{\sim k}]) - g(E[Y|X_k = x_k, X_{\sim k} = x_{\sim k}]) = \beta_k$

## 6.10 二元响应

- $\log\left(\frac{Pr(RW_i|RS_i, b_0, b_1)}{1-Pr(RW_i|RS_i, b_0, b_1)}\right) = b_0 + b_1 RS_i$
- $b_0$  预测变量为零时胜率对数
- $b_1$  预测变量变化一个单位胜率的改变对数
- $\exp(b_1)$  预测变量变化一个单位胜率的改变

## 6.11 计数或速率响应

- $\log(E[NH_i|JD_i, b_0, b_1]) = b_0 + b_1 JD_i$
- $e^{E[\log(Y)]}$   $Y$  的几何平均值
- $e^{\beta_0}$  第零天的几何平均值
- $e^{\beta_1}$  每天相对增加或减少的几何平均值
- 通过设置 `offset` 可用来估计增长率
- 注意方差膨胀与零膨胀问题

## 6.12 分段平滑

- 可用线性回归拟合曲线原理是分段拟合连接
- 断点平滑可用二次项
- 分段项可看作基进行组合

# 章 7

## 最优化

### 7.1 数学本质

当  $f_i(x) \leq b_i, (i = 1, \dots, m)$  时，最小化  $f_0(x)$ 。也就是满足限制条件下最小化某函数时其变量  $x$  的取值。

### 7.2 简史

- 1947, Dantzig 提出线性规划的 simplex 算法
- 1960s, 早期内点法
- 1970s, 椭球法与其他亚梯度方法
- 1980s, 线性规划的多项式时间内点法
- 1990 之前主要运筹学里用，后来用到工程里
- 现在，非线性凸优化的多项式时间内点法

### 7.3 凸集

- 仿射集  $x = \theta x_1 + (1 - \theta)x_2$ , 包含所有经过任意两个集合内点的线
- 凸集包含所有集合内任意两点间线段
- 凸组合  $x = \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_k x_k$ , 其中  $\theta_1 + \dots + \theta_k = 1, \theta_i \geq 0$

### 7.4 最小二乘法

最小化  $\|Ax - b\|_2^2$ , 其解析解为  $x^* = (A^T A)^{-1} A^T b$ , 该算法比较成熟, 计算时间正比于  $n^2 k (A \in R^{k \times n})$

### 7.5 线性规划

线性规划问题没有解析解, 求解算法比较成熟, 如果  $m \geq n$ , 求解时间正比于  $n^2 m$

## 7.6 凸优化

将问题转化为凸函数  $f_i(\alpha x + \beta y) \geq \alpha f_i(x) + \beta f_i(y)$ ，如果  $\alpha + \beta = 1, \alpha \geq 0, \beta \geq 0$ ，最小二乘法与线性规划是凸优化的特殊形式。

求解凸优化问题没有解析解，求解时间正比于  $\max\{n^3, n^2m, F\}$ ， $F$  是求函数一阶与二阶导数的时间，实际问题转化为凸优化问题不容易发现但确实可以求解。

# 章 8

## 统计模型

### 8.1 统计学习概论

- 统计学习：理解数据的工具集
- 监督学习：有因变量，根据自变量预测估计因变量
- 非监督学习：无因变量，探索自变量间的关系与结构

### 8.2 统计学习简史

- 19世纪初，Legendre 与 Gauss 发表了最小二乘法的论文，该方法首先应用在天文学领域
- 1936年，Fisher 提出线性判别分析来解决定性分析问题
- 1940s，logistic 回归提出
- 1970s，Nelder 与 Wedderburn 提出广义线性模型，将线性回归与 logistic 回归统一到一个体系
- 1980s，计算机技术进步，非线性问题开始得到解决
- Breiman, Friedman, Olshen 与 Stone 提出回归树与聚类，提供交叉检验方法
- 1986年，Hastie 与 Tibshirani 提出广义加性模型，将广义线性模型与一些非线性模型统一到一个体系
- 伴随软件，机器学习与其他理论的发展，统计学习作为统计学学科快速发展

### 8.3 统计学习定义

- $Y = f(X) + \epsilon$
- 统计学习本质上是在寻找最合适的  $f$  来进行预测与推断

### 8.4 预测

- $\hat{Y} = \hat{f}(X)$ ,  $\hat{f}(X)$  通常看作黑箱
- $\hat{Y}$  预测  $Y$  需要考虑两部分误差：可约误差与不可约误差
- 可约误差指  $\hat{f}$  推断  $f$  上的偏差
- 不可约误差指由  $\epsilon$  引入的误差
- 误差的期望  $E(Y - \hat{Y})^2 = [f(x) - \hat{f}(x)]^2 + Var(\epsilon)$  (证明用到  $E(Y)$ )

### 8.5 推断

- 关注  $X$  与  $Y$  的关系， $\hat{f}(X)$  通常有明确的形式

- 自变量因变量是否相关
- 如何相关
- 关系的数学描述

## 8.6 估计模型

- 使用训练集与验证集
- 参数方法与非参数方法
- 模型的欠拟合与过拟合
- 权衡模型的准确性（预测）与可解释性（推断）
- 模型的奥卡姆剃刀与黑箱

## 8.7 评价模型

### 8.7.1 拟合质量测量

- 训练集均方误  $MSE_{Tr} = Ave_{i \in Tr} [y_i \hat{f}(x_i)]^2$
- 测试集均方误  $MSE_{Te} = Ave_{i \in Te} [y_i \hat{f}(x_i)]^2$
- 测试集均方误源于训练集拟合模型的方差，误差项  $\epsilon$  的方差及模型误差的平方三部分

### 8.7.2 聚类评价

- 错误率  $Err_{Te} = Ave_{i \in Te} I[y_i \neq \hat{C}(x_i)]$
- 贝叶斯分类器：错误率最小的分类器，使  $x$  属于某个分类的概率最大
- k 临近值聚类：距离最小的 k 个为一类所产生的分类器
- 问题 -> 数据 -> 特征 -> 算法 -> 参数 -> 评价

The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data. John Tukey

- 数据质量优先于模型
- 不要自动特征选择
- 算法的可扩展性与计算性能要考虑
- 数据过拟合问题数据总是由信号与噪音组成但会被算法无差别对待
- 数据要与问题相关低相关度的组合可能产生高相关度

## 8.8 研究设计

- 定义错误率
- 将数据分割为训练集预测集验证集
- 在训练集上使用交叉检验选择特征与预测算法
- 在预测集或验证集上使用一次数据
- 预测效果起码要优于瞎猜
- 避免使用小样本
- 比例为 60% 训练集 20% 预测集 20% 验证集或 60% 训练集 40% 预测集或小样本交叉检验
- 注意数据结构时序分析要对数据分段采样

## 8.9 错误率

- 真阳性真的是对的 TP
- 假阳性真的是错的 FP Type I
- 真阴性假的是错的 TN
- 假阴性假的是对的 FN Type II
- 灵敏度  $TP/(TP+FP)$
- 特异性  $TN/(TN+FN)$
- 均方差  $MSE \frac{1}{n} \sum_{i=1}^n (Prediction_i - Truth_i)^2$
- 均方误  $RMSE \sqrt{\frac{1}{n} \sum_{i=1}^n (Prediction_i - Truth_i)^2}$
- 中位差 Median absolute deviation
- 准确性  $(TP+TN)/(TP+FP+TN+FP)$
- 一致性 kappa 值

## 8.10 ROC 曲线

- 分类问题寻找判别阈值满足一定 TP 下最小 FP 的模型
- FP v.s.TP 作图
- AUC 曲线下面积表示选择标准一般超过 80%
- 对角线是随机猜的结果

## 8.11 重采样技术

### 8.11.1 交叉检验

- 训练集上的操作
- 训练集上再分为训练集与测试集
- 在测试集上评价重复并平均化测试集错误
- 用来进行变量模型参数选择
- 随机分组留一
- 分组多方差大分组少有偏差
- 有放回的为 bootstrap 不建议用
- 核心思想：通过保留一部份训练集数据作为检验集来估计真实检验集的错误率与模型拟合效果
- 验证集方法：将训练集数据分为两部分，一部份拟合模型，一部份检验模型，这样得到的错误率为真实检验集的一个估计，选取错误率较低的模型建模
- 验证集方法缺点：错误率依赖于采样变动较大，训练集少，低估了错误率
- 留一法 (LOOCV)：每次建模留一个数据点作为验证集， $MSE_i = (y_i - \hat{y}_i)^2$  重复 n 次，得到一个 CV 值作为对错误率的估计： $CV_{(n)} = \frac{1}{n} \sum_{i=1}^n MSE_i$
- 留一法优点：使用数据量大，偏差小；结果唯一，不受随机化影响
- 留一法缺点：计算量大，公式插入杠杆统计量调节杠杆点对方程拟合的影响，得到  $CV_{(n)} = \frac{1}{n} \sum_{i=1}^n (\frac{y_i - \hat{y}_i}{1 - h_i})^2$
- k 叠交叉检验：将训练集分为 k 叠，每次建模用 (k-1) 叠，用 1 叠检验
- k 叠交叉检验优点：计算量小，结果与留一法相差不多- 交叉检验的结果用来寻找 CV 值最小的点来选择模型，通常与真实检验集最小点结果相差不大，但交叉检验给出的 MSE 会偏低

- 偏差方差权衡：使用的训练集数据越多，估计偏差越小，方差越大（相关性越高的方差越大）
- 分类问题使用错误率计算  $CV: CV(n) = \frac{1}{n} \sum_{i=1}^n Err_i$
- 少  $n$  多  $p$  问题上使用交叉检验，不可先进行全模型变量选择再交叉检验，应该对整个过程交叉检验

### 8.11.2 bootstrap

- 在训练集里有放回的重采样等长的数据形成新的数据集并计算相关参数，重复  $n$  次得到对参数的估计，计算标准误
- 生成 Bootstrap Percentile 置信区间
- 适用于独立样本，样本间有相关如时间序列数据可采用 block 法分组屏蔽掉进行 bootstrap
- 因为存在重复，使用 bootstrap 建立训练集与预测集会有非独立样本，造成检验集模型方差的低估，去掉重复使模型复杂，不如交叉检验对检验集误差估计的准确
- slipper 包

## 8.12 caret 包

- 数据清洗预处理
- 数据分割 `createDataPartition` 数据比例重采样产生时间片段
- 训练检验整合函数 `train predict`
- 模型对比
- 算法整合为选项线性判别回归朴素贝叶斯支持向量机分类与回归树随机森林 Boosting 等

## 8.13 数据分割

- `train <- createDataPartition(y=spam$type, p=0.75, list=FALSE)` 数据三一分得到 index
- `folds <- createFolds(y=spam$type, k=10, list=TRUE, returnTrain=TRUE)` 数据分 10 份返回每一份列表
- `folds <- createResample(y=spam$type, times=10, list=TRUE)` 数据 bootstrap 重采样返回每一份列表
- `folds <- createTimeSlices(y=tme, initialWindow=20, horizon=10)` 时序数据重采样产生 20 为窗口时序片段的训练集与预测集

## 8.14 训练选项

- `args(train.default)` 通过 `method` 控制算法 `metric` 控制算法评价 `trainControl` 控制训练方法
- `trainControl` 中 `method` 选择模型选择方法如 bootstrap 交叉检验留一法 `number` 控制次数 `repeats` 控制重采样次数 `seed` 控制可重复性总体设置一个具体每一次用列表设置控制具体过程特别是并行模型

## 8.15 预测变量作图

- `featurePlot`
- `ggplot2`

## 8.16 数据预处理

- `train` 中的 `preProcess=c("center", "scale")` 标准化

- `spatialSign` 该转化可提高计算效率有偏
- `preProcess(training[,-58],method=c("BoxCox"))` 正态化转化
- `method="knnImpute"` 用最小邻近法填补缺失值
- `nearZeroVar` 去除零方差变量
- `findCorrelation` 去除相关变量
- `findLinearCombos` 去除线性组合变量
- `classDist` 测定分类变量的距离生成新变量
- 测试集也要预处理

## 8.17 协变量生成

- 原始数据提取特征
- 提取特征后生成新变量
- 因子变量要转为虚拟变量
- 样条基变量 `splines` 包中的 `bs`
- 数据压缩 `preProcess` 中 `method` 设置为 `pca` `pcaComp` 指定主成分个数

## 8.18 线性回归 & 多元线性回归

- $ED_i = b_0 + b_1 WT_i + e_i$  基本模型
- 参见前面回归部分

### 8.18.1 简单线性回归

- $Y \approx \beta_0 + \beta_1 X$
- 用最小二乘法估计  $\beta_0$  与  $\beta_1$  得到估计值  $\hat{\beta}_0$  与  $\hat{\beta}_1$ , 代入  $X$ , 得到模型估计值  $\hat{Y}$
- 残差平方和:  $RSS = e_1^2 + e_2^2 + \dots + e_n^2$ , 使 RSS 最小, 求导可得参数
- 回归线不等于最小二乘线, 最小二乘线是通过采样对回归线的估计
- 估计会存在偏差, 均值的偏差用标准误来描述  $Var(\hat{\mu}) = SE(\mu)^2 = \frac{\sigma^2}{n}$
- 回归参数的估计也涉及标准误的计算  $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$
- $\sigma^2$  可用残差标准误  $RSE(RSE = RSS/(n-2))$  来估计  $\hat{\sigma}^2 = \frac{n-p}{n} s^2$
- 据此可得回归参数的 95% 置信区间  $\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$
- 参数的评价可通过假设检验进行, 零假设为  $\beta_1$  为 0, 也就是自变量对因变量无影响, 构建 t 统计量  $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$ , 然后可根据 p 值判断参数的显著性
- 评价参数后需要评价模型, 主要通过  $RSE$  与  $R^2$  来进行
- $R^2$  表示模型所解释总体方差的比例, 与  $RSE$  不同, 独立于 Y,  $R^2 = \frac{TSS - RSS}{TSS}$
- $R^2$  与两变量间的相关系数是一致的, 但  $R^2$  统计量的应用面要广于相关系数
- 相关系数也可进行假设检验进而判断相关的显著性

### 8.18.2 多元线性回归

- 通过统计量 F 检验确定回归是否显著, 零假设为所有自变量系数为 0  $F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$
- 变量选择: 向前选择 (从 0 个到 p 个, 显著则包含), 向后选择 (从 p 个到 0 个, 不显著则剔除), 混合选择 (通过 p 的阈值调节)
- 因为 RSS 会减少,  $R^2$  会伴随自变量数目的增加而增加
- $RSE$  在多元线性回归中为  $RSE = RSS/(np - 1)$ , 伴随自变量个数增加影响超过 RSS 减少的影响,  $RSE$  会增大
- 自变量间的影响会导致相比单一变量预测更容易出现不显著, 这说明自变量间有可能可相互解释
- 预测的置信区间与预测区间, 前者指模型的变动范围, 后者指某个预测值的变动范围, 考虑真值本身的变动, 后者大于前者

- 因子变量通过对每个水平添加系数 0, 1 来回归, 也可根据需要赋值

### 8.18.3 线性模型延拓

- 线性模型基本假设: 可加性与线性
- 去掉可加性: 考虑交互作用
- 层级原理: 交互作用项显著而主作用不显著时不可去掉主作用项
- 去掉线性: 多项式回归

### 8.18.4 常见问题

- 关系非线性: 残差图判断
- 误差项共相关: 误差项的相关会导致标准误估计偏低, 低估参数的区间使不显著差异变得显著, 考虑时间序列数据, 观察误差项轨迹判断
- 误差项方差非常数: 喇叭状残差图, 通过对因变量进行对数或开方来收敛方差, 或者用加权最小二乘
- 异常值: 通过标准化残差图判断
- 杠杆点: 加入后会影响模型拟合, 通过杠杆统计量判断:  $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^n (x'_{i'} - \bar{x})^2}$  多元回归中该统计量均值为  $(p+1)/n$ , 超过很多则可能为杠杆点
- 在标准残差-杠杆值图中, 右上或右下方为危险值, 左方数值对回归影响不大
- 共线性: 共线性的变量相互可替代, 取值范围扩大, 标准误加大, 对因变量影响相互抵消, 降低参数假设检验的功效
- 多重共线性: 引入方差膨胀因子, 自变量引入全模型与单一模型方差的比值, 超过 5 或 10 说明存在共相关,  $VIF(\hat{\beta}_j) = \frac{1}{1-R^2_{X_j|X_{-j}}}$
- 解决共线性: 丢弃变量或合并变量
- 共线性不同于交互作用

### 8.18.5 线性回归与 kmeans 算法比较

- k 临近算法:  $\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$  核心是选择 k
- KNN 算法在解决非线性问题上有优势, 但一样的面对高维诅咒
- 线性回归可给出可解释的形式与简单的描述

### 8.18.6 logistic 回归

- 因变量以概率形式出现
- $p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$
- 变形后  $\frac{p(X)}{1-p(X)}$  为胜率, 比概率应用更实际些, 去对数后为对数胜率 (logit)
- 因变量  $p(X)$  与自变量间关系非线性
- 用极大似然估计确定参数, 似然函数为  $l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$ , 该函数取最大值
- 线性回归中, 最小二乘法为极大似然估计的特例
- 混杂因素的解释上要考虑单因素回归与多元回归
- 多响应 logistic 回归一般被判别分析取代

### 8.18.7 线性判别分析

- 使用原因: 分类离散时 logistic 回归不稳定, n 小 X 正态时更稳定, 适用于多响应
- 贝页斯理论:  $Pr(Y = k | X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}$  其中  $\pi$  代表先验概率, 估计  $f_k(x)$  需要对  $x$  的分布作出假设

- 自变量为 1 时，假定  $f_k(x)$  分布为正态的，有  $f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2)$ ，代入可得  $p_k(x)$ ，取对数有  $\sigma_k(x) = x \cdot \frac{\mu_k}{\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2} + \log(\pi_k)$ ，使  $\sigma_k(x)$  最大的分类方法为判定边界
- 贝页斯分类器需要知道所有分布参数，实际中会采用线性判别分析 (LDA)，通过以下训练集估计方法来插入贝页斯分类器： $\hat{\pi}_k = n_k/n$ 、 $\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$  与  $\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^K \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$
- 线性体现在判别函数  $\hat{\sigma}_k(x)$  的形式是线性的
- 自变量多于 1 时，假设自变量均来自多元正态分布的分类
- 列连表，表示假阳性，假阴性，可计算灵敏度与特异性
- LDA 是对贝页斯分类的模拟，旨在降低总错误率，因此灵敏度与特异性区分并不明显，可根据实际需要调节
- ROC 曲线用来展示两种错误，横坐标假阳性，纵坐标真阳性

### 8.18.8 二次判别分析 (QDA) 及其它

- 不同于 LDA，二次判别分析考虑各分类参数中方差不同而不是相同，引入了二次项
- 对分类描述更为精细，但容易过拟合，样本较少，LDA 优先
- 对比 logistic 回归，两者数学形式相近，取值上 logistic 回归使用极大似然法，LDA 使用共方差的高斯分布假设，结论多数条件一致，但随假设不同而不同
- KNN 更适用于非线性关系，标准化很有必要，QDA 相对温和

### 8.18.9 线性模型选择与正则化

- 最小二乘法 (OLS) 容易解释，预测性能好，但不万能
- 预测准确性上，当  $p>n$  时，模型方差变大
- 模型解释上， $p$  过多需要去除，进行模型选择

#### 8.18.9.1 子集选择

- 从  $p$  个自变量中选出与模型响应相关的进行建模
- 使用 deviance，最大化为最优子集
- 最佳子集选择： $p$  个自变量  $\binom{p}{k}$ ，计算  $RSS$  与  $R^2$ ， $RSS$  要小， $R^2$  要大，选择最佳的
- 步进法： $p$  值过大，计算负担重，采用逐步改进法进行模型选择
- 向前步进选择：从 0 个自变量开始加，第  $k$  个自变量选择  $p-k$  个模型，如果  $RSS$  与  $R^2$  表现好就保留，递近选择变量，不保证选择最佳模型， $p$  值较大优先考虑
- 向后步进选择：从  $p$  个自变量开始减，如果第  $k$  个自变量在模型  $RSS$  与  $R^2$  中没表现，就剔除进行变量选择，不保证选择最佳模型，适用于  $p$  值较小的情况（较大可能无法拟合）
- 步进选择构建  $1+p(p+1)/2$  个模型，最佳子集法需要构建  $2^p$  个模型
- 混合模型：向前选择，之后向后验证，剔除不再提高效果的模型

#### 8.18.9.2 测试集误差估计

- $RSS$  与  $R^2$  评价的是训练集拟合状况，不适用于估计测试集误差
- 估计测试集误差可以构建统计量调节训练集误差或直接通过验证集来估计
- Mallow's  $C_p$ :  $C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$   $d$  代表使用的变量数， $\hat{\sigma}^2$  是对模型方差的估计
- AIC:  $AIC = -2\log L + 2 \cdot d$  极大似然估计，线性模型下  $C_p$  与 AIC 实质等同
- BIC:  $BIC = \frac{1}{n}(RSS + \log(n)d\hat{\sigma}^2)$   $n$  是样本数，大于 7 时 BIC 会比  $C_p$  选择更轻量的模型
- 调节  $R^2$ :  $Adjusted R^2 = 1 - \frac{RSS/n-d-1}{TSS/(n-1)}$  值越大，测试集误差越小
- 不同于  $C_p$ , AIC, BIC 有严格的统计学意义，调节  $R^2$  虽然直观，但理论基础相对薄弱，单纯考虑了对无关变量的惩罚
- 验证与交叉验证：直接估计测试集误差而不用估计模型方差
- 单标准误原则：先计算不同规模测试集  $MSE$  的标准差，选择曲线中最小测试集误差一个标准误内最简单的模型

### 8.18.9.3 收缩

- 对系数估计进行收缩，接近 0 或等于 0 进行变量选择

#### 8.18.9.3.1 岭回归

- 不同于最小二乘估计对  $RSS$  的最小化，岭回归最小化  $RSS + \lambda \sum_{j=1}^p \beta_j^2$ ，其中  $\lambda$  为调谐参数，后面一项为收缩惩罚，是个  $l_2$  范数，使参数估计逼近 0，选择合适  $\lambda$  很重要，可用交叉检验来实现
- 因为范数大小影响模型惩罚项，所以进行岭回归前要做标准化处理

$$\bar{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}$$

- 岭回归的参数  $\lambda$  与范数收缩状况可看作最小  $MSE$  的函数来表现偏差-误差均衡
- 岭回归适用于最小二乘回归产生方差较大的情况，同时，计算负担较小，只伴随  $\lambda$  取值范围变化而变化

#### 8.18.9.3.2 Lasso

- 形式与岭回归一致，最小化  $RSS + \lambda \sum_{j=1}^p |\beta_j|$ ，使用  $l_1$  范数
- 岭回归参数同步收缩接近 0，Lasso 可以通过软边界直接收缩到 0 实现变量选择，产生稀疏模型，想像超球体与超多面体与超球面的接触
- 贝页斯视角下，岭回归与 lasso 关于线性模型系数的先验分布是不同的：前者为高斯分布，接近 0 时平坦，后验概率等同最优解；后者为拉普拉斯分布，接近 0 时尖锐，先验概率系数接近 0，后验概率不一定为稀疏向量
- 岭回归与 Lasso 分别适用于真实模型自变量多或少的情况，并不广谱，考虑交叉检验来进行选择
- 交叉检验也可用来选择  $\lambda$ ，通过选择的自变量参与建模

#### 8.18.9.4 降维

- 前提是自变量间不独立，将  $p$  个自变量向量投影到  $M$  维空间 ( $M < p$ )，使用投影  $M$  拟合线性回归模型  $\sum_{m=1}^M \theta_m z_{im} = \sum_{m=1}^M \theta_m \sum_{j=1}^p \phi_{jm} x_{ij} = \sum_{j=1}^p \sum_{m=1}^M \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^p \beta_j x_{ij}$
- 主成分：各自变量在主成分方向上方差最大
- 主成分回归 (PCA)：实际为无监督算法，得到主成分后作为新变量进行最小二乘回归，认为因变量与自变量变异最大的方向一致，需要仔细检验这个假设，主成分个数的选择影响模型效果
- 岭回归疑似为主成分回归的连续版，两者都需要标准化，效果也相近
- 偏最小二乘 (PLA)：第一个投影方向为因变量与自变量回归方向，后续投影是对残差投影方向的回归，重复得到监督学习的效果
- PLA 通常并不比 PCA 更好，引入了监督算法提高了偏差

#### 8.18.9.5 高维数据

- $n$  远远少于  $p$  或接近的数据
- 最小二乘估计在  $n$  小于  $p$  时残差为 0，太过精细
- $C_p$ , AIC, BIC 方法因为有参数  $\hat{\sigma}^2$  需要估计，而这个参数会在高维数据下变成 0，调节  $R^2$  也会变成 1
- 高维诅咒：正则化或收缩对高维方法产生影响，合适调谐参数十分重要，测试集误差必然增长
- 引入新变量会对预测产生不可知影响，选出的自变量并非不可替代，结果用独立验证集误差或交叉检验误差描述

## 8.19 非线性

### 8.19.1 多项式回归

- 模型基本形式为单一自变量在不同幂指数下的多项式，最小二乘拟合

- 模型在特定点的方差受系数方差与协方差影响，幂越高，模型越精细，方差越大
- 幂次一般不超过 3 或 4
- 可进行 logistic 回归

### 8.19.2 阶梯函数

- 阶梯函数将自变量由连续变成有序分类变量
- 函数形式为引入指标函数  $C_K(x)$  进行自变量分段，然后进行最小二乘拟合
- 依赖找间隔点
- 可进行 logistic 回归

### 8.19.3 基函数

- 固定线性系数  $\beta$ ，自变量的形式由  $b(x)$  决定， $b(x)$  为基函数
- 多项式回归与阶梯函数均为基函数的特例

### 8.19.4 回归样条

- 设定分段点，分段点前后进行多项式回归
- $K$  个点分割  $(K+1)$  段，存在  $(K+1)$  个多项式回归，自由度过高
- 进行边界约束，对  $n$  次方程而言，约束分段点 0 阶，1 阶，2 阶导数连续，减少 3 个自由度，共有  $K$  个点，则有  $(n+1-3)k+n+1$  个自由度，相比无约束的  $(n+1)k$ ，自由度减少，更稳健
- 一般而言约束限制为 (自由度-1) 阶连续，这样自由度比分界点略多些，够用
- 分段样条最好在两端加入线性限制，收敛自由度，这样在边界稳健，为自然样条
- 分段点位置一般均匀分布，个数（本质上是自由度）通过交叉检验来确定
- 分段多项式回归限定了自由度，因此结果一般比多项式回归更稳定

### 8.19.5 平滑样条

- 如果以 RSS 衡量不加入限制，很容易产生过拟合，因此考虑加入平滑项
- 最小化

$$\sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

其中， $g(x)$  为平滑样条，由损失函数与惩罚项组成，二次导数表示在  $t$  处的平坦度，越平坦，惩罚越小，越崎岖，惩罚越大，因而平滑

- 对三次函数而言，平滑样条会将函数两端收敛的跟自然样条一样，实际上，平滑样条是自然样条的收缩版
- 参数  $\lambda$  也影响平滑效果，越大越平滑，因为  $k$  固定，只涉及  $\lambda$  的选择
- 参数  $\lambda$  的选择基于有效自由度，可以用留一法进行估计，形式与杠杆点统计量差不多，可以很方便的进行数值求解
- 平滑样条的自由度比多项式要小，更稳健

### 8.19.6 本地回归

- 首先分段，然后分段内进行加权回归，离某点越近，权重越高，进行最小二乘拟合，得到每个点的函数，联合模型拟合
- 自变量较多，可考虑本地有选择的选取自变量进行本地回归
- 同样遭受高维诅咒带来的临近值少或稀疏问题

### 8.19.7 广义加性模型

- 

$$y_i = \beta_0 + \sum_{i=1}^n f_j(x_{ij}) + \epsilon$$

每个自变量都有自己的函数形式，加合求解

- 每个自变量影响都可以展示
- 可分段，也可使用平滑，平滑方法中使用了反馈拟合策略对不易用最小二乘拟合求解的问题进行求解，效果差不多，分段不必要
- 可用于分类回归问题，解释性好
- 优点：非线性，更准确，易解释，可进行统计推断，可用自由度衡量平滑性
- 缺点：不易考虑交互影响

## 8.20 树

### 8.20.1 回归树

- 将因变量按自变量分区间，每个区间内预测值一致，直观易解释
- $\sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$
- 计算困难，使用自上而下的贪心算法
- 递归二元分割：构建树过程每个节点都选最佳分割点，也就是分割后残差最小的变量与数值
- 算法在叶样本数为 5 时结束
- 树修剪，选择训练集误差最小的子树，引入调谐因子  $\alpha$
- 最小化  $\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T|$  类似 lasso 算法
- 确定  $\alpha$  要用交叉检验，之后选出特定模型

### 8.20.2 分类树

- 因变量为分类变量，RSS 用分类错误率代替
- $E = 1 - \max_k(\hat{p}_{mk})$  但分类错误率对树生长并不敏感，应采用其他指标
- Gini 系数： $G = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk})$ ，分类越准，值越小，衡量端纯度
- cross-entropy： $D = -\sum_{k=1}^K \hat{p}_{mk} \log \hat{p}_{mk}$ ，与 Gini 系数相似，描述一致
- 如果以修剪树为目标，指标应选择分类错误率
- 节点产生相同预测说明预测纯度不同，可靠性不同
- 与线性模型相比，适用数据种类不同，借助可视化判断
- 优点：容易解释，适用于决策，容易出图，处理分类问题简单
- 缺点：预测准确率低于其他常见回归与分类方法

### 8.20.3 Bagging

- 决策树方法相比线性回归模型方差很大
- 引入 Bootstrap，通过平均构建低方差模型
- $\hat{f}_{avg}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{*b}(x)$
- 不修剪，通过平均降低方差
- 对于分类变量，通过投票，少数服从多数得到答案
- 误差估计通过包外样本（OOB）进行交叉检验并进行树的选择，降低计算成本
- 变量权重在 Bagging 中不易衡量，可通过衡量每棵树的 RSS 或者 Gini 系数在进行一次变量分割后 RSS 下降程度并进行排序取得

- 该方法可应用于其他统计模型

- 重采样重新计算预测值

- 平均或投票给出结果

- 减少方差偏差类似适用于非线性过程

- bagged trees

- 重采样

- 重建树

- 结果重评价

- 更稳健效果不如 RF

- Bagged loess 可用来处理细节

#### 8.20.4 随机森林

- bagging 中使用所有的变量进行选择，但是会更易出现共相关变量，方差降低不多

- 随机森林的核心在于强制使用较少的自变量，为其他自变量提供预测空间进而提高模型表现

- 变量数一般选择为  $\sqrt{p}$

- 表现会比 bagging 好一些

- bootstrap 采样

- 每一个节点 bootstrap 选取变量

- 多棵树投票

- 准确度高速度慢不好解释容易过拟合

#### 8.20.5 Boosting

- 通用的统计学习方法

- 树生长基于先前的树，不使用 bootstrap，使用修改过的原始数据

- 先生成有  $d$  个节点的树，之后通过加入收缩的新树来拟合残差，收缩因子为  $\lambda$ ，呈现层级模式，最后模型为  $\hat{f}(x) = \sum_{b=1}^B \lambda \hat{f}^b(x)$

- boosting 学习缓慢，一般学习较慢的学习效果更好

- 三个参数：树个数  $B$ （交叉检验），收缩因子  $\lambda$ （控制学习速率），树节点数（一般为 1，为交互作用深度，控制涉及变量）

- 深度  $d$  为 1 时是加性模型

- 随机森林与 Boosting 产生的模型都不好解释

- 迭代分割变量

- 在最大化预测时分割

- 评估分支的同质性

- 多个树的预测更好

- 优点容易解释应用可用在神经网络上

- 缺点不容易交叉验证不确定性不宜估计结果可能变化 - 算法

- 先在一个组里用所有的变量计算

- 寻找最容易分离结果的变量

- 把数据按照该变量节点分为两组

- 在每一个组中寻找最好的分离变量

- 迭代直到过程结束
- 节点纯度用 Gini 系数或交叉熵来衡量
- rattle 包的 fancyRpartPlot 出图漂亮
- 可用来处理非线性模型与变量选择
- 弱预测变量加权后构建强预测变量
- 从一组预测变量开始
- 添加有惩罚项的预测变量来训练模型
- 以降低训练集误差为目的
- 通用方法

## 8.21 支持向量机

### 8.21.1 最大边界分类器

#### 8.21.1.1 超平面

- $p$  维空间里  $(p-1)$  维子空间
- $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p = 0$  定义一个  $p$  维超平面,  $X$  落在超平面上
- $p$  维空间中点  $X$  不在超平面上就在其两侧

#### 8.21.1.2 超平面分类

- $n \times p$  矩阵  $X$  分为两类  $-1$  或  $1$
- 代入超平面大于  $0$  为  $1$ , 小于  $0$  为  $-1$ , 有  $Y * \beta * X > 0$  表示分类正确
- 构建训练函数  $f(x^*) = \beta_0 + \beta_1 X_1^* + \beta_2 X_2^* + \dots + \beta_p X_p^*$  正数表示为  $1$ , 负数为  $-1$ , 距离  $0$  越远表示距离超平面越远, 越近表示分类越不确定, 判定边界为线性

#### 8.21.1.3 最大边界分类器

- 最大边界超平面: 距离边界最近的距离的所有超平面中距离边界点最远的那个超平面
- 分类良好但容易在  $p$  大时过拟合
- 形成最大边界分类器所需要的边界点为支持向量, 用以支持最大边界超平面
- $f(x^*) * y_i$  在系数平方和为  $1$  时为点到平面的垂直距离, 最小化后最大化这个距离是求最大边界超平面的关键

## 8.21.2 支持向量分类器

- 有些情况不存在超平面, 需要求一个软边界来适配最多的分类, 这就是支持向量分类器
- 因为是软边界所以允许在超平面或边界一边出现误判
- 计算上还是为最小化最大化距离, 但分类上距离要乘以  $1 - \epsilon_i$  项, 也就是松弛变量
- 松弛变量大于  $0$  表示边界误判, 大于  $1$  表示超平面误判, 总和为  $C$ , 表示边界的容忍度, 越大分类越模糊
- $C$  可通过交叉检验获得, 控制 bias-variance 权衡
- 只有边界内观察点影响超平面的选择, 这些点为支持向量, 是形成模型的关键
- 与 LDA 不同, 使用部分数据, 与 logistic 回归类似

### 8.21.3 支持向量机原理

- 非线性条件下可以考虑将超平面理解为非线性超平面，提高样本维度换取分类效果
- 加入多项式等非线性描述后计算量不可控
- 支持向量机通过核来控制非线性边界
- 通过样本内积来解决支持向量分类问题
- 线性支持向量分类器  $f(x) = \beta_0 + \sum_{i=1}^n \alpha_i \langle x, x_i \rangle$  只有支持向量在解中非 0，现在只需要支持向量的内积就可以求解
- 内积可以推广为核函数，核函数可以采用非线性模式
- $f(x) = \beta_0 + \sum_{i=1}^n \alpha_i K(x, x_i)$  径向基核函数较为常用
- 使用内积的核函数计算上简单且等价与高维空间超平面分类

#### 8.21.3.1 多于二分类

- 一对一分类：对比  $\binom{K}{2}$  个分类器在检验集中的效果，通过计数来选择分类结果
- 一对多分类：对比 K 个与剩下的 K-1 个分类，分类结果最远的认为属于那个分类

### 8.21.4 svm 与 logistic 回归关系

- 中枢损失，对关键点敏感
- 传统方法也可以借鉴核函数观点视同
- 支持向量无法提供参数概率信息，采用核函数的 logistic 回归可以，计算量大
- 分类距离较远，支持向量机会比 logistic 回归好一点
- 支持向量机是计算机背景，logistic 回归是概率背景

## 8.22 无监督学习

### 8.22.1 主成分分析

- 用较少的变量代表较多的变量，方便可视化与理解数据
- 第一主成分  $Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p$  方差最大，正则化后有  $\sum_{j=1}^p \phi_{j1}^2 = 1$ ，则  $\phi$  为变量在第一主成分上的载荷
- 求解上第一主成分最大化  $\frac{1}{n} \sum_{i=1}^n z_{i1}^2$  求解载荷值， $z_{ni}$  是第一个样本在第一个主成分上的得分
- 载荷表示变量重要程度，得分表示样本重要程度
- 第二主成分与第一主成分正交求解
- biplot 同时表示载荷与得分，载荷向量接近表示有相关性，方向不一表示相关性弱，变量在主成分得分差异表示其状态
- 第一个主成分表示在 p 维空间里距离 n 个观察最近的超平面，因此具备代表性
- 取 M 个主成分可代表所有数据  $x_{ij} \approx \sum_{m=1}^M z_{im} \phi_{jm}$
- 变量单位要统一，已经统一就不要标准化了
- 主成分是唯一的，符号可能有变化，载荷与得分值也唯一
- 主成分的重要性通过方差解释比例 (PVE) 来衡量，用碎石图来可视化

$$\frac{\sum_{i=1}^n (\sum_{j=1}^p \phi_{jm} x_{ij})^2}{\sum_{j=1}^p x_{ij}^2}$$

- 寻找碎石图的肘部来确定选取主成分的个数，方法不固定
- 可用来进行 M 小于 p 的主成分回归
- SVD 算法

### 8.22.2 聚类方法

- 寻找子分类或簇的方法，从异质性寻找同质性

### 8.22.2.1 k 均值聚类

- 子类中方差小，子类间方差大
- 事先指定子类个数
- 最小化所有 K 个平均欧式距离  $W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2$
- 先对所有样本随机分类，然后每种分类取中心，选取里中心距离最近的点重新分类，重新计算中心，迭代得到聚类结果

### 8.22.2.2 分层聚类

- 不需要指定先前聚类数，形成冰柱图
- 冰柱图要垂直分层解释，水平解释容易出现误导- 修剪冰柱图可给出聚类数
- 计算所有样本间距离，越相近就融合为一类，重新计算距离，反复这一过程
- 计算两者间相似度是很关键的，不同场景应用不同算法
- 变量的标准化处理上也很重要，考虑实际场景

## 8.23 人工神经网络

- RNN 神经网络算法
- LSTM 神经网络算法
- tensorflow keras 与深度学习
- 通过正交变量监督黑箱模型的敏感度

## 8.24 模型联合

- 通过平均与投票结合模型
- 联合分类器提高准确率
- caretEnsemble 包
- 案例广义加性模型

```
library(ISLR); data(Wage); library(ggplot2); library(caret);
Wage <- subset(Wage, select=-c(logwage))
# Create a building data set and validation set
inBuild <- createDataPartition(y=Wage$wage, p=0.7, list=FALSE)
validation <- Wage[-inBuild,]; buildData <- Wage[inBuild,]
inTrain <- createDataPartition(y=buildData$wage, p=0.7, list=FALSE)
training <- buildData[inTrain,]; testing <- buildData[-inTrain,]
mod1 <- train(wage ~ ., method="glm", data=training)
mod2 <- train(wage ~ ., method="rf", data=training, trControl = trainControl(method="cv"), number=3)
pred1 <- predict(mod1, testing); pred2 <- predict(mod2, testing)
qplot(pred1, pred2, colour=wage, data=testing)
predDF <- data.frame(pred1, pred2, wage=testing$wage)
combModFit <- train(wage ~ ., method="gam", data=predDF)
combPred <- predict(combModFit, predDF)
sqrt(sum((pred1-testing$wage)^2))
sqrt(sum((pred2-testing$wage)^2))
sqrt(sum((combPred-testing$wage)^2))
```

## 8.25 无监督预测

- 先聚类后预测

- `clue` 包 `cl_predict` 函数
- 推荐系统

## 8.26 模型预测

- 时序数据包含趋势季节变化循环
  - 效应分解 `decompose`
  - `window` 窗口
  - `ma` 平滑
  - `ets` 指数平滑
  - `forecast` 预测
- 空间数据同样有这种问题临近依赖地域效应
- `quantmod` 包或 `quandl` 包处理金融数据
- 外推要谨慎

## 8.27 模型可视化

- 统计模型可视化



# 章 9

## 开发数据产品

### 9.1 shiny

- 源自 R-studio
- 动态网络应用
- 入门版 OpenCPU
- 高级版 Manipulate
- `install.packages("shiny");library(shiny)`
- `ui.R` 控制外观 `sever.R` 控制计算
- `runApp()` 启动应用
- `sever.R` 中 `shinyServer` 之前的代码只在启动应用时执行一次适合读入数据
- `shinyServer(function(input, output){` 之内的非互动函数只被每个用户执行一次
- `Render*` 为互动函数数值改变就执行一次
- `runApp(display.mode='showcase')` 可用来同时高亮显示执行代码
- `reactive` 用来加速互动函数外的信息交换
- `actionButton` 用来一次提交输入数据 `if (input$goButton == 1){ Conditional statements }` 用  
来定义条件语句
- `cat browser()` 调试
- `fluidRow` 产生表格
- `shinydashboard`
- `flexdashboard`
- `docker image`
- `prettydoc`

### 9.2 rCharts

- 主页
- 动态交互可视化工具
- `require(devtools);install_github('rCharts', 'ramnathv')`

### 9.3 GoogleVis

- 主页
- R 代码产生图表生成 html
- `install.packages('googleVis');library(googleVis)`

- 教程

## 9.4 Slidify

- 主页
- html5 幻灯片
- ```
install.packages("devtools");library(devtools);install_github('slidify', 'ramnathv');install_github('ramnathv');
```
- `author("yufree")`
- YAML 配置幻灯片结构
- ## 幻灯片开始 --- 加空行表结束 .class #id 自定义 css 文件 id
- `slidify("index.Rmd")` 生成 `browseURL("index.html")` 观看
- `publish_github(user, repo)` github 发布

## 9.5 yhat

- 主页
- 本地提交算法或模型生成可调用 API 支持 R 与 python

## 9.6 swagger

- 主页
- 生产 API

## 9.7 案例

- 算法先发现女儿怀孕
- netflix 为什么不用获奖算法
- 伪装品酒师
- facebook 算法压抑多样性
- 利用点评数据预测经济发展可能比政府数据更新更及时，也更准确，刚搜了一下发现淘宝就有卖这类数据的...
- #biorxiv 京都大学的一篇预印本论文基于深度神经网络与功能性核磁共振技术重构了视觉图像，看来读心术跟梦境重现术用不了多久就能见到产品了
- #qz 茶是一种很特殊的跨国交易品，Nikhil Sonnad 发现丝绸之路上国家对茶的发音接近 cha，而地理大发现后沿岸（大概可理解为海上丝绸之路）上国家对茶的发音接近 tea（闽南语的茶），这种利用特产发音研究贸易史的方法很有启发性
- 飓风玛丽亚的官方死亡人数是 64，但纽约时报对比了往年数据认为应该是 1052，这个往年对比对方法对灾害评价更有意义，可以发现一些非灾害直接导致但潜在相关的死亡现象
- 某数据科学家收集并可视化了 17 年的买菜收据研究其购买行为的潜在模式，很神奇地发现他每次买东西的顺序都是相似的，然而后来被证明是电脑根据货物分类的默认排序
- 写一个 markdown 编译器
- facebook 与普林斯顿的互掐

- 船跟教堂都曾是很直观的测量单位
- twitter 上喝酒的性别分析
- 美国各地对饮料的提法，南部人说 coke，东北与西海岸说 soda，其余地方说 pop
- 百度指数抓取
- twitter 上数据现实真相传播速度比谣言要慢，机器传播真假速率一致，人的传播起主要作用
- # 科学美国人雪花的生长模拟与分类，脑洞很大，这是我头一次看到有人把 t-sne 跟 rnm 算法用到科普里
- 地震记录可视化
- 国际象棋生存概率
- 各国人民想买什么
- 如何修厕所跟如何用筷子具有搜索相关性很有画面感
- 09-17 年推特的语义幸福度走势图，这两天因枪击达到了历史最低点，整体呈现出了 6-7 年的周期性，目前处于下行大趋势
- 谷歌新闻实验室出品的可视化图形、书籍及工具的流行趋势，可作为入门可视化的“光环效应”指南，没想到甘特图排那么靠前，另外漏掉了今年的新秀 joyplot
- 新西兰的新生儿名字在最近 50 年里呈现了长尾化，父母在给孩子起名字时有了更强的多样性，要知道在 1850 年英国新生儿里有 13.7% 的 William 和 14.6% 的 Mary，突然想明白了为啥有个历史悠久的威廉玛丽学院了...
- 图解机器学习系列
- 训练一个种族歧视的 AI



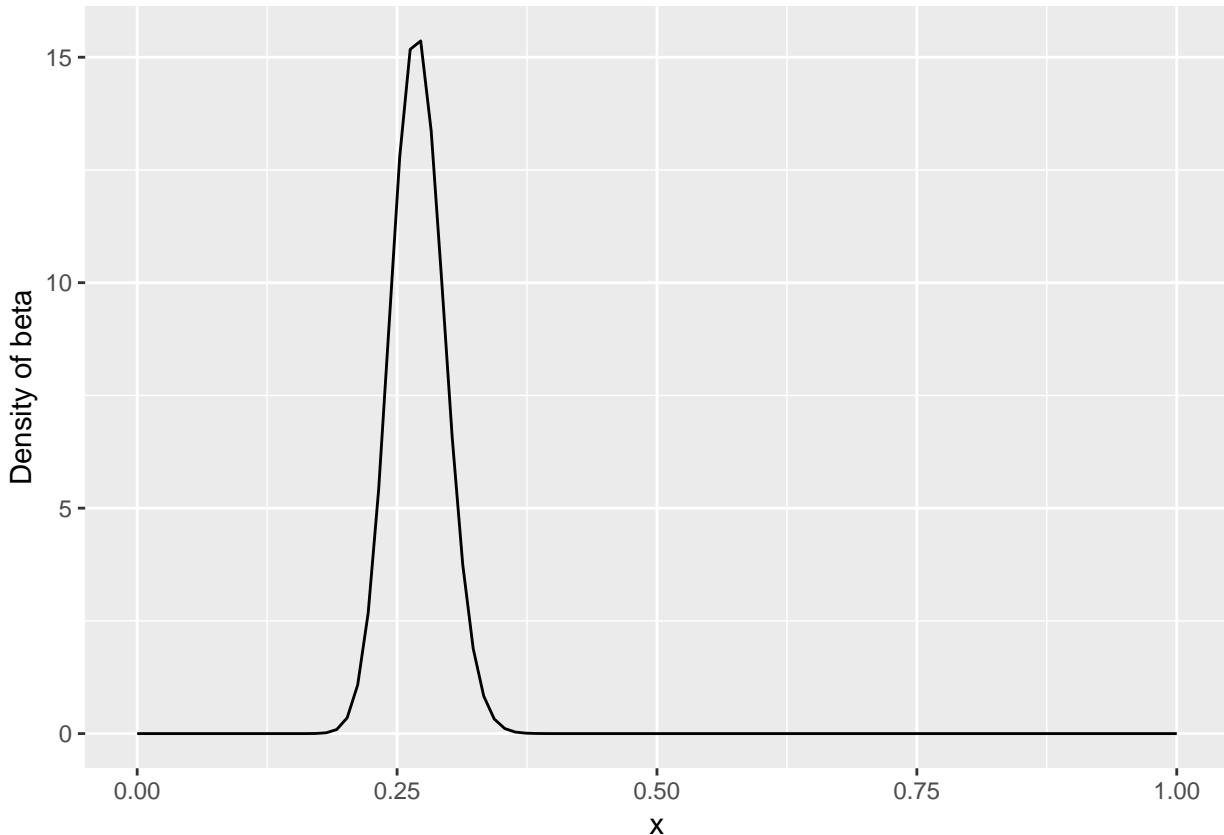
# 章 10

## 贝叶斯统计

### 10.1 贝塔分布

- 贝塔分布的本质是概率分布的分布
- 棒球击球率的预测问题，你不可能预测一个刚打出本垒下一个也击中，会有一个先验概率
- 这个概率可以用一个参数  $\alpha$  与  $\beta$  的贝塔分布来描述，例如一共打了 300 个球，81 个击中，219 个击空，那么  $\alpha$  为 81， $\beta$  为 219
- 均值为  $\frac{\alpha}{\alpha+\beta} = \frac{81}{81+219} = 0.27$
- 概率密度分布图，从图上我们可以看出一个大约在 0.2-0.35 的概率区间，表示击球的先验概率空间可能的取值

```
library(ggplot2)
x <- seq(0,1,length=100)
db <- dbeta(x, 81, 219)
ggplot() + geom_line(aes(x,db)) + ylab("Density of beta")
```



## 10.2 为什么击球的概率分布符合贝塔分布?

- 设想球员 A 打了一个球打中了，那么在没有先验知识的情况下我会认为他击中概率为 1
- 这个球员又打中了一个球，那么还是 1
- 但第三个没打中，我们会认为他击中概率是 0 吗？
- 一般而言，这类连续击球问题可以用二项分布来描述，例如 10 个球打中 8 个的概率，我们假设这个击球概率为  $q$ ，那么这个概率应该是个  $q$  的函数：

$$f(q) \propto q^a (1-q)^b$$

- $q$  对于一个实际问题是确定的常数，所以出现这个场景的概率实际上是  $a$  与  $b$  的函数
- 为了保障这个概率函数累积为 1，需要除一个跟  $a$  与  $b$  有关的数
- 这个数可以用贝塔函数  $B(a, b)$  来表示，数学证明略
- 如果接着打了一个中了，那么如何更新这个概率？
- 根据贝叶斯公式，最后推导出的结果如下：

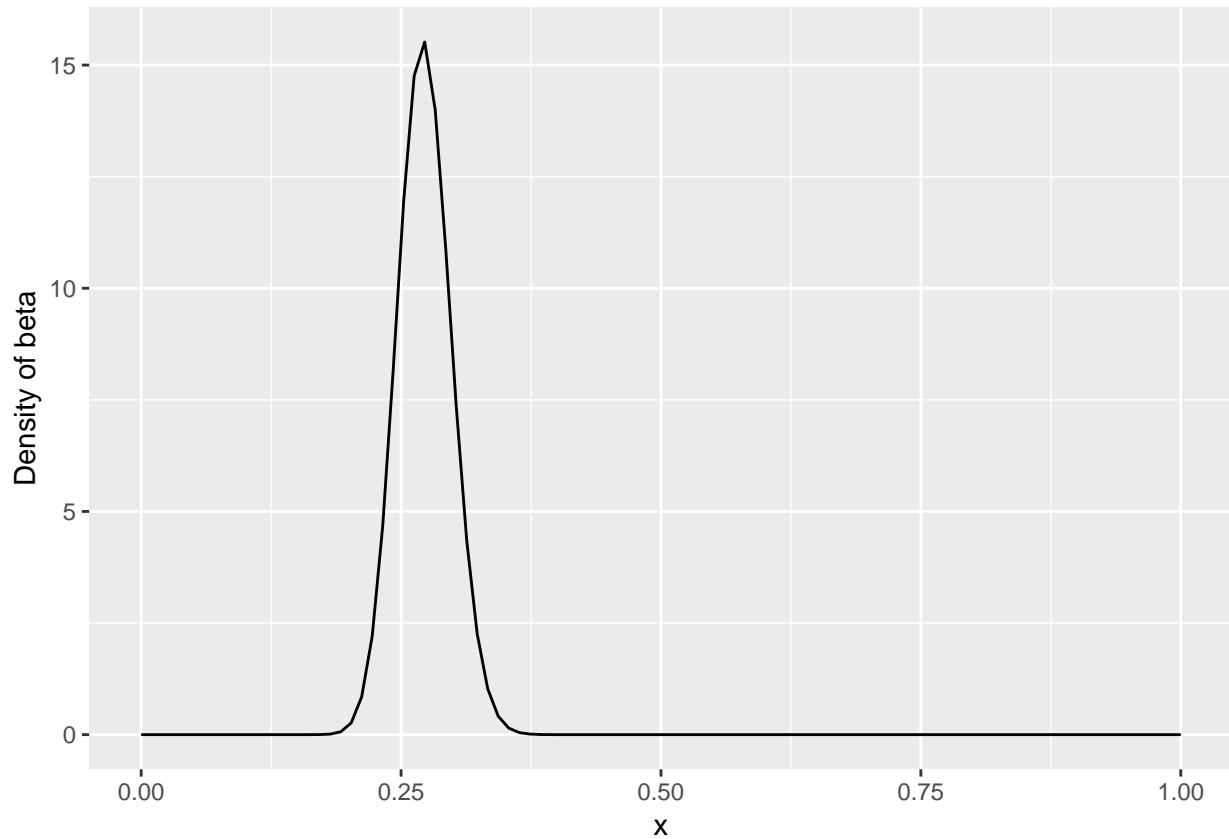
$$\text{Beta}(\alpha + 1, \beta + 0)$$

- 那么我们对这个击球率的估计就略高了一点，这是贝塔分布的神奇之处，形式非常简单，理解也很直观

## 10.3 先验与后验

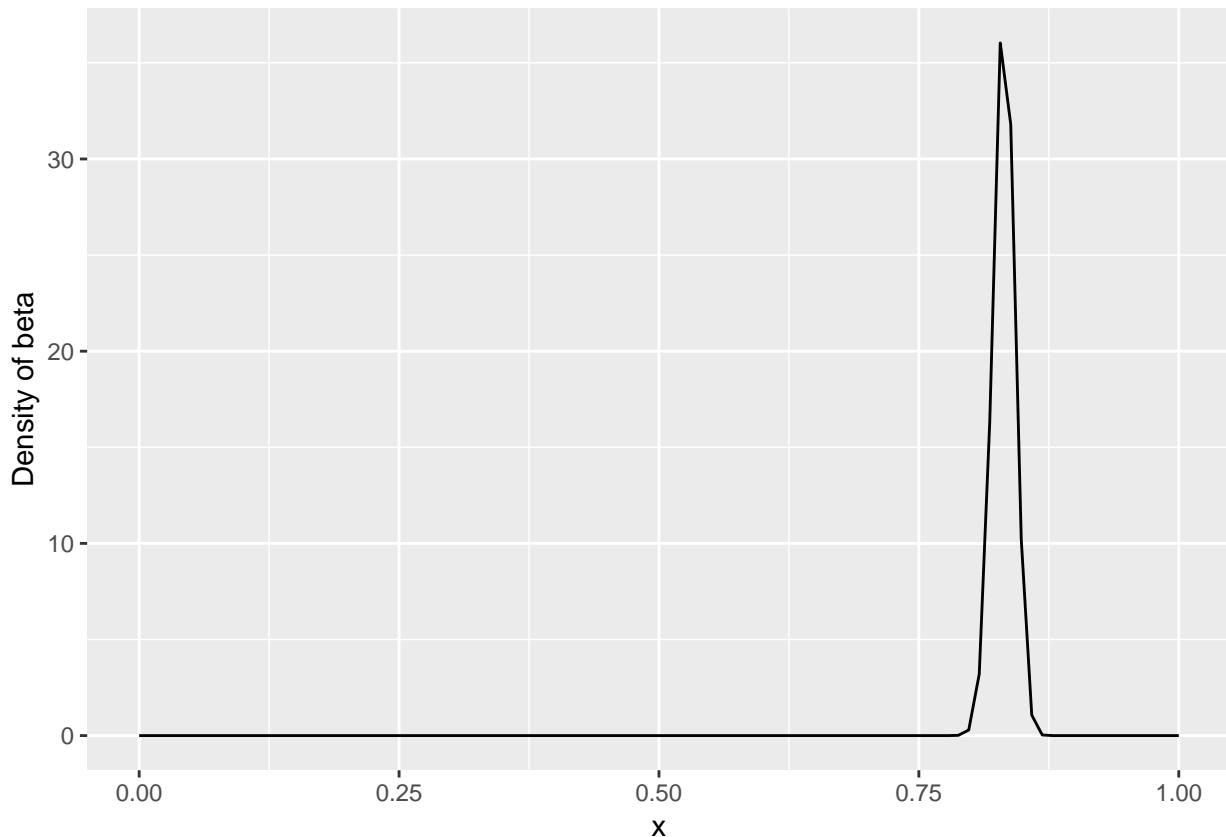
- 如果我们后续观察的击球少，那么不太容易影响到对概率的先验估计

```
x <- seq(0,1,length=100)
db <- dbeta(x, 81+1, 219)
ggplot() + geom_line(aes(x,db)) + ylab("Density of beta")
```



- 如果后续观察了大量的击球都中了，那么概率会偏向后面数据量的那一部分

```
x <- seq(0,1,length=100)
db <- dbeta(x, 81+1000, 219)
ggplot() + geom_line(aes(x,db)) + ylab("Density of beta")
```



- 这是贝叶斯分析的核心思想，通过证据更新经验
- 最后得到的均值（后验 0.83）一定是介于经验值（先验 0.27）与证据值（全击中就是 1）之间
- 贝塔分布天然适合描述一个对概率的估计场景
- 另一种不那么严谨的理解方法是如果一个概率是稳定的，那么多次实验的结果差别不会太大，则有：

$$\frac{a}{b} = \frac{c}{d} = \frac{a+b}{c+d}$$

- 如果每次实验的概率持平，那么不存在不确定度；但如果前面实验的次数少而后面实验的次数多，那么概率会偏重于后面，这就是贝塔分布想说明的事

## 10.4 经验贝叶斯

- 对于两个球员，一个打了 10 个球中了 4 个，另一个打了 1000 个球中了 300 个，一般击中概率 0.2，你会选哪一个？
- 我们对于小样本量的统计推断会有天然的不信任，如何通过统计量来描述？
- 下面用 MLB 的数据说明，首先提取出球员的击球数据：

```
library(dplyr)
library(tidyr)
library(Lahman)
# 拿到击球数据
career <- Batting %>%
  filter(AB > 0) %>%
```

```

anti_join(Pitching, by = "playerID") %>%
group_by(playerID) %>%
summarize(H = sum(H), AB = sum(AB)) %>%
mutate(average = H / AB)

# 把 ID 换成球员名字
career <- Master %>%
tbl_df() %>%
dplyr::select(playerID, nameFirst, nameLast) %>%
unite(name, nameFirst, nameLast, sep = " ") %>%
inner_join(career, by = "playerID") %>%
dplyr::select(-playerID)
# 展示数据
career

```

```

## # A tibble: 9,509 x 4
##   name             H   AB average
##   <chr>     <int> <int>    <dbl>
## 1 Hank Aaron      3771 12364  0.305
## 2 Tommie Aaron     216   944  0.229
## 3 Andy Abad        2    21  0.0952
## 4 John Abadie      11   49  0.224
## 5 Ed Abbaticchio   772  3044  0.254
## 6 Fred Abbott       107   513  0.209
## 7 Jeff Abbott       157   596  0.263
## 8 Kurt Abbott       523  2044  0.256
## 9 Ody Abbott        13    70  0.186
## 10 Frank Abercrombie 0     4   0
## # ... with 9,499 more rows

```

```

# 击球前 5
career %>%
  arrange(desc(average)) %>%
  head(5) %>%
  kable()

```

| name             | H | AB | average |
|------------------|---|----|---------|
| Jeff Banister    | 1 | 1  | 1       |
| Doc Bass         | 1 | 1  | 1       |
| Steve Biras      | 2 | 2  | 1       |
| C. B. Burns      | 1 | 1  | 1       |
| Jackie Gallagher | 1 | 1  | 1       |

```

# 击球后 5
career %>%
  arrange(average) %>%
  head(5) %>%
  kable()

```

| name              | H | AB | average |
|-------------------|---|----|---------|
| Frank Abercrombie | 0 | 4  | 0       |
| Lane Adams        | 0 | 3  | 0       |
| Horace Allen      | 0 | 7  | 0       |
| Pete Allen        | 0 | 4  | 0       |
| Walter Alston     | 0 | 1  | 0       |

- 如果仅考虑击球率会把很多板凳球员与运气球员包括进来，一个先验概率分布很有必要
- 那么考虑下如何得到，经验贝叶斯方法认为如果估计一个个体的参数，那么这个个体所在的整体的概率分布可作为先验概率分布
- 这个先验概率分布可以直接从数据中得到，然后我们要用极大似然或矩估计的方法拿到贝塔分布的两个参数：

```

career_filtered <- career %>%
  filter(AB >= 500)

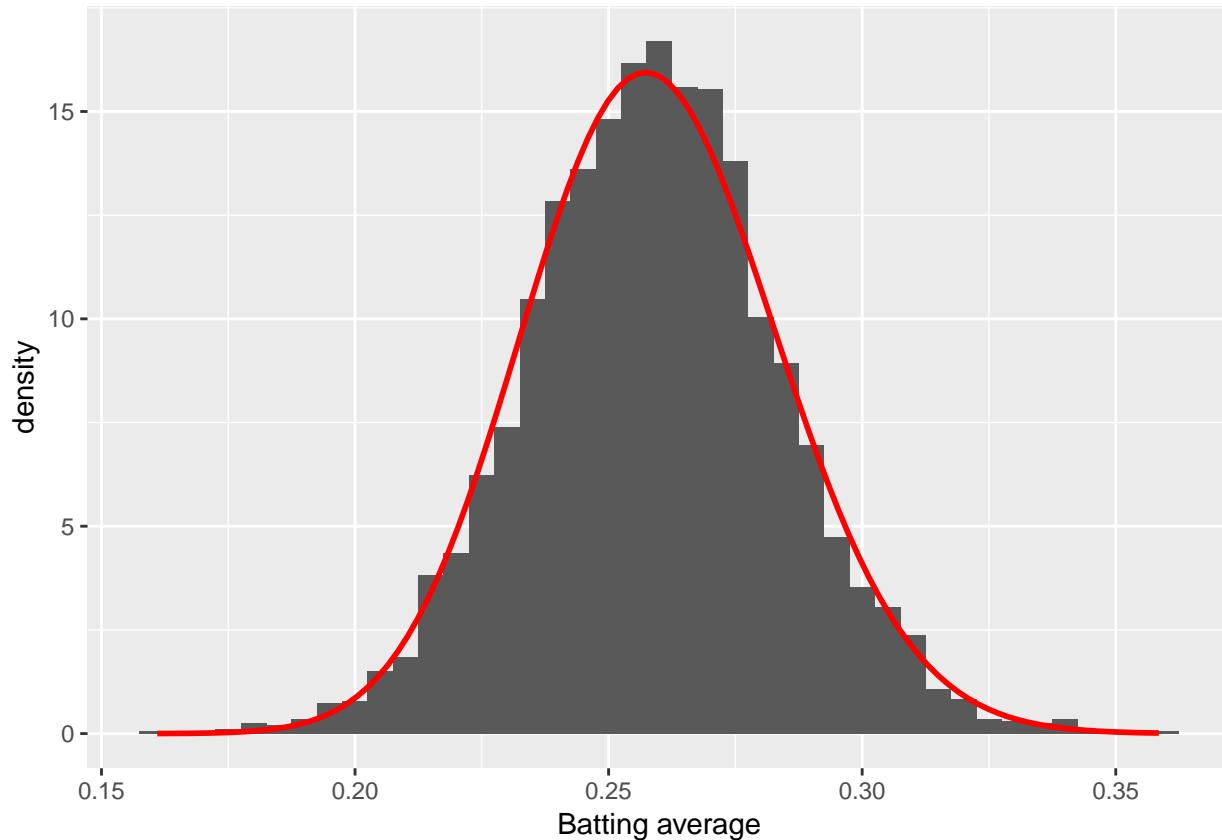
m <- MASS::fitdistr(career_filtered$average, dbeta,
                     start = list(shape1 = 1, shape2 = 10))

alpha0 <- m$estimate[1]
beta0 <- m$estimate[2]

# 看下拟合效果

ggplot(career_filtered) +
  geom_histogram(aes(average, y = ..density..), binwidth = .005) +
  stat_function(fun = function(x) dbeta(x, alpha0, beta0), color = "red",
                size = 1) +
  xlab("Batting average")

```



## 10.5 从整体到个人

- 当我们估计个人的击球率时，整体可以作为先验函数，个人的数据可以通过贝塔分布更新到个体
- 那么如果一个人数据少，我们倾向于认为他是平均水平；数据多则认为符合个人表现
- 这事实上是一个分层结构，经验贝叶斯推断里隐含了这么一个从整体到个人的过程

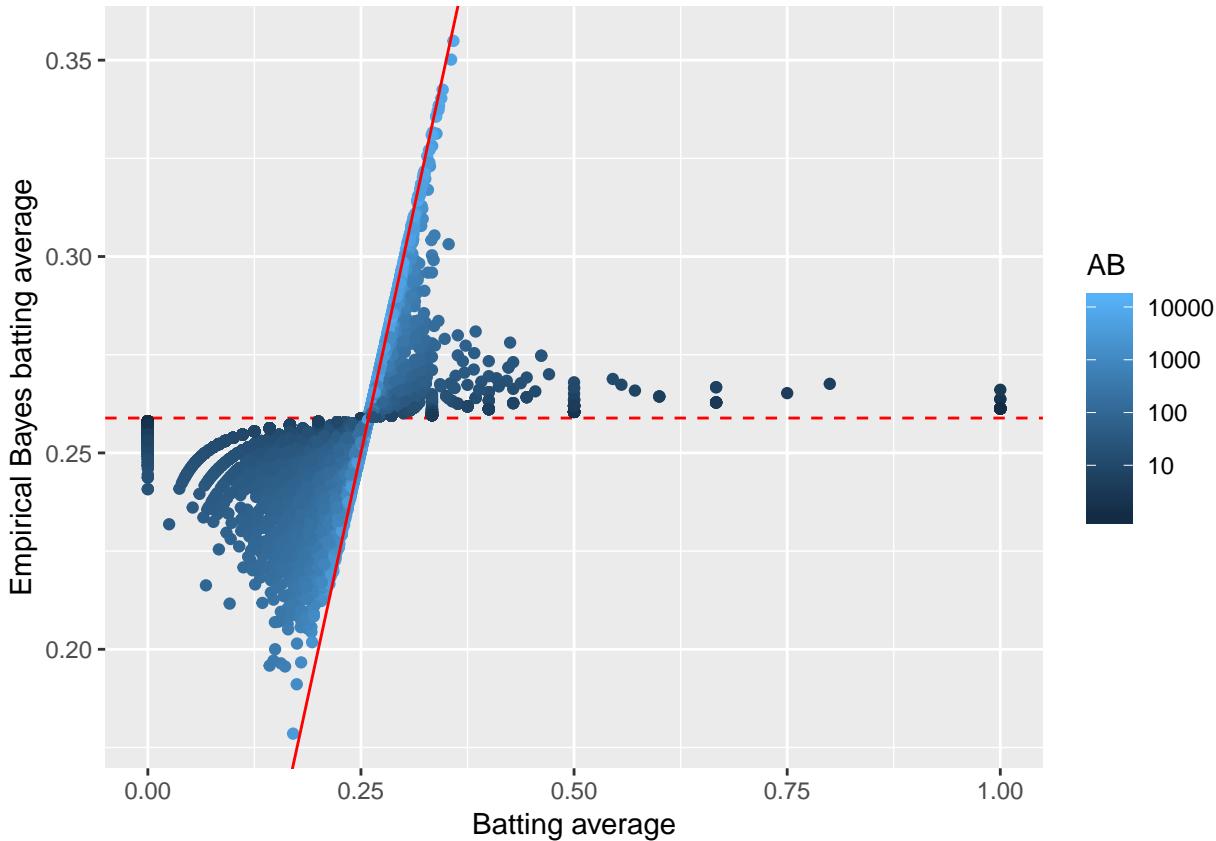
```
career_eb <- career %>%
  mutate(eb_estimate = (H + alpha0) / (AB + alpha0 + beta0))
# 击球率高
career_eb %>%
  arrange(desc(eb_estimate)) %>%
  head(5) %>%
  kable()
```

| name                 | H    | AB   | average | eb_estimate |
|----------------------|------|------|---------|-------------|
| Rogers Hornsby       | 2930 | 8173 | 0.358   | 0.355       |
| Shoeless Joe Jackson | 1772 | 4981 | 0.356   | 0.350       |
| Ed Delahanty         | 2596 | 7505 | 0.346   | 0.342       |
| Billy Hamilton       | 2158 | 6268 | 0.344   | 0.340       |
| Harry Heilmann       | 2660 | 7787 | 0.342   | 0.338       |

```
# 击球率低
career_eb %>%
  arrange(eb_estimate) %>%
  head(5) %>%
  kable()
```

| name           | H   | AB   | average | eb_estimate |
|----------------|-----|------|---------|-------------|
| Bill Bergen    | 516 | 3028 | 0.170   | 0.179       |
| Ray Oyler      | 221 | 1265 | 0.175   | 0.191       |
| John Vukovich  | 90  | 559  | 0.161   | 0.196       |
| John Humphries | 52  | 364  | 0.143   | 0.196       |
| George Baker   | 74  | 474  | 0.156   | 0.196       |

```
# 整体估计
ggplot(career_eb, aes(average, eb_estimate, color = AB)) +
  geom_hline(yintercept = alpha0 / (alpha0 + beta0), color = "red", lty = 2) +
  geom_point() +
  geom_abline(color = "red") +
  scale_colour_gradient(trans = "log", breaks = 10 ^ (1:5)) +
  xlab("Batting average") +
  ylab("Empirical Bayes batting average")
```



- 数据点多会收缩到  $x = y$ , 也就是个人的击球率; 数据点少则回归到整体击球率
- 这就是经验贝叶斯方法的全貌: 先估计整体的参数, 然后把整体参数作为先验概率估计个人参数

## 10.6 可信区间与置信区间

- 经验贝叶斯可以给出点估计, 但现实中我们可能更关心区间估计
- 一般这类区间估计可以用二项式比例估计来进行, 不过没有先验经验的限制置信区间大到没意义
- 经验贝叶斯会给出一个后验分布, 这个分布可以用来求可信区间

```
library(broom)
# 给出后验分布
career <- Batting %>%
  filter(AB > 0) %>%
  anti_join(Pitching, by = "playerID") %>%
  group_by(playerID) %>%
  summarize(H = sum(H), AB = sum(AB)) %>%
  mutate(average = H / AB)

career <- Master %>%
 tbl_df() %>%
  dplyr::select(playerID, nameFirst, nameLast) %>%
  unite(name, nameFirst, nameLast, sep = " ") %>%
  inner_join(career, by = "playerID")
```

```

career0 <- Batting %>%
  filter(AB > 0) %>%
  anti_join(Pitching, by = "playerID") %>%
  group_by(playerID) %>%
  summarize(H = sum(H), AB = sum(AB), year = mean(yearID)) %>%
  mutate(average = H / AB)

career2 <- Master %>%
 tbl_df() %>%
  dplyr::select(playerID, nameFirst, nameLast, bats) %>%
  unite(name, nameFirst, nameLast, sep = " ") %>%
  inner_join(career0, by = "playerID")

career_eb <- career %>%
  mutate(eb_estimate = (H + alpha0) / (AB + alpha0 + beta0))
career_eb <- career_eb %>%
  mutate(alpha1 = H + alpha0,
         beta1 = AB - H + beta0)
# 提取洋基队的数据
yankee_1998 <- c("brosisc01", "jeterde01", "knoblch01", "martiti02", "posadjo01", "strawda01", "willibe02")

yankee_1998_career <- career_eb %>%
  filter(playerID %in% yankee_1998)

# 提取可信区间
yankee_1998_career <- yankee_1998_career %>%
  mutate(low = qbeta(.025, alpha1, beta1),
         high = qbeta(.975, alpha1, beta1))
yankee_1998_career %>%
  dplyr::select(-alpha1, -beta1, -eb_estimate) %>%
  knitr::kable()

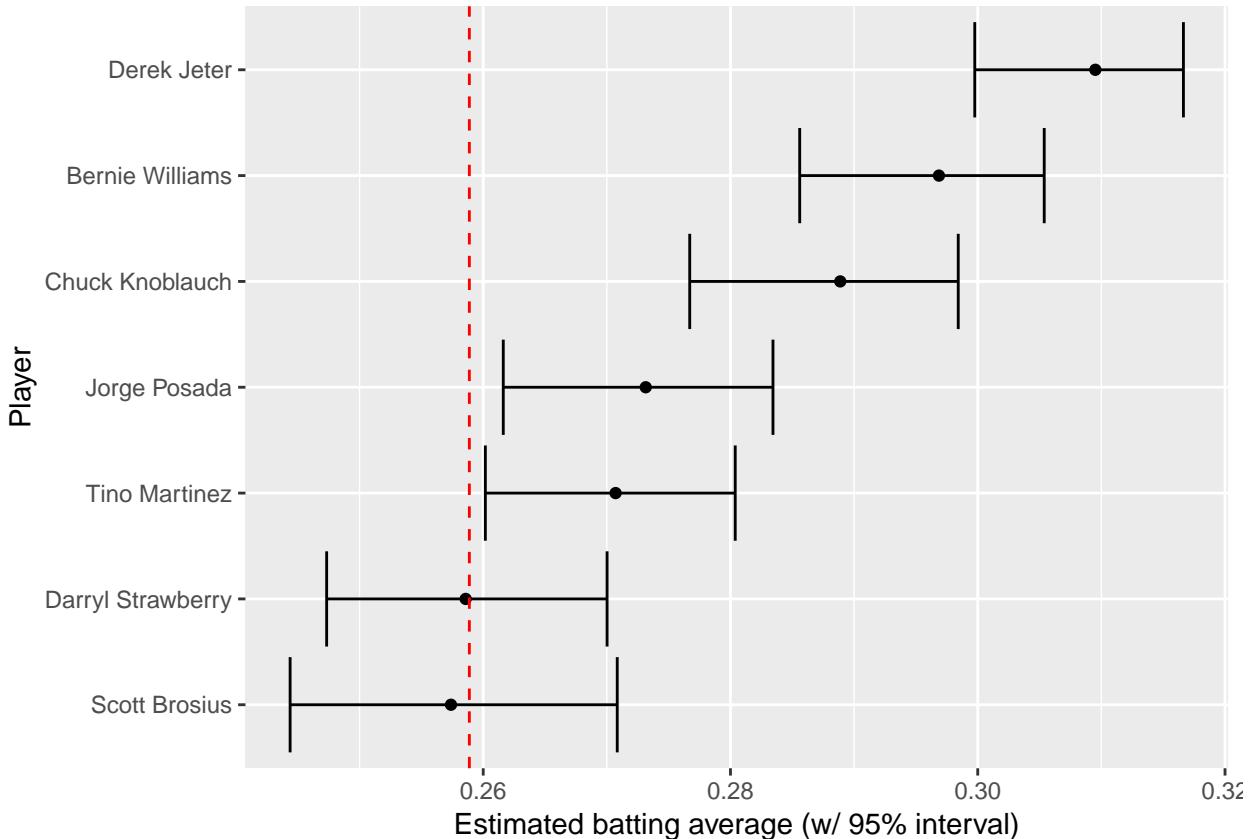
```

| playerID  | name              | H    | AB    | average | low   | high  |
|-----------|-------------------|------|-------|---------|-------|-------|
| brosisc01 | Scott Brosius     | 1001 | 3889  | 0.257   | 0.244 | 0.271 |
| jeterde01 | Derek Jeter       | 3465 | 11195 | 0.310   | 0.300 | 0.317 |
| knoblch01 | Chuck Knoblauch   | 1839 | 6366  | 0.289   | 0.277 | 0.298 |
| martiti02 | Tino Martinez     | 1925 | 7111  | 0.271   | 0.260 | 0.280 |
| posadjo01 | Jorge Posada      | 1664 | 6092  | 0.273   | 0.262 | 0.283 |
| strawda01 | Darryl Strawberry | 1401 | 5418  | 0.259   | 0.247 | 0.270 |
| willibe02 | Bernie Williams   | 2336 | 7869  | 0.297   | 0.286 | 0.305 |

```

# 绘制可信区间
yankee_1998_career %>%
  mutate(name = reorder(name, average)) %>%
  ggplot(aes(average, name)) +
  geom_point() +
  geom_errorbarh(aes(xmin = low, xmax = high)) +
  geom_vline(xintercept = alpha0 / (alpha0 + beta0), color = "red", lty = 2) +
  xlab("Estimated batting average (w/ 95% interval)") +
  ylab("Player")

```



```
# 对比置信区间与可信区间
career_eb <- career_eb %>%
  mutate(low = qbeta(.025, alpha1, beta1),
        high = qbeta(.975, alpha1, beta1))

set.seed(2016)

some <- career_eb %>%
  sample_n(20) %>%
  mutate(name = paste0(name, " (", H, "/", AB, ")"))

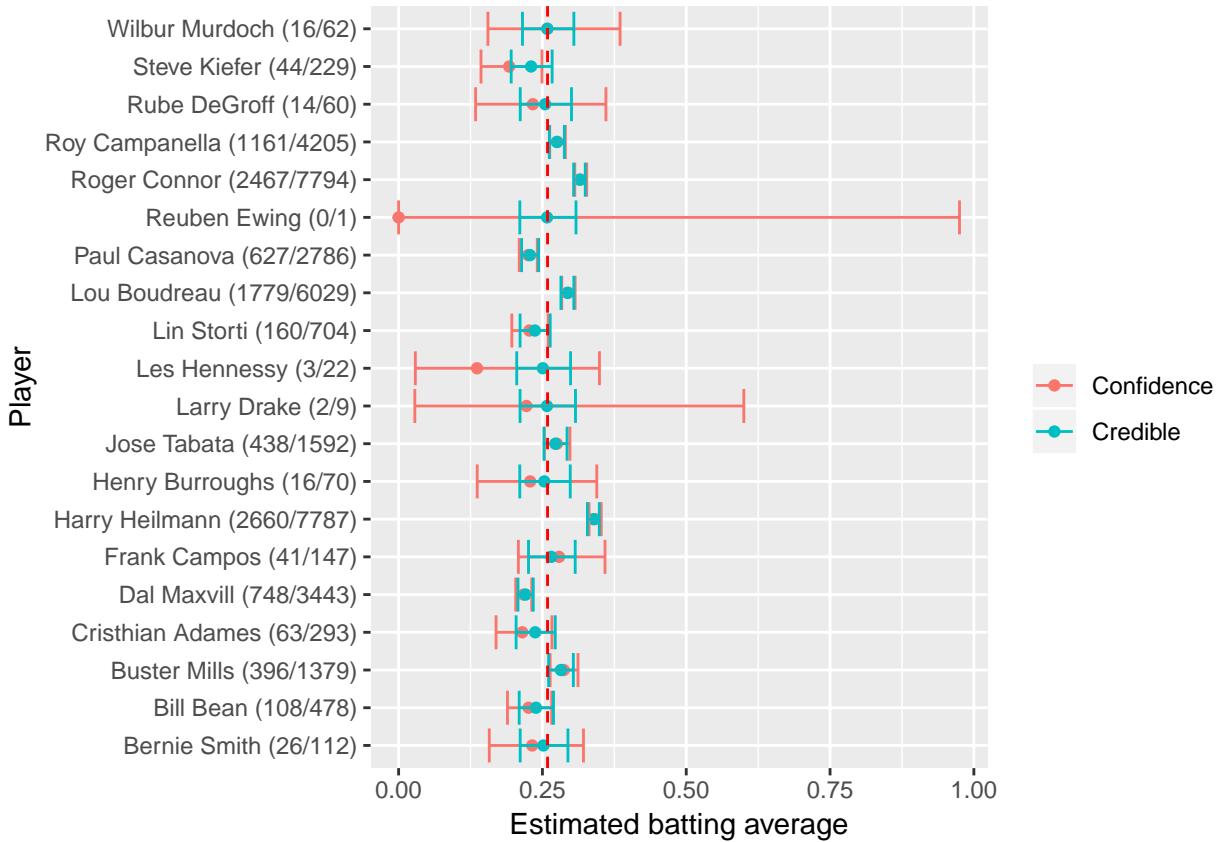
frequentist <- some %>%
  group_by(playerID, name, AB) %>%
  do(tidy(binom.test(.\$H, .\$AB))) %>%
  dplyr::select(playerID, name, estimate, low = conf.low, high = conf.high) %>%
  mutate(method = "Confidence")

bayesian <- some %>%
  dplyr::select(playerID, name, AB, estimate = eb_estimate,
                low = low, high = high) %>%
  mutate(method = "Credible")

combined <- bind_rows(frequentist, bayesian)

combined %>%
  mutate(name2 = reorder(name, -AB)) %>%
  ggplot(aes(estimate, name2, color = method, group = method)) +
```

```
geom_point() +
  geom_errorbarh(aes(xmin = low, xmax = high)) +
  geom_vline(xintercept = alpha0 / (alpha0 + beta0), color = "red", lty = 2) +
  xlab("Estimated batting average") +
  ylab("Player") +
  labs(color = "")
```

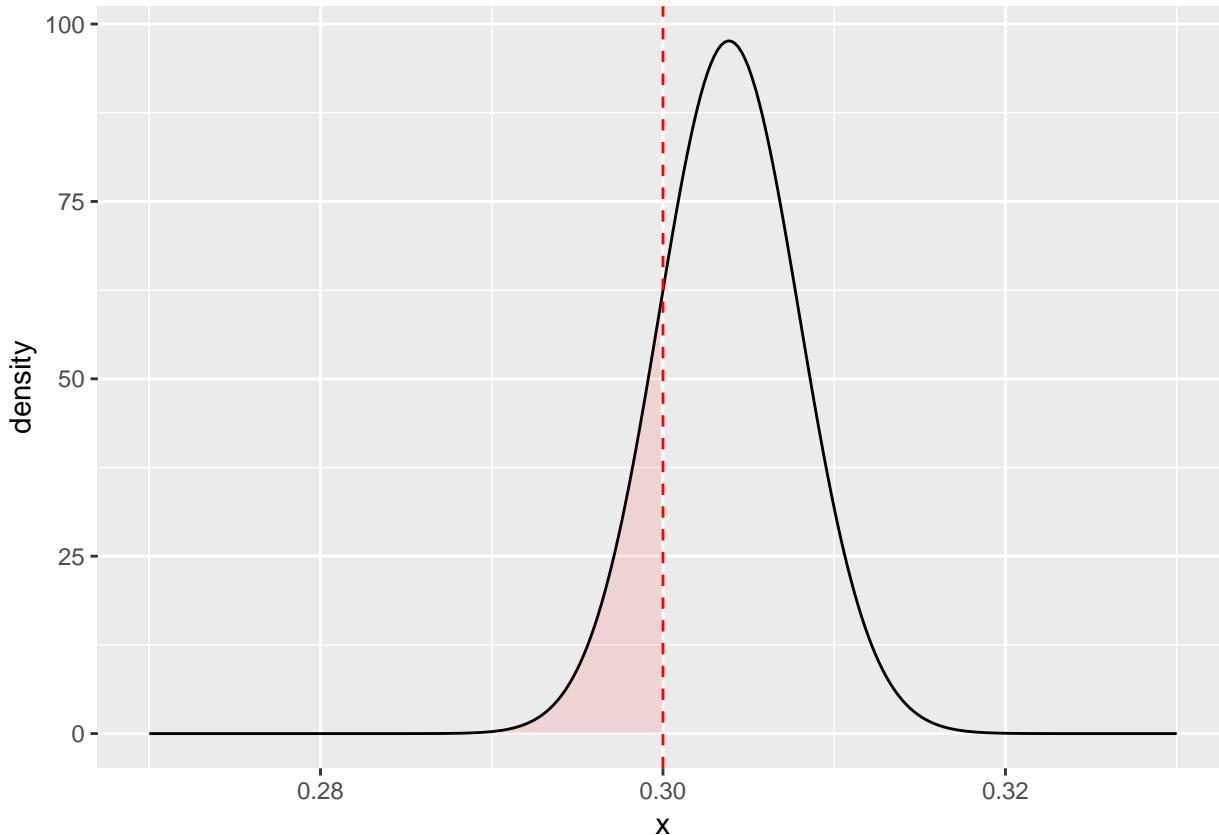


- 可信区间与置信区间很大的区别在于前者考虑了先验概率进而实现了区间的收缩，后者则可看作无先验贝塔分布给出的区间估计，频率学派目前没有很好的收缩区间估计的方法

## 10.7 后验错误率

- 现实问题经常不局限于估计，而是侧重决策，例如如果一个球员的击球率高于某个值，他就可以进入名人堂（击球率大于 0.3），这个决策常常伴随区间估计而不是简单的点估计

```
# 以 Hank Aaron 为例
career_eb %>%
  filter(name == "Hank Aaron") %>%
  do(data_frame(x = seq(.27, .33, .0002),
                density = dbeta(x, .\$alpha1, .\$beta1))) %>%
  ggplot(aes(x, density)) +
  geom_line() +
  geom_ribbon(aes(ymin = 0, ymax = density * (x < .3)),
              alpha = .1, fill = "red") +
  geom_vline(color = "red", lty = 2, xintercept = .3)
```



```
# 提取该球员数据
career_eb %>% filter(name == "Hank Aaron")
```

```
## # A tibble: 1 x 10
##   playerID  name      H     AB average eb_estimate alpha1 beta1    low   high
##   <chr>      <chr> <int> <int>   <dbl>       <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 aaronha01 Hank ~  3771 12364   0.305       0.304 3850. 8820. 0.296 0.312
```

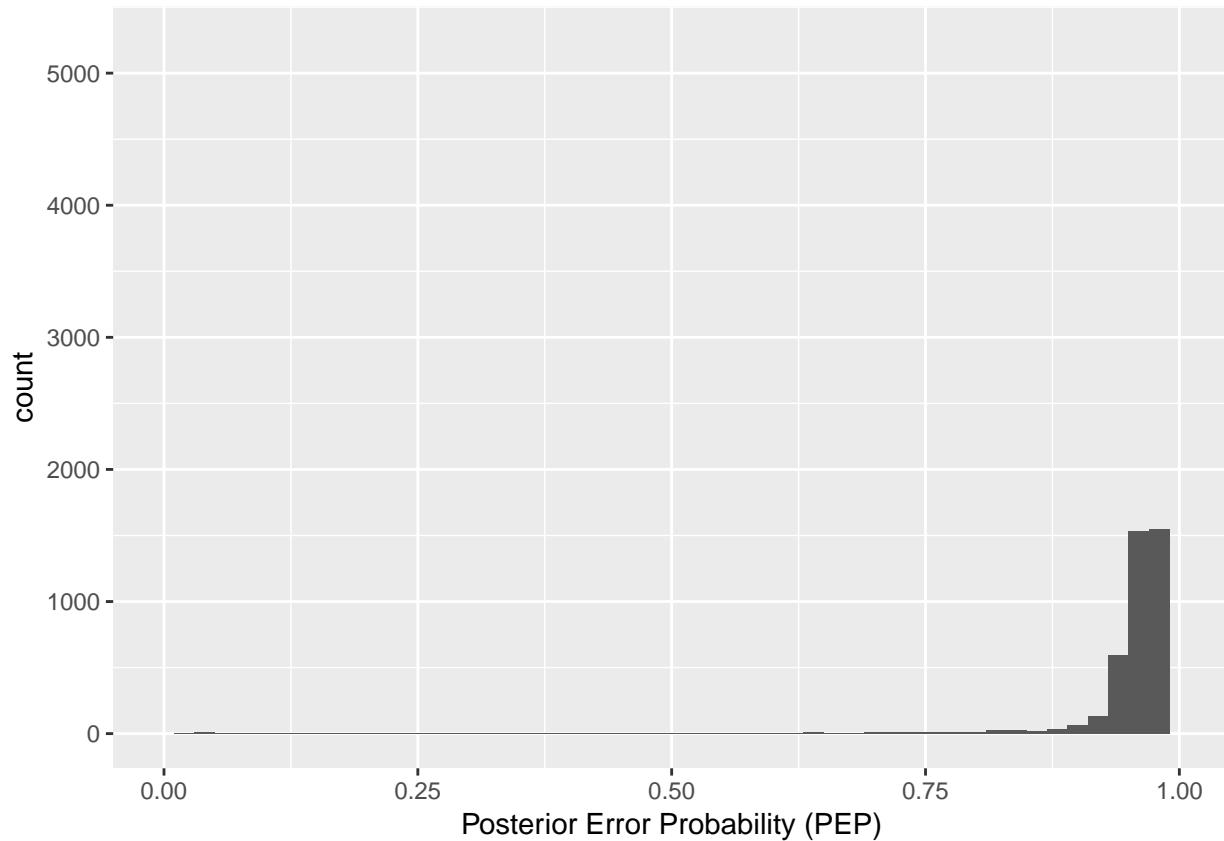
```
# 计算其不进入名人堂的概率
pbeta(.3, 3850, 8818)
```

```
## [1] 0.169
```

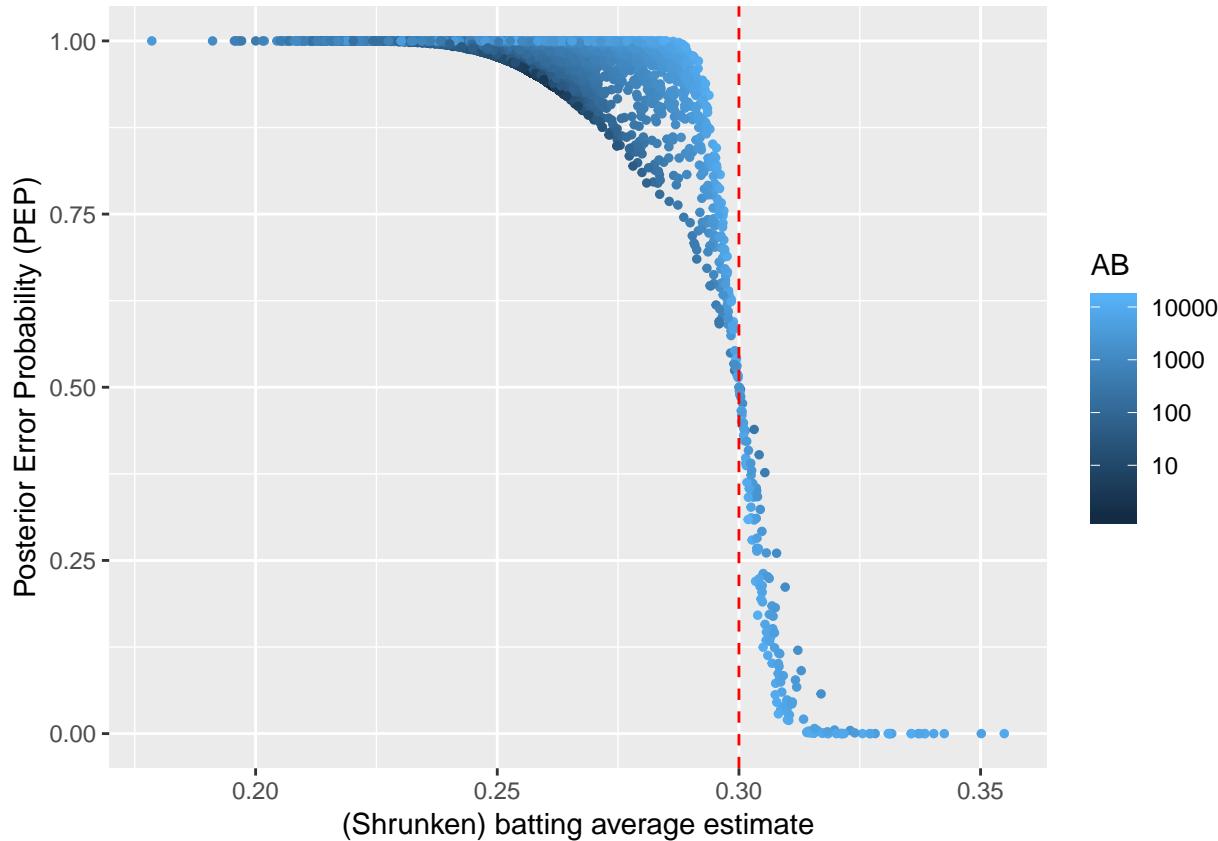
- 后验错误率 (Posterior Error Probability) 可类比经典假设检验中的显著性水平  $\alpha$

- 后验包括率 (Posterior Inclusion Probability) 可类比经典假设检验中的置信水平  $1 - \alpha$

```
# 所有球员的后验错误率分布，大部分不超过 0.3
career_eb <- career_eb %>%
  mutate(PEP = pbeta(.3, alpha1, beta1))
ggplot(career_eb, aes(PEP)) +
  geom_histogram(binwidth = .02) +
  xlab("Posterior Error Probability (PEP)") +
  xlim(0, 1)
```



```
# 后验错误率与击球率的关系
career_eb %>%
  ggplot(aes(eb_estimate, PEP, color = AB)) +
  geom_point(size = 1) +
  xlab("(Shrunken) batting average estimate") +
  ylab("Posterior Error Probability (PEP)") +
  geom_vline(color = "red", lty = 2, xintercept = .3) +
  scale_colour_gradient(trans = "log", breaks = 10^(1:5))
```



- 后验错误率高于 0.3 的多数是击球率与击球数都高的人，因为贝叶斯方法惩罚了击球数低的人

## 10.8 错误发现率

- 错误发现率 (FDR) 可用来控制一个整体决策，保证整体犯错的概率低于某个数值，错误发现率越高，越可能把假阳性包括进来
- 假如我们把进入名人堂的决策作为一个整体，则可允许一定的整体错误率，因为每个人的后验错误率可以计算且期望值线性可加和，我们可以得到一个整体的错误率

```
# 取前 100 个球员
top_players <- career_eb %>%
  arrange(PEP) %>%
  head(100)
# 总错率率
sum(top_players$PEP)
```

```
## [1] 5.07
```

```
# 平均错误率
mean(top_players$PEP)
```

```
## [1] 0.0507
```

```
# 错误率随所取球员的变化
sorted_PEP <- career_eb %>%
  arrange(PEP)
```

```
mean(head(sorted_PEP$PEP, 50))

## [1] 0.00186

mean(head(sorted_PEP$PEP, 200))

## [1] 0.246
```

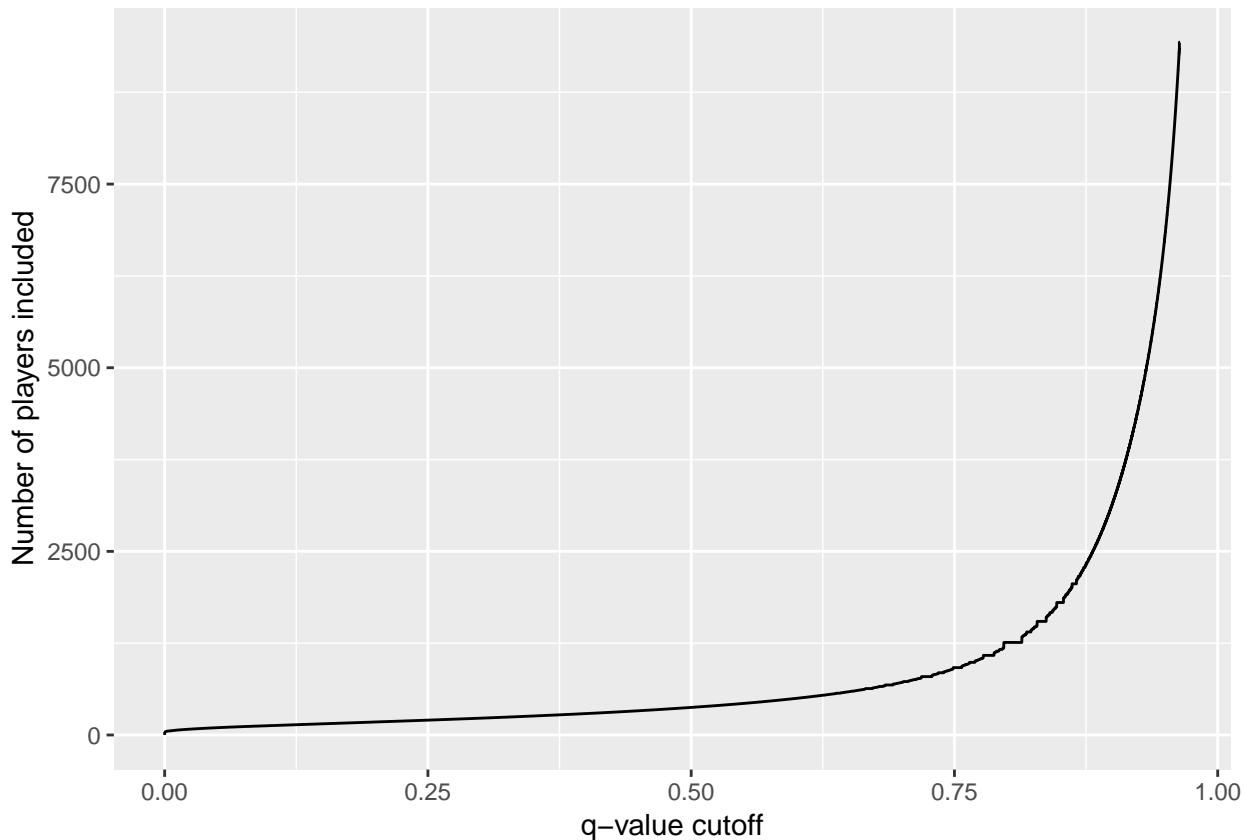
- 错误率在排序后前面低后面高，但这个错误率不特指某个球员，而是包含到某个球员的整体犯错的概率

## 10.9 q 值

- q 值定义为排序后累积到某个样本的整体平均错误率，类似多重比较中对整体错误率控制的 p 值

```
# 生成每个球员的 q 值
career_eb <- career_eb %>%
  arrange(PEP) %>%
  mutate(qvalue = cummean(PEP))

# 观察不同 q 值对名人堂球员数的影响
career_eb %>%
  ggplot(aes(qvalue, rank(PEP))) +
  geom_line() +
  xlab("q-value cutoff") +
  ylab("Number of players included")
```

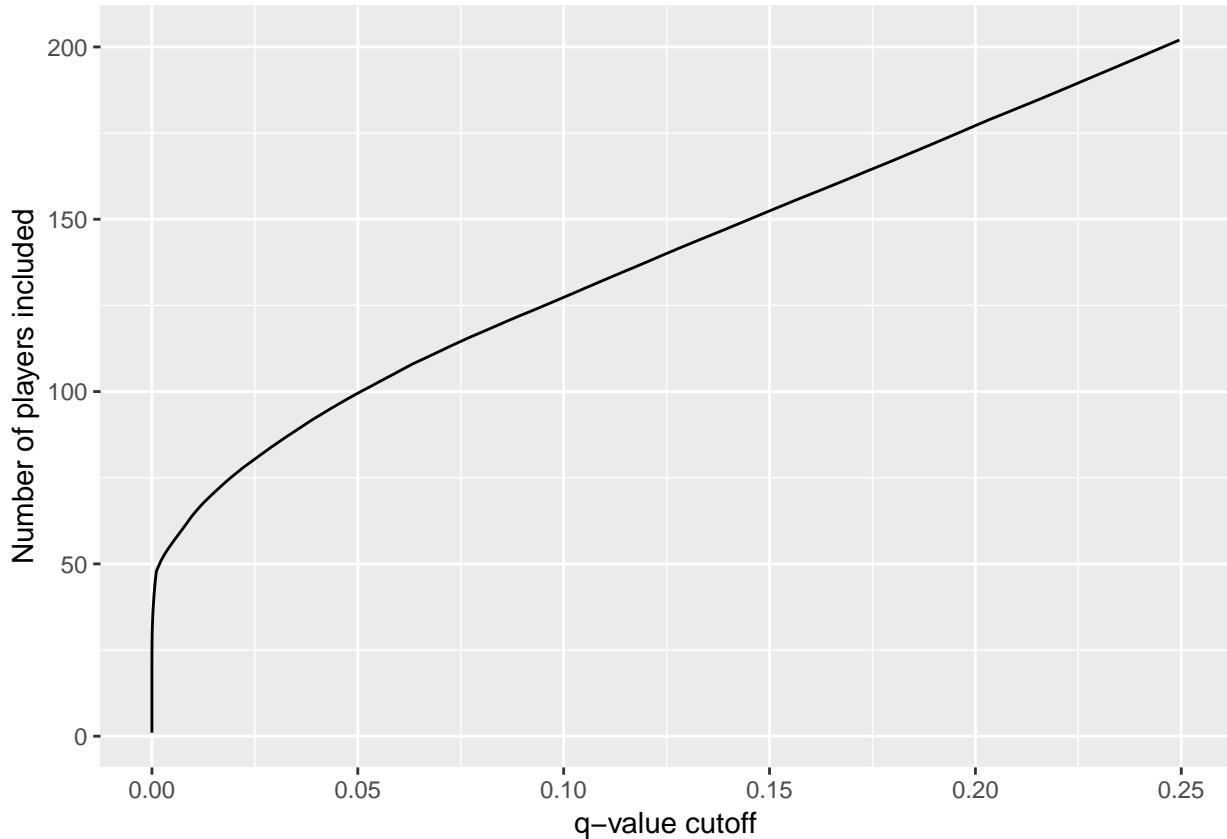


```
# 观察小 q 值部分
career_eb %>%
```

```

filter(qvalue < .25) %>%
ggplot(aes(qvalue, rank(PEP))) +
geom_line() +
xlab("q-value cutoff") +
ylab("Number of players included")

```



- 200 个人进入名人堂可能有约 1/4 的球员不合适，如果是 50 个人进入名人堂那么基本不会犯错
- q 值是一个整体而非个体的平均错误率，具有累积性，不代表 q 值大的那一个就是错的
- q 值在频率学派的多重比较里也有定义，虽然没有空假设（有先验概率），但实质等同

## 10.10 贝叶斯视角的假设检验

- 前面描述的是击球率如何求，如何进行区间估计与多个体的错误率控制，面向的个体或整体，那么如何解决比较问题
- 设想多个球员，我们考虑如何去比较他们击球率。如果两个球员击球率的概率密度曲线比较接近，那么即便均值有不同我们也无法进行区分；如果重叠比较少，那么我们有理由认为他们之间的差异显著
- 贝叶斯视角下如何定量描述这个差异是否显著？

### 10.10.1 模拟验证

- 单纯取样比大小然后计算比例

```
# 提取两人数据
aaron <- career_eb %>% filter(name == "Hank Aaron")
piazza <- career_eb %>% filter(name == "Mike Piazza")
# 模拟取样 10 万次
piazza_simulation <- rbeta(1e6, piazza$alpha1, piazza$beta1)
aaron_simulation <- rbeta(1e6, aaron$alpha1, aaron$beta1)
# 计算一个人超过另一个人的概率
sim <- mean(piazza_simulation > aaron_simulation)
sim

## [1] 0.606
```

### 10.10.2 数值积分

- 两个概率的联合概率分布，然后积分一个队员大于另一个的概率

```
d <- .00002
limits <- seq(.29, .33, d)
sum(outer(limits, limits, function(x, y) {
  (x > y) *
    dbeta(x, piazza$alpha1, piazza$beta1) *
    dbeta(y, aaron$alpha1, aaron$beta1) *
    d ^ 2
}))

## [1] 0.604
```

### 10.10.3 解析解

- 两个贝塔分布一个比另一个高是有含有贝塔函数的解析解的：

$$p_A \sim \text{Beta}(\alpha_A, \beta_A)$$

$$p_B \sim \text{Beta}(\alpha_B, \beta_B)$$

$$\Pr(p_B > p_A) = \sum_{i=0}^{\alpha_B-1} \frac{B(\alpha_A + i, \beta_A + \beta_B)}{(\beta_B + i)B(1+i, \beta_B)B(\alpha_A, \beta_A)}$$

```
h <- function(alpha_a, beta_a,
               alpha_b, beta_b) {
  j <- seq.int(0, round(alpha_b) - 1)
  log_vals <- (lbeta(alpha_a + j, beta_a + beta_b) - log(beta_b + j) -
    lbeta(1 + j, beta_b) - lbeta(alpha_a, beta_a))
  1 - sum(exp(log_vals))
}

h(piazza$alpha1, piazza$beta1,
  aaron$alpha1, aaron$beta1)

## [1] 0.605
```

#### 10.10.4 正态近似求解

- 贝塔分布在  $\alpha$  与  $\beta$  比较大时接近正态分布，可以直接用正态分布的解析解求，速度快很多

```

h_approx <- function(alpha_a, beta_a,
                      alpha_b, beta_b) {
  u1 <- alpha_a / (alpha_a + beta_a)
  u2 <- alpha_b / (alpha_b + beta_b)
  var1 <- alpha_a * beta_a / ((alpha_a + beta_a) ^ 2 * (alpha_a + beta_a + 1))
  var2 <- alpha_b * beta_b / ((alpha_b + beta_b) ^ 2 * (alpha_b + beta_b + 1))
  pnorm(0, u2 - u1, sqrt(var1 + var2))
}

h_approx(piazza$alpha1, piazza$beta1, aaron$alpha1, aaron$beta1)

## [1] 0.605

```

### 10.11 比例检验

- 这是个列联表问题，频率学派对比两个比例

```

two_players <- bind_rows(aaron, piazza)

two_players %>%
  transmute(Player = name, Hits = H, Misses = AB - H) %>%
  knitr::kable()

Player   Hits Misses
Hank Aaron 3771  8593
Mike Piazza 2127  4784

prop.test(two_players$H, two_players$AB)

```

```

##
## 2-sample test for equality of proportions with continuity
## correction
##
## data: two_players$H out of two_players$AB
## X-squared = 0.1, df = 1, p-value = 0.7
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## -0.0165 0.0109
## sample estimates:
## prop 1 prop 2
## 0.305 0.308

```

- 贝叶斯学派对比两个比例

```

credible_interval_approx <- function(a, b, c, d) {
  u1 <- a / (a + b)
  u2 <- c / (c + d)
  var1 <- a * b / ((a + b) ^ 2 * (a + b + 1))
  var2 <- c * d / ((c + d) ^ 2 * (c + d + 1))

  mu_diff <- u2 - u1
  sd_diff <- sqrt(var1 + var2)
}

```

```

data_frame(posterior = pnorm(0, mu_diff, sd_diff),
           estimate = mu_diff,
           conf.low = qnorm(.025, mu_diff, sd_diff),
           conf.high = qnorm(.975, mu_diff, sd_diff))
}

credible_interval_approx(piazza$alpha1, piazza$beta1, aaron$alpha1, aaron$beta1)

## # A tibble: 1 x 4
##   posterior estimate conf.low conf.high
##       <dbl>     <dbl>    <dbl>     <dbl>
## 1      0.605 -0.00181 -0.0151    0.0115

```

- 多个球员对比一个

```

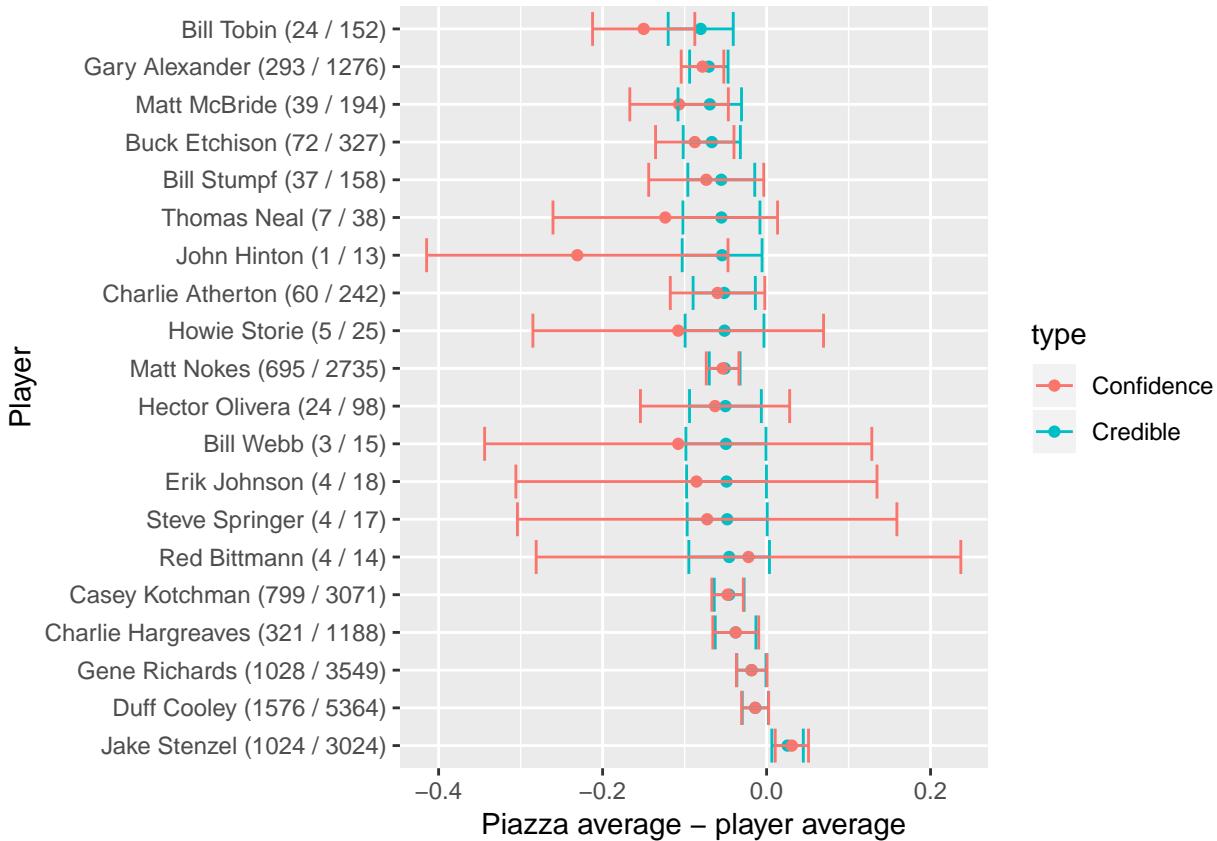
set.seed(2016)

intervals <- career_eb %>%
  filter(AB > 10) %>%
  sample_n(20) %>%
  group_by(name, H, AB) %>%
  do(credible_interval_approx(piazza$alpha1, piazza$beta1, .$alpha1, .$beta1)) %>%
  ungroup() %>%
  mutate(name = reorder(paste0(name, " (", H, " / ", AB, ")"), -estimate))
f <- function(H, AB) broom::tidy(prop.test(c(H, piazza$H), c(AB, piazza$AB)))
prop_tests <- purrr::map2_df(intervals$H, intervals$AB, f) %>%
  mutate(estimate = estimate1 - estimate2,
        name = intervals$name)

all_intervals <- bind_rows(
  mutate(intervals, type = "Credible"),
  mutate(prop_tests, type = "Confidence")
)

ggplot(all_intervals, aes(x = estimate, y = name, color = type)) +
  geom_point() +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high)) +
  xlab("Piazza average - player average") +
  ylab("Player")

```



- 置信区间与可信区间的最大差异来自于经验贝叶斯的区间收敛

## 10.12 错误率控制

- 如果我打算交易一个球员，那么如何筛选候选人？
- 先选那些击球率更好的球员

```
# 对比打算交易的球员与其他球员
career_eb_vs_piazza <- bind_cols(
  career_eb,
  credible_interval_approx(piazza$alpha1, piazza$beta1,
                            career_eb$alpha1, career_eb$beta1)) %>%
  dplyr::select(name, posterior, conf.low, conf.high)

career_eb_vs_piazza
```

```
## # A tibble: 9,509 x 4
##   name           posterior conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>
## 1 Rogers Hornsby 2.84e-11  0.0345    0.0639
## 2 Ed Delahanty  7.11e- 7  0.0218    0.0518
## 3 Shoeless Joe Jackson 8.82e- 8  0.0278    0.0611
## 4 Willie Keeler  4.62e- 6  0.0183    0.0472
## 5 Nap Lajoie      1.62e- 5  0.0158    0.0441
## 6 Tony Gwynn     1.83e- 5  0.0157    0.0442
```

```

## 7 Harry Heilmann      7.19e- 6  0.0180  0.0476
## 8 Lou Gehrig          1.43e- 5  0.0167  0.0461
## 9 Billy Hamilton     7.05e- 6  0.0190  0.0502
## 10 Eddie Collins      2.00e- 4  0.0113  0.0393
## # ... with 9,499 more rows

# 计算 q 值
career_eb_vs_piazza <- career_eb_vs_piazza %>%
  arrange(posterior) %>%
  mutate(qvalue = cummean(posterior))

# 筛选那些 q 值小于 0.05 的
better <- career_eb_vs_piazza %>%
  filter(qvalue < .05)

better

## # A tibble: 49 x 5
##   name           posterior conf.low conf.high    qvalue
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 Rogers Hornsby 2.84e-11  0.0345   0.0639  2.84e-11
## 2 Shoeless Joe Jackson 8.82e- 8  0.0278   0.0611  4.41e- 8
## 3 Ed Delahanty  7.11e- 7  0.0218   0.0518  2.66e- 7
## 4 Willie Keeler  4.62e- 6  0.0183   0.0472  1.35e- 6
## 5 Billy Hamilton 7.05e- 6  0.0190   0.0502  2.49e- 6
## 6 Harry Heilmann 7.19e- 6  0.0180   0.0476  3.28e- 6
## 7 Lou Gehrig      1.43e- 5  0.0167   0.0461  4.86e- 6
## 8 Nap Lajoie       1.62e- 5  0.0158   0.0441  6.28e- 6
## 9 Tony Gwynn       1.83e- 5  0.0157   0.0442  7.61e- 6
## 10 Bill Terry      3.04e- 5  0.0162   0.0472  9.89e- 6
## # ... with 39 more rows

```

- 这样我们筛选到一个可交易的群体，总和错误率不超过 5%

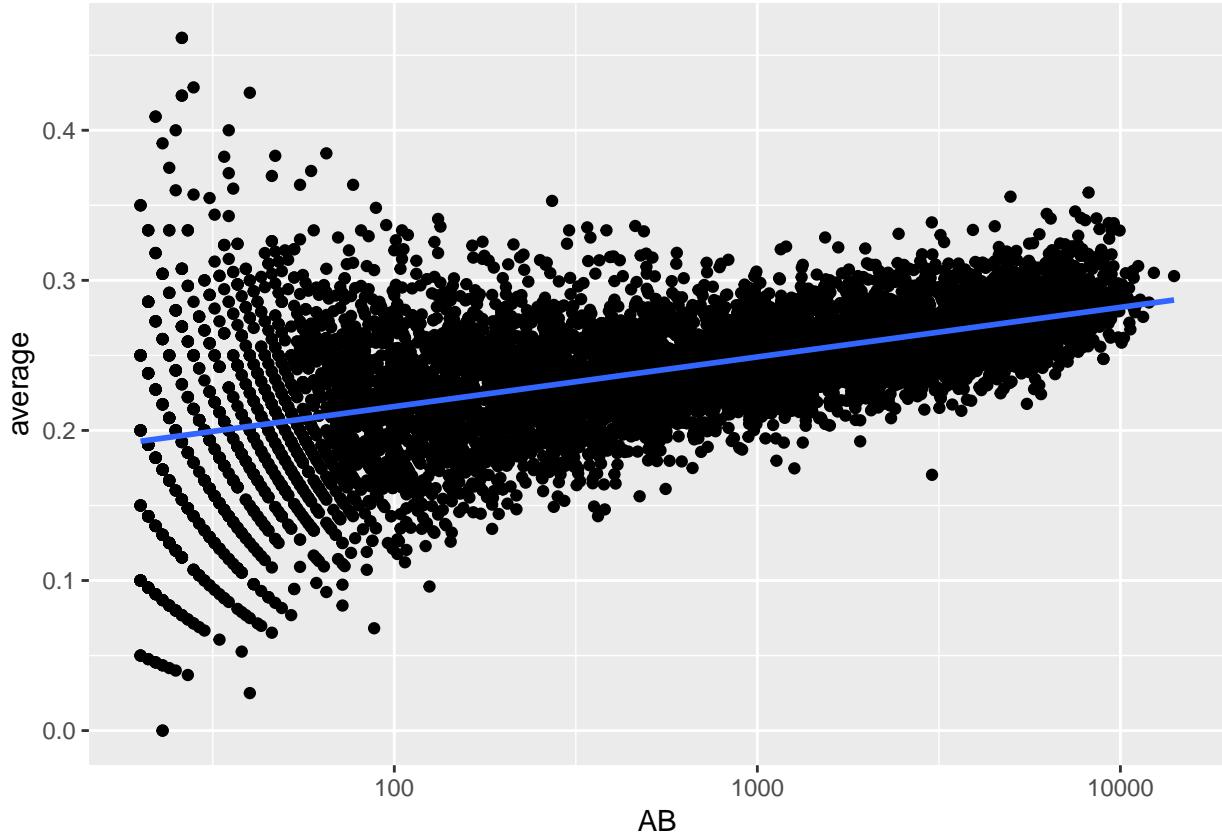
## 10.13 影响因子

- 击球率高除了能力影响外还有可能是因为得到的机会多或者光环效应，例如一开始凭运气打得好，后面给机会多，通过经验累积提高了击球率

```

career %>%
  filter(AB >= 20) %>%
  ggplot(aes(AB, average)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  scale_x_log10()

```



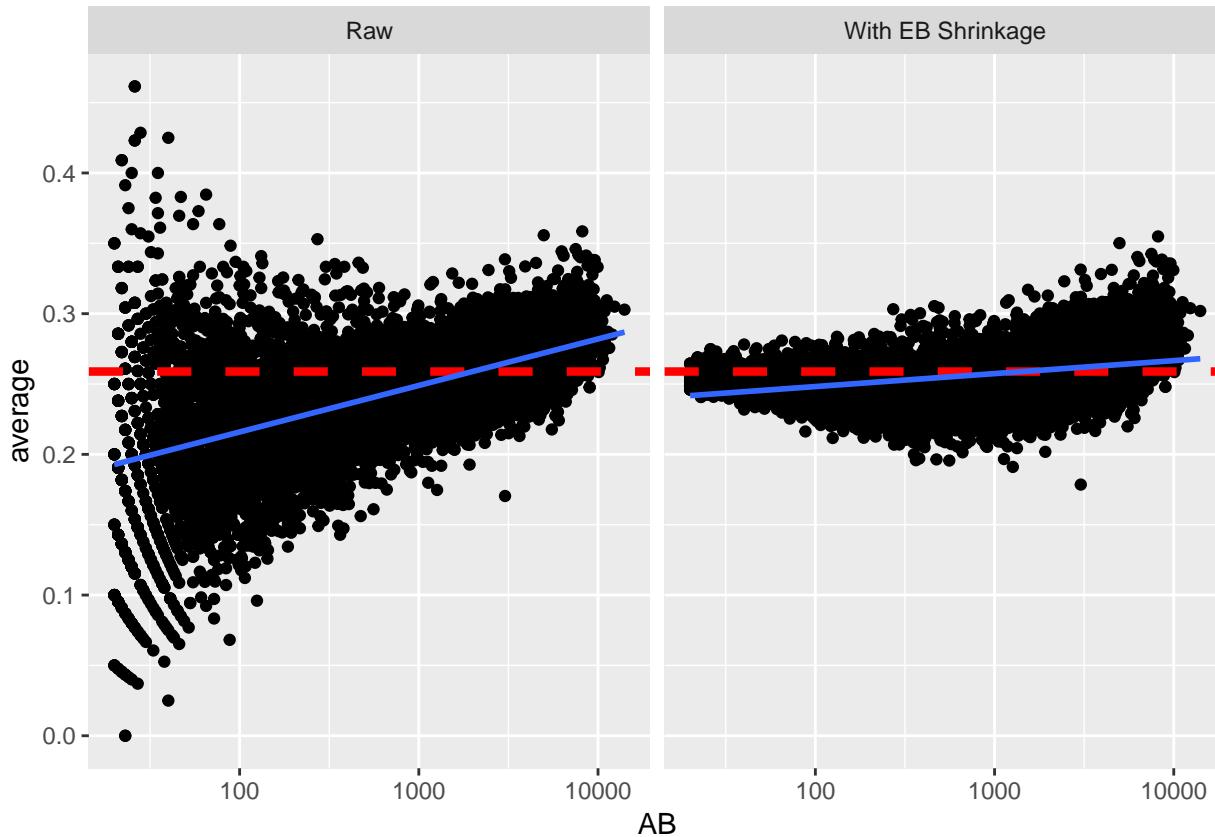
- 击球数低方差会大，这比较正常，很多人挂在起跑线上了

- 直接使用经验贝叶斯方法会导致整体向均值收敛，这高估了新手的数据

```

prior_mu <- alpha0 / (alpha0 + beta0)
career_eb %>%
  filter(AB >= 20) %>%
  gather(type, value, average, eb_estimate) %>%
  mutate(type = plyr::revalue(type, c(average = "Raw",
   eb_estimate = "With EB Shrinkage"))) %>%
  ggplot(aes(AB, value)) +
  geom_point() +
  scale_x_log10() +
  geom_hline(color = "red", lty = 2, size = 1.5, yintercept = prior_mu) +
  facet_wrap(~type) +
  ylab("average") +
  geom_smooth(method = "lm")

```



- 为了如实反应这种情况，我们应该认为击球率符合贝塔分布，但同时贝塔分布的两个参数受击球数的影响，击球数越多，越可能击中
- 这个模型可以用贝塔-二项式回归来描述

$$\mu_i = \mu_0 + \mu_{AB} \cdot \log(AB)$$

$$\alpha_{0,i} = \mu_i / \sigma_0$$

$$\beta_{0,i} = (1 - \mu_i) / \sigma_0$$

$$p_i \sim \text{Beta}(\alpha_{0,i}, \beta_{0,i})$$

$$H_i \sim \text{Binom}(AB_i, p_i)$$

### 10.13.1 拟合模型

- 寻找拟合后的模型参数，构建新的先验概率

```
library(gamlss)
# 拟合模型
fit <- gamlss(cbind(H, AB - H) ~ log(AB),
               data = career_eb,
               family = BB(mu.link = "identity"))
```

```
## GAMLSS-RS iteration 1: Global Deviance = 92881
## GAMLSS-RS iteration 2: Global Deviance = 73483
## GAMLSS-RS iteration 3: Global Deviance = 69321
## GAMLSS-RS iteration 4: Global Deviance = 69315
## GAMLSS-RS iteration 5: Global Deviance = 69315
```

# 展示拟合参数

```
td <- tidy(fit)
td
```

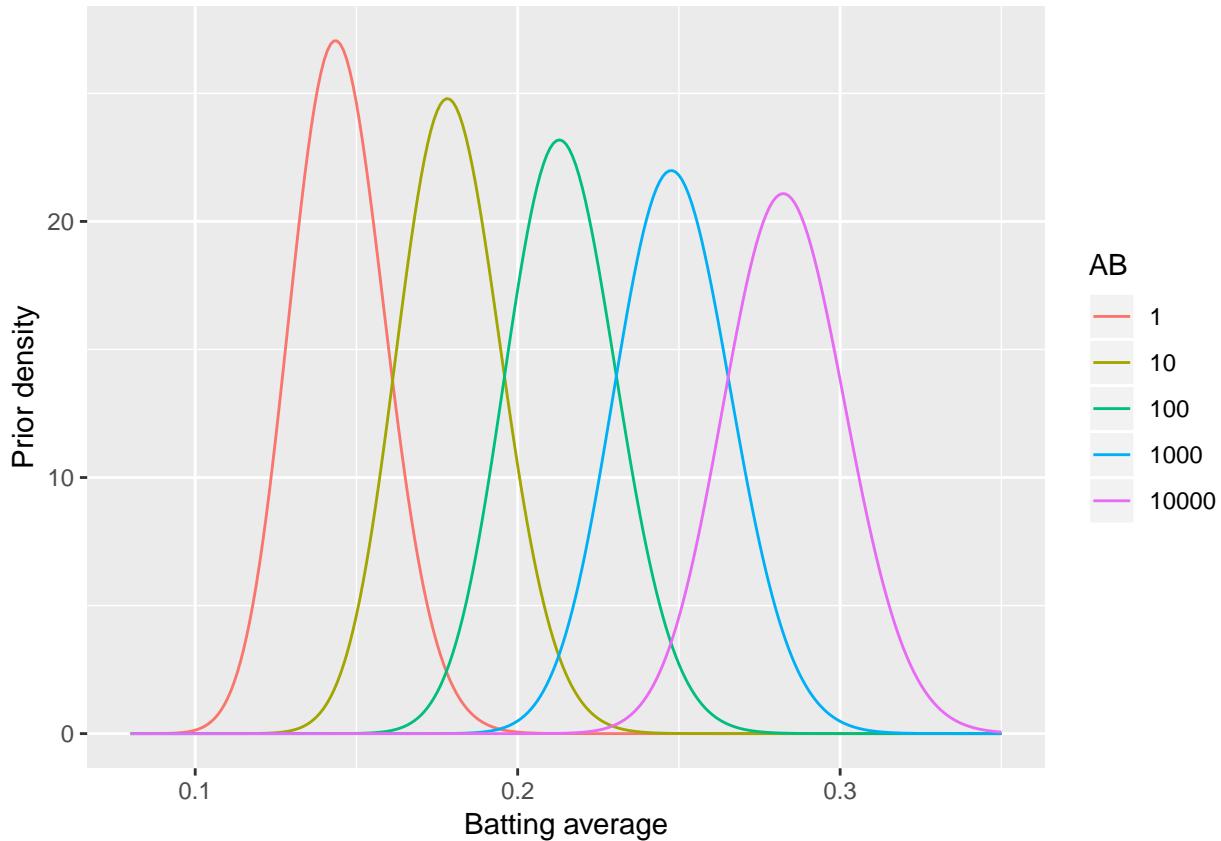
```
## # A tibble: 3 x 6
##   parameter term      estimate std.error statistic p.value
##   <chr>     <chr>      <dbl>     <dbl>     <dbl>    <dbl>
## 1 mu        (Intercept)  0.145    0.00160    90.5     0
## 2 mu        log(AB)      0.0150   0.000218    68.8     0
## 3 sigma     (Intercept) -6.34     0.0247   -257.     0
```

# 构建新的先验概率

```
mu_0 <- td$estimate[1]
mu_AB <- td$estimate[2]
sigma <- exp(td$estimate[3])
```

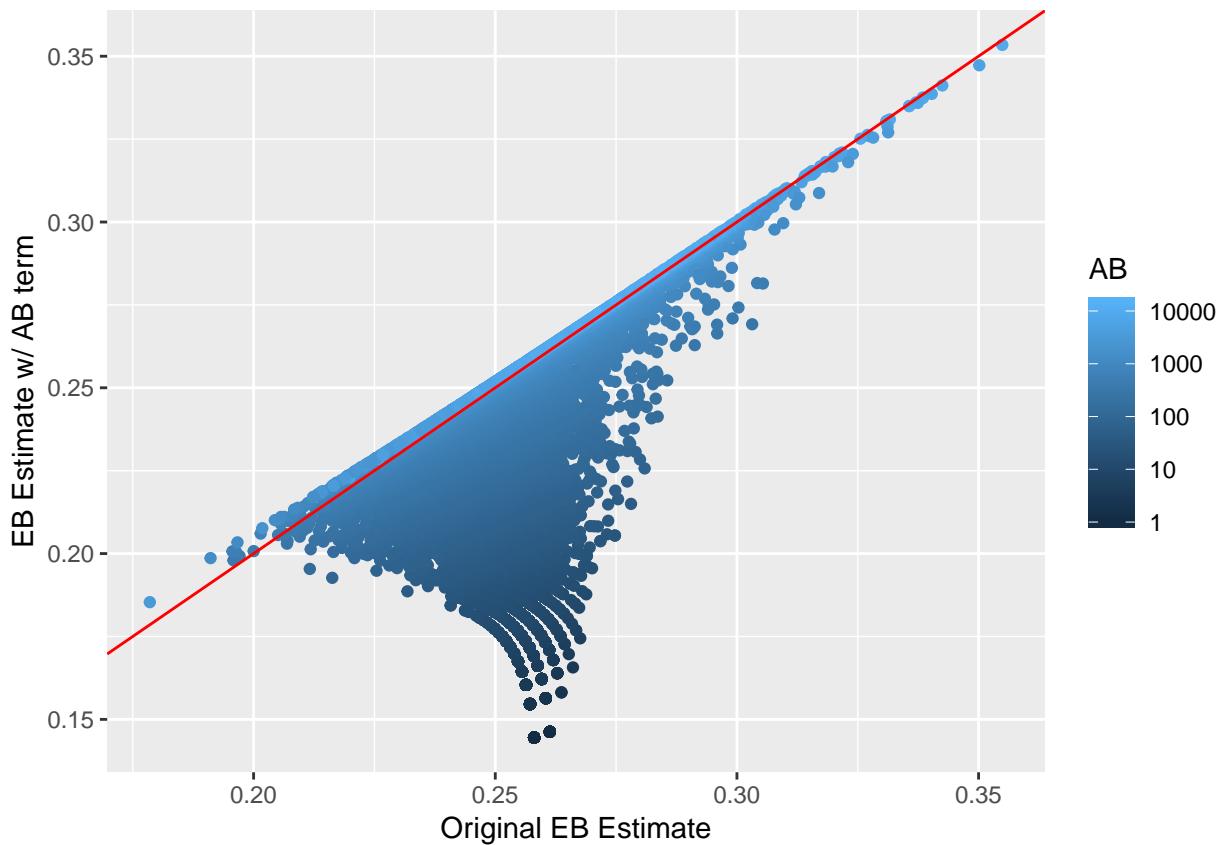
# 看看 AB 对先验概率的影响

```
crossing(x = seq(0.08, .35, .001), AB = c(1, 10, 100, 1000, 10000)) %>%
  mutate(density = dbeta(x, (mu_0 + mu_AB * log(AB)) / sigma,
                         (1 - (mu_0 + mu_AB * log(AB))) / sigma)) %>%
  mutate(AB = factor(AB)) %>%
  ggplot(aes(x, density, color = AB, group = AB)) +
  geom_line() +
  xlab("Batting average") +
  ylab("Prior density")
```



### 10.13.2 求后验概率

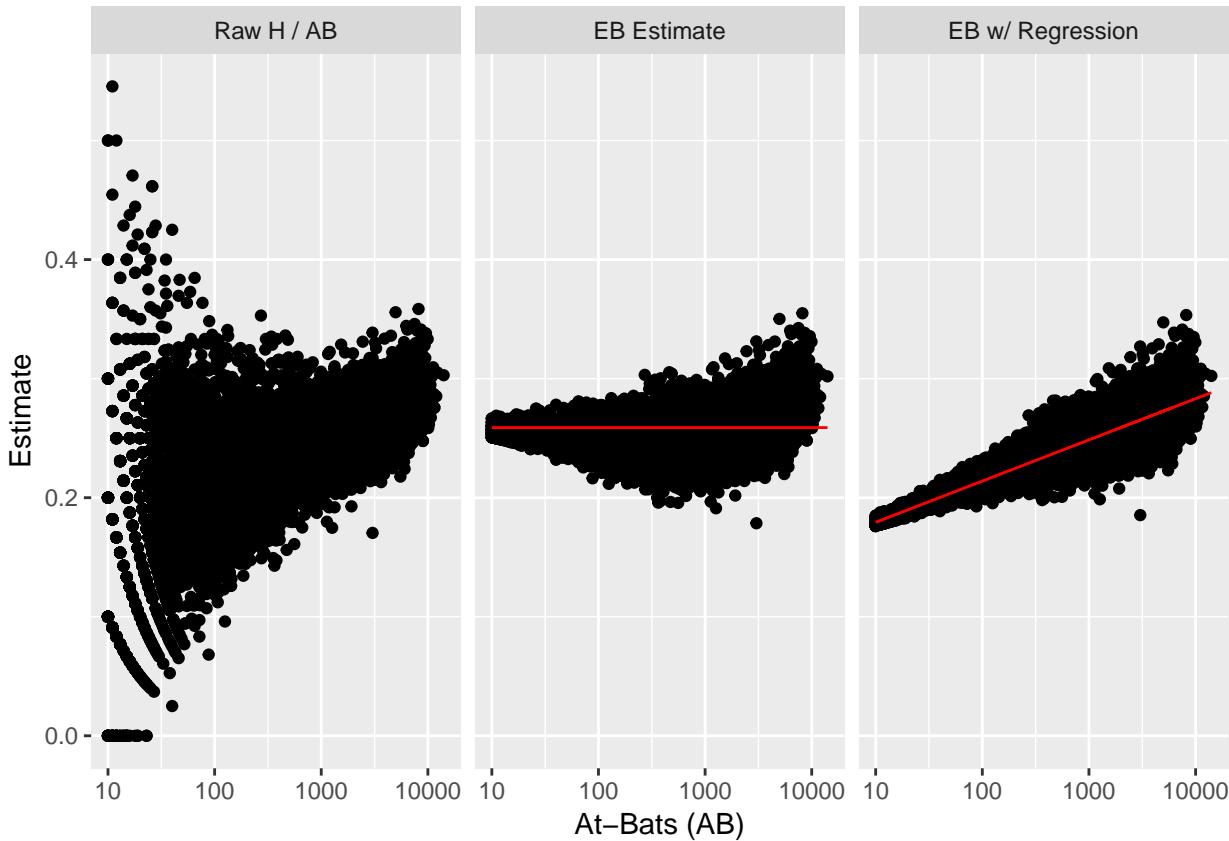
```
# 计算所有拟合值
mu <- fitted(fit, parameter = "mu")
sigma <- fitted(fit, parameter = "sigma")
# 计算所有后验概率
career_eb_wAB <- career_eb %>%
  dplyr::select(name, H, AB, original_eb = eb_estimate) %>%
  mutate(mu = mu,
        alpha0 = mu / sigma,
        beta0 = (1 - mu) / sigma,
        alpha1 = alpha0 + H,
        beta1 = beta0 + AB - H,
        new_eb = alpha1 / (alpha1 + beta1))
# 展示拟合后的击球率
ggplot(career_eb_wAB, aes(original_eb, new_eb, color = AB)) +
  geom_point() +
  geom_abline(color = "red") +
  xlab("Original EB Estimate") +
  ylab("EB Estimate w/ AB term") +
  scale_color_continuous(trans = "log", breaks = 10^(0:4))
```



```
# 对比
library(tidyr)

lev <- c(raw = "Raw H / AB", original_eb = "EB Estimate", new_eb = "EB w/ Regression")

career_eb_wAB %>%
  filter(AB >= 10) %>%
  mutate(raw = H / AB) %>%
  gather(type, value, raw, original_eb, new_eb) %>%
  mutate(mu = ifelse(type == "original_eb", prior_mu,
                     ifelse(type == "new_eb", mu, NA))) %>%
  mutate(type = factor(plyr::revalue(type, lev), lev)) %>%
  ggplot(aes(AB, value)) +
  geom_point() +
  geom_line(aes(y = mu), color = "red") +
  scale_x_log10() +
  facet_wrap(~type) +
  xlab("At-Bats (AB)") +
  ylab("Estimate")
```



- 纠正后我们的数据更复合现实了，其实这是贝叶斯分层模型的一个简单版本，通过考虑更多因素，我们可以构建更复杂的模型来挖掘出我们所需要的信息

### 10.13.3 考虑更多因素

- 现在我们听说左利手跟右利手的表现可能不一样，所以我们要对模型进行完善，考虑把左右手参数加入模型

```
# 展示数据
career2 %>%
  count(bats)

## # A tibble: 4 x 2
##   bats      n
##   <fct> <int>
## 1 B        777
## 2 L       2680
## 3 R       5397
## 4 <NA>     655

# 排除 NA
career3 <- career2 %>%
  filter(!is.na(bats)) %>%
  mutate(bats = relevel(bats, "R"))

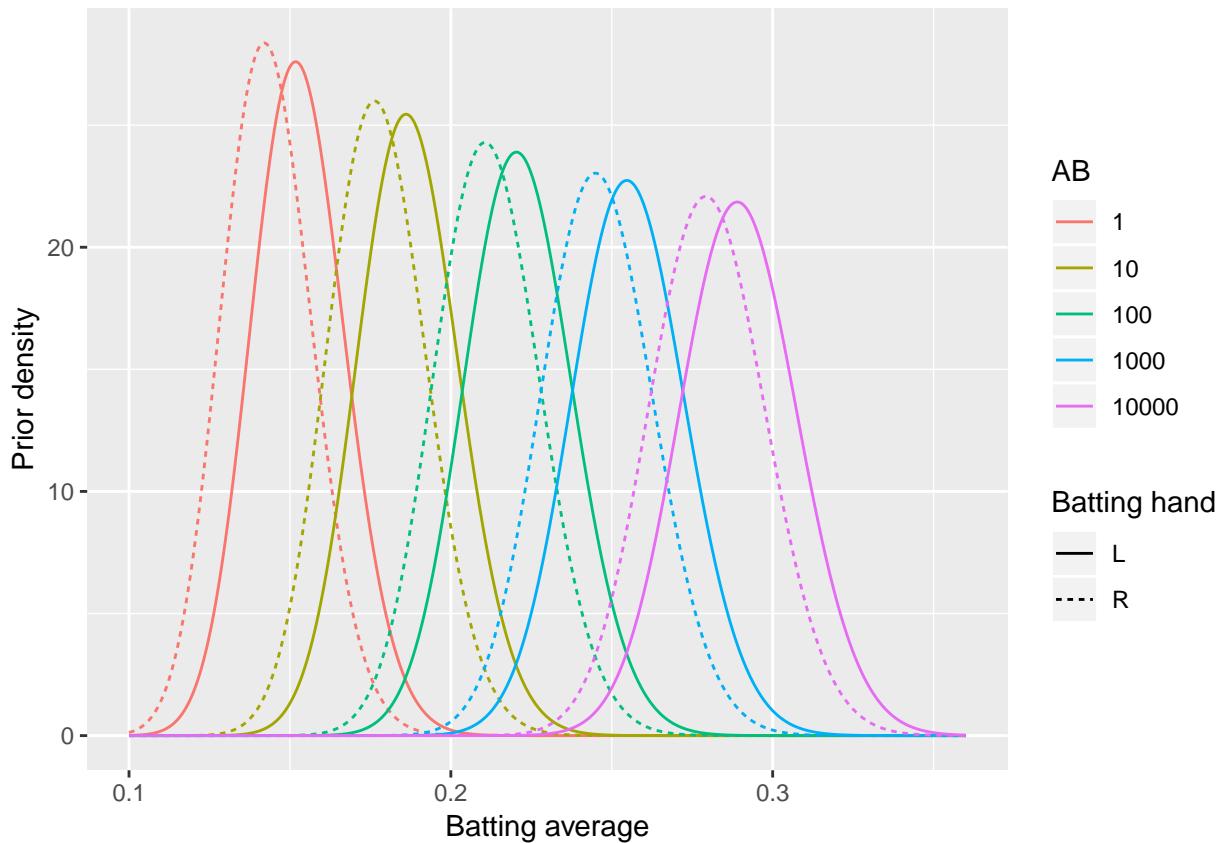
# 重建模型
fit2 <- gamlss(cbind(H, AB - H) ~ log(AB) + bats,
                data = career3,
                family = BB(mu.link = "identity"))
```

```
## GAMLSS-RS iteration 1: Global Deviance = 89139
## GAMLSS-RS iteration 2: Global Deviance = 70275
## GAMLSS-RS iteration 3: Global Deviance = 66100
## GAMLSS-RS iteration 4: Global Deviance = 66094
## GAMLSS-RS iteration 5: Global Deviance = 66094

# 观察参数
tidy(fit2)

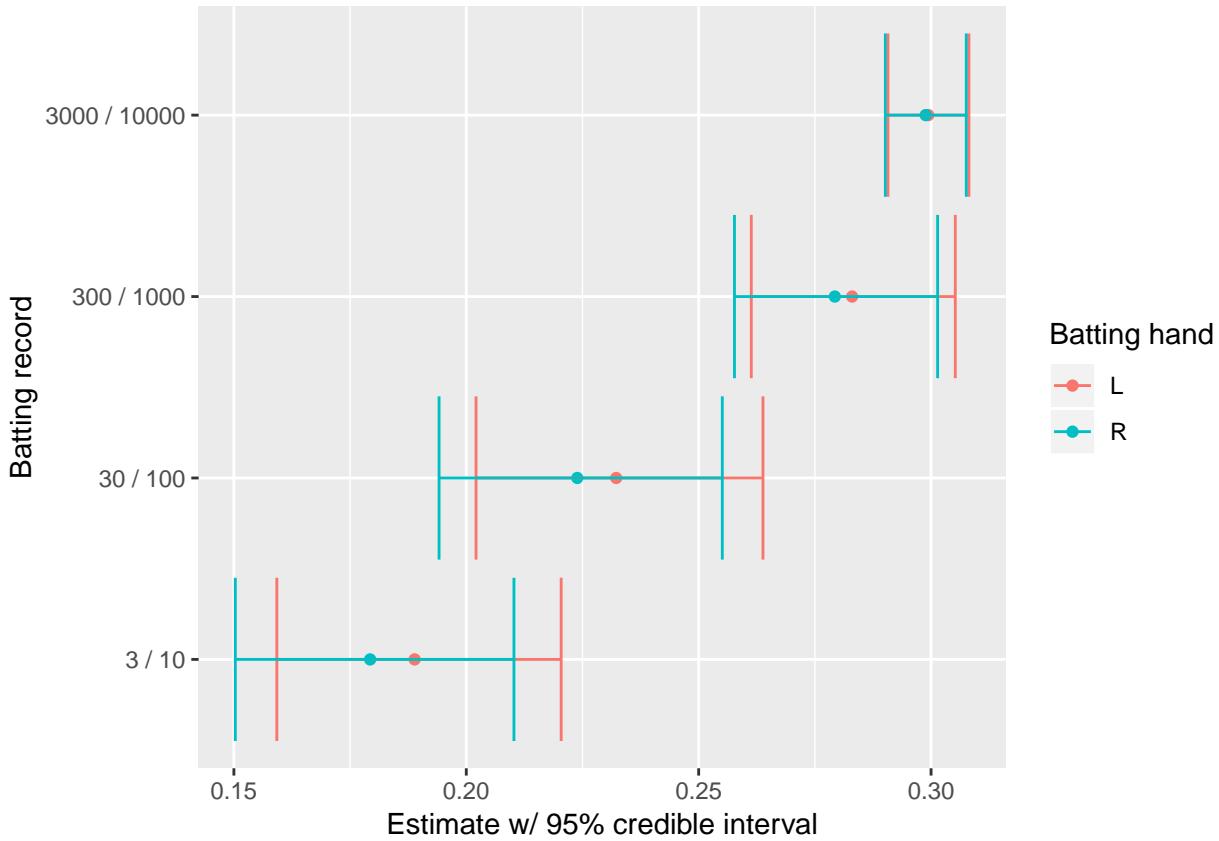
## # A tibble: 5 x 6
##   parameter term      estimate std.error statistic p.value
##   <chr>     <chr>      <dbl>     <dbl>     <dbl>    <dbl>
## 1 mu        (Intercept)  0.143    0.00164    87.2    0.
## 2 mu        log(AB)      0.0149   0.000222   67.0    0.
## 3 mu        batsB       -0.00138  0.000991   -1.39   1.65e- 1
## 4 mu        batsL       0.00973  0.000637   15.3    5.03e-52
## 5 sigma     (Intercept) -6.43     0.0252    -255.    0.

sigma <- fitted(fit2, "sigma")[1]
crossing(bats = c("L", "R"),
         AB = c(1, 10, 100, 1000, 10000)) %>%
  augment(fit2, newdata = .) %>%
  rename(mu = .fitted) %>%
  crossing(x = seq(.1, .36, .0005)) %>%
  mutate(alpha = mu / sigma,
         beta = (1 - mu) / sigma,
         density = dbeta(x, alpha, beta)) %>%
  ggplot(aes(x, density, color = factor(AB), lty = bats)) +
  geom_line() +
  labs(x = "Batting average",
       y = "Prior density",
       color = "AB",
       lty = "Batting hand")
```



- 存在先验概率的情况下，可以考虑考察随着击球数增长左右手的不同

```
crossing(bats = c("L", "R"),
         AB = c(10, 100, 1000, 10000)) %>%
  augment(fit2, newdata = .) %>%
  mutate(H = .3 * AB,
        alpha0 = .fitted / sigma,
        beta0 = (1 - .fitted) / sigma,
        alpha1 = alpha0 + H,
        beta1 = beta0 + AB - H,
        estimate = alpha1 / (alpha1 + beta1),
        conf.low = qbeta(.025, alpha1, beta1),
        conf.high = qbeta(.975, alpha1, beta1),
        record = paste(H, AB, sep = " / ")) %>%
  ggplot(aes(estimate, record, color = bats)) +
  geom_point() +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high)) +
  labs(x = "Estimate w/ 95% credible interval",
       y = "Batting record",
       color = "Batting hand")
```



- 另一个要考虑的因素是不同年份的平均击球率可能也有起伏

```

career3 %>%
  mutate(decade = factor(round(year - 5, -1))) %>%
  filter(AB >= 500) %>%
  ggplot(aes(decade, average)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ylab("Batting average")
# 用样条插值来进行拟合
library(splines)
fit3 <- gamlss(cbind(H, AB - H) ~ 0 + ns(year, df = 5) + bats + log(AB),
  data = career3,
  family = BB(mu.link = "identity"))

# 观察在击球数 1000 上先验概率的变化
plot_gamlss_fit <- function(f) {
  career3 %>%
    dplyr::select(year, bats) %>%
    distinct() %>%
    filter(bats != "B") %>%
    mutate(AB = 1000) %>%
    augment(f, newdata = .) %>%
    rename(mu = .fitted) %>%
    mutate(sigma = fitted(fit3, "sigma")[1],
      alpha0 = mu / sigma,
      beta0 = (1 - mu) / sigma,
      prior = 1 / (alpha0 + beta0))
}

```

```

    conf_low = qbeta(.025, alpha0, beta0),
    conf_high = qbeta(.975, alpha0, beta0)) %>%
ggplot(aes(year, mu, color = bats, group = bats)) +
  geom_line() +
  geom_ribbon(aes(ymin = conf_low, ymax = conf_high), linetype = 2, alpha = .1) +
  labs(x = "Year",
       y = "Prior distribution (median + 95% quantiles)",
       color = "Batting hand")
}
plot_gamlss_fit(fit3)

```

- 同时另一个问题是这些因素会交互影响

```

fit4 <- gamlss(cbind(H, AB - H) ~ 0 + ns(year, 5) * bats + log(AB),
                 data = career3,
                 family = BB(mu.link = "identity"))
plot_gamlss_fit(fit4)

```

```

Pitching %>%
dplyr::select(playerID, yearID, GS) %>%
distinct() %>%
inner_join(dplyr::select(Master, playerID, throws)) %>%
count(yearID, throws, wt = GS) %>%
filter(!is.na(throws)) %>%
mutate(percent = n / sum(n)) %>%
filter(throws == "L") %>%
ggplot(aes(yearID, percent)) +
  geom_line() +
  geom_smooth() +
  scale_y_continuous(labels = scales::percent_format()) +
  xlab("Year") +
  ylab("% of games with left-handed pitcher")

```

- 左右手之间的差距伴随年份在逐渐减少

```

players <- crossing(year = c(1915, 1965, 2015),
                     bats = c("L", "R"),
                     H = 30,
                     AB = 100)

players_posterior <- players %>%
  mutate(mu = predict(fit4, what = "mu", newdata = players),
         sigma = predict(fit4, what = "sigma", newdata = players, type = "response"),
         alpha0 = mu / sigma,
         beta0 = (1 - mu) / sigma,
         alpha1 = alpha0 + H,
         beta1 = beta0 + AB - H)
players_posterior %>%
  crossing(x = seq(.15, .3, .001)) %>%
  mutate(density = dbeta(x, alpha1, beta1)) %>%
  ggplot(aes(x, density, color = bats)) +
  geom_line() +
  facet_wrap(~ year) +
  xlab("Batting average") +
  ylab("Posterior density") +

```

```
ggtitle("Posterior distributions for batters with 30 / 100")
```

- 经验贝叶斯对先验概率的估计类似频率学派，但进行的又是贝叶斯分析

## 10.14 混合概率模型

- 用击球概率为例，击球手跟非击球手的概率分布是不一样的，那么实际看到的总体球员概率分布应该是一个混合在一起的两个独立分布

```
# 找出投球 3 次以上的人
pitchers <- Pitching %>%
  group_by(playerID) %>%
  summarize(gamesPitched = sum(G)) %>%
  filter(gamesPitched > 3)

# 参考上一章节的发现找出击球率稳定的选手

career <- Batting %>%
  filter(AB > 0, lgID == "NL", yearID >= 1980) %>%
  group_by(playerID) %>%
  summarize(H = sum(H), AB = sum(AB), year = mean(yearID)) %>%
  mutate(average = H / AB,
    isPitcher = playerID %in% pitchers$playerID)

# 链接上名字
career <- Master %>%
 tbl_df() %>%
  dplyr::select(playerID, nameFirst, nameLast, bats) %>%
  unite(name, nameFirst, nameLast, sep = " ") %>%
  inner_join(career, by = "playerID")
```

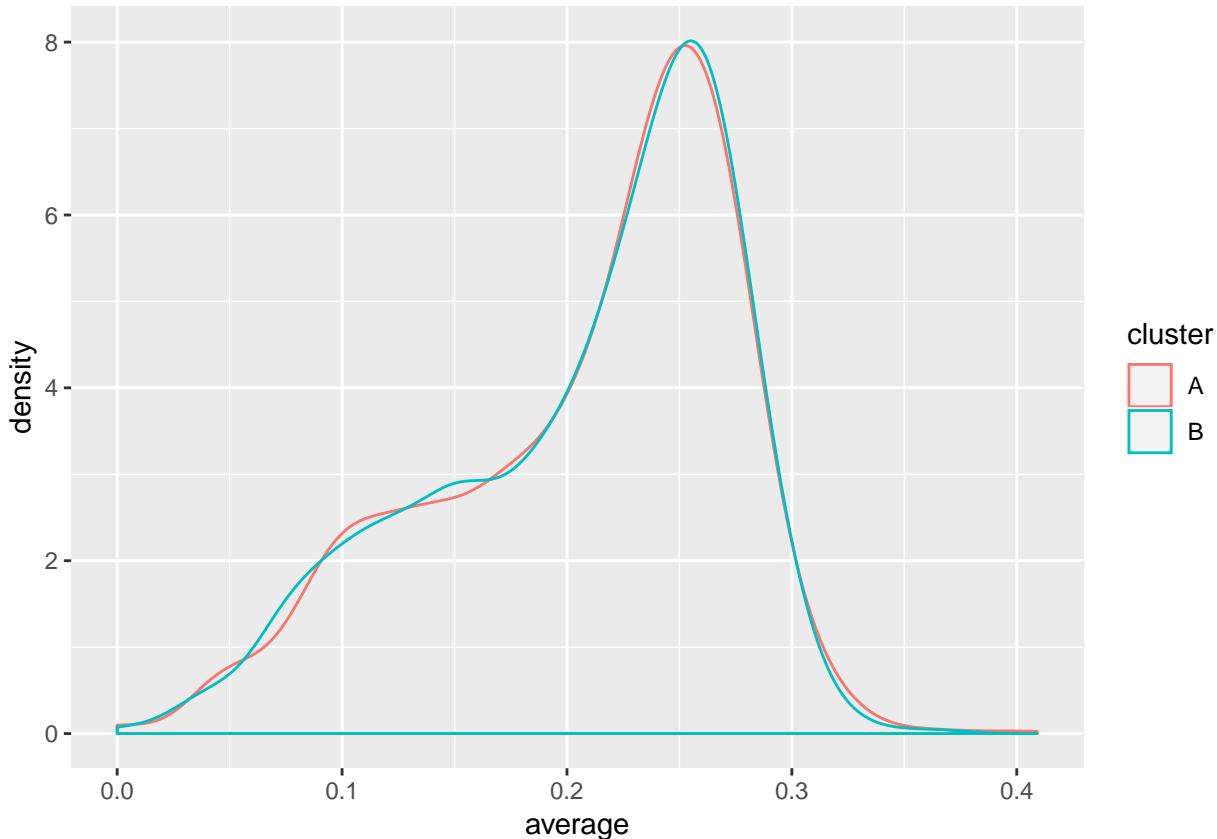
### 10.14.1 期望最大算法

- 将一个分布拆成两个，可以使用期望最大算法

```
set.seed(2017)

# 先随机分为两组
starting_data <- career %>%
  filter(AB >= 20) %>%
  dplyr::select(-year, -bats, -isPitcher) %>%
  mutate(cluster = factor(sample(c("A", "B"), n(), replace = TRUE)))

# 观察效果
starting_data %>%
  ggplot(aes(average, color = cluster)) +
  geom_density()
```



```
library(VGAM)
fit_bb_mle <- function(x, n) {
  # dbetabinom.ab 是用 n、alpha 与 beta 作为参数的二项贝塔分布的似然度函数
  ll <- function(alpha, beta) {
    -sum(dbetabinom.ab(x, n, alpha, beta, log = TRUE))
  }
  m <- stats4::mle(ll, start = list(alpha = 3, beta = 10), method = "L-BFGS-B",
                    lower = c(0.001, .001))
  ab <- stats4::coef(m)
  data_frame(alpha = ab[1], beta = ab[2], number = length(x))
}
# 看下初始参数
fit_bb_mle(starting_data$H, starting_data$AB)
```

```
## # A tibble: 1 x 3
##   alpha   beta number
##   <dbl> <dbl>  <int>
## 1  12.7  45.1   3420
# 看下随机分拆后的参数并生成各分组样本数的先验概率
fits <- starting_data %>%
  group_by(cluster) %>%
  do(fit_bb_mle(. $H, . $AB)) %>%
  ungroup() %>%
  mutate(prior = number / sum(number))

fits
```

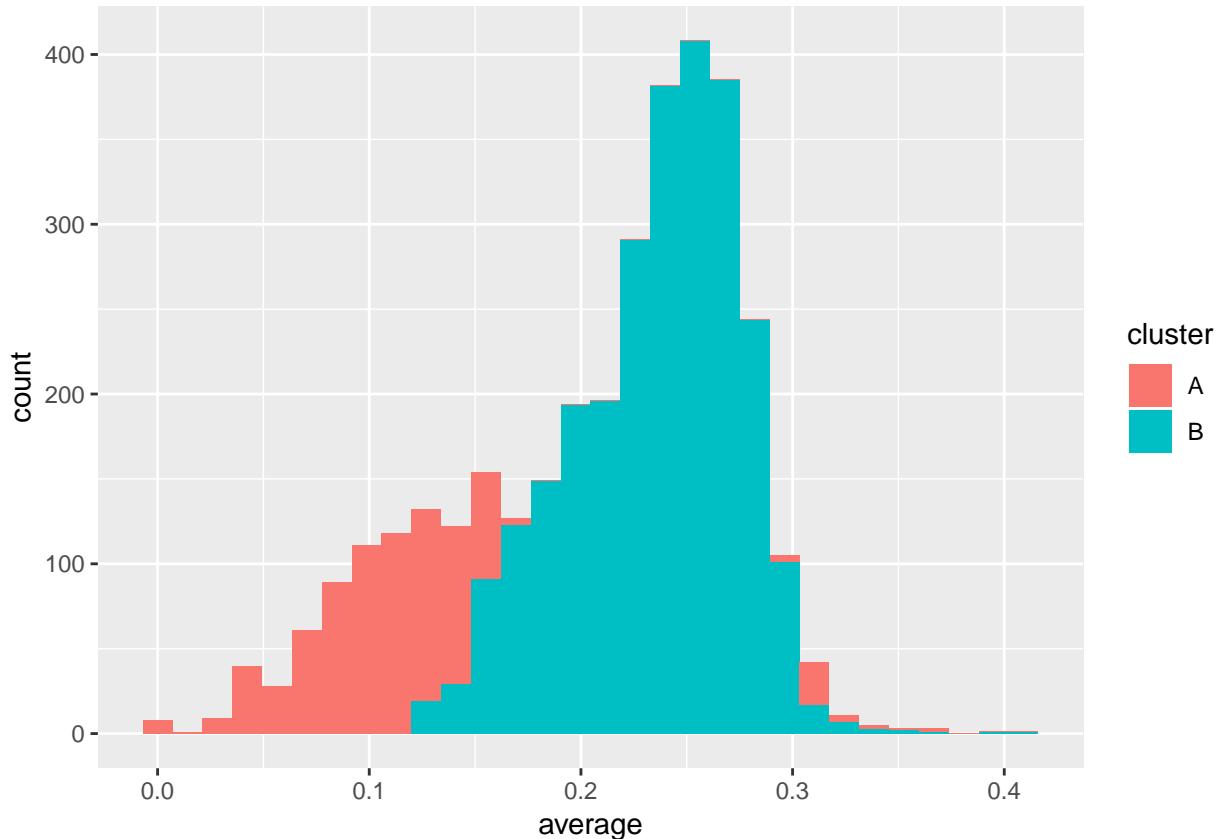
```
## # A tibble: 2 x 5
##   cluster alpha beta number prior
##   <fct>    <dbl> <dbl> <int> <dbl>
## 1 A        12.4  44.3  1704 0.498
## 2 B        12.9  46.0  1716 0.502
```

- 算法优化的期望是将这两个分布分拆开，前面一次分拆已经产生微弱差异，下面就通过贝叶斯思想对数据更新这个差异重新分组让两者分开

```
assignments <- starting_data %>%
  dplyr::select(-cluster) %>%
  crossing(fits) %>%
  mutate(likelihood = prior * VGAM::dbetabinom.ab(H, AB, alpha, beta)) %>%
  group_by(playerID) %>%
  top_n(1, likelihood) %>%
  ungroup()
# 去除掉原有分组，根据更新后的后验概率重新分组
assignments
```

```
## # A tibble: 3,420 x 11
##   playerID name     H     AB average cluster alpha  beta number prior
##   <chr>     <chr> <int> <int>    <dbl> <fct>    <dbl> <dbl> <int> <dbl>
## 1 abbotje~ Jeff~    11     42  0.262 B      12.9  46.0  1716 0.502
## 2 abbotji~ Jim ~    2      21  0.0952 A      12.4  44.3  1704 0.498
## 3 abbotku~ Kurt~   475    1860 0.255 B      12.9  46.0  1716 0.502
## 4 abbotky~ Kyle~   3      31  0.0968 A      12.4  44.3  1704 0.498
## 5 abercre~ Regg~   86     386 0.223 B      12.9  46.0  1716 0.502
## 6 abnersh~ Shaw~  110    531 0.207 B      12.9  46.0  1716 0.502
## 7 abreubo~ Bobb~  1607   5395 0.298 B      12.9  46.0  1716 0.502
## 8 abreuto~ Tony~  129    509 0.253 B      12.9  46.0  1716 0.502
## 9 acevejo~ Jose~   8     101  0.0792 A      12.4  44.3  1704 0.498
## 10 aceveju~ Juan~  6      65  0.0923 A     12.4  44.3  1704 0.498
## # ... with 3,410 more rows, and 1 more variable: likelihood <dbl>
```

```
# 观察更新后概率分布
ggplot(assignments, aes(average, fill = cluster)) +
  geom_histogram()
```



- 不断重复这个过程，最终分拆数据（其实就是第一步分拆最重要，后面直接收敛了）

```
set.seed(1987)

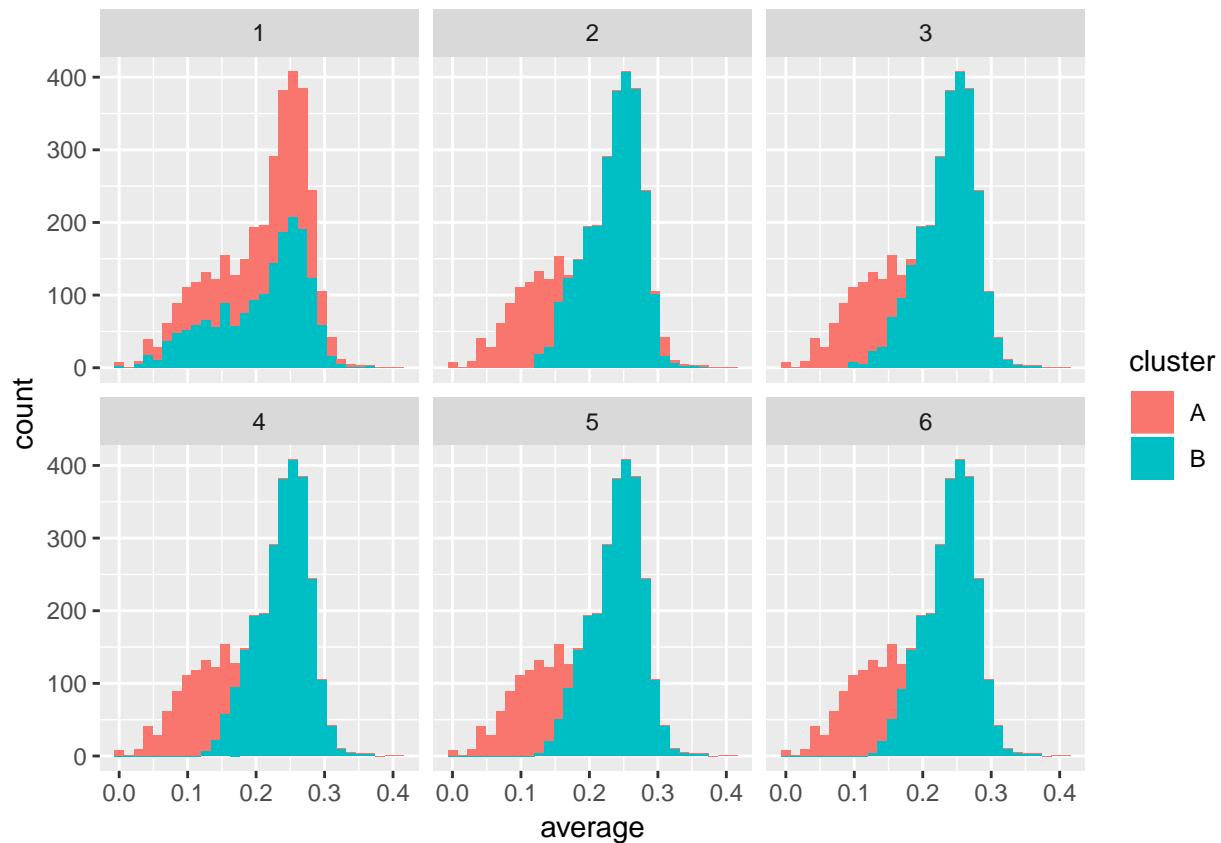
iterate_em <- function(state, ...) {
  fits <- state$assignments %>%
    group_by(cluster) %>%
    do(mutate(fit_bb_mle(. $H, . $AB), number = nrow(.))) %>%
    ungroup() %>%
    mutate(prior = number / sum(number))

  assignments <- assignments %>%
    dplyr::select(playerID:average) %>%
    crossing(fits) %>%
    mutate(likelihood = prior * VGAM::dbetabinom.ab(H, AB, alpha, beta)) %>%
    group_by(playerID) %>%
    top_n(1, likelihood) %>%
    ungroup()

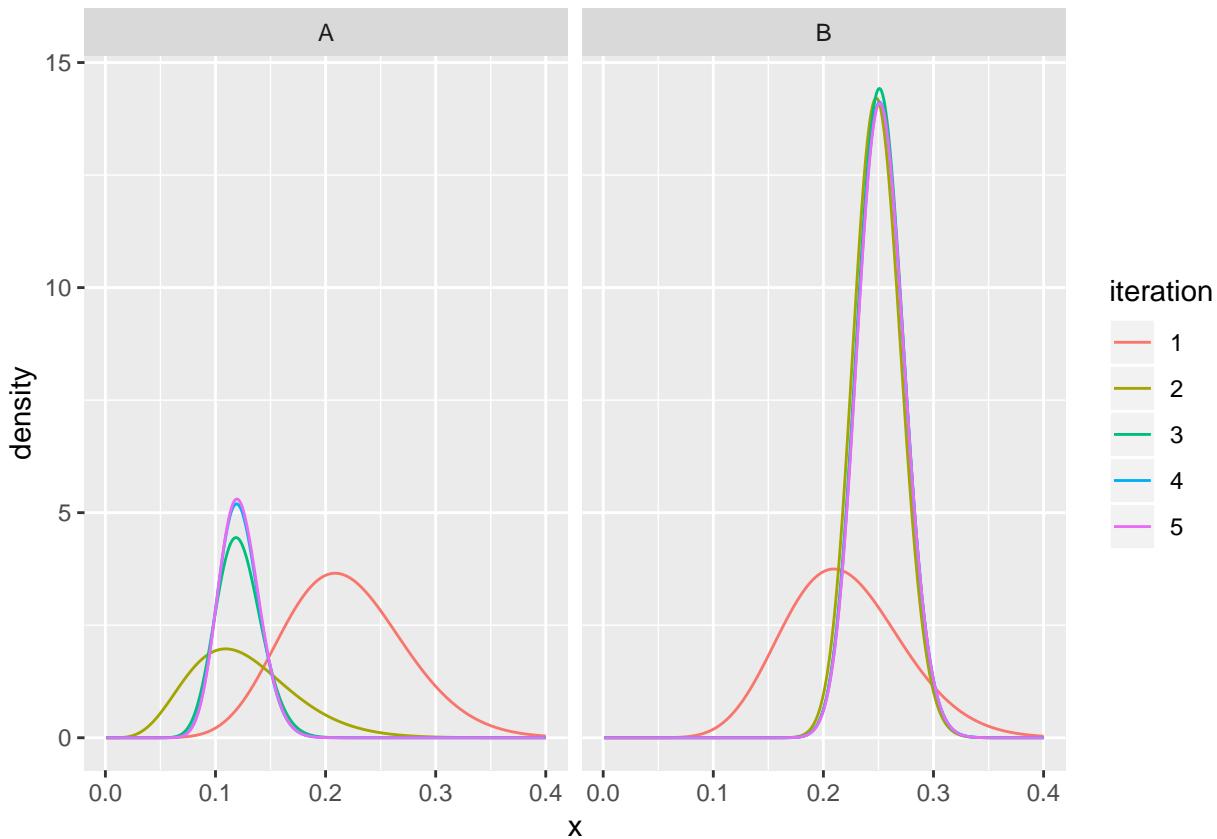
  list(assignments = assignments,
       fits = fits)
}

library(purrr)
# 使用 purrr 包存储中间结果
iterations <- accumulate(1:5, iterate_em, .init = list(assignments = starting_data))
```

```
assignment_iterations <- iterations %>%
  map_df("assignments", .id = "iteration")
# 观察收敛过程
assignment_iterations %>%
  ggplot(aes(average, fill = cluster)) +
  geom_histogram() +
  facet_wrap(~ iteration)
```



```
fit_iterations <- iterations %>%
  map_df("fits", .id = "iteration")
# 两个分布的收敛过程
fit_iterations %>%
  crossing(x = seq(.001, .4, .001)) %>%
  mutate(density = prior * dbeta(x, alpha, beta)) %>%
  ggplot(aes(x, density, color = iteration, group = iteration)) +
  geom_line() +
  facet_wrap(~ cluster)
```



### 10.14.2 分配

- 得到每个选手在两个分布中后验概率后要对其进行分配，这里我们认为拆分出的两个分布其实就是是否是击球手的两个分组，由于两组重叠较多，直接分配会有困难

```
# 找 6 个击球数 100 的选手进行分配
batter_100 <- career %>%
  filter(AB == 100) %>%
  arrange(average)
batter_100

## # A tibble: 5 x 8
##   playerID  name      bats     H     AB year average isPitcher
##   <chr>      <chr>    <fct> <int> <int> <dbl>  <dbl> <lgl>
## 1 dejesjo01 Jose de Jesus R       11    100 1990.   0.11 TRUE
## 2 mahonmi02 Mike Mahoney R       18    100 2002.   0.18 FALSE
## 3 cancero01 Robinson Cancel R       20    100 2007.   0.2  FALSE
## 4 buschmi01 Mike Busch     R       22    100 1996.   0.22 FALSE
## 5 shealry01 Ryan Shealy   R       32    100 2006.   0.32 FALSE

# 前面算法得到的最终结果
final_parameters <- fit_iterations %>%
  filter(iteration == max(iteration))

final_parameters

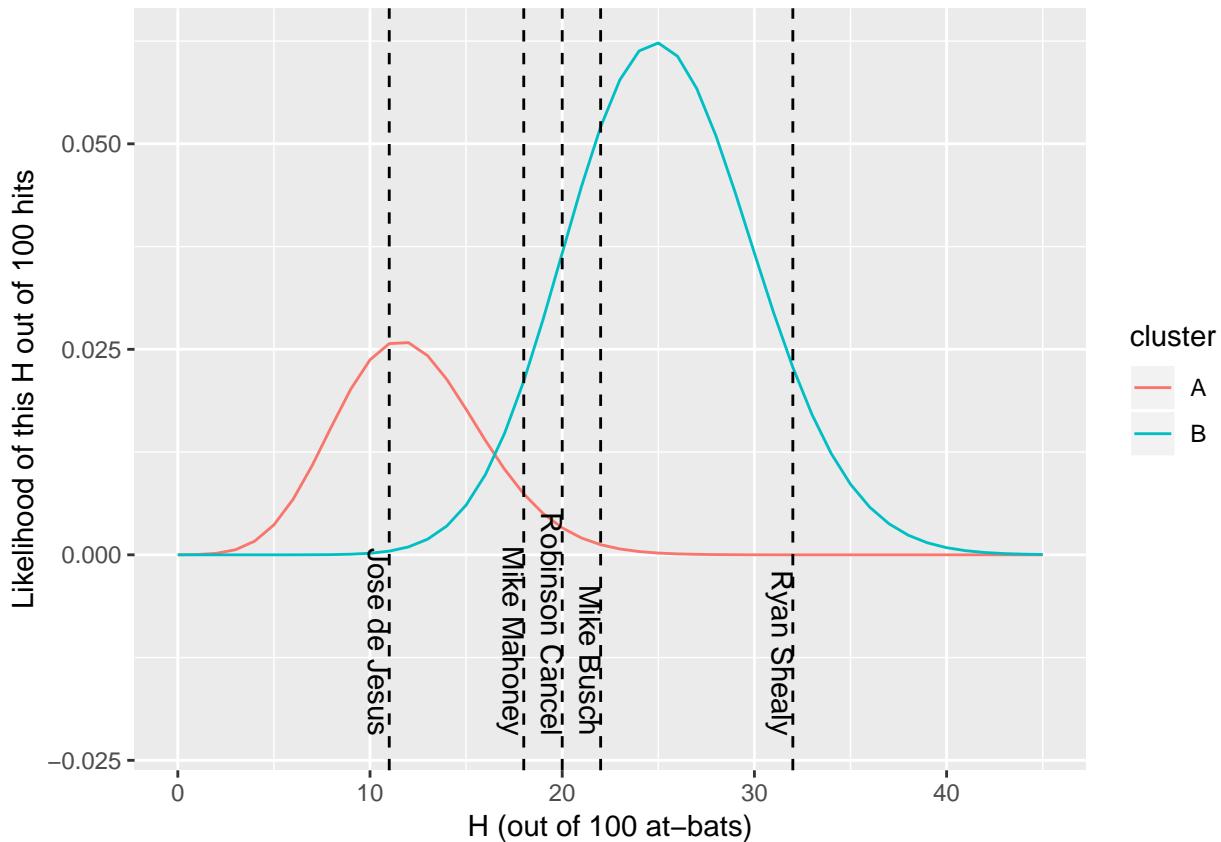
## # A tibble: 2 x 6
##   iteration cluster alpha beta number prior
##       <dbl>   <dbl> <dbl> <dbl>   <dbl> <chr>
```

```

##   <chr>     <fct>    <dbl> <dbl>  <int> <dbl>
## 1 5          A         38.6  278.    831  0.243
## 2 5          B         104.   307.   2589  0.757

# 观察球员位置
final_parameters %>%
  crossing(x = 0:45) %>%
  mutate(density = prior * VGAM::dbetabinom.ab(x, 100, alpha, beta)) %>%
  ggplot(aes(x, density)) +
  geom_line(aes(color = cluster)) +
  geom_vline(aes(xintercept = H), data = batter_100, lty = 2) +
  geom_text(aes(x = H, y = -0.022, label = name), data = batter_100, hjust = 1, vjust = 1, angle = 270) +
  labs(x = "H (out of 100 at-bats)",
       y = "Likelihood of this H out of 100 hits")

```



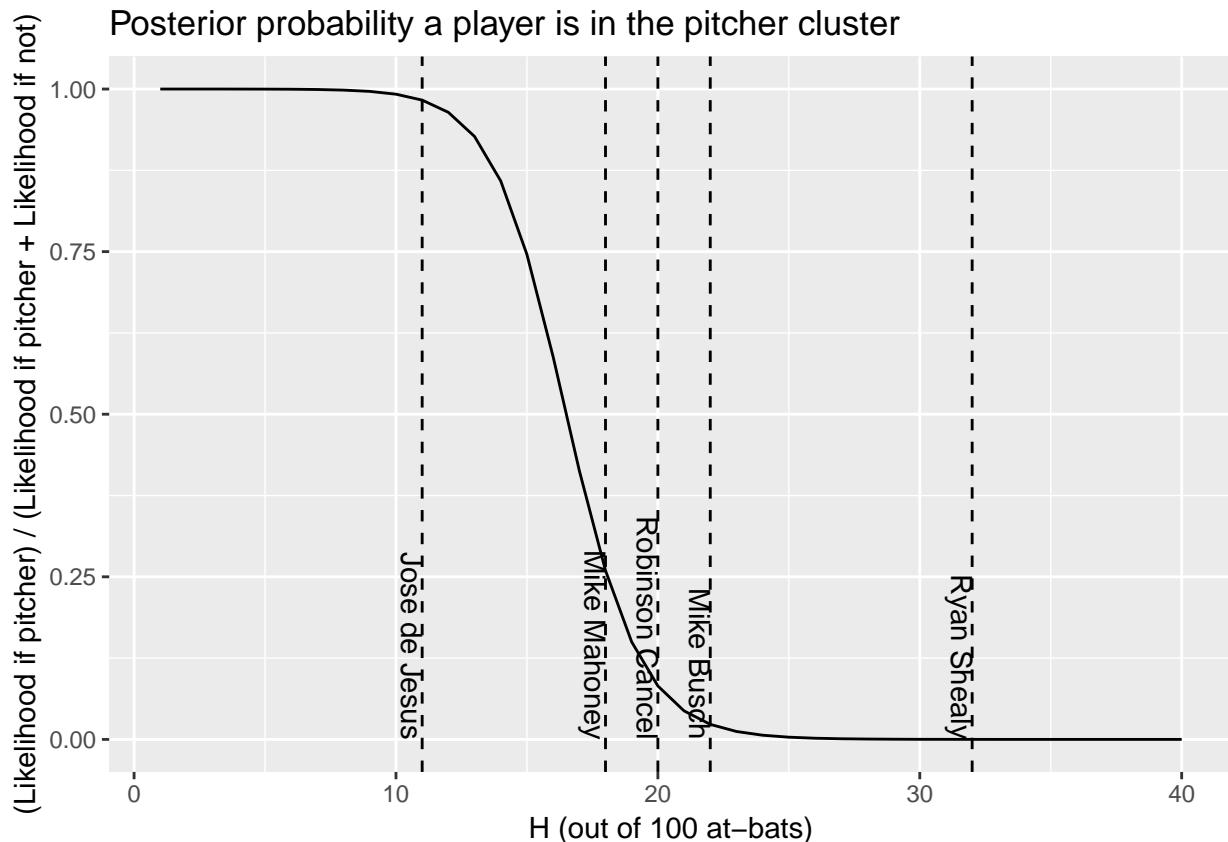
```
# 根据贝叶斯理论，我们可以用在 A 分组的似然度比上两个分组似然度的和得到后验概率
```

```

final_parameters %>%
  crossing(H = 1:40) %>%
  transmute(H, cluster, likelihood = prior * VGAM::dbetabinom.ab(H, 100, alpha, beta)) %>%
  spread(cluster, likelihood) %>%
  mutate(probability_A = A / (A + B)) %>%
  ggplot(aes(H, probability_A)) +
  geom_line() +
  geom_vline(aes(xintercept = H), data = batter_100, lty = 2) +
  geom_text(aes(x = H, y = 0, label = name), data = batter_100, hjust = 1, vjust = 1, angle = 270) +
  labs(x = "H (out of 100 at-bats)",
       y = "Posterior Probability of H")

```

```
y = "(Likelihood if pitcher) / (Likelihood if pitcher + Likelihood if not)",
title = "Posterior probability a player is in the pitcher cluster")
```



- 通过构建后验概率，我们可以直接对结果基于概率进行分组

```
career_likelihoods <- career %>%
  filter(AB > 20) %>%
  crossing(final_parameters) %>%
  mutate(likelihood = prior * VGAM::dbetabinom.ab(H, AB, alpha, beta)) %>%
  group_by(playerID) %>%
  mutate(posterior = likelihood / sum(likelihood))

career_assignments <- career_likelihoods %>%
  top_n(1, posterior) %>%
  ungroup()
# 对比这种分组与实际数据的结果
career_assignments %>%
  filter(posterior > .8) %>%
  count(isPitcher, cluster) %>%
  spread(cluster, n)

## # A tibble: 2 x 3
##   isPitcher     A     B
##   <lgl>     <int> <int>
## 1 FALSE        26   2135
## 2 TRUE         542   160
```

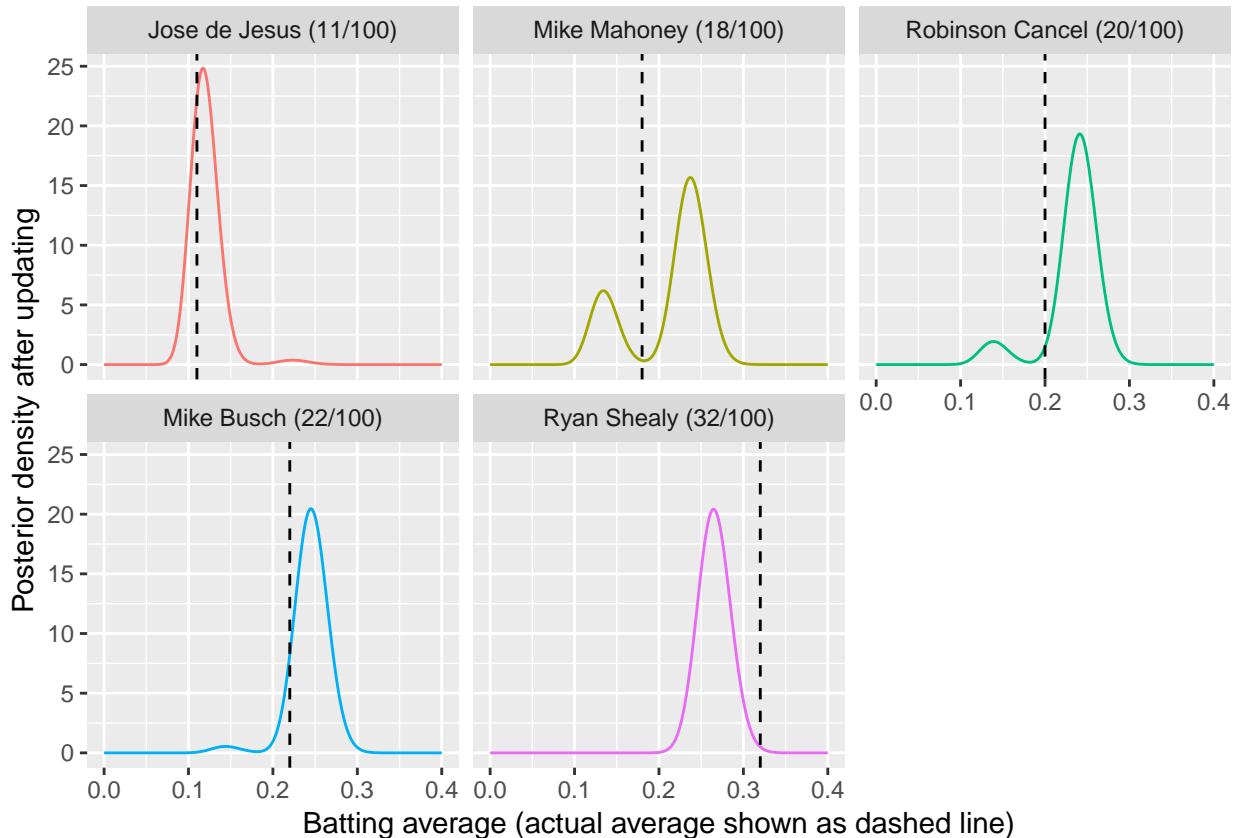
- 这样基于对概率分布的观察，我们可以实现有现实意义的分组，对分组的改进则需要对数据的进一步理解

### 10.14.3 经验贝叶斯收缩

- 混合模型下前面所做的工作都需要重新考虑

```
# 观察击球数 100 选手的后验概率分布
batting_data <- career_likelihoods %>%
  ungroup() %>%
  filter(AB == 100) %>%
  mutate(name = paste0(name, " (", H, "/", AB, ")"),
         name = reorder(name, H),
         alpha1 = H + alpha,
         beta1 = AB - H + beta)

batting_data %>%
  crossing(x = seq(0, .4, .001)) %>%
  mutate(posterior_density = posterior * dbeta(x, alpha1, beta1)) %>%
  group_by(name, x) %>%
  summarize(posterior_density = sum(posterior_density)) %>%
  ggplot(aes(x, posterior_density, color = name)) +
  geom_line(show.legend = FALSE) +
  geom_vline(aes(xintercept = average), data = batting_data, lty = 2) +
  facet_wrap(~ name) +
  labs(x = "Batting average (actual average shown as dashed line)",
       y = "Posterior density after updating")
```

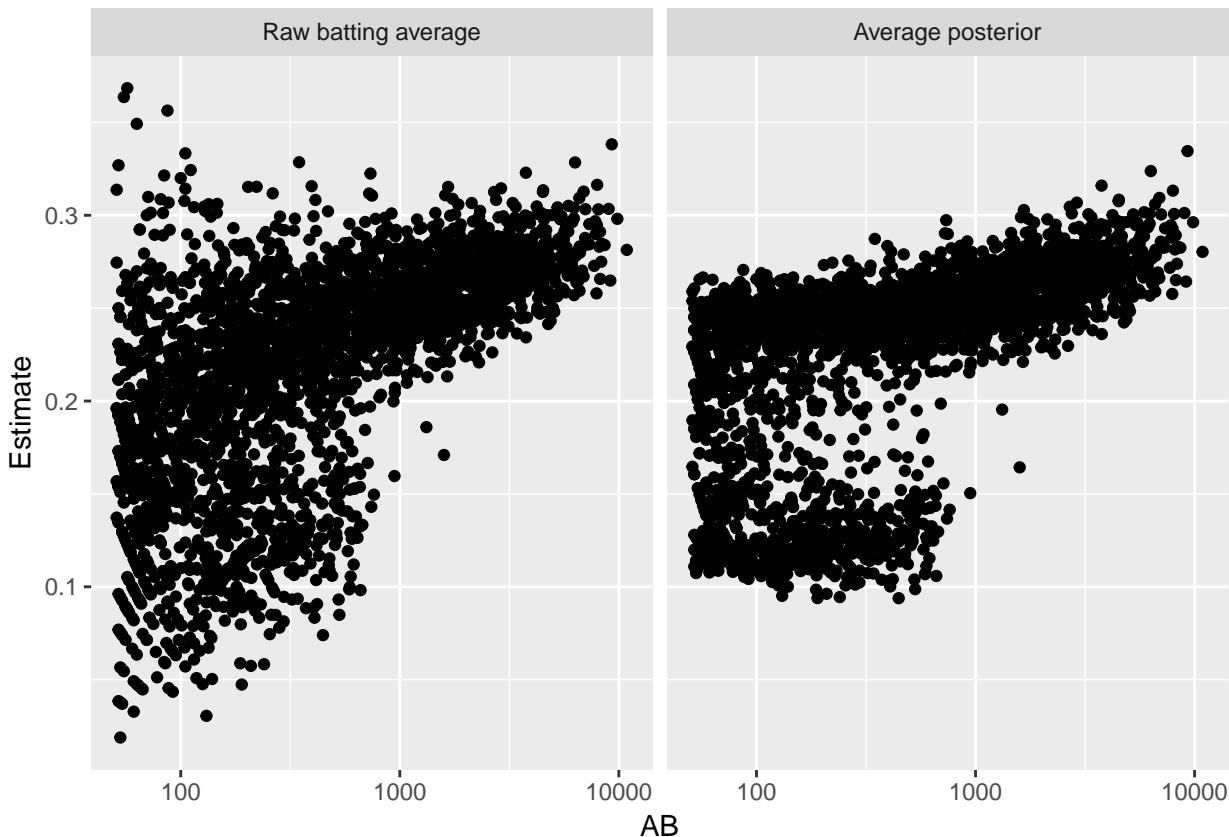


- 此时不太好判断属于哪一分布，可采用后验概率对平均分布进行加权

```

eb_shrinkage <- career_likelihoods %>%
  mutate(shrunken_average = (H + alpha) / (AB + alpha + beta)) %>%
  group_by(playerID) %>%
  summarize(shrunken_average = sum(posterior * shrunken_average))
# 观察加权分布
eb_shrinkage %>%
  inner_join(career) %>%
  filter(AB > 50) %>%
  gather(type, value, average, shrunken_average) %>%
  mutate(type = ifelse(type == "average", "Raw batting average", "Average posterior"),
         type = relevel(factor(type), "Raw batting average")) %>%
  ggplot(aes(AB, value)) +
  geom_point() +
  facet_wrap(~ type) +
  scale_x_log10() +
  ylab("Estimate")

```



收敛后的分布会朝向两个中心而不是一个，并非所有之前的方法（例如区间估计）都可以适用到混合模型里，需要根据实际情况进行分析

## 10.15 模拟验证结果

- 上面的经验贝叶斯推断大都是给出的结果，我们需要对其进行模拟验证

```

pitchers <- Pitching %>%
  group_by(playerID) %>%
  summarize(gamesPitched = sum(G)) %>%

```

```

filter(gamesPitched > 3)
career <- Batting %>%
  filter(AB > 0) %>%
  anti_join(pitchers, by = "playerID") %>%
  group_by(playerID) %>%
  summarize(H = sum(H), AB = sum(AB))
# 从数据中找到贝塔分布的两个参数
library(ebbr)
prior <- career %>%
  ebb_fit_prior(H, AB)

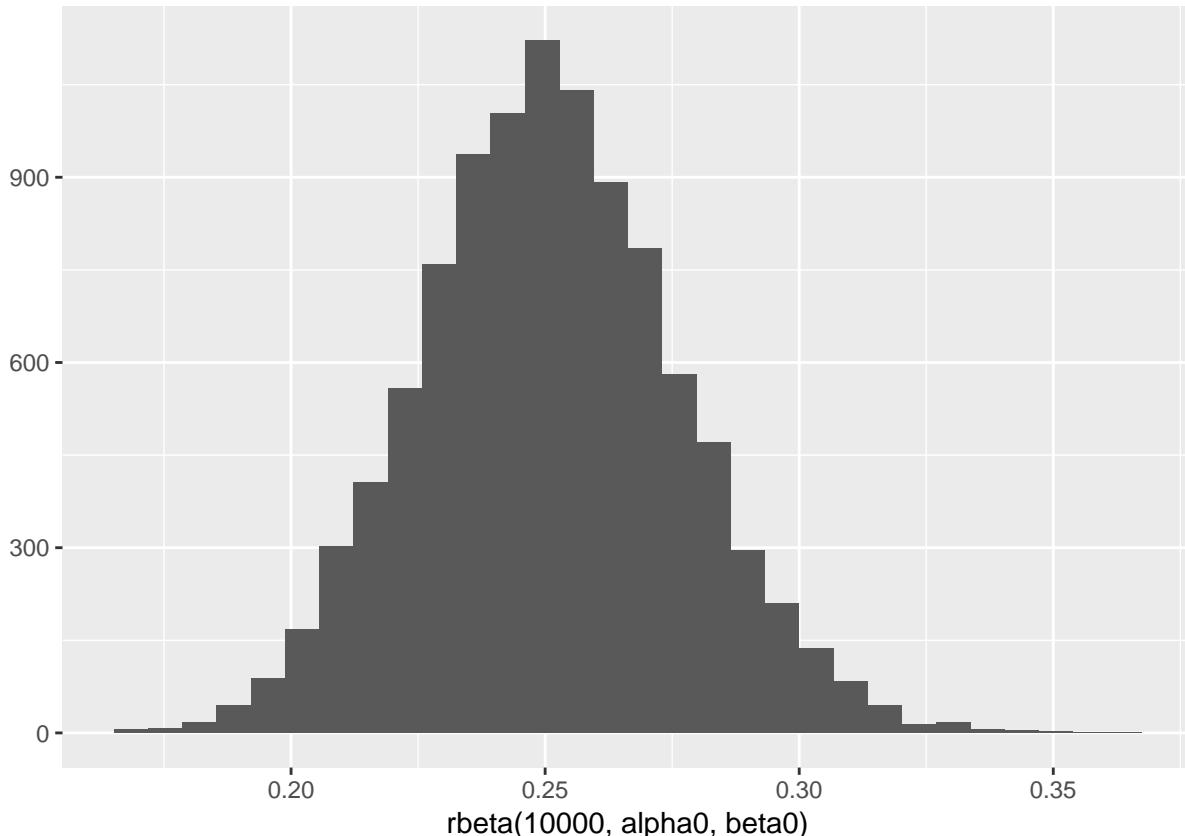
prior

## Empirical Bayes binomial fit with method mle
## Parameters:
## # A tibble: 1 x 2
##   alpha   beta
##   <dbl> <dbl>
## 1    72.7  217.

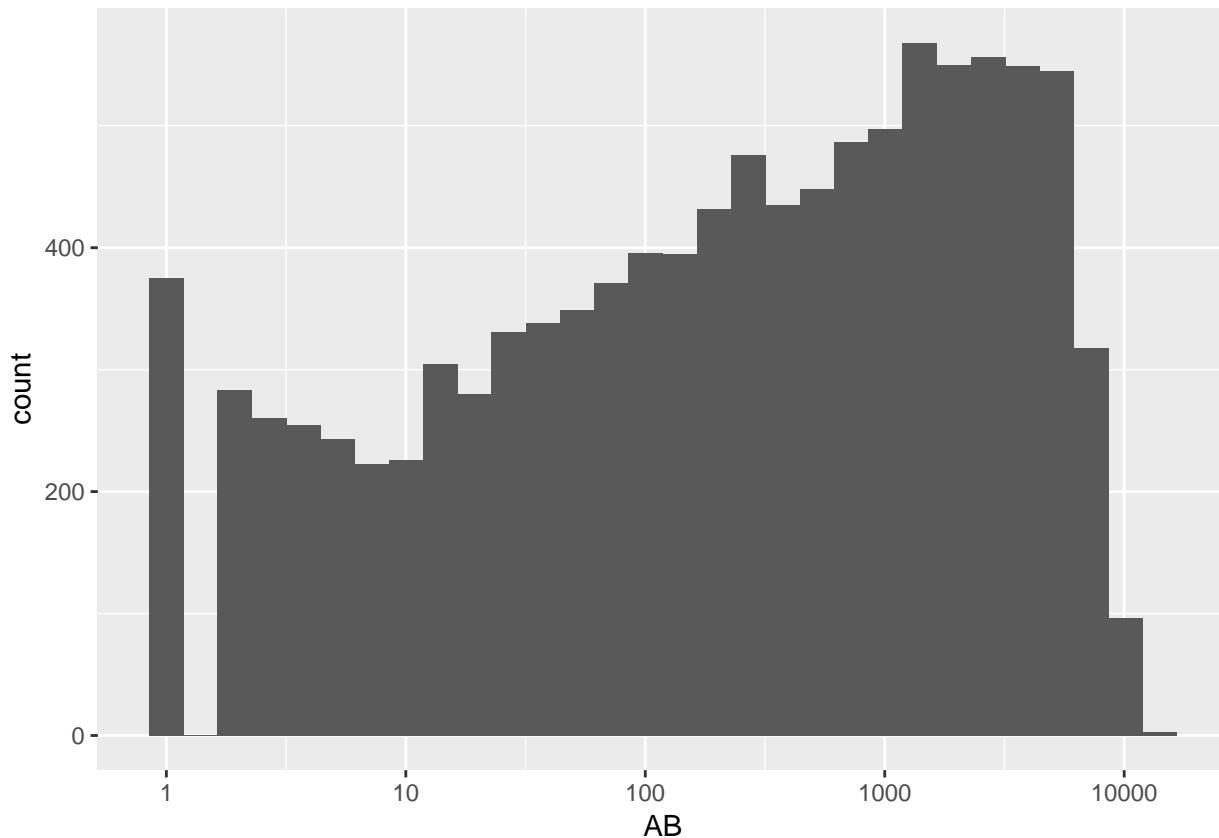
# 用这两个参数生成球员的击球概率
alpha0 <- tidy(prior)$alpha
beta0 <- tidy(prior)$beta

qplot(rbeta(10000, alpha0, beta0))

```



```
# 击球数使用原始数据
ggplot(career, aes(AB)) +
  geom_histogram() +
  scale_x_log10()
```



```
# 构建仿真数据
set.seed(2017)

career_sim <- career %>%
  mutate(p = rbeta(n(), alpha0, beta0),
        H = rbinom(n(), AB, p))
```

```
career_sim
```

```
## # A tibble: 10,590 x 4
##   playerID     H    AB      p
##   <chr>     <int> <int>  <dbl>
## 1 aaronha01  3661 12364 0.299
## 2 aaronto01   229   944 0.249
## 3 abadan01     2    21 0.273
## 4 abadijo01    6    49 0.198
## 5 abbated01   746  3044 0.249
## 6 abbeych01   445  1751 0.264
## 7 abbotda01    2     7 0.191
## 8 abbotfr01   122   513 0.251
## 9 abbotje01   154   596 0.243
## 10 abbotku01   583  2044 0.261
```

```
## # ... with 10,580 more rows
```

### 10.15.1 模拟对分布参数的估计

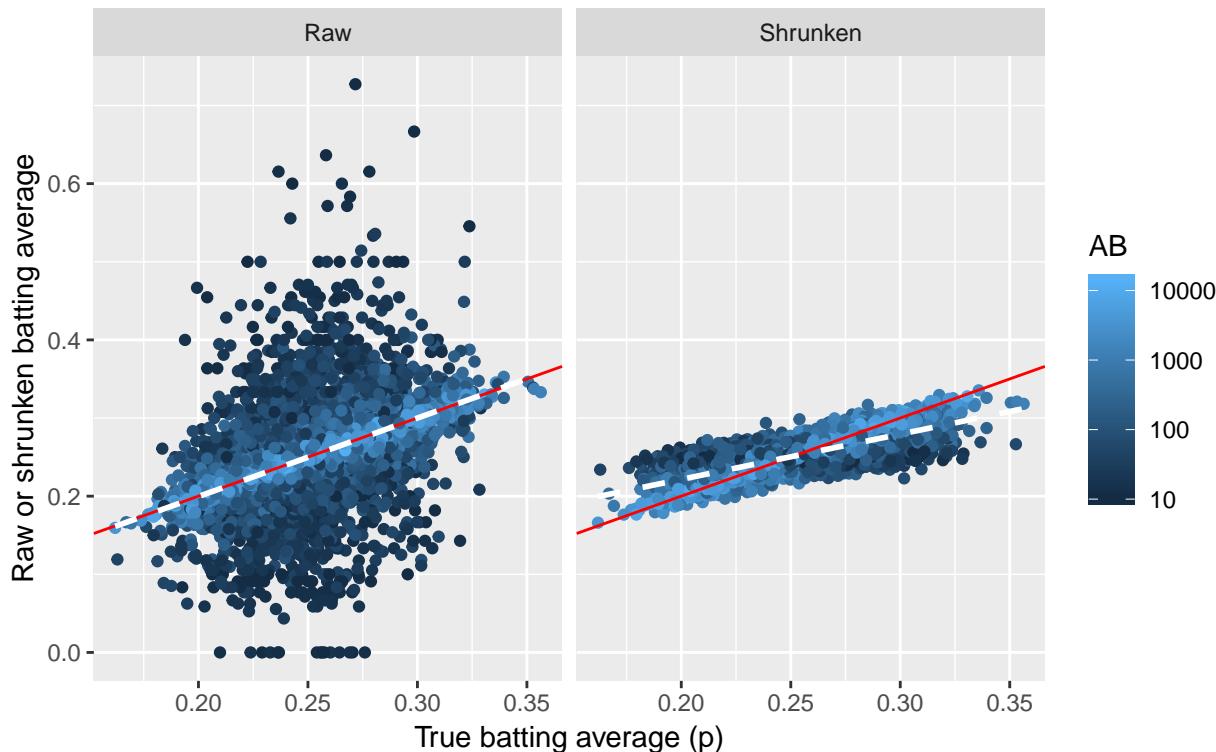
- 生产数据后我们可以估计分布参数，看能否与模拟值对应

```
career_sim_eb <- career_sim %>%
  add_ebb_estimate(H, AB)

career_sim_gathered <- career_sim_eb %>%
  rename(Shrunken = .fitted, Raw = .raw) %>%
  gather(type, estimate, Shrunken, Raw)
# 观察是否能收敛数据
career_sim_gathered %>%
  filter(AB >= 10) %>%
  ggplot(aes(p, estimate, color = AB)) +
  geom_point() +
  geom_abline(color = "red") +
  geom_smooth(method = "lm", color = "white", lty = 2, se = FALSE) +
  scale_color_continuous(trans = "log", breaks = c(10, 100, 1000, 10000)) +
  facet_wrap(~ type) +
  labs(x = "True batting average (p)",
       y = "Raw or shrunken batting average",
       title = "Empirical Bayes shrinkage reduces variance, but causes bias",
       subtitle = "Red line is x = y; dashed white line is a linear fit")
```

Empirical Bayes shrinkage reduces variance, but causes bias

Red line is  $x = y$ ; dashed white line is a linear fit



- 我们可以看到，估计方差有了一定收敛，但出现了一定偏差，可参考统计学习中方差-偏差权衡的描述

- 可以用均方误来衡量，此处虽牺牲了偏差，但整体误差降低了

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (p - \hat{p})^2$$

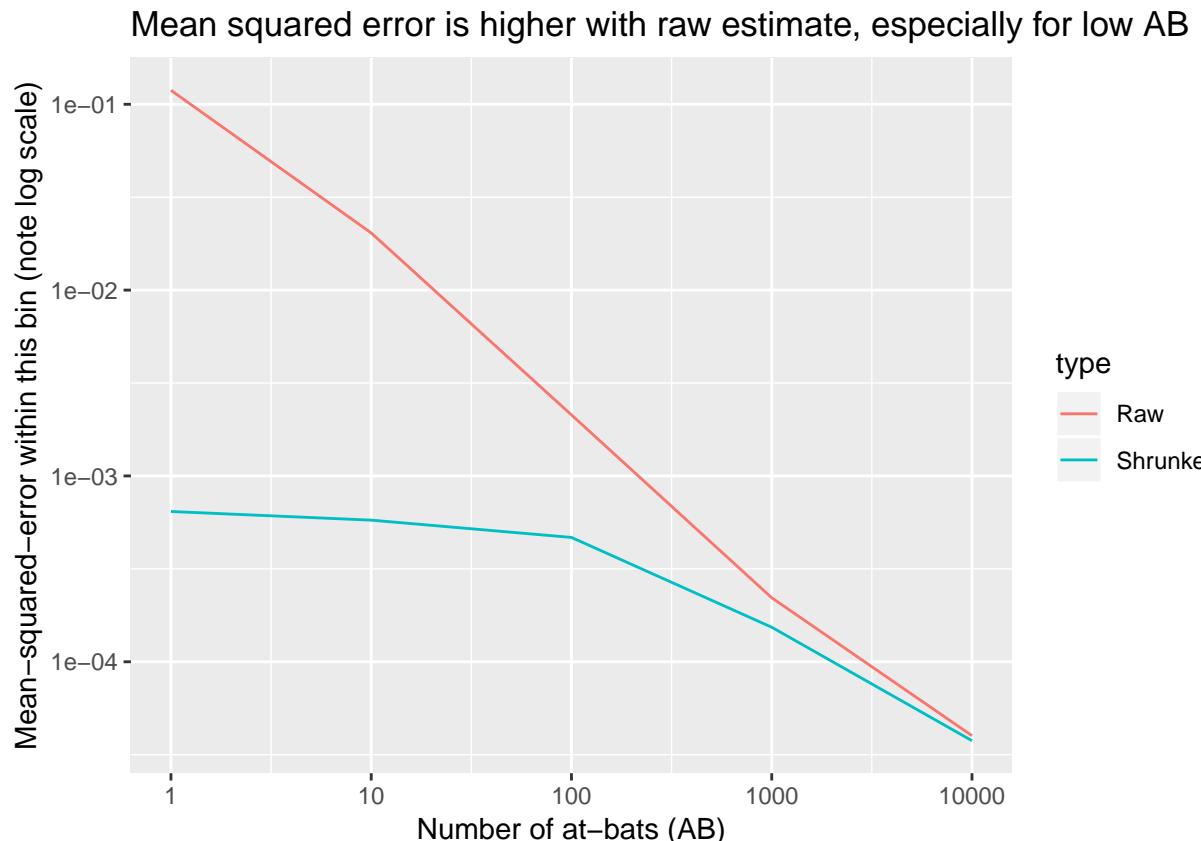
```
career_sim_gathered %>%
  group_by(type) %>%
  summarize(mse = mean((estimate - p)^2))
```

```
## # A tibble: 2 x 2
##   type      mse
##   <chr>    <dbl>
## 1 Raw     0.0145
## 2 Shrunken 0.000335
```

- 注意到击球数可能影响收敛，所以可以探索其对均方误的影响

```
metric_by_bin <- career_sim_gathered %>%
  group_by(type, AB = 10^(round(log10(AB)))) %>%
  summarize(mse = mean((estimate - p)^2))

ggplot(metric_by_bin, aes(AB, mse, color = type)) +
  geom_line() +
  scale_x_log10() +
  scale_y_log10() +
  labs(x = "Number of at-bats (AB)",
       y = "Mean-squared-error within this bin (note log scale)",
       title = "Mean squared error is higher with raw estimate, especially for low AB")
```

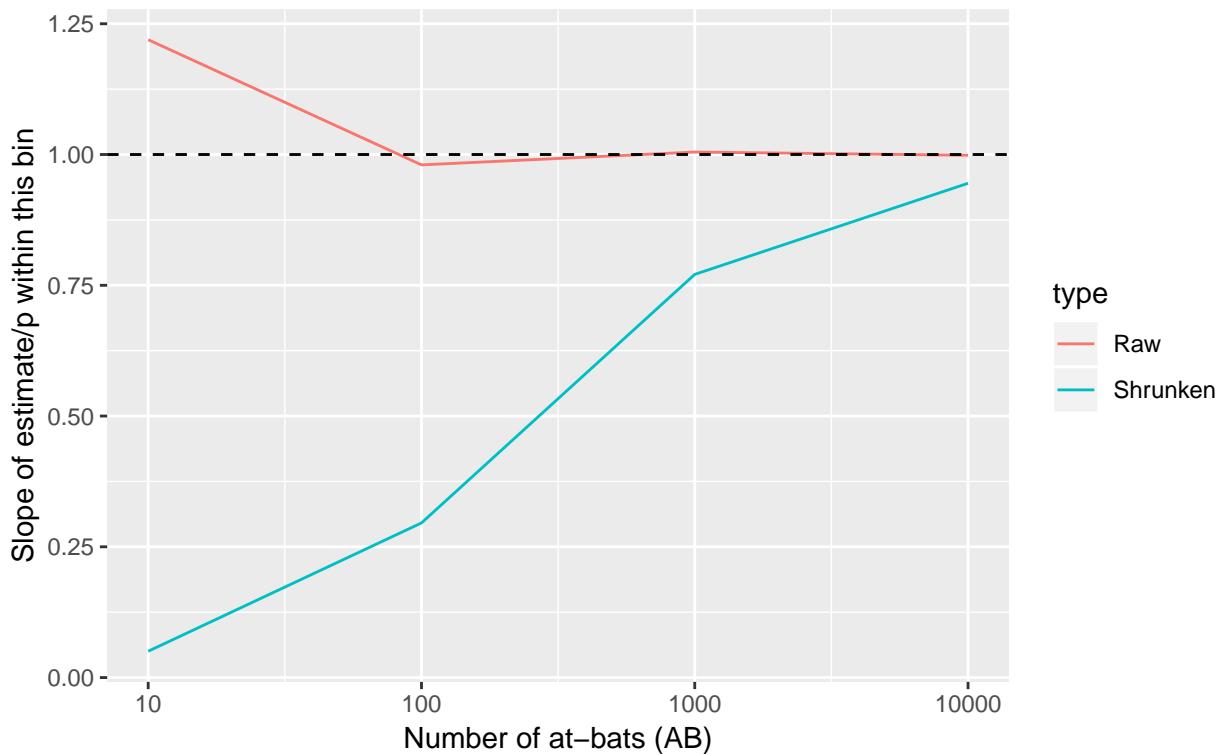


- 击球数越多，均方误越低，此时可进一步探索

```
library(scales)
# 观察斜率 p 值变化
career_sim_gathered %>%
  mutate(AB = 10 ^ (round(log10(AB)))) %>%
  filter(AB > 1) %>%
  nest(-type, -AB) %>%
  unnest(data, ~ tidy(lm(estimate ~ p, .))) %>%
  filter(term == "p") %>%
  ggplot(aes(AB, estimate, color = type)) +
  geom_line() +
  scale_x_log10(breaks = c(10, 100, 1000, 10000)) +
  geom_hline(yintercept = 1, lty = 2) +
  labs(x = "Number of at-bats (AB)",
       y = "Slope of estimate/p within this bin",
       title = "Shrunken estimates introduce bias for low AB",
       subtitle = "Note that an unbiased estimate would have a slope of 0")
```

### Shrunken estimates introduce bias for low AB

Note that an unbiased estimate would have a slope of 0

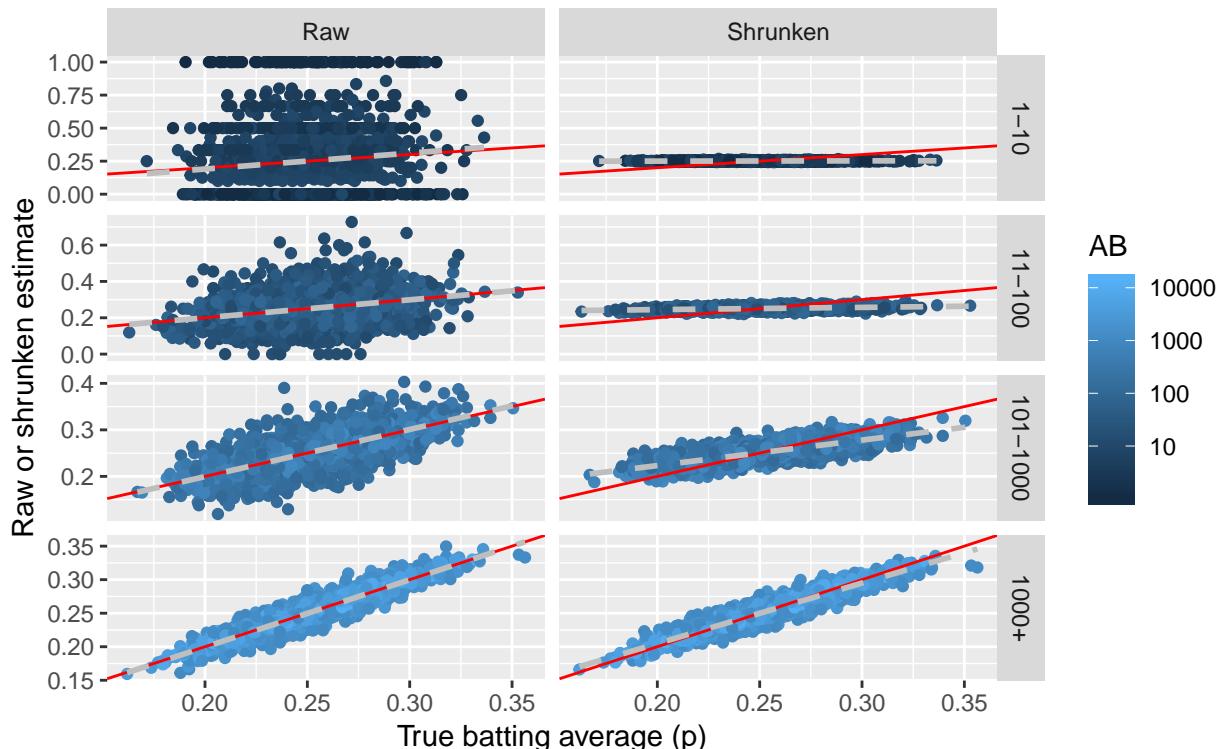


```
# 分层
career_sim_gathered %>%
  mutate(ab_bin = cut(AB, c(0, 10, 100, 1000, Inf),
                     labels = c("1-10", "11-100", "101-1000", "1000+")) %>%
  ggplot(aes(p, estimate, color = AB)) +
  geom_point() +
  geom_abline(color = "red") +
  geom_smooth(method = "lm", color = "gray", lty = 2, se = FALSE) +
  scale_color_continuous(trans = "log", breaks = c(10, 100, 1000, 10000)) +
```

```
facet_grid(ab_bin ~ type, scales = "free_y") +
  labs(x = "True batting average (p)",
       y = "Raw or shrunken estimate",
       title = "Empirical Bayes shrinkage reduces variance, but introduces bias",
       subtitle = "Red line is x = y; dashed white line is a linear fit")
```

### Empirical Bayes shrinkage reduces variance, but introduces bias

Red line is  $x = y$ ; dashed white line is a linear fit



- 击球数越多，越接近真相

#### 10.15.2 区间估计

- 检验区间估计是否覆盖 95% 的真值

```
career_sim_eb %>%
  summarize(coverage = mean(.low <= p & p <= .high))
```

```
## # A tibble: 1 x 1
##   coverage
##       <dbl>
## 1     0.951
```

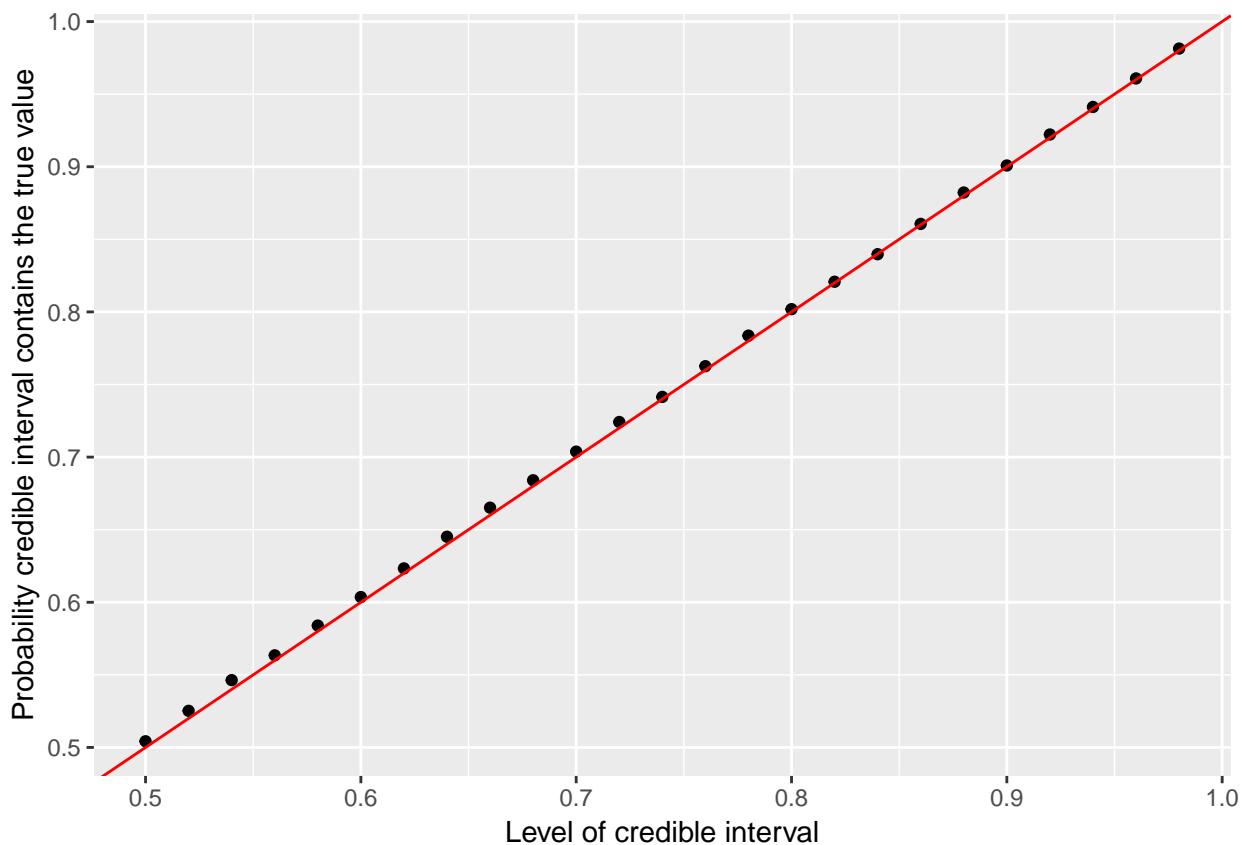
- 观察不同区间的覆盖范围

```
sim_prior <- ebb_fit_prior(career_sim, H, AB)
estimate_by_cred_level <- data_frame(level = seq(.5, .98, .02)) %>%
  unnest(map(level, ~ augment(sim_prior, career_sim, cred_level = .)))
estimate_by_cred_level %>%
  group_by(level) %>%
```

```

mutate(cover = .low <= p & p <= .high) %>%
summarize(coverage = mean(cover)) %>%
ggplot(aes(level, coverage)) +
geom_point() +
geom_abline(color = "red") +
labs(x = "Level of credible interval",
y = "Probability credible interval contains the true value")

```



- 结果基本吻合，说明区间估计也比较准

### 10.15.3 错误发现率

- 看一下进入名人堂的人

```

pt <- career_sim_eb %>%
  add_ebb_prop_test(.3, sort = TRUE)

# 错误发现率控制为 10%
hall_of_fame <- pt %>%
  filter(.qvalue <= .1)

mean(hall_of_fame$p < .3)

## [1] 0.128

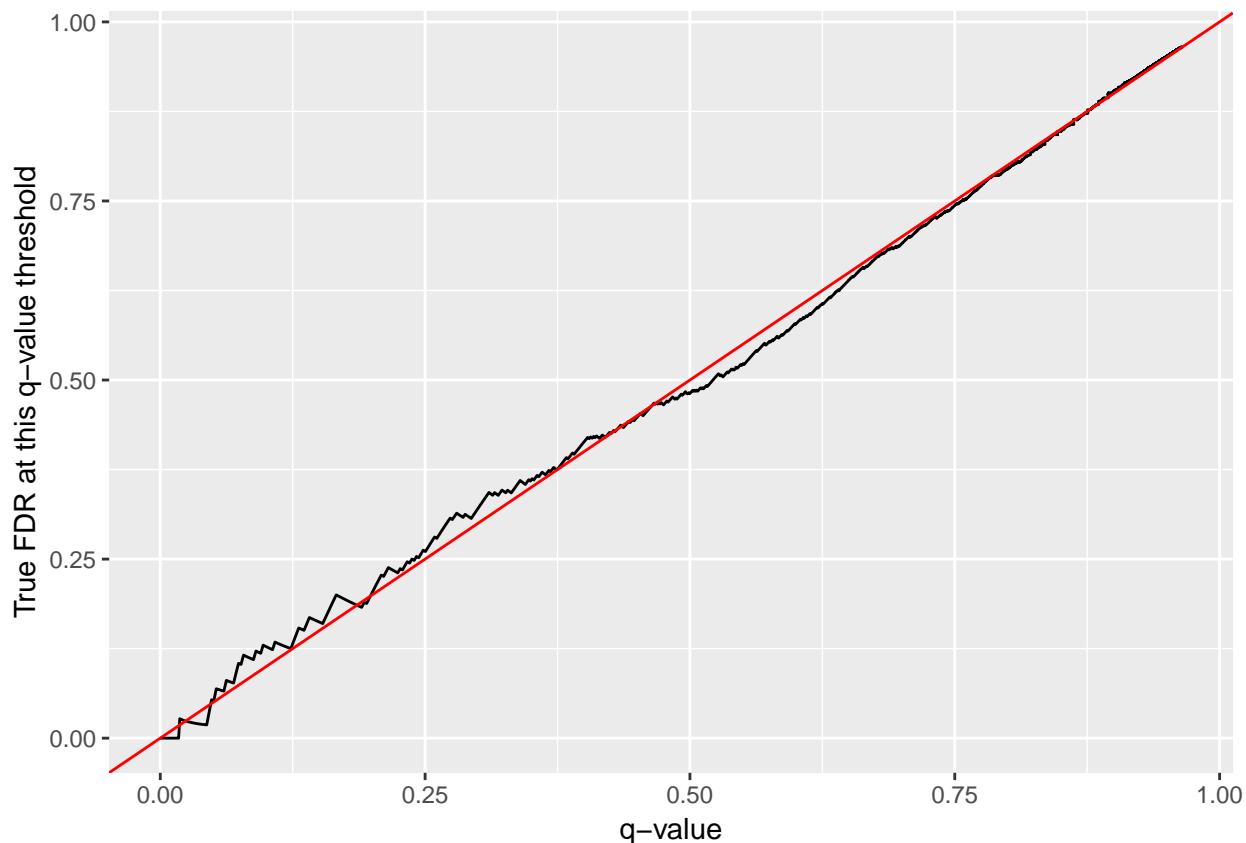
# 观察整体错误发现率的变动
pt %>%

```

```

mutate(true_fdr = cummean(p < .3)) %>%
ggplot(aes(.qvalue, true_fdr)) +
geom_line() +
geom_abline(color = "red") +
labs(x = "q-value",
y = "True FDR at this q-value threshold")

```



#### 10.15.4 贝塔二项回归

- 看下影响因素

```

# 回归值
bb_reg <- career %>%
  ebb_fit_prior(H, AB, method = "gamlss", mu_predictors = ~ log10(AB))

tidy(bb_reg)

## # A tibble: 3 x 6
##   parameter term      estimate std.error statistic p.value
##   <chr>     <chr>      <dbl>     <dbl>     <dbl>    <dbl>
## 1 mu        (Intercept) -1.68     0.00896   -188.     0
## 2 mu        log10(AB)    0.191    0.00277     69.1     0
## 3 sigma     (Intercept) -6.30     0.0229    -275.     0
set.seed(2017)

career_sim_ab <- augment(bb_reg, career) %>%

```

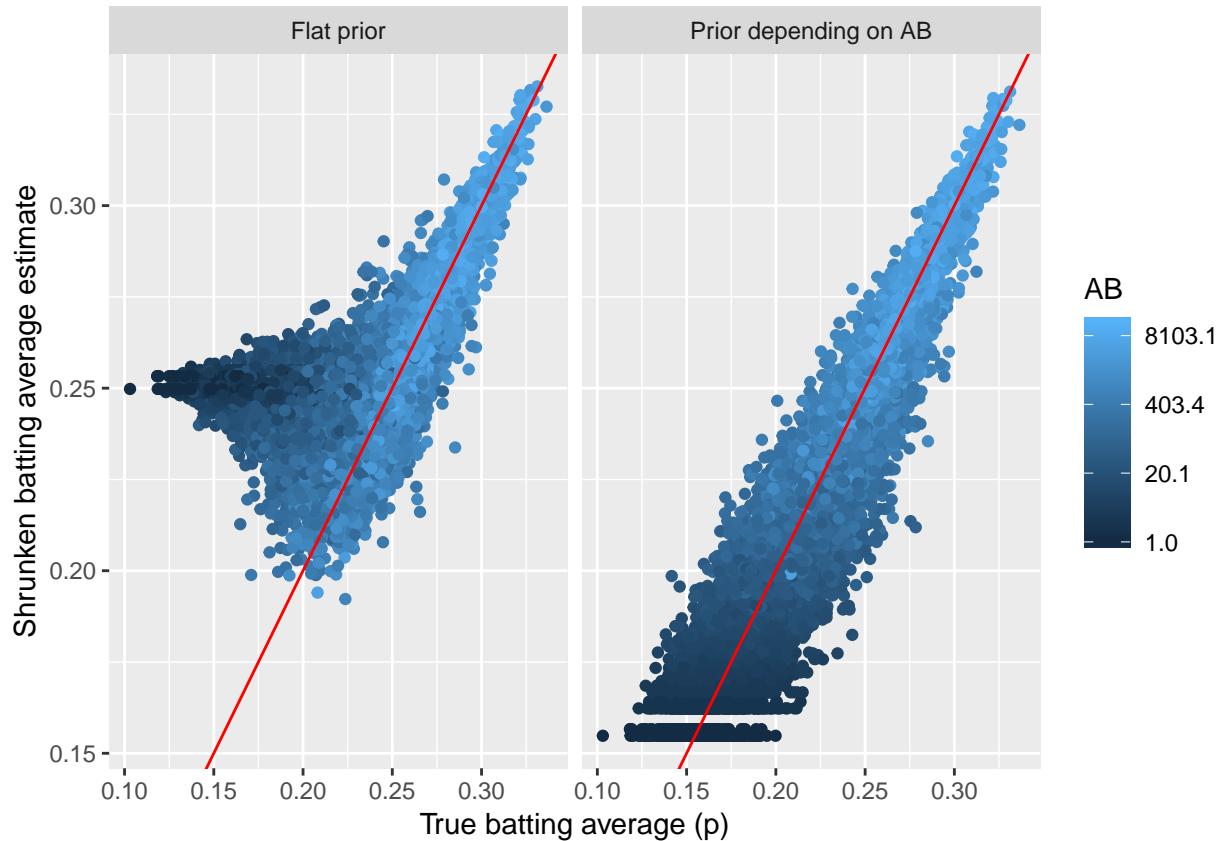
```
dplyr::select(playerID, AB, true_alpha0 = .alpha0, true_beta0 = .beta0) %>%
  mutate(p = rbeta(n(), true_alpha0, true_beta0),
        H = rbinom(n(), AB, p))
# 真实值
career_ab_prior <- career_sim_ab %>%
  ebb_fit_prior(H, AB, method = "gamlss", mu_predictors = ~ log10(AB))

# 对比
tidy(career_ab_prior)
```

```
## # A tibble: 3 x 6
##   parameter term      estimate std.error statistic p.value
##   <chr>     <chr>      <dbl>     <dbl>     <dbl>    <dbl>
## 1 mu        (Intercept) -1.70     0.00893   -190.     0
## 2 mu        log10(AB)    0.194    0.00276    70.4     0
## 3 sigma     (Intercept) -6.31     0.0260    -243.     0

# 观察击球数影响
career_flat_prior <- career_sim_ab %>%
  ebb_fit_prior(H, AB)

data_frame(method = c("Flat prior", "Prior depending on AB"),
           model = list(career_flat_prior, career_ab_prior)) %>%
  unnest(map(model, augment, data = career_sim_ab)) %>%
  ggplot(aes(p, .fitted, color = AB)) +
  geom_point() +
  scale_color_continuous(trans = "log") +
  geom_abline(color = "red") +
  facet_wrap(~ method) +
  labs(x = "True batting average (p)",
       y = "Shrunken batting average estimate")
```



### 10.15.5 重复模拟

- 为防止意外或运气可以重复模拟看看

```
set.seed(2017)
```

```
sim_replications <- career %>%
  crossing(replication = 1:50) %>%
  mutate(p = rbeta(n(), alpha0, beta0),
        H = rbinom(n(), AB, p))

sim_replications

## # A tibble: 529,500 x 5
##   playerID     H    AB replication     p
##   <chr>    <int> <int>       <int> <dbl>
## 1 aaronha01  3736 12364         1 0.299
## 2 aaronha01  3079 12364         2 0.249
## 3 aaronha01  3323 12364         3 0.273
## 4 aaronha01  2405 12364         4 0.198
## 5 aaronha01  3109 12364         5 0.249
## 6 aaronha01  3283 12364         6 0.264
## 7 aaronha01  2353 12364         7 0.191
## 8 aaronha01  3053 12364         8 0.251
## 9 aaronha01  3043 12364         9 0.243
## 10 aaronha01 3268 12364        10 0.261
## # ... with 529,490 more rows
```

```

sim_replication_models <- sim_replications %>%
  nest(-replication) %>%
  mutate(prior = map(data, ~ ebb_fit_prior(., H, AB)))

# 估计参数
sim_replication_priors <- sim_replication_models %>%
  unnest(map(prior, tidy), .drop = TRUE)

sim_replication_priors

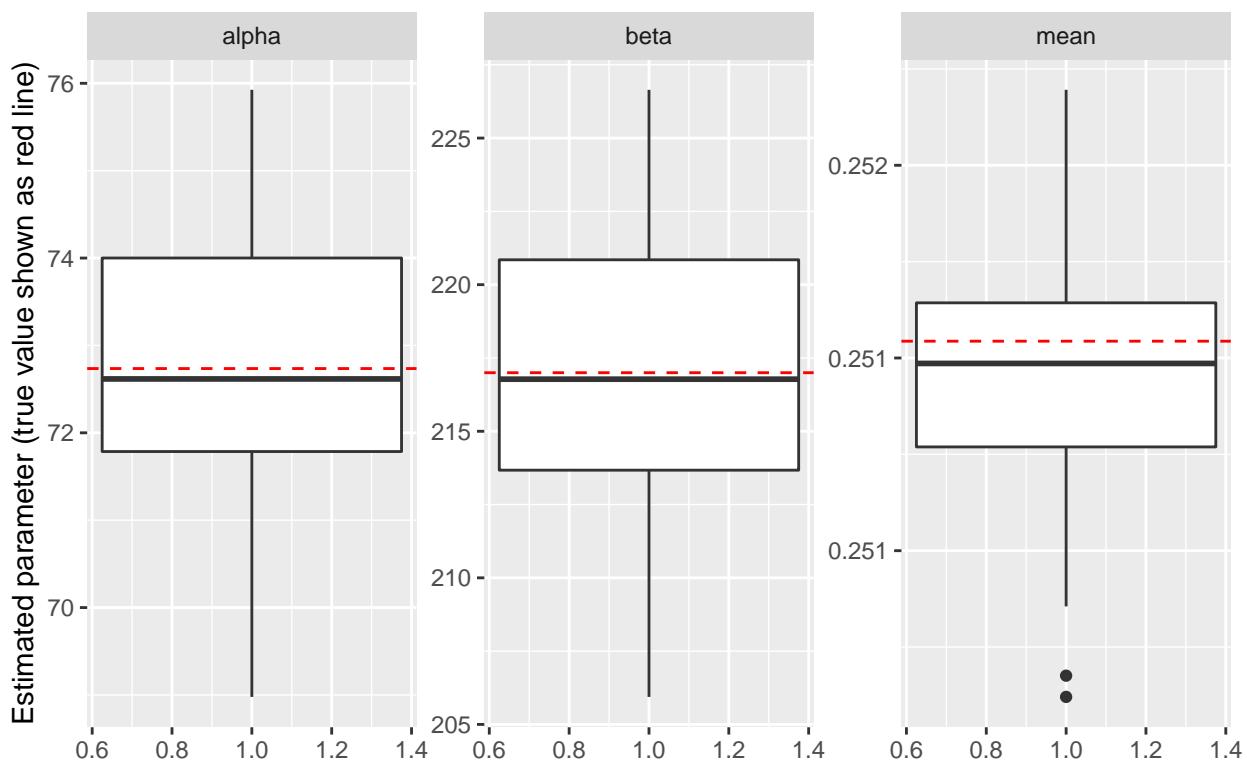
## # A tibble: 50 x 4
##       replication alpha   beta   mean
##       <int>     <dbl>  <dbl>  <dbl>
## 1             1    69.0  206.  0.251
## 2             2    72.5  216.  0.252
## 3             3    72.7  217.  0.250
## 4             4    74.3  222.  0.251
## 5             5    71.9  215.  0.250
## 6             6    71.8  214.  0.251
## 7             7    72.7  217.  0.251
## 8             8    71.3  213.  0.251
## 9             9    75.9  227.  0.251
## 10            10    71.5  213.  0.251
## # ... with 40 more rows

true_values <- data_frame(parameter = c("alpha", "beta", "mean"),
                           true = c(alpha0, beta0, alpha0 / (alpha0 + beta0)))

sim_replication_priors %>%
  gather(parameter, value, -replication) %>%
  inner_join(true_values, by = "parameter") %>%
  ggplot(aes(1, value)) +
  geom_boxplot() +
  geom_hline(aes(yintercept = true), color = "red", lty = 2) +
  facet_wrap(~ parameter, scales = "free_y") +
  labs(x = "",
       y = "Estimated parameter (true value shown as red line)",
       title = "Estimated hyperparameters across 50 replications")

```

### Estimated hyperparameters across 50 replications

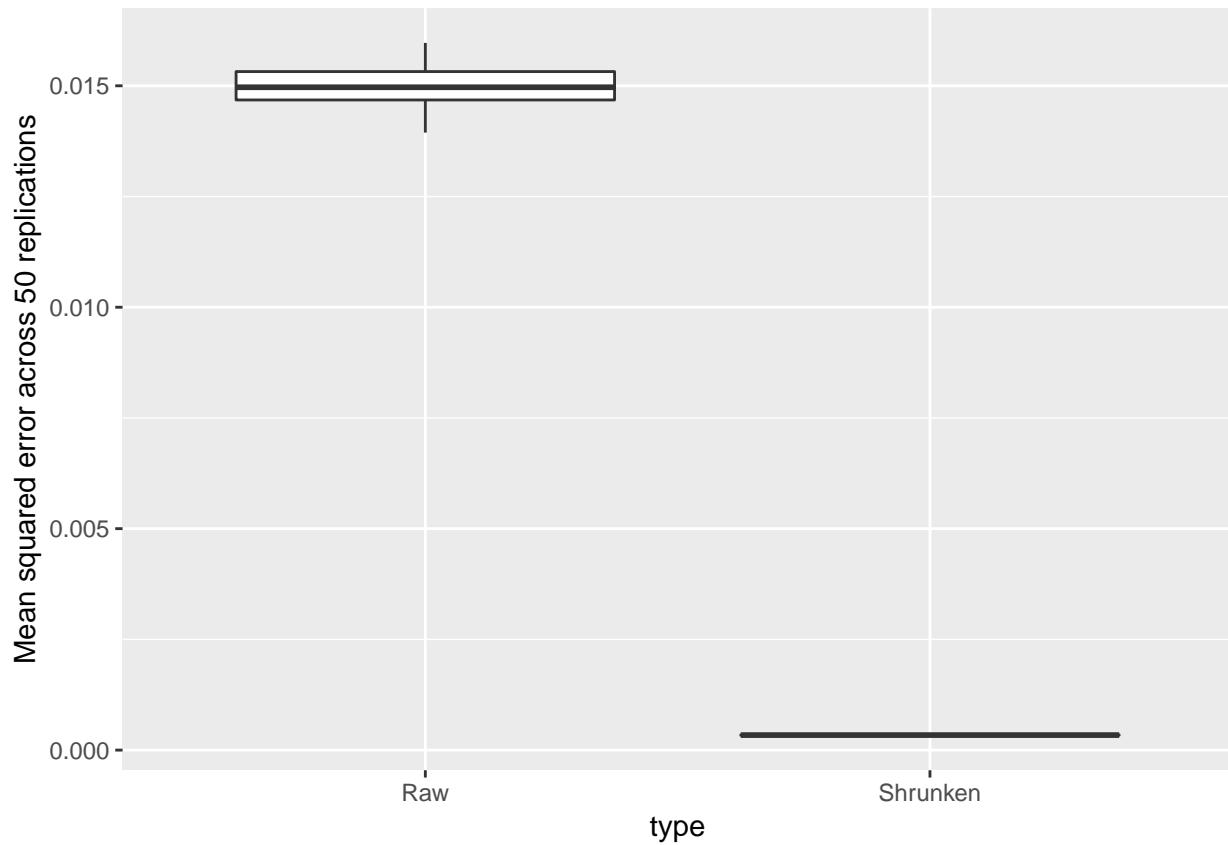


```
# 估计区间与假设检验

## 估计均方误
sim_replication_ae <- sim_replication_models %>%
  unnest(map2(prior, data, augment))

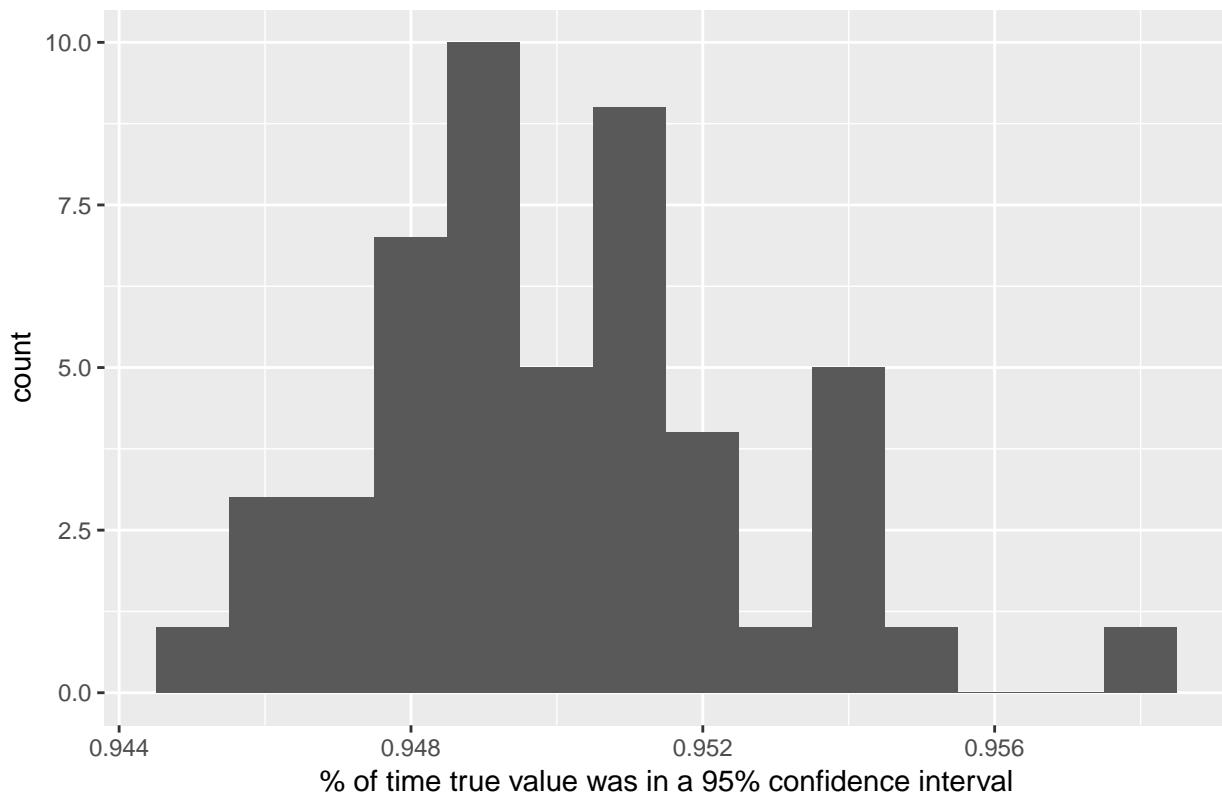
sim_replication_mse <- sim_replication_ae %>%
  rename(Raw = .raw, Shrunken = .fitted) %>%
  gather(type, estimate, Raw, Shrunken) %>%
  group_by(type, replication) %>%
  summarize(mse = mean((estimate - p)^ 2))

ggplot(sim_replication_mse, aes(type, mse)) +
  geom_boxplot() +
  ylab("Mean squared error across 50 replications")
```



```
## 估计区间
sim_replication_ae %>%
  mutate(cover = .low <= p & p <= .high) %>%
  group_by(replication) %>%
  summarize(coverage = mean(cover)) %>%
  ggplot(aes(coverage)) +
  geom_histogram(binwidth = .001) +
  labs(x = "% of time true value was in a 95% confidence interval",
       title = "95% credible interval is well calibrated across replications")
```

### 95% credible interval is well calibrated across replications



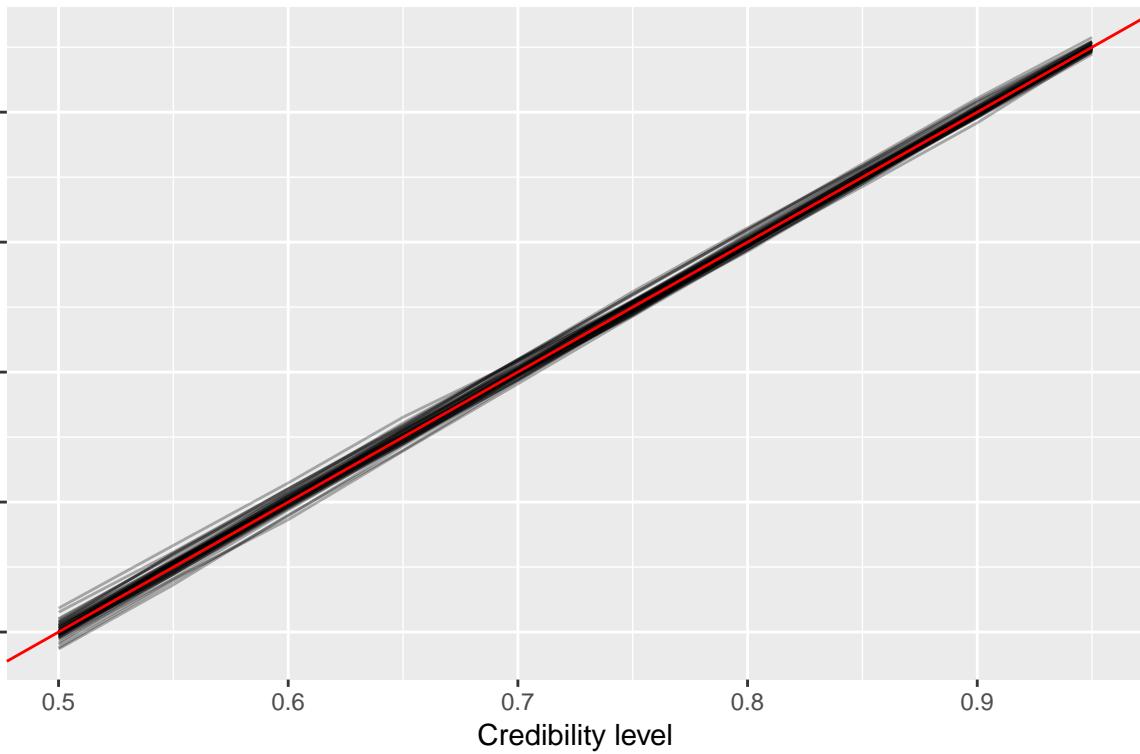
```
sim_replication_intervals <- sim_replication_models %>%
  crossing(cred_level = c(seq(.5, .9, .05), .95)) %>%
  unnest(pmap(list(prior, data, cred_level = cred_level), augment)) %>%
  dplyr::select(replication, cred_level, p, .low, .high)

sim_replication_intervals %>%
  mutate(cover = .low <= p & p <= .high) %>%
  group_by(replication, cred_level) %>%
  summarize(coverage = mean(cover)) %>%
  ggplot(aes(cred_level, coverage, group = replication)) +
  geom_line(alpha = .3) +
  geom_abline(color = "red") +
  labs(x = "Credibility level",
       y = "% of credible intervals in this replication that contain the true parameter",
       title = "Credible intervals are well calibrated across 50 replications",
       subtitle = "Red line is x = y")
```

of credible intervals in this replication that contain the true parameter

Credible intervals are well calibrated across 50 replications

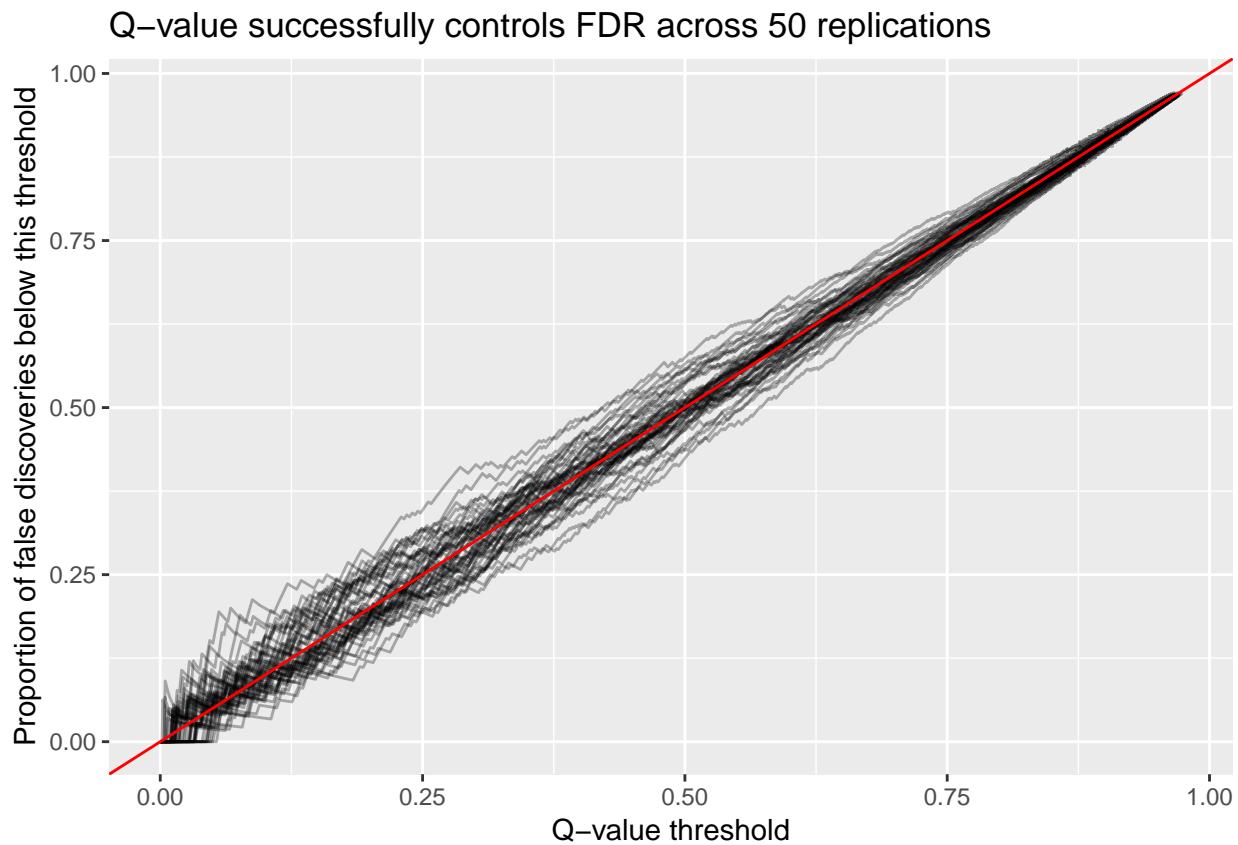
Red line is  $x = y$



```
## q 值的稳定性
```

```
sim_replication_prop_tests <- sim_replication_aus %>%
  nest(-replication) %>%
  unnest(map(data, add_ebb_prop_test, threshold = .3, sort = TRUE))

sim_replication_prop_tests %>%
  group_by(replication) %>%
  mutate(fdr = cummean(p < .3)) %>%
  ggplot(aes(.qvalue, fdr, group = replication)) +
  geom_line(alpha = .3) +
  geom_abline(color = "red") +
  labs(x = "Q-value threshold",
       y = "Proportion of false discoveries below this threshold",
       title = "Q-value successfully controls FDR across 50 replications")
```



## 10.16 网络资源

- 贝叶斯方法
- 贝叶斯镜像



# 章 11

## 生存分析

### 11.1 Concepts

- examines and models the time it takes for events to occur
- the distribution of survival times
- Popular: Cox proportional-hazards regression model

### 11.2 Notation

- T as a random variable with cumulative distribution function  $P(t) = Pr(T \leq t)$  and probability density function  $p(t) = \frac{dP(t)}{dt}$
- survival function  $S(t)$  is the complement of the distribution function,  $S(t) = Pr(T > t) = 1 - P(t)$
- hazard function  $logh(t) = -\ln[S(t)]$

### 11.3 Cox proportional-hazards regression model

- $logh_i(t) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}$
- Cox model
  - $logh_i(t) = (\eta_i(t)) + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}$
- the Cox model is a proportional-hazards model
  - $\frac{h_i(t)}{h_{i'}(t)} = \frac{e^{\eta_i}}{e^{\eta_{i'}}}$

### 11.4 Case: Recidivism

- Target: recidivism of 432 male prisoners, who were observed for a year after being released from prison
- arrest means the male prisoners who rearrested
- 52 weeks
- factors: financial aid after release from prison, affected, release ages, race, work experience, marriage, parole, prior convictions, education

```
library(survival)
library(car)
# perform survival analysis
```

```

Rossi <- read.table('http://ftp.auckland.ac.nz/software/CRAN/doc/contrib/Fox-Companion/Rossi.txt', head=TRUE)
Rossi[1:5, 1:10]

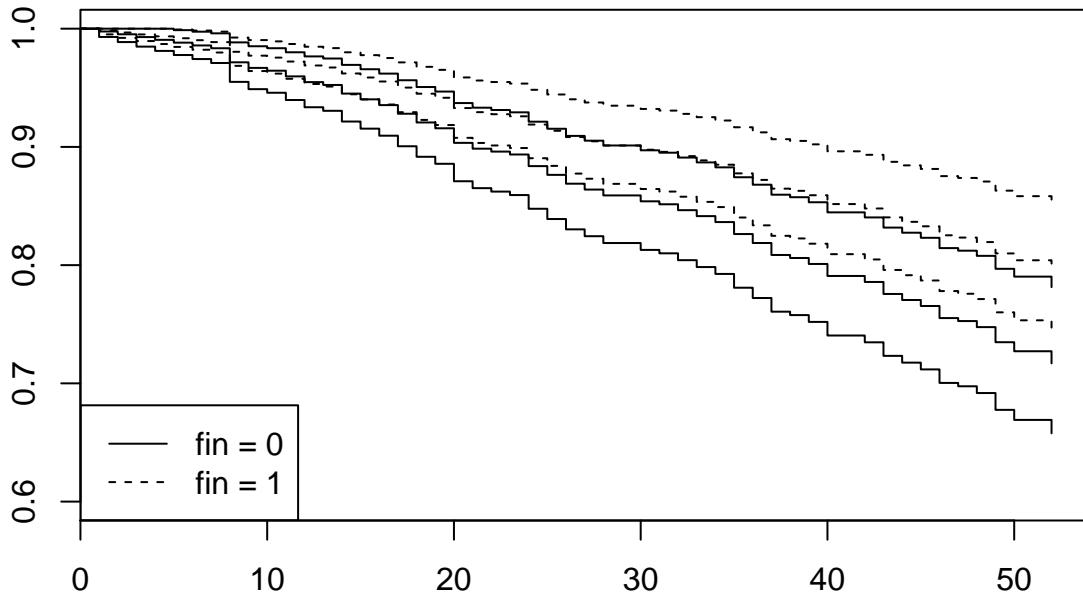
##   week arrest fin age race wexp mar paro prio educ
## 1    20      1   0  27    1   0   0    1    3    3
## 2    17      1   0  18    1   0   0    1    8    4
## 3    25      1   0  19    0   1   0    1   13    3
## 4    52      0   1  23    1   1   1    1    1    5
## 5    52      0   0  19    0   1   0    1    3    3

mod.allison <- coxph(Surv(week, arrest) ~ fin + age + race + wexp + mar + paro + prio, data=Rossi)
summary(mod.allison)

## Call:
## coxph(formula = Surv(week, arrest) ~ fin + age + race + wexp +
##        mar + paro + prio, data = Rossi)
##
## n= 432, number of events= 114
##
##          coef exp(coef) se(coef)     z Pr(>|z|)
## fin   -0.3794  0.6843  0.1914 -1.98  0.0474 *
## age   -0.0574  0.9442  0.0220 -2.61  0.0090 **
## race   0.3139  1.3688  0.3080  1.02  0.3081
## wexp  -0.1498  0.8609  0.2122 -0.71  0.4803
## mar   -0.4337  0.6481  0.3819 -1.14  0.2561
## paro  -0.0849  0.9186  0.1958 -0.43  0.6646
## prio   0.0915  1.0958  0.0286  3.19  0.0014 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## fin       0.684      1.461     0.470     0.996
## age       0.944      1.059     0.904     0.986
## race      1.369      0.731     0.748     2.503
## wexp      0.861      1.162     0.568     1.305
## mar       0.648      1.543     0.307     1.370
## paro      0.919      1.089     0.626     1.348
## prio      1.096      0.913     1.036     1.159
##
## Concordance= 0.64  (se = 0.027 )
## Rsquare= 0.074  (max possible= 0.956 )
## Likelihood ratio test= 33.3 on 7 df,  p=2e-05
## Wald test            = 32.1 on 7 df,  p=4e-05
## Score (logrank) test = 33.5 on 7 df,  p=2e-05

# plot time vs survival prob
plot(survfit(mod.allison), ylim=c(.7, 1), xlab='Weeks', ylab='Proportion Not Rearrested')

```



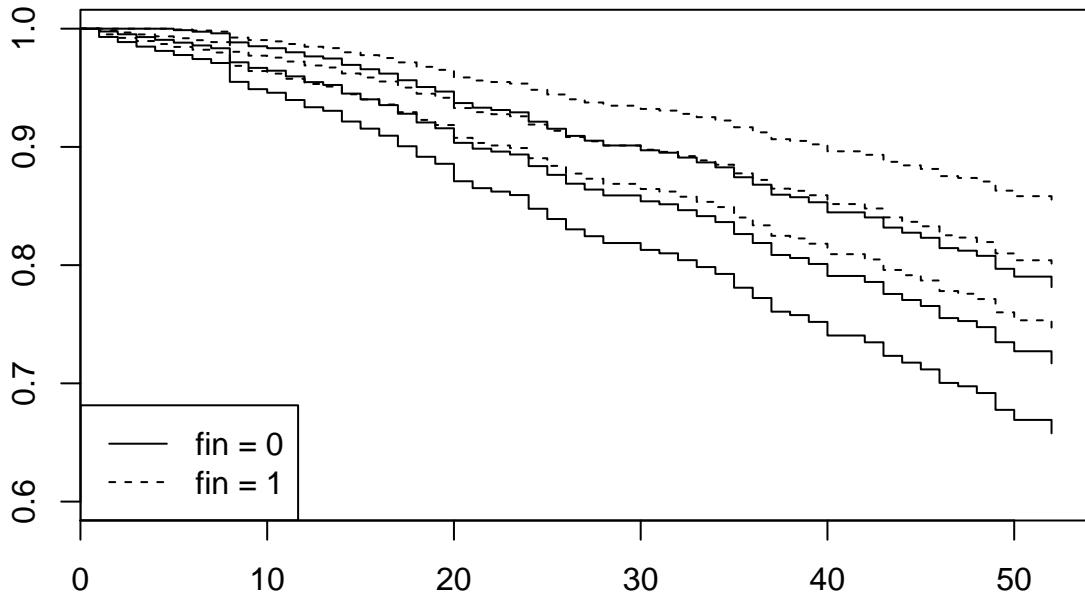
#### 11.4.1 result

- The covariates age and prio (prior convictions) have highly statistically significant coefficients, while the coefficient for fin (financial aid) is marginally significant
- holding the other covariates constant, an additional year of age reduces the weekly hazard of rearrest by a factor of  $e^b = 0.944$  on average – that is, by 5.6
- likelihood-ratio, Wald, and score chi-square statistics: null hypothesis all of the  $\beta$ 's are zero.

## 11.5 further

- assess the impact of financial aid on rearrest
- new data frame with two rows, one for each value of fin; the other covariates are fixed to their average values

```
attach(Rossi)
Rossi.fin <- data.frame(fin=c(0,1), age=rep(mean(age),2), race=rep(mean(race),2), wexp=rep(mean(wexp),2))
detach()
plot(survfit(mod.allison, newdata=Rossi.fin), conf.int=T, lty=c(1,2), ylim=c(.6, 1))
legend("bottomleft", legend=c('fin = 0', 'fin = 1'), lty=c(1,2))
```



- the higher estimated ‘survival’ of those receiving financial aid, but the two confidence envelopes overlap substantially, even after 52 weeks

## 11.6 Time-Dependent Covariates

- treat the employed variable as a tim-dependent covariates with 52 weeks' record

```
sum(!is.na(Rossi[,11:62])) # record count
## [1] 19809

Rossi2 <- matrix(0, 19809, 14) # to hold new data set
colnames(Rossi2) <- c('start', 'stop', 'arresttime', names(Rossi)[1:10], 'employed')

row<-0
for (i in 1:nrow(Rossi)) {
  for (j in 11:62) {
    if (is.na(Rossi[i, j])) next
    else {
      row <- row + 1 # increment row counter
      start <- j - 11 # start time (previous week)
      stop <- start + 1 # stop time (current week)
      arresttime <- if (stop == Rossi[i, 1] && Rossi[i, 2] == 1) 1 else 0
      Rossi2[row,] <- c(start, stop, arresttime, unlist(Rossi[i, c(1:10, j)]))
    }
  }
}
Rossi2 <- as.data.frame(Rossi2)
remove(i, j, row, start, stop, arresttime)
modallison2 <- coxph(Surv(start, stop, arresttime) ~ fin + age + race + wexp + mar + paro + prio + employed)
summary(modallison2)

## Call:
## coxph(formula = Surv(start, stop, arresttime) ~ fin + age + race +
##       wexp + mar + paro + prio + employed, data = Rossi2)
```

```

## 
##   n= 19809, number of events= 114
##
##           coef exp(coef)  se(coef)      z Pr(>|z|)
## fin     -0.3567    0.7000   0.1911 -1.87   0.0620 .
## age     -0.0463    0.9547   0.0217 -2.13   0.0330 *
## race     0.3387    1.4031   0.3096  1.09   0.2740
## wexp     -0.0256    0.9748   0.2114 -0.12   0.9038
## mar     -0.2937    0.7455   0.3830 -0.77   0.4431
## paro    -0.0642    0.9378   0.1947 -0.33   0.7416
## prio     0.0851    1.0889   0.0290  2.94   0.0033 **
## employed -1.3283    0.2649   0.2507 -5.30  1.2e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## fin        0.700     1.429     0.481     1.018
## age        0.955     1.047     0.915     0.996
## race       1.403     0.713     0.765     2.574
## wexp       0.975     1.026     0.644     1.475
## mar        0.745     1.341     0.352     1.579
## paro       0.938     1.066     0.640     1.374
## prio       1.089     0.918     1.029     1.152
## employed    0.265     3.775     0.162     0.433
##
## Concordance= 0.708  (se = 0.023 )
## Rsquare= 0.003  (max possible= 0.066 )
## Likelihood ratio test= 68.7 on 8 df,  p=9e-12
## Wald test          = 56.1 on 8 df,  p=3e-09
## Score (logrank) test = 64.5 on 8 df,  p=6e-11

```

## 11.7 Model Diagnostics

- Checking Proportional Hazards

```
modallison3 <- coxph(Surv(week, arrest) ~ fin + age + prio, data=Rossi)
modallison3
```

```

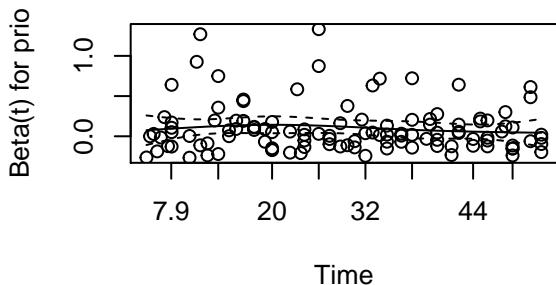
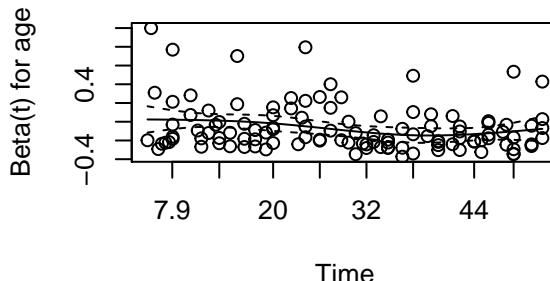
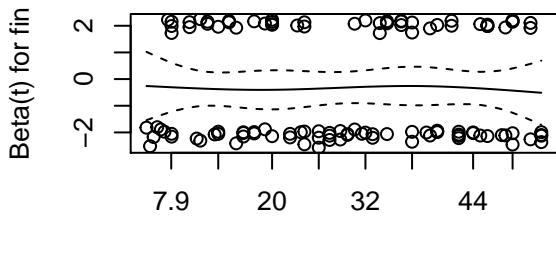
## Call:
## coxph(formula = Surv(week, arrest) ~ fin + age + prio, data = Rossi)
##
##           coef exp(coef)  se(coef)      z      p
## fin     -0.35      0.71      0.19 -2 0.068
## age     -0.07      0.94      0.02 -3 0.001
## prio    0.10      1.10      0.03  4 4e-04
##
## Likelihood ratio test=29 on 3 df, p=2e-06
## n= 432, number of events= 114
```

```
cox.zph(modallison3)
```

```

##           rho    chisq      p
## fin     -0.00657 0.00507 0.9433
## age     -0.20976 6.54147 0.0105
```

```
## prio -0.08004 0.77288 0.3793
## GLOBAL NA 7.13046 0.0679
par(mfrow=c(2,2))
plot(cox.zph(modallison3))
```



- there appears to be a trend in the plot for age, with the age effect declining with time

```
modallison4 <- coxph(Surv(start, stop, arresttime) ~ fin + age + age:stop:stop + prio, data = Rossi2)
modallison4
```

```
## Call:
## coxph(formula = Surv(start, stop, arresttime) ~ fin + age + age:stop:stop +
##        prio, data = Rossi2)
##
##          coef  exp(coef)  se(coef)    z     p
## fin      -0.349    0.706    0.190 -1.8 0.067
## age       0.032    1.033    0.039  0.8 0.413
## prio      0.098    1.103    0.027  3.6 3e-04
## age:stop -0.004    0.996    0.001 -2.6 0.009
##
## Likelihood ratio test=36 on 4 df, p=3e-07
## n= 19809, number of events= 114
```

- the coefficient for the interaction is negative and highly statistically significant: The effect of age declines with time
- use residual to find influential observations

## 11.8 Reference

- Cox Proportional-Hazards Regression for Survival Data

- Real Cases



# 章 12

## 生物信息

### 12.1 数据结构

- 列代表特征行代表条目
- 每个条目有一个唯一性特征
- 数据表可通过列链接成为关系数据库

### 12.2 Pubmed 搜索

- PubMed search tags
  - [AD] – Affiliation (company or school)
  - [ALL] – All fields (eliminates defaults)
  - [AU] or [AUTH] – Author
  - [1AU] – First author
  - [ECNO] – Enzyme Commission Numbers
  - [EDAT] – Entry date (YYYY/MM/DD)
  - [ISS] - Issue # of journal
  - [JOUR] - Journal (Title, Abbreviation , ISSN)
  - [LA] – Language
  - [PDAT] – Publication date (YYYY/MM/DD)
  - [PT] – Publication type
  - [SUBS] – Substance name
  - [TIAB] – Title/Abstract
  - [TW] – Text words
  - [UID] – Unique identifiers (primary keys)
  - [VOL] or [VI] – Volume of journal
- MeSH terms [MH][MAJR][SH]
  - 被 MeSH 索引的关系数据库
  - 保守性检索有层级关系
- 时间段搜索冒号分割 YYYY/MM/DD:YYYY/MM/DD
- 序列长度搜索 [SLEN] 可以是蛋白可以是核酸
- 蛋白分子量搜索 [MOLWT]
- 物种搜索 [ORGN]
- Nucleotide 序列蛋白数据库
- MMDB 3D 结构数据库
- Genome 基因组数据库
- OMIM 人类孟德尔遗传数据库用来探索等位基因问题

- 分类数据库 用来界定分类
- GEO 基因芯片的实验数据
- SNP 基因指纹数据库

### 12.3 动态规划

- 用于序列比对
- 对角线得分按总分评价比对结果
- 可全局可局部
- 序列比对指标是特异性与相似性
- 特异性指精确匹配比率
- 相似性指精确匹配加化学相似性比率结构相近则相似
- FASTA 慢准 BLAST 快
- 三种情况匹配不匹配间隔
- 间隔罚分

### 12.4 得分矩阵

- 考虑突变的比对
- 蛋白的自然突变率矩阵 PM1
- 矩阵自相乘得到外推矩阵 PM10 PM250 取对数为打分矩阵
- 取不同矩阵源于研究目的对多样性的判断

### 12.5 E 值

- 表示序列的同源性比对得分的稀有性
- 两个参数数据库大小 (N) 比对得分 (S)  $E = N/S$
- 数据库越大越可能随机碰到相同序列得分越高越可能同源
- E 值很小说明同源性很高 E 值很大什么说明不了
- 一般阈值  $1e-04$

### 12.6 PSI-BLAST

- 先用 BLAST 在一定 E 值上建库
- 计算新库的氨基酸概率再与全库比对得分得到统计显著性
- 可以发现 BLAST 未发现的序列建立蛋白家族

### 12.7 蛋白

- Profiles 定量描述
- Patterns 定性描述
- Signature 蛋白保守序列
- motif 少于 20 个氨基酸指示二级结构
- Domains 超过 40 个氨基酸蛋白的球状区
- 共同点保守
- 正则表达式表示保守区
  - $E-X(2,4)-[FHM]-X(4)-\{P\}-L$
  - E 后随意两个, 三个, 四个然后 FHM 其中一个, 然后随意四个, 然后一个不是 P, 最后为 L

- 可以精确可以模糊
- 没有 E 值

## 12.8 蛋白结构预测

- 分子量道尔顿 (Da) 描述质量
- 等电点蛋白不带电的 pH 值
  - 小于 7 酸性中性带负电
  - 大于 7 碱性中性带正点
- 网站计算
- 蛋白定位分泌胞内核内
  - MITOPRED 预测线粒体蛋白

## 12.9 细菌基因组

- 细菌是环形 DNA 真核是线性染色体
- 细菌不加工 mRNA
- 细菌一段 mRNA 上有多个顺反子也就是多个编码 DNA 序列
- 操纵子在 mRNA 编码的上游或下游调控转录
- GLIMMER 与 FGENESB 用来预测一段序列的转录情况

## 12.10 病毒

- 三种 RNA DNA 逆转录病毒突变快
- RNA 病毒三种双链正链负链
- 逆转录基因组简单 Gag Pol Env
- 凝集素等决定病毒亚型

## 12.11 单核苷酸多态性 (SNP)

- 至少 1% 种群中存在的 DNA 单核苷酸变化
- 后果
  - 编码区改变影响表型
  - 不改变蛋白序列的编码区可能影响 mRNA 加工
  - 启动子或调控区可能影响表达
  - 其他区没有影响可作为染色体标记- 类型
  - 不改变氨基酸
  - 改变氨基酸
  - 非编码区
- 数据库
  - dbSNP
  - SNPEffect SNPs 对蛋白的影响
  - SNPedia SNPs 的临床效应
  - 1000 基因组外显子计划 第二代测序的发展

## 12.12 真核基因预测

- CDS 是 mRNA 的子集

- CDS 可能比 mRNA 外显子少
- 基因预测只能发现编码区外显子
- 有些转录变化不改变蛋白序列：UTR 区与同义密码子

## 12.13 DNA 指纹

- 重复突变会影响限制性片段长度
- VNTR 用来排除嫌犯
- PCR 用来扩增相关片段
- CODIS 区域在美国用来鉴定身份

## 12.14 Ensembl

- 外显子基因组学数据库
- 可选择人类鼠斑马鱼等常见物种

## 12.15 基因组学数据分析

- 主页

### 12.15.1 microarrays

#### 12.15.1.1 原理

- 生成互补 DNA 探针
- 标记样品中的 DNA 单链可以对不同样品标记不同颜色
- 特异性互补反应
- 测定标记物光信号

#### 12.15.1.2 应用

- 测定基因表达
  - 已知序列
  - 3' 端（降解从 5' 端开始）选取 11 个片段作为探针
  - 样品对 11 个片段都是高表达则基因高表达
- 寻找 SNP
  - SNP 单核苷酸多态性用来探索基因型
  - 合成 SNP 探针
  - 测定对不同探针的响应判断 AA AG GG 类型
- 寻找转录因子结合位点
  - 样品处理为含蛋白与不含蛋白两份去除蛋白后扩增
  - 探针是基因组感兴趣的片段
  - 瓦片分析可知探针与含转录因子 DNA 结合位点
  - 总 DNA 作为对照

## 12.15.2 NGS

### 12.15.2.1 原理

- DNA 打成 50~70 片段一个样品片段上亿
- 加上 adaptor 固定在板上后原位扩增成束
- 使用标记过的单核苷酸逐碱基对测光强同时测序大量片段
- 得到测序结果与强度

### 12.15.2.2 应用

- 寻找 SNP
- RNA-seq 测定 RNA 表达量
- 寻找结合位点表达量

## 12.15.3 数据分析应用背景

### 12.15.3.1 DNA 甲基化

- CpG 5' 端到 3' 端 CG
- C 上甲基化复制时该特性会保留
- 临近 CpG 位点的基因不会被表达
- CpG 成簇存在称为 CpG islands
- bisulfite treatment 可以用来测定 CpG 是否被甲基化通过将未甲基化的 CpG 中的 C 改为 T
- 测序中测定改变率就可知 CpG 位点甲基化程度与位置

### 12.15.3.2 CHIP-SEQ

- 蛋白结合后固定，洗掉其余片段，然后洗掉蛋白，对序列片段测序得到结合位点

### 12.15.3.3 RNA 测序

- RNA 反转录为 cDNA 测序
- 只有外显子
- 同一基因多种 RNA 片段
- 均值与方差有相关性需要进行 log 变换后分析

## 12.15.4 Bioconductor

- 官方说明
- 使用 biocLite() 安装，安装后仍需要 library() 才能使用

```
source("http://bioconductor.org/biocLite.R")
biocLite()
```

### 12.15.4.1 数据结构

#### 12.15.4.1.1 分析数据

- F 行 S 列 F 代表芯片特征数，S 代表样本数

#### 12.15.4.1.2 表型数据

- S 行 V 列 V 代表样本特征，为分类或连续变量
- 如果表型数据解释不清，可以建立一个解释样本特征的 labelDescription 数据框，通过 `phenoData <- new("AnnotatedDataFrame", data=pData, varMetadata=metadata)` 建立 AnnotatedDataFrame 类型数据

#### 12.15.4.1.3 实验描述

- MIAME 类型对象
- 描述实验参数

#### 12.15.4.1.4 组装数据

- 将分析数据、表型数据、实验描述组装为一个 ExpressionSet 类型的对象
- `exampleSet <- ExpressionSet(assayData=exprs, phenoData=phenoData, experimentData=experimentData, annotation=annotation)`
- annotation 代表了一组相似实验设计的芯片数据的代号，通过相关代号可以索引到芯片特征信息并将其与其他数据如基因型、染色体位置等连接以便于分析
- 从 ExpressionSet 里可以按照表型数据提取子集，也就是对 S 截取 V 中特定子集 `exampleSet[, exampleSet$gender == "Male"]`
- `esApply` 用来针对 ExpressionSet 应用函数

#### 12.15.4.1.5 数据集应用

- `library(BioBase)`
- `library(GEOquery)`
- `geoq <- getGEO("GSE9514")` 从基因表达精选集（GEO）上得到数据表达集
- `names(geoq)` 得到文件名
- `e <- geoq[[1]]` 得到数据集
- `dim(e)` 查看表达集维度给出样本数与特征值，也就是测定序列数
- `dim(exprs(e))` 与上面等同，给出基因分析数据
- `dim(pData(e))` 给出 8 个样本的信息，信息头用 `names(pData(e))` 给出
- `dim(fData(e))` 给出特征与信息头列表
- `exprs` 为特征数 × 样本数矩阵 `pdata` 为样本数 × 信息头 `fdata` 为特征数 × 信息
- `experimentData(e)` 给出实验信息
- `annotation(e)` 特征注释
- `exptData(se)$MIAME` 给出实验相关关键信息
- `Y <- log2(exprs(bottomly.eset) + 0.5)` 对 NGS 数据加 0.5 取 2 为底的对数（防 0）得 -1，排除掉 0 后可得 MAplot 观察数值分布，一般为均值小差异大，均值大相对稳定
- `formula` 用来定义公式
- `model.matrix` 用定义的公式生成矩阵
- `rowttests(y[, smallset], group[smallset])` 定义分组，设定模型可进行 t-test，用火山图来表示

#### 12.15.4.1.5.1 IRanges

- `library(IRanges)` 序列范围
- `ir <- IRanges(start = c(3, 5, 17), end = c(10, 8, 20))` 定义序列
- `IRanges(5, 10)` 表示 5 到 10 这 6 个碱基对，可以 shift
- `range(ir)` 表示存在 ir 中序列的起止范围
- `gaps(ir)` 表示寻找 ir 中间隔片段
- `disjoin(ir)` 表示将 ir 中序列碎片化后互不重叠的片段

### 12.15.4.1.5.2 GRanges and GRangesList

- library(GenomicRanges) 基因范围
- gr <- GRanges("chrZ", IRanges(start = c(5, 10), end = c(35, 45)), strand = "+", seqlengths = c(chrZ = 100L)) 定义位于染色体 chrZ 上几个序列范围, 认为这些范围共同定义一个基因
- 可以 shift, 可以定义长度后 trim
- mcols(gr)\$value <- c(-1, 4) 定义该基因类型中的列并赋值
- grl <- GRangesList(gr, gr2) 多个 Granges 定义一个基因库
- length(grl) 给出基因库里基因个数
- mcols(grl)\$value <- c(5, 7) 定义该基因库类型中的列并赋值

### 12.15.4.1.5.3 findOverlaps

- gr1 gr2 为两个基因范围对象
- fo <- findOverlaps(gr1, gr2) 寻找两个基因重叠序列
- queryHits(fo) 与 subjectHits(fo) 提取两个基因重叠序号成对出现
- gr1[gr1 %over% gr2] 提取对应序列范围

### 12.15.4.1.5.4 Rle

- Rle(c(1, 1, 1, 0, 0, -2, -2, -2, rep(-1, 20))) 表示 4 组处理, 每组各有 3 2 3 20 个重复
- Rle 是一种压缩存储实验设计的方式, 可以用 as.numeric() 提取原始数据
- Views(r, start = c(4, 2), end = c(7, 6)) 提取对应实验组

### 12.15.4.2 数据读取

- microarray 或 NGS 数据由芯片厂商提供, 常见读取原始信息的包有 affyPLM、affy、oligo、limma
- 在 Bioconductor 里, 这些原始数据要转为 ExpressionSet 格式

### 12.15.4.2.1 Affymterix CEL files

- library(affy)
- tab <- read.delim("sampleinfo.txt", check.names = FALSE, as.is = TRUE) 读取样本信息
- ab <- ReadAffy(phenoData = tab) 读取样本数据, 探针层次
- ejust <- justRMA(filenames = tab[, 1], phenoData = tab) 直接读取为基因层数据
- e <- rma(ab) 对样本进行背景校正与正则化, 从探针层转化为基因层数据

### 12.15.4.2.2 背景干扰

- spikein 方法梯度加入已知浓度的基因片段阵列上进行 shift 类似拉丁方设计
- 可以看到同一基因不同片段大致符合先平后增模式开始阶段是噪声主导后面是浓度主导
- 使用类似基因模拟噪声主导相减后得到去干扰浓度效应但低值部分会导致方差过大
- 也可以使用统计建模方法模拟背景值与响应得到还原度更高的信号

### 12.15.4.2.3 正则化

- 基因组数据大多数为 0 加标样品变化正则化是为了还原这一结果
- 分位数正则化
- 局部回归正则化
- 稳方差正则化
- 当重复实验时直接用分位数正则会掩盖样品差异可以考虑只对加标基因正则化然后推广到全局

#### 12.15.4.2.4 探索分析作图

##### 12.15.4.2.4.1 MA-plot

- x 轴为两组基因组的均值，y 轴为两组基因组的均值差
- 用来表示两组平行间的差异

##### 12.15.4.2.4.2 Volcano plot

- 横坐标为处理间基因表达差异，纵坐标为差异的 $-\log_{10}(p.value)$
- 一般为火山喷发状，差异越大，p 值越小

#### 12.15.5 示例：甲基化数据分析

##### 12.15.5.1 读取数据

```
devtools::install_github("coloncancermeth","genomicsclass")
library(coloncancermeth)
data(coloncancermeth)
```

该数据集为结肠癌病人与对照的 DNA 甲基化数据集。

##### 12.15.5.2 数据说明

```
dim(meth)
dim(pd)
length(gr)
```

meth 为测序数据，pd 为样本信息，gr 测序片段信息。

```
colnames(pd)
table(pd$Status)
X = model.matrix(~pd$Status)
```

查看病患与正常人的分组并构建模型。

```
chr = as.factor(seqnames(gr))
pos = start(gr)

library(bumphunter)
cl = clusterMaker(chr, pos, maxGap=500)
res = bumphunter(meth, X, chr=chr, pos=pos, cluster=cl, cutoff=0.1, B=0)
```

按染色体生成因子变量，找出基因起始位点，然后利用 bumphunter 包寻找甲基化数据中某个阈值（0.1）下甲基化基因聚类的后出现的位置，聚类号，聚类相关性等信息寻找问题基因，可从中提取相关信息

```
cols=ifelse(pd$Status=="normal",1,2)
Index=(res$table[6,7]-3):(res$table[6,8]+3)
matplot(pos[Index],meth[Index,,drop=TRUE],col=cols,pch=1,xlab="genomic location",ylab="Methylation",ylim=c(0,1))

Index=(res$table[6,7]):(res$table[6,8])
```

```

test <- meth[Index,,drop=T]
colnames(test) <- pd$bcr_patient_barcode
test1 <- test[,cols==1]
test2 <- test[,cols==2]

test3 <- apply(test2, 2, mean)
apply(matrix, 1, rank)

```

从上面可以得到有差异的甲基化数据所在的基因位置并提取相关样本数据信息。可根据差异作图，得到两组数据甲基化水平差异所在的基因位置。可对差异进行平滑操作，得到位置。这样就可以知道甲基化发生的序列位置与水平差异的信息了。

下面的例子是用人类基因组数据探索潜在的 CpG 岛。

```

library(BSgenome.Hsapiens.UCSC.hg19)

Hsapiens[["chr1"]]

# 计算某染色体上潜在位点个数

countPattern('CG',Hsapiens[["chr1"]])

# 计算某染色体上特定序列比例 观察与期望出现的比例

CG <- countPattern('CG',Hsapiens[["chr1"]]) / length(Hsapiens[["chr1"]])
GC <- countPattern('GC',Hsapiens[["chr1"]]) / length(Hsapiens[["chr1"]])

table <- alphabetFrequency(Hsapiens[["chr1"]])
expect <- table['C'] * table['G'] / (length(Hsapiens[["chr1"]]))^2

CG / expect

```

## 12.16 链接

- Michael Love 的教案
- 生信前沿信息集



# 章 13

## 流行病学

### 13.1 声明

- 本笔记来自于波士顿大学在线教程与北卡大学教堂山分校的公开课
- 二手知识，谨防消化不良
- 翻译有误之处见谅，烦请告知，谢谢！

### 13.2 早期疾病的概念

- 渔猎时期主要问题是食物的供应与营养均衡问题
- 农耕时期开始群居，出现疾病的流行问题
  - 神秘主义，迷信与神的惩罚
  - 希波克拉底：理性思考疾病起源，提出体液学说
  - 欧洲流行 300 多年的黑死病，开始认为病因是“瘴气”，其实是老鼠跳蚤上细菌
  - 没有验证病因与疾病关系的方法与预防措施
- 工业革命时期城市规模迅速扩大，分工细化，出现职业暴露健康问题

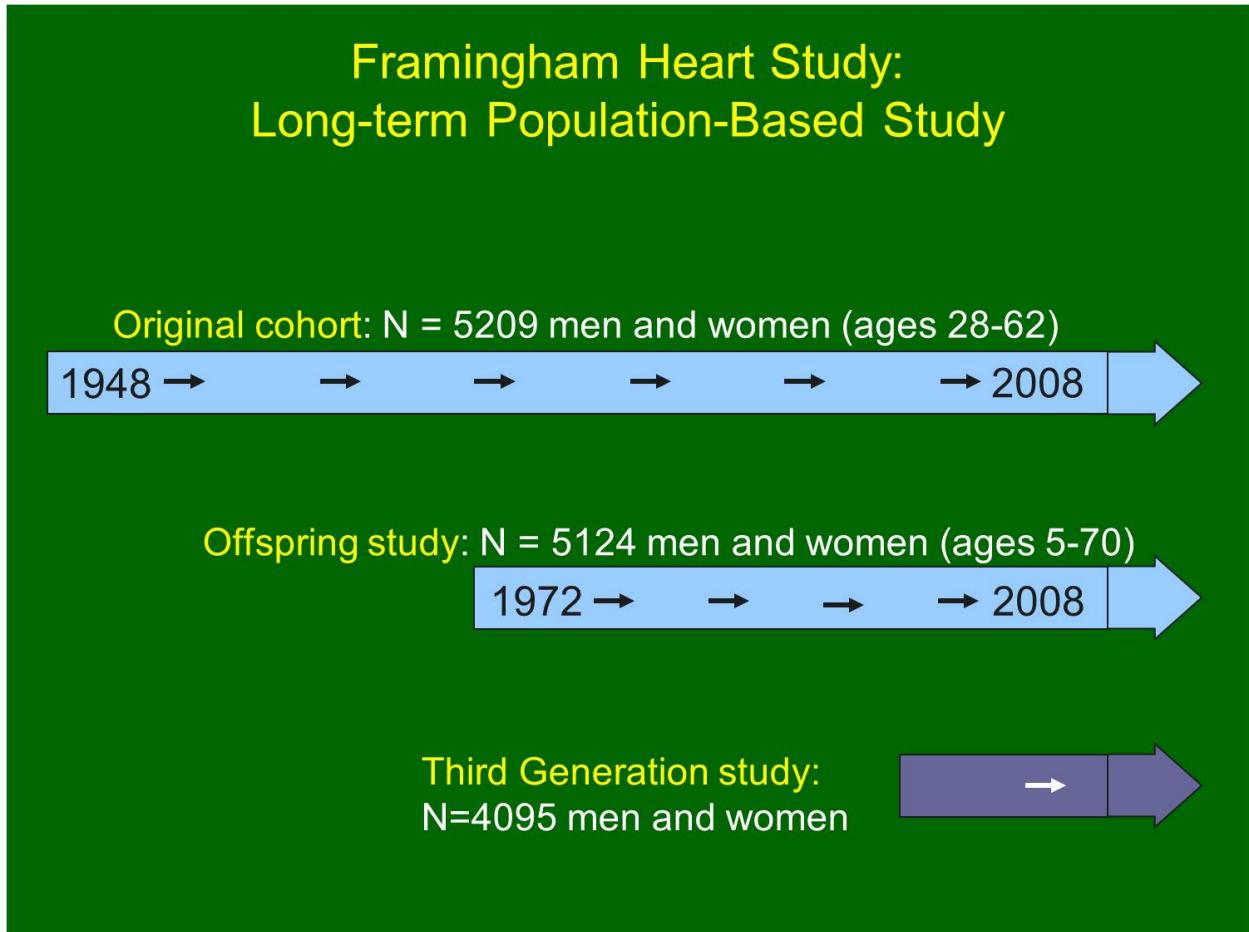
### 13.3 近代流行病学关键人物

- Girolamo Fracastoro (1546) 认为疾病来自种子
- John Graunt - The Bills of Mortality (1662) 记录伦敦死亡率，出生率数据并进行分析
- Anton van Leeuwenhouk (1670s) 显微镜之父，首先观察到细胞
- John Pringle and “Jail Fever” (1740s) 研究军队与监狱卫生与疾病并讨论了与伤寒的关系
- James Lind and Scurvy (1754) 进行了第一例临床控制实验，验证柑橘对坏血病的治疗作用
- Francois Broussais & Pierre Louis (1832) 提出放血疗法，没有验证但沿用几个世纪
- Ignaz Semmelweis and Oliver Wendell Holmes (1840s) 前者发现了某种产科疾病是由刚解剖完尸体的学生引入的，后者推行了疾病可能来源于医护人员的理念而饱受争议
- John Snow - The Father of Epidemiology (1850s) 流行病学之父，研究并验证了霍乱与城市供水的关系
- Louis Pasteur (late 1800) 提出巴氏消毒法与疫苗理论
- 公共卫生的概念 (1850-1875) 起源于人口统计及 18 世纪的启蒙运动例如功利主义的兴起对公众健康的关注

## 13.4 现代慢性病流行病学

- 肺癌：工业革命后的肺癌发病率很高，普遍认为是工厂公路导致，但后续研究表明吸烟可能是主要因素，该研究促进了病例对照研究的发展
- 佛雷明翰心脏病研究：48 年起追踪心脏病研究，已持续三代人

```
knitr::include_graphics('images/Framingham1.jpg')
```



## 13.5 流行病学基本概念

### 13.5.1 基本假设

- 病有病因非随机
- 病因可察可研究

### 13.5.2 定义

- 研究人群中疾病分布与成因的学科

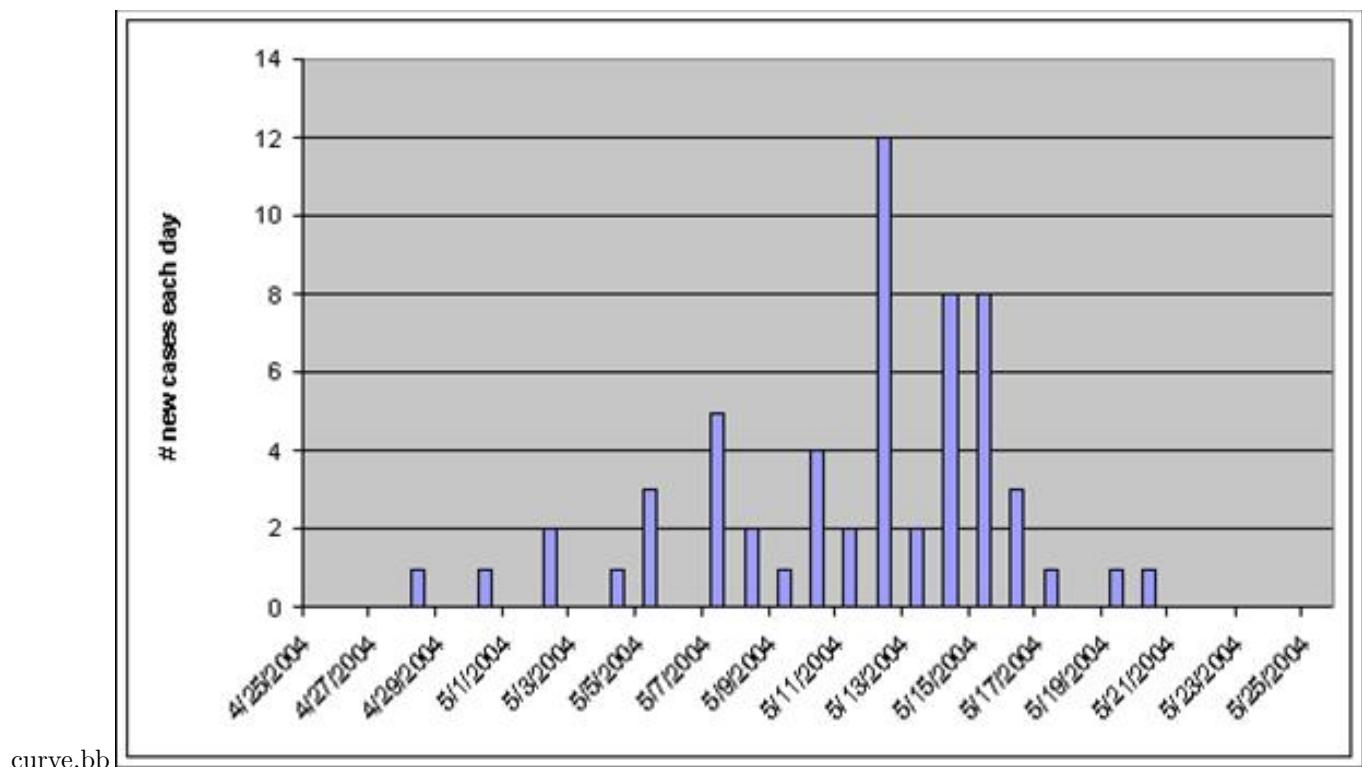
| 阶段        | 研究类型            |
|-----------|-----------------|
| 观察 & 形成假说 | 案例研究／断面研究／生态学研究 |
| 观察研究假设检验  | 病例控制研究与队列研究     |

| 阶段       | 研究类型 |
|----------|------|
| 临床研究假设检验 | 临床实验 |

## 13.6 描述性流行病学

- 关注不同时间、地点、人群的差异、相似性与相关性，形成假说
- 案例：传染病暴发
  - 早期关注案例采访，寻找共同点
  - 对地理位置作图寻找空间关系
  - 对时间趋势作图寻找变化规律
- 流行曲线：横轴日期，纵轴新增病例
  - 点源爆发：单峰，潜伏期相对一致

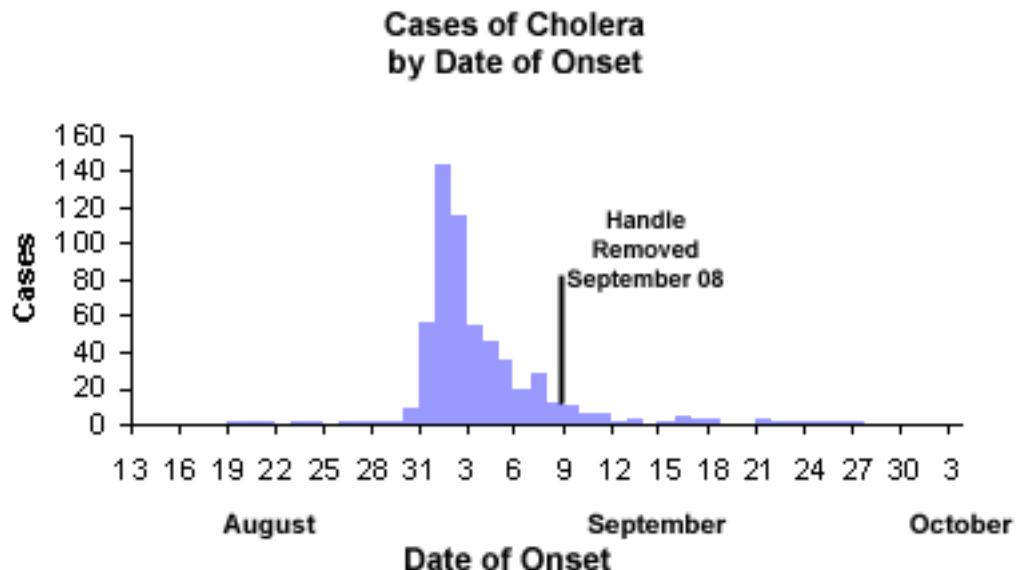
```
knitr:::include_graphics('images/epidemic%20curve.jpg')
```



curve.bb

- 持续源爆发：单峰，潜伏期持续出现

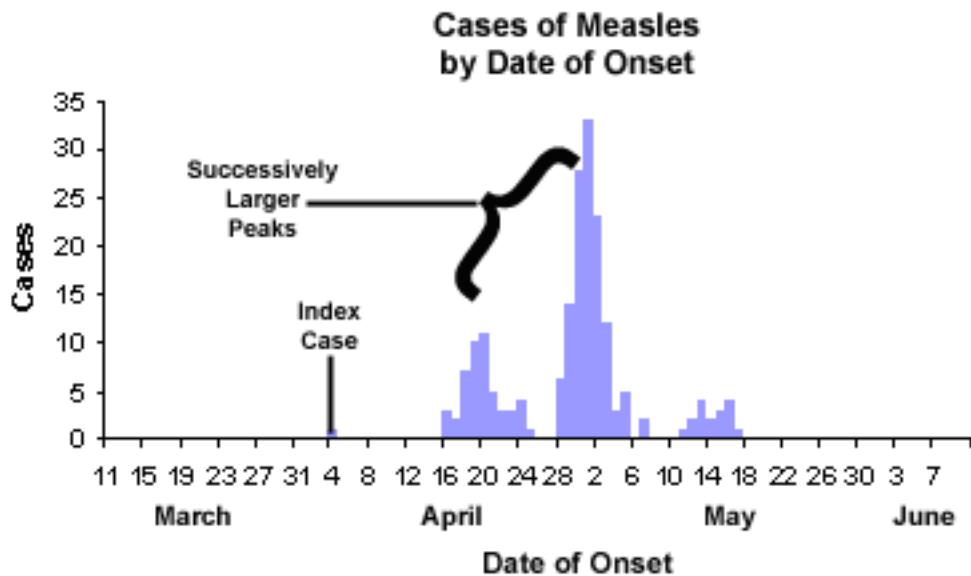
```
knitr:::include_graphics('images/EpidemicCurve_Cholera.png')
```



- 逐步流行: 多

峰, 存在人对人传染

```
knitr:::include_graphics('images/EpidemicCurve_Measles.png')
```



### 13.6.1 疾病爆发的研究步骤

- 准备研究
- 验证诊断与爆发的存在
- 定义案例并寻找案例
- 进行描述性流行病学研究确定时间、地点与人群的差异
- 生成爆发原因与来源的假设
- 假设检验
- 制定控制与预防措施
- 交流研究发现

### 13.6.2 慢性病的描述性流行病学

- 人群特质：年龄，性别，种族，职业，饮食习惯，宗教习惯，业余活动等
- 地点：慢性病地域差距
- 时间：大趋势，季节性，片断性
- 其他：环境变化，诊断精度，医疗水平，人群年龄分布

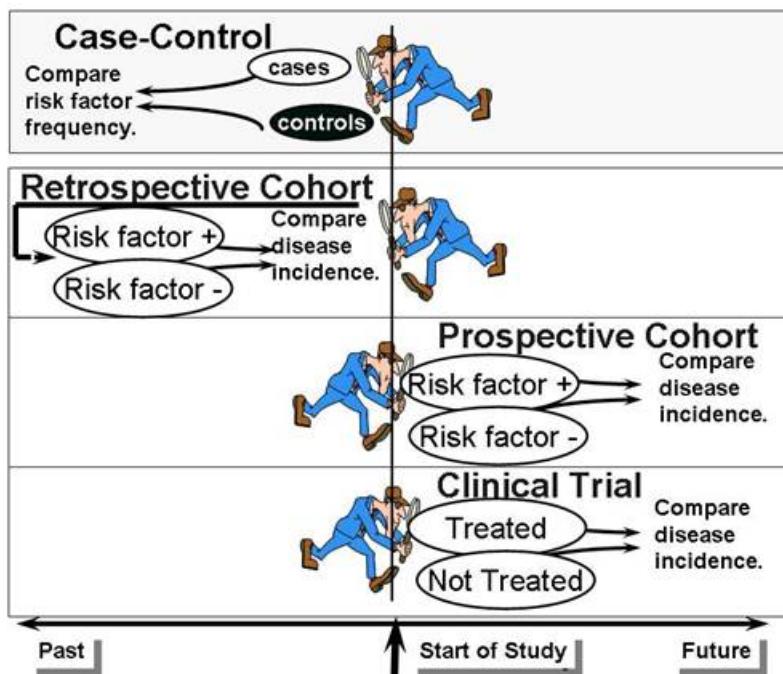
### 13.6.3 描述流行病学分类

- 病例报道单个案例
  - 艾滋病血液传播的发现
  - 无注射狂犬疫苗后痊愈
- 系列病例多个案例
  - 男同性恋间获得性免疫缺陷传染
- 断面研究同一时间对特定人群的健康状况与风险因子进行调查
  - HIS
  - NHANES
- 生态学研究以群体为单位研究区域平均暴露状况

## 13.7 分析流行病学

- 不同于描述流行病学提出假设，分析流行病学进行假设检验
- 队列研究：定义基线与风险人群
  - 前瞻性队列研究：参与者参加的时候不出现健康效应
  - 回顾性队列研究：根据已经出现的健康效应反向追查风险因子
- 临床实验：风险因子由研究人员指定
- 病例对照研究：不用来研究发病率，侧重风险比，不追踪，根据已有状况回溯，对幸存者采样，适合稀有病症的研究

```
knitr:::include_graphics('images/paste_image47.jpg')
```



- 判断流程

- 是否个人（生态学研究）
- 是否有对照（系列案例）
- 是否追踪（断面研究）
- 是否不先选取出现健康状况的组（病例对照研究）
- 是否不出现健康状况（回顾队列研究）
- 是否指定对照组（前瞻队列研究）
- 临床实验

## 13.8 疾病监控

- 早期教堂记录出生率与死亡率
- 1662 John Graunt “Bills of Mortality”
- 1837 英国建立 General Registrar’s Office 记录市民出生、死亡与婚姻
- John Snow 对霍乱数据的分析
- 1842 马萨诸塞州开始记录出生死亡状况
- 1901 全美开始记录疾病流行状况
- 1925 强制执行疾病监控
- 目前基本是 CDC 控制，除了强制汇报，也有主动收集
- 综合疾控，不等确诊收集症状，例如 google flu

## 13.9 疾病频率的测量

### 13.9.1 人群

- 固定人群（相对固定，或由事件定义）
- 动态人群（由当前状态决定的人群）

### 13.9.2 患病率 (prevalence)

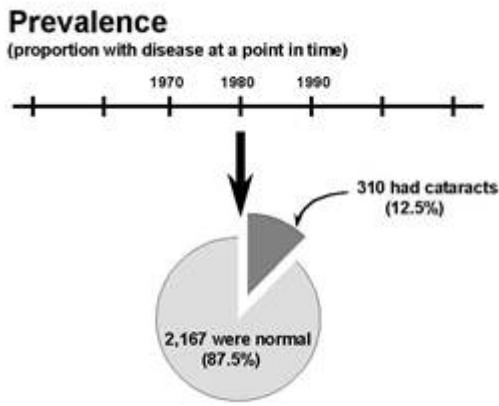
表示在指定时间里具有某种健康效应的人群比例，非新增

$$prevalence = \frac{affected individuals}{total individuals in the population}$$

举例：患病率 0.25 表示人群中有 25% 的人在指定的时间段里受某种健康效应影响

经常会导致因果推断不准，因为影响因素产生的效应被本来的效应覆盖了

```
knitr::include_graphics("images/paste_image18.jpg")
```



### 13.9.3 风险 (risk)

也称作发病率 (incidence 或 cumulative incidence)，表示在一段给定时间里新增某种健康效应的比例

$$risk = \frac{newcases}{totalindividualsatrisk}$$

举例：5 年风险 0.1 表示在 5 年里某个个体有 10% 的几率出现某种健康效应

前瞻性研究 (prospective studies) 常用，但控制性研究 (case-control studies) 里总体风险无法确定，不能使用。

### 13.9.4 比率 (rate)

也称作发病率比率 (incidence rates) 表示在一个人群中某种健康效应出现的速度。单位为每个人年，个人年表示风险个体参与研究到出现健康效应的总时间

$$rate = \frac{newcases}{totalperson - timeatrisk}$$

举例：0.1 案例每个人年表示对于每 10 个人追踪 1 年或 2 个人追踪 5 年将会有一个案例出现

人群无限状态下，有

$$risk = rate \cdot time$$

考虑到人口的指数衰减，有

$$risk = 1 - exp(-rate \cdot times)$$

当风险率非恒定时，或者对时间分段计算，或者进行生存分析

人群出现稳态时，有

$$\frac{prevalence}{1 - prevalence} = rate \cdot Avg.Duration$$

当患病率很低时，有

$$prevalence = rate \cdot Avg.\text{Duration}$$

疾病的持续期可计算为

$$Avg.\text{Duration} = \frac{prevalence}{rate}$$

### 13.9.5 其他频率测量

- 分类比率：如年龄，性别，种族等
- 病态比率：不致命的状态
- 死亡率
- 致死率：患病中导致死亡的比率
- 攻击率：短期食物中毒
- 生育率：育龄妇女一年内生育新生儿的比率
- 新生儿死亡率：一岁以下新生儿死亡率
- 特殊患病率：体检率，新生儿感染率，非新增

## 13.10 联系测量

- 测定频率不涉及对比，探索关系需要对比
- 不同暴露状态下的频率差或者比表征

### 13.10.1 风险比与比率比 (risk ratio rate ratio)

$$riskratio = \frac{risk_{exposed}}{risk_{unexposed}}$$

$$rateratio = \frac{rate_{exposed}}{rate_{unexposed}}$$

表示暴露与健康效应的关系强度，1 表示无关，但比例关系不能给出绝对差异表述风险比不要使用更多或更少，如果更多或更少需要减一除以风险比：相比不服用，服用阿司匹林有 0.57 倍的心肌梗死风险或 43% 的风险下降

### 13.10.2 对照组

暴露量最小的一组通常作为风险比计算中的对照组

### 13.10.3 风险差 (risk difference)

$$risk = risk_{exposed} - risk_{unexposed} = \frac{casesinexposedgroup}{totalatriskinexposedgroup} - \frac{casesincontrolgroup}{totalatriskincontrolgroup}$$

正数表明对某种健康效应有促进作用，负数表示有抑制作用，要指明时间区段

### 13.10.4 归因比例 (Attributable Proportion Among the Exposed)

$$\text{attributableproportion} = \frac{\text{riskratio} - 1}{\text{riskratio}}$$

暴露组风险中归因于该原因的比例

### 13.10.5 人群归因比例 (Population Attributable Fraction)

$$\text{populationAttributableFraction} = (\text{proportionofcasesexposed}) \cdot (\text{attributableproportionintheexposed})$$

人群中风险归于该原因的比例

### 13.10.6 胜率比 (odds ratio)

$$\text{oddsratios} = \frac{\text{odds}_{\text{exposed}}}{\text{odds}_{\text{unexposed}}}$$

常用在控制性研究里替代风险比或比率比，这时风险比无法计算，但几率比可以在总体效应比较小与特殊采样技术使用的时候近似于风险比或比率比，解释起来与它们一致

## 13.11 随机误差

- 偏差，混杂与随机误差是流行病学采样中最常见问题
- 随机误差也是采样误差
- 置信区间用来表示随机误差而非混杂偏差等误差
- 95% 置信区间与 p 值计算方法一致，可用来判断是否统计显著

## 13.12 研究道德

- 无论目的如何，以人作为研究对象是不道德的
- 纳粹在二战期间集中营里使用人作为研究对象，1946 年审批时提出 Nuremberg Code：

Voluntary consent of the human subject is absolutely essential. The experiment must yield generalizable knowledge that could not be obtained in any other way and is not random and unnecessary in nature. Animal experimentation should precede human experimentation. All unnecessary physical and mental suffering and injury should be avoided. No experiment should be conducted if there is reason to believe that death or disabling injury will occur. The degree of risk to subjects should never exceed the humanitarian importance of the problem. Risks to the subjects should be minimized through proper preparations. Experiments should only be conducted by scientifically qualified investigators. Subjects should always be at liberty to withdraw from experiments. Investigators must be ready to end the experiment at any stage if there is cause to believe that continuing the experiment is likely to result in injury, disability or death to the subject.

- 1964 年，WMA 接受赫尔辛基宣言
- 塔斯基吉梅毒研究，没有征得患者同意，也没有进行有效治疗，1972 年泄漏，1974 年美国出台人类被试保护法案，所有涉及人的研究需要通过IRB审核
- 1978 年，议会出台人类被试研究指南
- 法律不强制，但基金一般有相关要求
- 相信研究有益等同于对其怀疑才可进行
- 安慰剂可以使用，但要保证患者最终能得到最好的治疗

## 13.13 临床实验

- 分为预防性与治疗性干涉研究
- 新药研发的四阶段
  - 8-80 人小规模评价安全性，副作用及副作用出现的剂量
  - 80-200 人中等规模测试有效性，副作用及与剂量的关系
  - 200-40,000 人大规模测试其与当前治疗方式的副作用强度
  - 推向市场后的监测，测试罕见但严重的副作用，例如H1N1 疫苗

### 13.13.1 研究对象

- 人群分层考虑是否为目标人群及是否愿意参加
- 内部验证准确性，外部验证广泛性
- 样本数由功效决定，由于疾病发病率低，很小的差异也需要很大的样本来保证功效

### 13.13.2 对照组与控制组

- 排除混杂因素需要考虑除考察因素外其他因素在研究客体中分配均匀
- 分配方法包括自我前后对比与随机非随机分配
- 屏蔽
  - 单盲：被试不知道是否是处理组
  - 双盲：研究人员与被试都不知道处理组
  - 三盲：进行处理的人也不知道是否是处理组
- 安慰剂（placebo），也就是无效药
- 装假（sham），假装进行某个操作流程（有道德风险）
- 安慰剂效应：接受治疗的人都认为会从中收益，即使知道是安慰剂也会产生该效应
- 服从度，处理组与控制组要区分明显，内部一致
  - 设计尽量简单
  - 被试生活规律
  - 通知明确
  - 实时追踪
  - 屏蔽处理信息
  - 对不服从的仔细询问，收集未使用药片，收集血液尿样进行评价
- 掉队会导致功效降低及存在偏误

### 13.13.3 临床分析中的问题

- 随即控制时要给出基线信息
- 混杂因素可能不在基线里而是直接影响结果
- 如果混杂因素在调整前后影响结果超过 10%，那么就要进行调整- 希望被处理分析，保证处理与对照中接受治疗的意愿接近
- 二次分析，只对接受的人进行结果分析，失去随机性与一定样本数及混杂控制
- 某项研究同时有利弊，利大于弊，是否继续研究？
- 预防花费如果很高，是否值得推荐？

## 13.14 队列研究

### 13.14.1 前瞻性队列研究

- 研究开始时没病，记录基线，追踪个人
- 案例：BMI 与心脏病关系

### 13.14.2 回顾性队列研究

- 适合职业暴露，回溯暴露状况
- 案例：游泳池污染事件

### 13.14.3 双向队列研究

- 同时进行前瞻性与回顾性队列研究较少见
- 案例：橙剂喷洒飞行员追踪，急性与慢性暴露

### 13.14.4 固定队列与开放队列

- 固定队列表示人数固定或只能减少，例如日本原子弹受害者追踪，多数研究是固定队列
- 开放队列表示人数动态，可随时加入

### 13.14.5 研究对象

- 一般人群队列或特殊暴露队列（事件幸存者）
- 对比组越接近越好，信息收集越全越好
  - 内部比对：同队列未暴露被试，肥胖调查中不肥胖的人
  - 外部比对：内部不存在未暴露时，例如化学品职业暴露
  - 一般人群比对：从国家抽取基础数据，但因为健康工人效应现在不常用，可使用标准死亡率（SMR）或标准流行指数（SIR）来测量联系，也就是用基础数据计算期望值与实际值的对比
- 健康工人效应：能工作的工人比一般人群要健康

### 13.14.6 队列追踪

- 队列研究一般开始时没有偏见
- 追踪率低于 60% 不可靠，丢失 20% 可能因掉队原因关联结果导致偏误
- 不愿参与会导致偏差
- 回溯研究会因保留疾病比例高于保留正常病例的原因导致选择偏误

### 13.14.7 优点

- 直接给出暴露与效应的时间序列关系
- 可直接计算疾病的风险率
- 可用来评价稀有暴露
- 可用来同时评价多种暴露
- 基本无选择偏误

### 13.14.8 前瞻性队列研究缺点

- 耗时长
- 费用高
- 不适用稀有疾病
- 不适用潜伏期长疾病
- 掉队会导致偏差

### 13.14.9 回顾性队列研究缺点

- 不适用稀有疾病
- 记录不匹配研究需要结果不好
- 过去记录丢失混杂因素信息
- 暴露组与对比组很难区分
- 掉队会导致偏差

### 13.14.10 偏误

- 选择人群无法代表群体
- 产生原因
  - 在病例对照研究中控制组没有代表性
  - 追踪丢失率在控制与处理组不同
  - 是否愿意参加影响暴露与结果
  - 健康工人效应 (职业暴露)
  - 诊断标准不同
  - 回顾性研究中回顾会放大暴露或结果
  - 观察偏误，如果是非特异性区分错误，那会指向空假设
  - 记录偏误，引导性问题
  - 暴露比结果难评价，结果比暴露稀有，因而暴露更容易产生导致结论错误的偏误
  - 灵敏度与特异性对风险比与风险差的影响不同

## 13.15 病例对照研究

- 现有案例，后回顾暴露状态，两者在研究前是独立的
- 适用于稀有疾病，只能计算胜率比而不能计算风险比
- 经常内置于已有的队列研究，适用于稀有疾病假设
- 当研究对象为人群而不是队列研究中的未发病人群时，允许发病者作为控制组
- 案例：DES 与子宫癌
- 病例来源：住院病人、死亡证明、死亡注册、断面研究
- 对照来源：代表群体的组、独立采样且采样策略一致避免代表性丧失
- 随机电话访问是之前一种选择对照的方法，由于存在偏误（无法区分居民与商业电话，固定电话使用率降低）而逐渐被替代
- 对照组的数量选择要考虑统计功效
- 采样方法：幸存者采样，基于队列采样（按队列开始时风险人群），风险组采样（出现案例时存在风险的人群）后两种可以不考虑稀有假设，因为他们对照可代表整体，胜率比可用来估计风险比
- 优点：对罕见病高效，节约成本，可动态研究
- 缺点：选择偏误，对罕见暴露低效，不能计算风险

## 13.16 标准化

- 粗比率 (crude rates) 忽略了人群组成差异，需要调整
- 死亡率上如果两组中有一组老年人占总体比率高，那么会使两组风险比较时有偏差
- 用整体人群作为基础分布，比率乘各分组人数之后求和得到标准比率，其实质是将各分组年龄分布归一来消除年龄偏误
- Standardized Incidence Ratios 标准发病率用整体发病概率作为基准，计算各分组发病人的期望值并对比观察值

### 13.17 混杂

- 混杂因素是同时对暴露与结果产生影响的因素，例如唐氏综合症研究中出生的顺序其实对病症无影响，孕妇年龄为该研究的混杂因素
- 混杂因素判据：对暴露与结果都有影响；在暴露组间分配不均；不能是暴露与结果的中间步骤（饮酒通过升高 HDL 来降低心血管病发病率，HDL 与两者相关但不是混杂因素）
- 混杂因素可能是另一个风险，也可以是预防因素，也可以是其他替代物
- 残差混杂表示在排除混杂因素后由于排除不全或分类错误或未知导致的混杂
- 现象或禁忌混杂，暴露与结果实际受结果的反馈影响，例如抗抑郁药与绝育的关系中抑郁本事会对绝育产生影响，这样在观察研究中不易区分
- 因果互换，例如母乳对婴幼儿有益，但有研究发现母乳可能造成营养不良，但后来人们发现其实是因为调查人群中婴儿出现体重偏轻或腹泻的家庭往往停止使用母乳喂养，案例；另一个案例是止痛药与肾衰的研究中并非服用止痛药导致肾衰而是因为糖尿病多导致肾衰而糖尿病人经常服用止痛药
- 研究设计中防止混杂可通过限制研究人群，个体匹配与随机化实现
- 数据分析中混杂控制 - 分层，例如年龄分组后原有差异可能就消失
- 多分层方法可采用 CMH 方法计算风险比，其实就是对分组比率加权来忽视分组因素的影响，影响因素多要采用多元分析

### 13.18 效应修饰 (EMM)

- 指由于另外的变量导致效应分类的状态，例如年龄可能造成某种药药效相反，可理解为线性模型中的交互作用项- 测定有混杂因素与无混杂因素下的两个风险比，如果差异很大且差异区间包括原始风险比，则存在修正测量效应；如果差异不大但影响原始风险比，则要同时考虑混杂因素；如果两者同时存在，则要考虑分层讨论混杂的情况
- 存在 EMM 时不能使用 CMH 方法，因为此时样本不适合混合，应该分层讨论，可用卡方检验 EMM 的存在与否
- 统计交互作用与生物学的交互作用需要区分

### 13.19 多变量方法

- 本质上是多元回归，通过参数判断变量影响
- 混杂变量用增加参数的方法排除，参数变化超过 10% 可认为明显混杂
- EMM 用交互作用项排除，观察是否显著
- 最终模型是否含有不显著相具体分析

### 13.20 筛选

- “detectable pre-clinical phase” 或 DPCP 表示在筛选与有症状后检测之间的时间
- 筛选的价值（高血压中测血压）
  - 疾病很严重（子宫癌）
  - 症状发生前的治疗效果要比发生后好
  - DPCP 疾病流行概率很高
- 筛选的限制
  - 胆结石中预先检测对治疗没意义，都是大了以后手术去除
  - 肺癌中检测到了也无法有效治疗
  - 疾病不流行
- 好的筛选应具有的标准
  - 便宜
  - 容易操作
  - 最小化不适

- 可靠
- 有区分
- 测试验证
  - 灵敏度（真阳性占阳性比例）
  - 特异性（真阴性占阴性比例）
  - 真阳性预测值（真阳性占阳性比例）这个值会随流行度变化而变化，即使灵敏度特异性都高，较低的流行度也会降低预测准确性，所以测试要针对易感人群并计算流行度
  - 真阴性预测值（真阴性占阴性比例）
  - gold standard “金标”
  - ROC 曲线左上方靠近
  - 多数情况可以接受假阳性而提高灵敏度
  - 前列腺癌筛查的案例
- 测试本身的缺点
  - 低流行率的假阳性
  - 假阴性
  - 前列腺相关报道
  - 过度测试
  - 癌症测试
- 评估筛选中需要注意的偏误
- 宫颈癌，乳腺癌也在常见筛选之中

## 13.21 因果推断

因果推断在流行病学中很重要，但目前没有标准来界定因果而仅仅有一些指南。

### 13.21.1 Hill 因果标准

由流行病学家 Austin Bradford Hill 提出的 9 条判断因果关系的标准，充分不必要条件，不能作为清单使用。

#### 13.21.1.1 联系强度

由风险比，比率比，胜率比来测量，越强代表因果联系越大，反之不成立。

#### 13.21.1.2 数据一致性 (Consistency)

一致性用来排除解释某健康效应的其他可能，缺少不代表没有，可能有其他共有因素，越强因果联系越大。

#### 13.21.1.3 特异性 (Specificity)

因素结果 1 对 1，该标准不是特别有效，有些因素会对应多种结果。

#### 13.21.1.4 时序性

因果必要条件，先有因后有果。

#### 13.21.1.5 剂量效应关系

剂量效应关系是充分不必要条件，例如阈值效应。

### 13.21.1.6 生物合理性

基础研究，没有流行病学研究前的实验室数据如毒理学研究。

### 13.21.1.7 相干性

相干性表示新数据不应该与现有证据矛盾。

### 13.21.1.8 实验证据

随机控制实验的结果，改变原因结果不同。

### 13.21.1.9 类比

最弱的标准，主观性较强。

## 13.21.2 部分原因理论

Kenneth Rothman 提出，认为健康效应的原因可看成一个饼图，缺少任一部分结果都不会发生，用来了解结果的发生过程。

### 13.21.3 逆向模型 (Counterfactual models)

考虑无暴露状态下是否产生效应的思路，群组水平考察。

### 13.21.4 有向无环图 (DAGs)

概念流程图，考虑混杂因素。

## 13.22 论文研读

- 科研论文类型包括原始研究（描述性与分析性）、方法、荟萃分析与评论
- 文章结构包括题目、作者、摘要、前言、方法、结果、讨论、结论、致谢、文献引用与图表
- 依次浏览摘要（概况），前言（问题的重要性），讨论（看结论与意义），方法（看实验设计），结果（看图表）并记录疑点

### 13.22.1 Introduction

- What was the primary question that the authors were trying to answer? Why were they asking this? Rationale? What was their goal?

### 13.22.2 Methods

- What type of study design was used?
  - Was this a logical choice, given the goals of the study?
  - What are the weaknesses of this study design?
  - What problems and biases might have occurred?
- How were subjects identified and enrolled? How successful was enrollment?
  - Could selection bias have occurred as a result of control selection bias, or differential non-participation in a case-control study?
  - Did selection of controls meet the “would” criterion?
  - If it was a cohort study, how complete was follow up
- How carefully was the exposure of interest defined?
  - How was the exposure assessed?
  - What was the quality of the exposure data?
  - Was exposure data validated?
- How carefully was the outcome of interest defined?
  - How was it assessed? Was it validated?
- Could selection bias have affected the results?
- What was the potential for information bias?
  - Non-differential misclassification? Errors in recording or coding of data? General inability of subjects to remember?
  - Differential misclassification? Recall bias? Interviewer bias? Recorder bias? Differential quality of data?
- What were the likely confounding variables?
  - Did the authors control for confounding in the design of the study, in the analysis, or both?
  - Did they fail to account for any potentially important confounders? Was control of confounding adequate? Could there have been residual confounding?
  - Did they perform stratified analysis? Did they use regression analysis?
- Would these problems bias toward the null or away from the null?

### 13.22.3 Results

- Do the results suggest an association?
  - If so, how was it assessed, and how strong was the association?
  - Did the authors estimate risk ratios or risk differences?
  - How precise were the measures of association? Was the sample size adequate? Did the authors report confidence intervals? p-values?
  - Did the authors adequately assess random error?

### 13.22.4 Discussion

- Was the interpretation appropriate?
- Are the results of this study consistent with other studies in this area? If there are differences with other study findings, what could they be due to?

### 13.22.5 Conclusion

What are the public health implications of the study?

### 13.22.6 Clinical Trial

- Were patients randomly assigned to the comparison groups?
- Was the study blinded? Did the patients or doctors know which group the patient was in?
- Was the randomization effective in creating two groups which were similar with respect to age, gender, race, and other potentially confounding variables?
- Did patients adhere to the treatment? Did patients drop out?
- Were appropriate statistical tests used to compare the groups?
- Were the groups analyzed based on their randomized assignment, i.e. a so-called “intention to treat analysis”
- Was the sample size large enough to detect a meaningful difference if it had existed?

### 13.22.7 Cohort Study

- How did they select the subjects in the comparison groups? Were the groups comparable with respect to other factors?
- How did they ascertain risk factor status? Was the data accurate?
- Could there have been bias?
- How complete was the follow up data?
- Was the statistical analysis appropriate?
- Did they control for possible confounding variables?
- Was the sample size adequate to detect clinically important differences if they existed?

### 13.22.8 Case-Control Study

- What was the source population? How were cases and controls defined?
- Was there selection bias? Was the ‘would’ criterion met?
- How was information collected? Was it accurate? Was it collected in a comparable way in both groups?
- Could there have been recall bias?
- Interviewer bias?
- Was the statistical analysis appropriate?
- Did they control for possible confounding variables?
- Was the sample size adequate to detect clinically important differences if they existed?

### 13.22.9 Screening Test

- If so, did they have an independent blind comparison with a reference diagnostic technique, i.e., a “gold standard”?
- Was the diagnostic test evaluated in an appropriate group of patients, similar to those you would find in your practice?
- Did they address the ability of the test to discriminate between normal and abnormal? How was abnormality defined? Did they calculate sensitivity and specificity or likelihood ratios or report their data in such a way that you could calculate them?

### 13.22.10 Additional Considerations

- Are the Findings Important?
- External Validity (Generalizability)



# 章 14

## 博弈论

### 14.1 术语

- 博弈论说白了就是讲两方势力在一件事上为了自己的最大利益所采取的行动或决策的理论
- 参与者 (players)  $N = 1, \dots, n$
- 参与者  $i$  有一组行动 (Actions), 行动的集合  $a = (a_1, \dots, a_n) \in A = A_1 \times A_2 \times \dots \times A_n$
- 参与者  $i$  的每个行为的收益 (Payoffs) 都可以用  $u_i : A \rightarrow \mathfrak{R}$  这个函数表示  $u_i(a)$  表示某个行为产生的收益,  $u = (u_1, \dots, u_n)$  是效用函数的集合
- $n$  个参与者的标准博弈 (normal form)  $\langle N, A, u \rangle$
- 两人博弈可以用矩阵 (matrix) 来描述, 行代表选手 1, 列代表选手 2, 行对应选手 1 的行动  $a_1 \in A_1$ , 列对应选手 2 的行动  $a_2 \in A_2$ , 每个单元列出每个参与者的收益, 先行选手, 后列选手
- 纯竞争博弈 (Games of Pure competition): 两个参与者收益对立且对于所有行动集合  $a \in A, u_1(a) + u_2(a) = c$ ,  $c$  是常数, 零和博弈是一个常数为 0 的纯竞争博弈, 此时我们只用考虑一个参与者的收益函数就可以了

### 14.2 支配策略 (dominate strategy)

- 行动单一称为纯策略 (pure strategies) 有一定概率分布的行动策略称为混合策略 (mixed strategies)
- 对于一个参与者, 不论其它参与者采取任何策略, 某策略都会得到最大的收益
- $a_i \in a_i$  为强支配策略当且仅当参与者  $i$  的收益  $u_i(a_i, a_{-i}) > u_i(a'_i, a_{-i})$ , 如果  $a'_i = a_i$ , 那么为弱支配策略

### 14.3 最佳回应 (Best response)

- 对于一个参与者, 当已知其他参与者的行动后, 收益最大的策略
- 参与者  $i$ , 对于其它参与者的策略  $a_{-i} \in a_{-i}$  的策略如果  $u_i(a_i, a_{-i}) \geq u_i(a'_i, a_{-i})$

### 14.4 纳什均衡 (Nash Equilibrium)

- $a = \langle a_1, \dots, a_n \rangle$  是一个纯策略纳什均衡当且仅当对于任何一个行为  $i$ , 有  $a_i \in BR(a_{-i})$
- 如果任何参与者改变行为的收益都不会增加, 那么此时进入纳什均衡
- 纳什均衡是一系列行为的列表, 这些行为都是稳定的
- 支配策略是纳什均衡但反过来不一定对
- 任何有限博弈都存在一个纳什均衡 (纳什 1950 年提出)

## 14.5 帕累托最优 (Pareto Optimality)

- 某个结果是帕累托最优当且仅当没有其他结果可以全局帕累托支配这个结果
- 帕累托最优表示某个博弈结果不差于其他博弈结果
- 纳什均衡不一定是帕累托最优（囚徒困境）

## 14.6 混合策略 (Mixed strategies)

- 策略  $S_i$  指对于每个参与者行动的概率分布集合
- 概率  $Pr(a|s) = \prod_{j \in N} s_j(a_j)$
- 期望收益函数  $u_i(s) = \sum_{a \in A} u_i(a) Pr(a|s)$
- 随机策略会使对手混乱进入动态，很多博弈只存在混合策略纳什均衡而没有纯策略纳什均衡（石头剪子布）
- 混合策略的目的在于不论你使用哪一种行动，对方的收益都不变

| 选手 | 甲   | 乙   |
|----|-----|-----|
| 丙  | a,b | c,d |
| 丁  | e,f | g,h |

- 对于选手 1 而言，采取丙行动的收益是  $aq + c(1 - q)$ ，采取丁行动的收益是  $eq + g(1 - q)$ ， $q$  代表选手 2 采取甲行动概率
- 如果要达到双方均衡，那么不论采取什么行动收益应该一致，不会因为对方概率的变化而偏离，那么我们就可以求解均衡时选手 2 的行动概率：

$$aq + c(1 - q) = eq + g(1 - q)$$

$$q = \frac{g - c}{a + g - c - e}$$

- 这个结果表明选手 2 采取行动主要参考选手 1 的行动收益差
- 同理，对选手 2 而言，采取甲行动收益是  $bp + f(1 - p)$ ，采取乙行动收益是  $dp + h(1 - p)$ ， $p$  为选手 1 采取丙行动的概率，求解的到：

$$p = \frac{h - f}{b + h - d - f}$$

- 要达到混合策略纳什均衡，博弈双方的行动概率主要参考对方的行动收益差；同时因为概率已知，我们也可以给出纳什均衡时双方的期望收益 - 应用：守门员博弈，当攻守双方达到纳什均衡时，其概率分布十分接近现实的统计数据，通过策略调整，博弈双方收益会逐渐收敛到纳什均衡，然后就不再变化

## 14.7 寻找纳什均衡

- 两人博弈可以用线性互补算法 (Linear Complementarity) 求解
- 严格说因为一定存在纳什均衡，寻找它不是 NPC 问题，但是是 PPAD 问题，后来人证明纳什均衡是 PPAD 问题
- PPAD 问题包括 P 问题的同时属于 NP 问题，指存在多项式的解，但不好找
- P 问题指多项式时间可解决的问题
- NP 问题指多项式时间可验证一个解的问题
- NPC 问题指 NP 问题的归约问题，同时也是 NP 问题，复杂度不断提高，NPC 问题的存在让  $P = NP$  问题很难有答案
- NP-hard 问题指 NP 问题的归约问题，但不一定是 NP 问题

## 14.8 被支配策略 (dominated strategy)

- 指不论其他参与者采取任何策略，该策略劣于其他策略
- 被支配策略永远不会是最佳回应
- 排除法：把被支配策略删除，从剩下的行动方案中选择，交互地去除掉每个选手的被支配策略

## 14.9 最大最小策略 (Maxmin strategies)

- 指其他参与者对某参与者最小收益策略下的最大收益策略  $\operatorname{argmax}_{s_i} \min_{s_{-i}} u_i(s_1, s_2)$
- 最小最大策略指让对方收益最大而自己收益最小的策略  $\operatorname{argmin}_{s_i} \max_{s_{-i}} u_{-i}(s_1, s_2)$
- 在有限两人零和博弈的纳什均衡中，参与者的最大最小值与最小最大值一致
- 最大最小可用来线性求解纳什均衡

## 14.10 扩展形式博弈

- 正常形式博弈不涉及行动顺序与时间
- 扩展形式博弈考虑时序影响，是一个层级结构，双方根据对方已经使用的策略来使用自己的策略，包括信息对称与不对称两种
- 有限信息对称博弈用  $(N, A, H, Z, \dots, u)$  来表示
- $N$  代表  $n$  个参与者
- $A$  代表一组行动
- $H$  代表一组非终点的选择节点
- 行动函数  $\chi : H \rightarrow 2^A$  表示每个选择节点的可能行动
- 参与者函数  $\rho : H \rightarrow N$  表示在节点  $h$  上采取行动的选手  $i \in N$
- $Z$  代表终止节点
- 后继者函数  $\sigma : H \times A \rightarrow H \cup Z$  映射一个选择节点和一个行动对于所有的节点与行动，如果后继者函数相同，那么节点与行动相同
- 效用函数  $u = (u_1, \dots, u_n); u_i : Z \rightarrow R$  表示在终止节点上参与者的效用
- 信息对称扩展形式博弈里参与者的纯策略是行动函数的乘积  $\prod_{h \in H, \rho(h)=i} \chi(h)$
- 扩展形式博弈可以转为正常形式博弈，但是有大量冗余，正常形式博弈不一定可转化为扩展形式博弈
- 信息对称扩展形式博弈都有纯策略纳什均衡
- 求解扩展形式博弈要从完美子博弈开始，从最小的分支倒推，记录策略，实际上也是最小最大值的求解

## 14.11 完美子博弈

- 在节点  $h$  的子博弈  $G$  是节点集合  $H$  对博弈  $G$  的限制
- 完美子博弈均衡也是纳什均衡，但不考虑无信用恐吓
- 倒推法：从最低层寻找纳什均衡，逐层反推排除掉其他选择得到策略
- 对于零和博弈，倒推法实际就是最小最大算法

## 14.12 信息不对称扩展形式博弈

- 参与者选择节点被分配到不同信息集合里，个体无法区分选择节点
- 信息不对称博弈用  $(N, A, H, Z, \dots, u, I)$  来表示
- 对于  $I = (I_1, \dots, I_n)$ ,  $I_i = (I_{i,1}, \dots, I_{i,k_i})$  是依赖于  $h \in H : \phi(h) = i$  的平衡，具有当存在  $j$  在节点  $h \in I_{i,j}$  与  $h' \in I_{i,j}$  时，有  $\chi(h) = \chi(h')$  与  $\rho(h) = \rho(h')$  的属性

### 14.13 混合与行为策略

- 混合策略随机化纯策略
- 行为策略是遇到每个信息集后的抛硬币

### 14.14 重复博弈

- 参与者  $i$  给定一个无限序列  $r_1, r_2, \dots$ , 其平均回报是  $\lim_{k \rightarrow \infty} \sum_{j=1}^k \frac{r_j}{k}$
- 考虑折扣因子  $\beta$ , 未来折扣回报是  $\sum_{j=1}^{\infty} \beta^j r_j$ , 一般人会更关注当下, 对未来关注不会超过当下, 但以  $1-\beta$  的概率终止博弈
- 重复博弈中, 当前收益跟未来收益权重不一致, 未来收益一般小于当前收益权重:

$$U = U_1 + \sigma U_2 + \dots$$

- $\sigma$  介于 0, 1 之间
- 有限重复博弈可以用倒推法得到解, 基本收敛于子博弈均衡
- 无限重复博弈要分别计算不同策略下收益, 当无限重复博弈概率不断增加, 有可能打破子博弈均衡, 此时会发生偏移

### 14.15 随机博弈 (stochastic game)

- 随机博弈是重复博弈的泛化, 每一次都取决于上一次博弈结果
- 用  $(Q, N, A, P, R)$  来表示
- $Q$  代表有限状态集
- $N$  代表有限参与者集合
- $A = A_1 \times \dots \times A_n$  其中  $A_i$  是参与者  $i$  的有限行动集
- $P : Q \times A \times Q \rightarrow [0, 1]$  表示转移概率函数, 从状态  $Q$  采取行动  $A$  变化另一个状态  $Q'$
- $R = r_1, \dots, r_n$  中  $r_i : Q \times A \rightarrow R$  表示参与者  $i$  的效用函数

### 14.16 虚拟行动 (fictitious play)

- 对于行动  $a \in A$ , 用  $w(a)$  表示对手行动次数, 可以非零初始化
- 用这个数字评价对手策略  $\sigma(a) = \frac{w(a)}{\sum_{a' \in A} w(a')}$
- 在虚拟行动中每一个参与者的策略经验分布收敛, 那么一定收敛到纳什均衡

### 14.17 无悔学习 (No-regret learning)

- 后悔表示参与者在时间  $t$  上没有采用策略  $sR^t(s) = \max(\alpha^t(s) - \alpha^t, 0)$
- 无悔学习表现出对任何纯策略有  $\Pr([\liminf R^t(s)] \leq 0)$
- 在每一步每个行为正比于其后悔  $\sigma_i^{t+1}(s) = \frac{R^t(s)}{\sum_{s' \in S_i} R^t(s')}$
- 对有限博弈收敛到均衡

### 14.18 无限重复博弈的平衡

- 著名的策略包括以牙还牙 (tit-for-tat) 跟扳机 (trigger)
- 纳什均衡只适用于有限博弈
- 无限策略里有无限个纯策略均衡

- 对于 n 个参与者的博弈  $G = (N, A, u)$  其收益向量  $r = (r_1, r_2, \dots, r_n)$ , 让  $v_i = \min_{s_{-i} \in S_{-i}} \max_{s_i \in S_i} u_i(s_{-i}, s_i)$  i 的最小最大值是对方使用最小最大策略时其收益
- 一个收益向量  $r$  是增强的如果  $r_i \geq v_i$
- 一个收益向量是可行的当存在非负值  $\alpha_a$  对于所有 i, 有  $\sum_{a \in A} \alpha_a u_i(a) = 1$  且  $\sum_{a \in A} \alpha_a = 1$
- 无名氏定理 (folk theorem): 如果收益向量对无限博弈的纳什均衡是平均回报, 那么对每个参与者都是增强的, 如果可行且增强, 收益向量就是有平均回报的无限纳什均衡

## 14.19 贝叶斯博弈

- 贝叶斯博弈  $(N, G, P, I)$  里, N 代表参与者集合, G 代表博弈集合, P 代表对某个博弈集合的先验概率, I 代表 G 里面对每个参与者的组成部分
- 也可以用认知类型来定义  $N, A, \Theta, p, u$ , A 表示行为集合,  $\Theta$  表示对参与者 i 的类型空间, p 表示没中类型的先验概率, u 表示某类型某行动对参与者 i 的收益
- 贝叶斯纳什均衡: 最大化每个参与者每种行动类型收益的策略, 可以是纯策略, 也可以是混合策略
- 三种类型: 对自己对方都不知道 (现存), 知道自己不知道对方 (过渡), 都知道 (过后)
- 过渡态期望收益:  $EU_i(s|\theta_i) = \sum_{\theta_{-i} \in \Theta_{-i}} p(\theta_{-i}|\theta_i) \sum_{a \in A} (\prod_{j \in N} s_j(a_j|\theta_j)) u_i(a, \theta_i, \theta_{-i})$
- 现存期望收益:  $EU_i(s) = \sum_{\theta_i \in \Theta_i} p(\theta_i) EU_i(s|\theta_i)$
- 贝叶斯均衡混合策略  $s_i \in \arg\max_{s'_i} EU_i(s'_i, s_{-i}|\theta_i)$
- 给定一方行为, 考虑先验概率另一方收益最大时的博弈平衡
- 双方信息不对称, 一方知道结果, 另一方只能通过概率猜测是否对方是某种类型
- 计算不同行动的收益期望, 求解概率, 如果认为概率高于某个值, 则选择对应行动, 此时达到收益与概率的均衡

## 14.20 联盟博弈

- 收益可转移的联盟博弈  $(N, v)$  N 代表有限的参与者,  $v : 2^N \rightarrow R$  里每个联盟  $S \subseteq N$  里的能够分配的收益, 假定  $v(\emptyset) = 0$
- 联盟博弈解决的问题是哪些联盟会生成及收益如何分配
- 超加性博弈  $G = (N, v)$  表示对于所有的  $S, T \subseteq N$ , 如果  $S \cap T = \emptyset$ , 那么有  $v(S \cup T) = v(S) + v(T)$ , 这种情况下整体绑定为一个联盟收益最高

## 14.21 夏普利值 (Shapley Value)

- 如何公平分割收益, Lloyd Shapley 认为参与者要按照边际贡献的比例获得收益
- 公理 - 对于每一种收益方法如果两个人可相互交换  $v(S \cup \{i\}) = v(S \cup \{j\})$ , 那么其收益相等  $\psi_i(N, v) = \psi_j(N, v)$  - 如果某个人不产生收益  $v(S \cup \{i\}) = v(S)$ , 那么不分成  $\psi_i(N, v) = 0$  - 对于  $v_1$  和  $v_2$ , 博弈  $(N, v_1 + v_2)$  用  $(v_1 + v_2)(S) = v_1(S) + v_2(S)$  定义, 有  $\psi_i(N, v_1 + v_2) = \psi_i(N, v_1) + \psi_i(N, v_2)$
- 夏普利值按照  $\psi_i(N, v) = \frac{1}{N!} \sum_{S \subseteq N_i} |S|!(|N| - |S| - 1)! [v(S \cup \{i\}) - v(S)]$  分配收益, 也就是满足上面三个公理的分配方式

## 14.22 核心

- 夏普利值分配比较公平, 但不一定稳定, 不容易形成大联盟
- 核心指对于  $S \subseteq N$ , 有  $\sum_{i \in S} x_i \geq v(S)$ , 总和收益至少不低于内部小联盟收益
- 核心可能是空的且不唯一
- 对于简单博弈核心是空的当且仅当没有否决参与者, 如果有否决参与者, 核心包括所有非否决参与者收益 0 的收益向量
- 凸博弈:  $v(S \cup T) \geq v(S) + v(T) - v(S \cap T)$ , 每个凸博弈有一个非空核心且是夏普利值

### 14.23 选举

- 选举结果  $O$
- 参与者的选择  $>$
- 线性排序  $L$  选择顺序, 可传递; 非强制选择  $L_{NS} \geq$  可传递
- 社会选择函数  $C: L_{NS}^n \rightarrow O$ , 其中  $L$  是非强制选择偏好
- 社会福利函数  $W: L_{NS}^n \rightarrow L_{NS}$
- 多数票制: 选择大多数人选最多的
- 累计投票: 每个人多张票, 可以重复投给一个人
- 同意投票: 每个人可随意投, 投多个人或不投都可以
- 淘汰制多数票: 得票最多获胜, 否则淘汰得票少的重新投票直到有结果
- 波达投票: 每个结果分配一个有排序的数字, 最后选择总和最高的那个
- 连续消除: 设定顺序, 然后前两个投票, 胜的淘汰, 之后赢的跟第三个再投票直到出结果
- 孔多塞连续性: 如果有个结果在所有对比中都胜出, 那么这个候选人可以当选 (不一定总是有这样的人, 有时会出现循环)
- 不同投票方法结果可能不一样
- 帕累托有效 (PE) 如果所有参与者都同意某两个结果的顺序, 那么社会福利函数也会使用那个顺序
- 无关选择独立性 (IIA) 两个结果间顺序指依赖于参与者给出的相对顺序
- 独裁者 (dictator) 某可决定社会排序的参与者
- Arrow 理论任何超过三人的具有 PE 与 IIA 的社会福利函数都是独裁的 证明
- 弱帕累托有效: 没有出现过被支配的结果
- 单调: 一个获胜的结果再拿到更高的排序也是获胜者
- 独裁: 存在某参与者的投票与总体选择顺序一致
- 弱帕累托有效且单调的社会选择函数是独裁的
- 单峰选择: 每个投票的人都有最想选跟最不想选的候选人, 选择上会避免极端

### 14.24 机制设计

- 直接形式: 参与者同时传递信息到中心
- 间接形式: 参与者传递一系列信息, 这些信息前后相关
- 启示原则 (Revelation Principle): 任何社会选择函数都可以用真实直接的机制来实施
- Gibbard-Satterthwaite 理论: 有限选择空间里三个元素以上的支配策略都是独裁的, 非独裁策略需要转移函数
- 转移函数: 集体对个体的税收或补贴, 个人私人收益与转移函数之和是个体的整体结果收益, 个人私人收益不考虑其他人的选择
- 真实性: 对于每个参与者平衡策略里转移收益机制是真实的且个体接受个人收益函数
- 有效性: 有效的机制会选择最大化个体的收益, 转移函数的和小于等于 0
- 预算平衡: 所有人转移函数的和为 0, 大于等于 0 是弱预算平衡
- 个体理性: 没有人会因为参与某个机制得到负收益
- 可处理: 收益与转移函数的计算可在多项式时间内完成
- 税收最大化: 满足所有限制条件下最大化转移函数总和期望的机制
- 税收最小化: 满足所有限制条件下最小化转移函数总和期望的机制
- 公平性: 让最不开心的参与者也开心最大化
- 设计政策时要考虑支出补贴税收对所有人都无害或达到均衡

### 14.25 VCG 机制

- 存在货币补偿时自私参与者选择社会福利最大化的通用方法
- 一个直接机制包括选择规则跟补偿规则, VCG 作为支配策略是真实有效的, 在额外假设下可以满足弱预算平衡跟个体理性

- Groves 机制  $(\chi, p)$

$$\chi(\hat{v}) \in \operatorname{argmax}_x \sum_i \hat{v}_i(x)$$

$$p_i(\hat{v}) = h_i(\hat{v}_{-i}) - \sum_{j \neq i} \hat{v}_j(\chi(\hat{v}))$$

- VCG 机制  $(\chi, p)$

$$\chi(\hat{v}) \in \operatorname{argmax}_x \sum_i \hat{v}_i(x)$$

$$p_i(\hat{v}) = \max_x \sum_{j \neq i} \hat{v}_j(x) - \sum_{j \neq i} \hat{v}_j(\chi(\hat{v}))$$

- 在 VCG 机制下，计算你自己存在时社会收益最大时其他人收益总和，然后计算没有你且社会收益最大时其他人收益总和，你的补偿是它们的差值
- 不影响最后结果的人不需要补偿，其存在导致他人收益减少的要付出，其存在导致他人收益增加的要补偿
- 其真实性有效性可以数学证明
- VCG 需要个体真实汇报个人信息，但个体间的勾结可以消除补偿，同时 VCG 会导致开支暴涨，增加参与人也会增加财政支出，个人可能伪装多个人，同时返还机制也不允许所有人返还所有收益
- 价值隐私信息需要损失效率，是动机效率的平衡

## 14.26 拍卖

- 自私参与者分配资源的机制
- 英国人拍卖：从保留值开始拍卖，参与人喊价，价高的得到物品并付出对应价格
- 日本人拍卖：所有参与者先站着，价格提升，当价格不合适就坐下，最后一个站着的人得到物品
- 荷兰人拍卖：设定一个高价然后开始逐渐降低，当有人说我接受时拍卖结束，价格就是当时价格
- 第一价格拍卖：参与者将价格写好封装，然后拍卖人开封，价高的人得到物品，付对应价格
- 第二价格拍卖：同上，但付第二高的价格
- 付费拍卖：同上，但所有人都要支出自已写的价格，彩票？
- 拍卖三种规则：竞拍规则，信息释放规则，清场规则
- 在第二价格拍卖里，讲真话是支配策略，符合 VCG 机制，也可以直观证明
- 在英国人日本人拍卖里，独立私有价值模型下的支配策略是用真实值
- 在第一价格拍卖与荷兰人拍卖实质等同，前者可以异步进行，后者交流快，参与者竞拍价要低于价值，无支配策略
- 两个中等风险的参与者参与第一价格竞拍，分布是均匀分布，贝叶斯纳什均衡是各自价格的一半
- 更多的参与人参加后，价格会不断提升接近真实价格，如果是均匀分布，系数是  $\frac{n-1}{n}$
- 收益均等理论：n 个风险中等的参与者对于单一物品有独立私有价值，参与竞拍时每个人都从风险分布 F 里报价，当均衡时分配总是一样的，价值为 0 其期望也是 0，所有有效拍卖产生的收益是一样的
- 最优化拍卖，参与人的虚拟价值  $\psi_i(v_i) = v_i - \frac{1-F_i(v_i)}{f_i'(v_i)}$  在保留价格处为 0，并非 VCG



# 章 15

## 量化投资

### 15.1 股票收益模型

- 股票指数可代表市场，收益率  $r_M$
- 某股票跟市场的相关性是市场贝塔（market beta） $\beta_M$
- 市场之外的贡献阿尔法  $\alpha$
- 特定股票收益率  $r = \alpha + \beta_M r_M$
- 资本资产定价模型 (capital asset pricing model, CAPM)：所有股票阿尔法部分期望是零
- 套利定价理论 (arbitrage pricing theory, APT)：除了市场因素，还有其他公用因素来解释股票收益
- 宏观经济因子模型 (macroeconomic factor model, MFM)： $r_i = u_i + \sum_{k=1}^K \beta_{ik} f_k$
- 成熟的商业模型可以确定一些共同因素，但也面临不确定性

### 15.2 风险

- 单只股票波动很大，一般会组合投资降低风险
- 单只股票波动率  $\sigma_i$ ，两只股票之间的相关系数  $\rho_{ij}$ ，股票组合波动率  $\sigma_p^2 = \sum_{i=1}^n w_i^2 \sigma_i^2 + \sum_{i \neq j} w_i \sigma_i \rho_{ij} \sigma_j w_j$ ，其中  $w$  是股票权重
- 两两间需要估计的参数多
- 因为组合的股票多而共同因素少，可以直接估计与共同因素的相关性来替代总体估计
- 但是除了共同部分还有股票特有特征，所以个股的波动是个体波动与整体波动的综合  $\sigma_i^2 = \sigma_{S,i}^2 + \sum_{l,m=1,\dots,K} \beta_{il} \sigma_i \rho_{lm} \sigma_m \beta_{im}$ ，这里我们认为股票间波动可以用共同波动来解释  $\sigma_i \rho_{ij} \sigma_j = \sum_{l,m=1,\dots,K} \beta_{il} \sigma_i \rho_{lm} \sigma_m \beta_{im}$
- 所谓风险控制就是通过对每只股票权重的调节优化来降低波动率，但结果不稳定
- 通过风险控制就可以计算收益率，但毕竟有些因素是不可控的，解决的方法就是通过权重调整让其对总收益贡献为零或做空该因素对冲风险

### 15.3 收益

- 共同因素的收益可通过投资市场指数来获得
- 某特定共同因素的收益需要自己构建，smart beta
- 这属于被动投资，因为股票市场长期看总是上涨，所以收益不差
- 也存在主动投资，根据自己预测投资并不断调整，smart beta 属于半被动投资
- 特异收益方面可以进行择时收益 (timing skill)，但一定注意风控

## 15.4 一般性投资

- 三要素：成本、收益、风险
- 时间序列分析：统计学
- 技术分析：根据价格曲线图形进行分析的一种方法，技术分析本质是经验分析，可以量化为模式进行识别，有些指标可以进行机理解释
- 收益的统计分析 - 环比：以月为周期的相对收益率 - 同比：以年为周期的相对收益率 - 收益率一般是正态分布，但也可能是肥尾的，极端事件发生概率高；甚至是长尾，比较极端事件也有可能发生；还有可能出现黑天鹅，非常极端事件 - 可预见事件会有过度反应然后回调，可进行事件驱动投资
- 影响价格因素的分析，要考虑专业知识、随机性、噪声、数据来源、混杂因素

## 15.5 风险管理原理

- 概率论是核心，17世纪出现，保险基础
- 概率乘法规则，每件事都是独立的，联合发生概率是乘积，保险公司认为被保的事相对独立不会同时发生，因为概率特别低
- 二项分布可用来估计准备金比率
- 均值表示集中，几何均值用来估计收益表现，全是正数，比算术均值小
- 方差表示离散
- 风险需要高均值低方差的收益率
- 协方差表示两个变量共同变动的能力
- 相关性是协方差除以各自的标准差
- 回归线截距就是阿尔法，斜率就是贝塔
- 现值就是某物当前的价格， $n$ 年后  $C$  元的现值是  $C/(1+r)^n$ ， $r$  是利率
- 永续国债每年承诺固定年金  $C$ ，其现值计算就是个等比数列，等于  $C/r$ ，利率越高国债不值钱，利率低时现值高
- 永续国债也可以承诺每年年金增长，年金现值等于  $C/(r-g)$ ， $g$  是增长利率，一定小于市场利率
- 固定期限年金  $C$  付款  $n$  年，现值是  $C \frac{1-(1/(1+r)^n)}{r}$ ，利率越高，年限越小，现值越高，贷款买房就是如此，按揭与利率年限有关，年限越长实际现在花的越少，流动性好
- 效用函数  $U$  表示钱的边际效果，钱越多，边际效果越低

## 15.6 金融技术与发明

- 长期风险，最大的是道德风险（共同利益与个人利益的冲突）
- 经济风险都是可以分担的，消费行为之间是有关联的，社会主义思想下消费行为趋同，通过计划分配来控制分摊风险，但道德风险无法规避
- 框架效应，含义相同但说法不同会导致不同的决策判断
- 公共财政设计税收与福利来控制社会风险，避免道德问题
- 心理账户多是货币框架，但实物框架要考虑消费指数与通胀
- 发明，技术推动金融发展，解决风险问题与框架效应
- 标准化、行政机构、社会保障、邮政服务促进了金融业的发展

## 15.7 投资组合

- 高收益低方差组合
- 资产通常不独立
- 寻找预期、标准差、协方差的投资组合，找出有效边界，构成共同基金
- 所有共同基金都试图通过多样化来进行高收益低方差投资并试图打败市场
- 美国股市受益长期高于债券，约 4%，发达国家类似，这个现象与政治有关系
- 税收，特别是受益税（企业利润税与个人所得税）对经济有重要调节作用
- 全球范围，企业利润约 1/3 会被课税（实际税率）

- 投资管理包括资产配置（多样化，占 90%）、入场时机与证券选择
- 交易行为本身会对市场产生影响出现系统损耗与负和博弈
- 对冲基金存在生存偏差与回填偏差，多元化可以让投资平稳
- 如果要打败市场，就要主动出击，在多元化保护下平衡风险，吸纳各种来源的非传统投资收益，获取超额收益

## 15.8 保险

- 保险与共同基金一样都是通过汇集分摊个体风险管理受益的金融工具，共同基金的受益目标是增值，保险的受益目标是保值与未来损失，个体好恶会导致保险与基金成本不同
- 面对道德风险与选择歧视（只有有风险的买保险）
- 当独立个体风险汇聚一体（二项分布接近正态分布）时，整体风险的期望不变但方差可测算控制
- 合同设计用来标注风险、例外（规避道德风险与选择歧视）、数学模型、公司类型、政府监管（准备金）
- 两种公司，综合险种保险公司与单一险种保险公司，后者风险大、监管严（次贷危机影响）
- 主要类型：财产保险/健康保险/人寿保险
- 寿险保护孩子（遗产），比财险规模大
- 汽车保险收入最高，然后是住房保险
- 始于 1600s，金融创新推动
  - 1840s，高薪雇佣保险推销员来推动人们天生的抗拒（农业保险）
  - 1900s，具有现金属性的保险，可升值，但取消后也损失，防止取消保险
  - 19 世纪卖给女性丈夫的保险，用宗教转述消除掉道德抵触
- 次贷危机中为债券提供保险的公司购买了大量次贷产品，当评级降低时，出现系统恐慌与抛售，政府举债募资困难，社会衰退
- 气候变化导致的风险是全球性的，也会导致社会衰退，这个风险在提高
- 巨灾债券：没有灾难时发行，有灾难时不偿付或少量偿付，一般是城市发行，用来均摊风险，收益率高与市场相关性低

## 15.9 有效市场假说

- 市场价格反映所有有效信息
- 股价反映未来股息价格的现值，否则就是击鼓传花，股息代表盈利能力
- 随机行走理论  $x_t = x_{t-1} + \text{error}$
- 一阶自相关回归因为自相关数据可能回归原点

## 15.10 行为经济学

- 股市是随机运行，现值一直平稳增长
- 期望效用曲线，在预期值附近厌恶损失喜欢收益
- 权重理论，对于确定性的事给更高的权重
- 祝愿性思考，总认为自己心仪的一方会表现好
- 注意力无法长期集中
- 锚定效应，指标锚定
- 赌博行为
- 迷信行为
- 营销手段：吹嘘、隐藏信息、病毒营销与过度交易

## 15.11 金融监管

- 信息纰漏
- 电话销售促进“锅炉房”交易所的出现

- SEC 成立
- 不上市公司无法向公众登广告例如私募对冲基金，但也要接受监管，例如资产达标等
- 局内人与局外人监管，防止内幕信息出现公开时间差
- 会计监管
- SIPC 证券投资保护公司

## 15.12 利率

- 短期折扣债券 (bills, 低于一年)，低于面值发售
- 中期债券 (notes, 一年到十年)，付息
- 长期债券 (bonds, 大于十年)，6 个月付息一次，老式债券附带利息券，付息时剪下来去银行兑换钱
- 物价指数债券，调整过通胀或 CPI 的债券，早期指定物品来实现
- 中间商赚取买入卖出价的价差，流动性好的资产价差小，差的或市场小的价差大，卖方主导
- 利率期限结构，不同期限收益率与到期期限的关系

## 15.13 银行

- 西方起源于金匠对黄金的保存，采用部分准备金制度进行借贷
- 银行分为商业银行、储蓄银行、存储协会与信用合作社
- 商业银行规模比较大，美国 20% 的银行资产是外资
- 信用合作社数量比较多，但资产规模小，多为区域内或行业内的
- 存储协会规模中等，数量不多
- 银行解决逆向选择问题，经营核心是关系，对借贷进行筛选
- 银行可以避免企业对债主不道德的资金使用
- 银行提供资金流动性，长贷短存，需要准备金应对挤兑但要保证企业有长期资金
- 发展中国家因为金融体系不健全，银行体系往往规模很大，关系国计民生
- 1988，巴塞尔条约规定银行风险标准
- 里根政府不限制存储协会存款利率，结果风险积聚，最后政府埋单
- 墨西哥政府银行私有化，结果银行过量发放贷款最后全部倒闭，墨西哥银行被海外资本控制
- 亚洲金融危机，外资集中撤资，大量银行倒闭
- 阿根廷银行挤兑风潮
- 美国影子银行发行商业短期借贷票据，银行通过内部组织 (SIV) 变相担保，但却不接受银行体系监管，出现风险积聚

## 15.14 信用评级

- 穆迪评级 1909
- S&P 1916 统计方法评级
- 经营关系
- 接收被评级费用，用制度避免道德风险

## 15.15 股票

- 公司是法人，要有董事会，公司就是为了盈利存在，向股东负责
- 非盈利公司不发行股票，但也有董事会
- 公司通常会进行股票分割，让股价维持在可买范围 (20-40 美元)，一般一手 100 股
- 买股票是为了股息，股息董事会决定
- 净值与股票收益比例为 PE，一般为 15，也就是投资者 15 年收益回本，收益会计算出来的
- 股息除收益现在都是小于 1，说明公司在留钱投资，大于 1 则出现在大萧条
- 公司可以进行股票回购，与分红本质一样，股利政策不影响价值

- 公司可以举债，提高杠杆率，举债不收税，但债务过高会导致破产成本高
- 林特纳模型认为股息的发放会影响价值与市场反应，股息与盈利间要进行平滑，稳定股息的发放

## 15.16 房产

- 美国房产规模与股票接近
- 房地产存在 DPP，有限合伙制，可避税
- 商业房地产公司 REITS 限制很多
- 商公用房也可算资产，但不算地产公司
- 贷款价值比不能超过 1
- 最早房产贷款期限很短，只有 5 年，贷款价值比不超过 60%，但只还利息，最后还本金，续约时如果房产贬值就会收回房子
- 美国出台房屋贷款公司，提供 15 年贷款，但本息一起还
- 后来贷款进一步延长到 30 年，但因为考虑退休后收入下降，一般不超过 30 年
- 后来出现面向低收入者的 2+28 浮动利率贷款，但后面利率很高
- 房贷按揭固定还款最开始还的是利息，越往后本金比例越高，公式等同年金计算公式
- 房地产业存在全球尺度的非理性繁荣



# 章 16

## 自然语言处理

早期依赖语言学家的文法，后来依赖统计模型，基础是语料库语言模型。 $n$  元语料库，一般是三元模型，计算当前词与前面两个词同时出现的及自己单独出现的概率。这个概率可以统计很多文章，通过 tf-idf 加权矩阵来计算每个词的加权概率。

为了防止出现零概率，可以用古德-图灵估计里为零概率的词赋很小的概率，这个概率来自低词频贡献，词频越低，对其他零词频词的概率贡献越多。也就是设定一个词频阈值，阈值之上不对概率扣减，阈值之下概率进行古德-图灵估计，零概率的均分前面打折省下来的概率。另一个思路是低元模型对高元模型的线性插值。

分词问题，可以用语料库计算出现某个序列的概率，概率最大的分词方法被选用。分词可以采用层级模型，颗粒度不同的分词存在各自最适用的应用场景。

隐马尔可夫模型（HMM），这是一个信号模型，给定一组状态序列，寻找对应的信号序列，例如翻译与语音识别。本质上是寻找产生最大概率的那一组信号，通过贝叶斯定理可以转换为寻找并最大化某信号下状态的条件概率与该信号合理出现概率的乘积。在 HMM 下，通过马尔可夫假设，我们只考虑每个信号前面的一个信号之间条件概率（也就是前面的语言模型）与当前信号状态条件概率乘积的连乘，然后用维特比算法（线性规划）找出最大值，也就是给定模型与输出信号，计算最可能出现的状态（语音识别问题）。同时，HMM 可以解决给定参数后，计算某个输出序列的概率（输入法）。但 HMM 本身也需要进行参数估计，可以采用标注数据但成本高，也可以用 EM 的方法，先假设一个随机模型可以产生输出序列，前后转移概率随机，两两含义对应概率也随机，然后用数据输入模型计算所有路径概率，此时相当于标注了数据，用其计算最大似然度下的最优参数，然后再输入新参数，不断迭代直到新参数比旧参数对数据没有更高的最大似然输出概率。

信息量比特数跟发生可能性的对数有关，信息量与发生概率的乘积的累积总和就是信息熵，信息熵越大，均质性越强，不确定性越高；反之，高概率信息的引入会降低不确定性，系统内额外相关的信息总会不增加不确定度（条件信息熵），由额外信息减少的信息熵是互信息，取值 0-1 之间，越大额外信息与原始信息越相关取值越高，可用来消除语言二意性。相对熵用来衡量两个文本信息熵的大小。自然语言处理考虑上下文是考虑了条件熵，考虑语境就是考虑了相对熵，语言复杂度就是给定上下文每个位置可选择的单词数量，语料库引入可提高模型准确度。

布尔逻辑可用来进行索引表的检索，图论与深度／广度优先算法，爬虫找到页面，提取链接，构建哈希表存储，可采取异步分批存储与定向发送（网页聚类），pagerank 通过计算网页链入链出的数量与质量迭代来确定网页的重要程度，在进行关键词检索时，首先依赖词频，但要去掉高频词，同时如果某个词词频高但出现在独立网页中比较多时（类高频词）可以用逆词频，也就是网页总数除以出现该词的页面数的对数来对词频加权，如果在独立网页中出现概率高，信息量大，其总排序要靠前，这样可以筛选出跟查询词比较接近的语境（TF-IDF）

语音识别需要用到动态规划与有限状态机来降低计算／搜索成本

高维相似度可以用余弦定理来计算，首先得到每个样本的特征向量，然后计算相似度，高的归为一类，然后同一类中计算余弦相似度不断归类。不同的相似度可以用不同的权重来表示

同时，也可以用矩阵计算来进行分类，对矩阵进行奇异值分解，左边酉矩阵可以提取特征跟主题相关性，右边则可提取主题与样本相关性，可快速进行较粗的分类，而余弦定理就需要不断迭代。

相似哈希表可用来快速比较相似度

费马小定理可用来进行基于质数的加密

最大熵模型可用来计算存在影响因素下（或者说特定语境下）的出现概率问题，这是一个指数模型，需要训练的参数非常多，但也可以用 em 方法求解训练。

输入法可通过动态规划方法求解备选词，同时可以用余弦定理来构建个人语言模型与整体语言模型的插值模型，这是一个最大熵与通用的折衷方案，其实也是最优的

布隆过滤器，对字符随机转换为 8 个数字，然后构建一个二进制长向量，将 8 个数字对应的位置设为 1，如果某个地址出现同样位置为 1，那么过滤掉，用较小的空间对比较大的数目，比计算哈希值省空间，可用来快速过滤垃圾邮件

贝叶斯网络，条件概率的网络，训练出的模型可以预测其中变量间关系，然后可以反过来调整模型结构，例如我们可以知道文本对应的主题，也可以知道主题对应的文本，这样在主题跟文本还有关键词间可以构建一个有向网络来探索其之间的相互关系

如果关系是无向的贝叶斯网络那就是条件随机场，可用来解决句法分析问题

- NLP 词向量法来快速线性搜索相关词
- tidytext 包
- quanteda 包
- jieba 中文分词

# 章 17

## 因果分析

### 17.1 Introduction

“Impact assessment, simply defined, is the process of identifying the future consequences of a current or proposed action.” (IAIA, 2009)

“Policy assessment seeks to inform decision-makers by predicting and evaluating the potential impacts of policy options.” (Adelle and Weiland, 2012)

“... I see no greater impediment to scientific progress than the prevailing practice of focusing all our mathematical resources on probabilistic and statistical inferences while leaving causal considerations to the mercy of intuition and good judgment.” (Pearl, 1999)

### 17.2 Causal Information

#### 17.2.1 Target

practical framework for causal effect estimation in the context of policy assessment and impact analysis, and in the absence of experimental data

#### 17.2.2 Causal sources

- Causal Inference by Experiment: Randomized experiments
- Causal Inference from Observational Data and Theory: Existing data or Big data

#### 17.2.3 Identification and Estimation Process

- Causal Identification: domain knowledge based
- Computing the Effect Size: Bayesian networks

## 17.3 Theoretical Background

### 17.3.1 Potential Outcomes Framework

- $Y_{i,1}$  Potential outcome of individual i given treatment  $T=1$  (e.g. taking two Aspirins)
- $Y_{i,0}$  Potential outcome of individual i given treatment  $T=0$  (e.g. drinking a glass of water)
- individual-level causal effect (ICE)
  - $ICE = Y_{i,1} - Y_{i,0}$
- average causal effect (ACE)
  - $ACE = E[Y_{i,1}] - E[Y_{i,0}]$

### 17.3.2 Causal Identification

- $Y_{i,1}$  (treatment) and  $Y_{i,0}$  (non-treatment) can never be both observed for the same individual at the same time
- Association S
  - $S = E[Y_1|T=1] - E[Y_0|T=0]$
- S is not the **same** with ACE
  - association does not imply causation
  - randomized experiment

#### 17.3.2.1 Ignorability

- $(Y_1, Y_0) \perp\!\!\!\perp T$ ,  $Y_1$  and  $Y_0$  must be jointly independent of the treatment assignment
- $(Y_1, Y_0) \perp\!\!\!\perp T|X$  for real observational studies, conditional on variables X,  $Y_1$ , and  $Y_0$  are jointly independent of T, the assignment mechanism
- $ACE|X = E[Y_1|X] - E[Y_0|X] = E[Y_1|T=1, X] - E[Y_0|T=0, X] = E[Y|T=1, X] - E[Y|T=0, X] = S|X$

#### 17.3.2.2 Assumptions

- Causal inference requires causal assumptions

## 17.4 Methods for Identification and Estimation

### 17.4.1 Directed Acyclic Graphs(DAG) for Identification

- DAGs Are Nonparametric
- A Node represents a variable in a domain, regardless of whether it is observable or unobservable
- A Directed Arc has the appearance of an arrow and represents a potential causal effect. The arc direction indicates the assumed causal direction, i.e. “A → B” means “A causes B.”
- A Missing Arc encodes the definitive absence of a direct causal effect, i.e. no arc between A and B means that there exists no direct causal relationship between A and B and vice versa. As such, a missing arc represents an assumption

### 17.4.2 Indirect Connection

- A causes B via node C
- $A \not\rightarrow B$  and  $A \perp\!\!\!\perp B|C$

### 17.4.3 Common Cause

- C causes both A and B
- $A \perp\!\!\!\perp B$  and  $A \perp\!\!\!\perp B|C$

### 17.4.4 Common Effect

- C is the common effect of A and B
- $A \perp\!\!\!\perp B$  and  $A \not\perp\!\!\!\perp B|C$

### 17.4.5 Example: Simpson's Paradox

## 17.5 链接

- 因果推断
- Google 的因果分析包