

面向数据科学家的实用统计学



[面向数据科学家的实用统计学_下载链接1](#)

著者:[美] 彼得·布鲁斯

出版者:人民邮电出版社

出版时间:2018-10-1

装帧:平装

isbn:9787115493668

本书解释了数据科学中至关重要的统计学概念，介绍如何将各种统计方法应用于数据科学。作者以易于理解、浏览和参考的方式，引出统计学中与数据科学相关的关键概念；解释各统计学概念在数据科学中的重要性及有用程度，并给出原因。

作者介绍:

彼得·布鲁斯 (Peter Bruce)，知名统计学家，Statistics.com统计学教育学院的创立者兼院长，重采样统计软件的开发者。曾在美国马里兰大学和各种短训班教授重采样统计课程。

安德鲁·布鲁斯 (Andrew Bruce)，华盛顿大学统计学博士，拥有30多年的统计学和数据科学经验，在多家知名学术期刊上发表过多篇论文。

目录: 前言 xiii

第 1 章 探索性数据分析 1

1.1 结构化数据的组成 2

1.2 矩形数据 4

1.2.1 数据框和索引 5

1.2.2 非矩形数据结构 5

1.2.3 拓展阅读 6

1.3 位置估计 6

1.3.1 均值 7

1.3.2 中位数和稳健估计量 8

1.3.3 位置估计的例子：人口和谋杀率 9

1.3.4 拓展阅读 10

1.4 变异性估计 10

1.4.1 标准偏差及相关估计值 11

1.4.2 基于百分位数的估计量 13

1.4.3 例子：美国各州人口的变异性估计量 14

1.4.4 拓展阅读 14

1.5 探索数据分布 14

1.5.1 百分位数和箱线图 15

1.5.2 频数表和直方图 16

1.5.3 密度估计 18

1.5.4 拓展阅读 20

1.6 探索二元数据和分类数据 20

1.6.1 众数 21

1.6.2 期望值 22

1.6.3 拓展阅读 22

1.7 相关性 22

1.7.1 散点图 25

1.7.2 拓展阅读 26

1.8 探索两个及以上变量 26

1.8.1 六边形图和等势线（适用于两个数值型变量） 26

1.8.2 两个分类变量 28

1.8.3 分类数据和数值型数据 29

1.8.4 多个变量的可视化 31

1.8.5 拓展阅读 33

1.9 小结	33
第2章 数据和抽样分布	34
2.1 随机抽样和样本偏差	35
2.1.1 偏差	36
2.1.2 随机选择	37
2.1.3 数据规模与数据质量：何时规模更重要	38
2.1.4 样本均值与总体均值	38
2.1.5 拓展阅读	39
2.2 选择偏差	39
2.2.1 趋均值回归	40
2.2.2 拓展阅读	41
2.3 统计量的抽样分布	42
2.3.1 中心极限定理	44
2.3.2 标准误差	44
2.3.3 拓展阅读	45
2.4 自助法	45
2.4.1 重抽样与自助法	47
2.4.2 拓展阅读	48
2.5 置信区间	48
2.6 正态分布	50
2.7 长尾分布	53
2.8 学生t分布	55
2.9 二项分布	57
2.10 泊松分布及其相关分布	58
2.10.1 泊松分布	59
2.10.2 指数分布	59
2.10.3 故障率估计	60
2.10.4 韦伯分布	60
2.10.5 拓展阅读	61
2.11 小结	61
第3章 统计实验与显著性检验	62
3.1 A/B 测试	62
3.1.1 为什么要有对照组	64
3.1.2 为什么只有处理A和B，没有C、D……	65
3.1.3 拓展阅读	66
3.2 假设检验	66
3.2.1 零假设	67
3.2.2 备择假设	67
3.2.3 单向假设检验和双向假设检验	68
3.2.4 拓展阅读	68
3.3 重抽样	68
3.3.1 置换检验	69
3.3.2 例子：Web 黏性	69
3.3.3 穷尽置换检验和自助置换检验	72
3.3.4 置换检验：数据科学的底线	72
3.3.5 拓展阅读	72
3.4 统计显著性和p 值	72
3.4.1 p 值	74
3.4.2 α 值	75
3.4.3 第一类错误和第二类错误	76
3.4.4 数据科学与p 值	76
3.4.5 拓展阅读	77
3.5 t 检验	77
3.6 多重检验	78

3.7 自由度	81
3.8 方差分析	82
3.8.1 F 统计量	84
3.8.2 双向方差分析	85
3.8.3 拓展阅读	86
3.9 卡方检验	86
3.9.1 卡方检验：一种重抽样方法	86
3.9.2 卡方检验：统计理论	88
3.9.3 费舍尔精确检验	88
3.9.4 与数据科学的关联	90
3.9.5 拓展阅读	91
3.10 多臂老虎机算法	91
3.11 检验效能和样本规模	93
3.11.1 样本规模	95
3.11.2 拓展阅读	96
3.12 小结	96
第4章 回归与预测	97
4.1 简单线性回归	97
4.1.1 回归方程	98
4.1.2 拟合值与残差	100
4.1.3 最小二乘法	101
4.1.4 预测与解释（剖析）	102
4.1.5 拓展阅读	103
4.2 多元线性回归	103
4.2.1 美国金县房屋数据案例	103
4.2.2 评估模型	104
4.2.3 交叉验证	106
4.2.4 模型选择和逐步回归法	107
4.2.5 加权回归	108
4.3 使用回归做预测	109
4.3.1 外推法的风险	109
4.3.2 置信区间和预测区间	110
4.4 回归中的因子变量	111
4.4.1 虚拟变量的表示	112
4.4.2 多层因子变量	113
4.4.3 有序因子变量	114
4.5 解释回归方程	115
4.5.1 相关的预测变量	116
4.5.2 多重共线性	117
4.5.3 混淆变量	117
4.5.4 交互作用和主效应	118
4.6 检验假设：回归诊断	119
4.6.1 离群值	120
4.6.2 强影响值	121
4.6.3 异方差性、非正态分布和相关误差	123
4.6.4 偏残差图和非线性	126
4.7 多项式回归和样条回归	127
4.7.1 多项式回归	128
4.7.2 样条回归	129
4.7.3 广义加性模型	131
4.7.4 拓展阅读	132
4.8 小结	133
第5章 分类	134
5.1 朴素贝叶斯算法	135

5.1.1 准确的贝叶斯分类是不切实际的	136
5.1.2 朴素解决方案	136
5.1.3 数值型预测变量	138
5.1.4 拓展阅读	138
5.2 判别分析	138
5.2.1 协方差矩阵	139
5.2.2 费希尔线性判别分析	139
5.2.3 一个简单的例子	140
5.2.4 拓展阅读	142
5.3 逻辑回归	142
5.3.1 逻辑响应函数和Logit 函数	143
5.3.2 逻辑回归和广义线性模型	144
5.3.3 广义线性模型	145
5.3.4 逻辑回归的预测值	145
5.3.5 解释系数和优势比	146
5.3.6 线性回归与逻辑回归：相似之处和不同之处	147
5.3.7 模型评估	148
5.3.8 拓展阅读	150
5.4 评估分类模型	150
5.4.1 混淆矩阵	151
5.4.2 稀有类问题	152
5.4.3 准确率、召回率和特异性	153
5.4.4 ROC 曲线	153
5.4.5 AUC	155
5.4.6 提升	156
5.4.7 拓展阅读	157
5.5 不平衡数据的处理策略	157
5.5.1 欠采样	158
5.5.2 过采样以及上权重和下权重	158
5.5.3 数据生成	159
5.5.4 基于代价的分类	160
5.5.5 探索预测值	160
5.5.6 拓展阅读	161
5.6 小结	161
第6章 统计机器学习	162
6.1 K 最近邻算法	163
6.1.1 预测贷款拖欠的示例	164
6.1.2 距离度量	165
6.1.3 独热编码	166
6.1.4 标准化	166
6.1.5 K 值的选取	168
6.1.6 KNN 作为特征引擎	169
6.2 树模型	170
6.2.1 一个简单的例子	171
6.2.2 递归分区算法	172
6.2.3 测量同质性或不纯度	174
6.2.4 阻止树模型继续生长	175
6.2.5 预测连续值	176
6.2.6 如何使用树模型	176
6.2.7 拓展阅读	177
6.3 Bagging 和随机森林	177
6.3.1 Bagging 方法	178
6.3.2 随机森林	178
6.3.3 变量的重要性	181

6.3.4 超参数	183
6.4 Boosting	184
6.4.1 Boosting 算法	184
6.4.2 XGBoost 软件	185
6.4.3 正则化：避免过拟合	186
6.4.4 超参数和交叉验证	189
6.5 小结	191
第7章 无监督学习	192
7.1 主成分分析	193
7.1.1 一个简单的例子	194
7.1.2 计算主成分	195
7.1.3 解释主成分	196
7.1.4 拓展阅读	198
7.2 K-Means 聚类	198
7.2.1 一个简单的例子	199
7.2.2 K-Means 算法	201
7.2.3 解释类	201
7.2.4 选择类的个数	203
7.3 层次聚类	204
7.3.1 一个简单的例子	205
7.3.2 树状图	205
7.3.3 凝聚算法	206
7.3.4 测量相异性	207
7.4 基于模型的聚类	208
7.4.1 多元正态分布	209
7.4.2 混合正态分布	210
7.4.3 类数的选取	212
7.4.4 拓展阅读	213
7.5 变量的缩放和分类变量	213
7.5.1 变量的缩放	214
7.5.2 控制变量	215
7.5.3 分类数据和高氏距离	216
7.5.4 混合数据的聚类问题	218
7.6 小结	219
作者简介	220
封面说明	220
• • • • •	(收起)

[面向数据科学家的实用统计学_下载链接1](#)

标签

统计学

统计

数据科学

数据分析

图灵

数据分析与机器学习

R

算法

评论

了解名词概念科普挺好的，覆盖了目前能用的机器学习部分，但是不深，适合按图索骥

适合对统计学和R都有了解且有实际需求的读者。因为这本书实际上像是一本说明书，指导要实现某种统计用途该使用那些R语言工具。但是对原理的讲解不多，对于工具也只是寥寥数语。好在对于实践中可能会出现的问题都cover到了。

人们在思想上倾向于低估天然随机行为的范围。
人们倾向于将随机事件曲解为具有某种显著性的模式。只要拷问数据做够长的时间，数据总会招供。

是一本定位查漏补缺的书，告诉你什么场景下可以用到统计推断或机器学习的方法

梳理了数据科学实践中需要用到的统计学知识。其优点在于，指明了统计理论对于数据科学的实践价值，而诸如参数估计与假设检验等内容，则可适当忽略。书是好书，但两星扣在翻译，有大量的翻译错误或不知所云的地方，严重影响阅读。

算是重要概念的汇总和流程概述吧，讲得挺好懂，适合我这种入门级读者。

补一下基础知识

明明都懂 都还看的很难受

书是好书，内容也是好内容，不过中文版被翻译毁了。
好多名词翻译之后完全莫名奇妙的。例如Regression to The Mean
被翻译为趋均值回归，搞得还以为是一种特定类别的回归呢。

[面向数据科学家的实用统计学_下载链接1](#)

书评

这本书的作者是统计学领域大咖，
Statistics.com统计学教育学院的创立者兼院长，重采样统计软件的开发者。
统计学的书市面上有不少了，但能从应用角度把统计学一些关键概念讲明白的不多。虽然书名说是”面向数据科学家“的，但适合所有人用来学习和巩固统计学基础。
最好了解一...

[面向数据科学家的实用统计学_下载链接1](#)