中级应用统计学

导言

张晨峰

华东理工大学商学院

2016年9月13日

教材

参考教材

- 多元统计分析（何晓群，第4版，中国人民大学出版社）
- 概率论及其应用（威廉·费勒，卷1·第3版）
- 统计推断（卡塞拉和贝耶，翻译版·原书第2版）
- 实用多元统计分析（约翰逊和威克恩，第6版）
- 应用多元统计分析（哈得勒和西马，翻译版，第2版）
- 多元数据分析（Joseph F.H, William C.B, etc. 英文版，第7版）

教师信息

联系方式

- 电邮: glen.zhang7@gmail.com
- 手机: 13918610536

课程评分

- 平时成绩：出勤+平时作业+小项目=30%
- 期末考试：闭卷考试=70%

课程大纲

1. 导论
2. 相关分析和回归分析
3. 主成分和因子分析
4. 聚类分析
5. 定性统计分析
6. 决策分析
7. 时间序列分析
8. 判别分析
9. 结构方程
10. 大数据和机器学习

课程信息

课程网站

github网站：https://github.com/plutoese/datascience

软件程序

- Spss和Stata
- R
- Python

1 导论

主要内容

- 统计学与生活
- 统计学简史
- 应用统计学简介
- 一些应用的实例
- 小项目——词云
- 概率统计简要回顾

数据搜集无处不在

You are being WATCHED!

1.1 统计学简史

统计学的历史时刻

公元前450年，伊利斯（Elis）的希皮亚斯利（Hippias）用国王统治时间的均值计算出了第一次奥林匹克运动会的时间，就是在他生活的时代的300年之前。

统计学的历史时刻

中国历史上第一次完整地记载中国各州、郡的户数和人口，是在公元5年（西汉平帝元始五年），据《汉书·地理志》记载，当时有1223.3万户，5959.4万人。

1.1 统计学简史

统计学的历史时刻

公元840年，数学家肯迪（Al-Kindi）应用频谱分析（frequency analysis）进行了密码破解。此外，他还把阿拉伯数字引入到了欧洲。

统计学的历史时刻

公元1570年，天文学家第谷·布拉赫应用算术平均减少了行星观测的误差。

1.1 统计学简史

统计学的历史时刻

公元1657年，物理学家惠更斯（Huygens）发表了第一部概率论著作《论赌博中的计算》（On Reasoning in Games of Chance）。

统计学的历史时刻

公元1805年，勒让德引入最小二乘法对观测数据集进行曲线拟合。

1.1 统计学简史

统计学的历史时刻

公元1894年，卡尔·皮尔逊（Karl Pearson）提出了“标准差”这一术语，并且发展了卡方检验。

统计学的历史时刻

公元1900年，劳伦斯·巴施里耶（Louis Bachelier）将股票价格的涨跌看作是一种随机布朗运动，并据此进行了严格的数学描述，成为了金融数学领域的开创性研究。

1.1 统计学简史

统计学的历史时刻

公元1935年，费雪（R.A. Fisher）出版的《实验设计》（The Design of Experiments）建立了实验设计法的基础，革新了现代统计学。

统计学的历史时刻

公元1979年，布拉德利·埃弗龙（Bradley Efron）提出了自助法（Bootstrap），这是一个很通用的算法，可以计算任意估计的标准误差。

1.1 统计学简史

统计学的历史时刻

公元1997年，大数据（Big Data）这个词首先见诸于出版物中。

统计学的历史时刻

公元2012年，内特·希尔（Nate Silver）成功预测了美国大选中50个州的投票结果，成为了媒体明星。

统计学的创立时期（17世纪中叶至18世纪中叶）

- **国势学派** 产生于17世纪的德国，其主要代表人物是海尔曼·康令和阿亨华尔。该学派在进行国势比较分析中，偏重事物性质的解释，而不注重数量对比和数量计算，但却为统计学的发展奠定了经济理论基础。
- **政治算术学派** 产生于19世纪中叶的英国，创始人是威廉·配第。他在《政治算术》一书中，利用实际资料，运用数字、重量和尺度等统计方法对英国、法国和荷兰三国的国情国力，作了系统的数量对比分析，从而为统计学的形成和发展奠定了方法论基础。

1.1 统计学简史

统计学的发展时期（18世纪末至19世纪末）

- **数理统计学派** 以比利时的阿道夫·凯特勒为首的统计学家，主张用研究自然科学的方法研究社会现象，正式把古典概率论引进统计学，使统计学进入一个新的发展阶段。
- **社会统计学派** 主要代表人物主要有恩格尔、梅尔等人。他们融合了国势学派与政治算术学派的观点，沿着凯特勒的“基本统计理论”向前发展，但在学科性质上认为统计学是一门社会科学，是研究社会现象变动原因和规律性的实质性科学，以此同数理统计学派通用方法相对立。

现代统计学的发展时期（20世纪初到现在）

- 统计学的主流从描述统计学转向推断统计学
- 向多分支学科发展
- 统计预测和决策科学的发展
- 信息论、控制论、系统论与统计学的相互渗透和结合
- 计算技术和一系列新技术、新方法在统计领域不断得到开发和应用

1.2 应用统计学简介

统计学的概念

统计学是一门收集数据、分析数据，并根据数据进行推断的艺术和科学。（大英百科全书）

统计学的分支

- 描述统计学
- 推断统计学

1.2 应用统计学简介

描述统计学

致力于数据集的整理、概括以及描述的统计学分支

推断统计学

利用样本数据集对总体作出推断的统计学分支

1.2 应用统计学简介

现代统计分析方法

- **分类分析方法** 聚类分析、判别分析、定性资料分析等
- **结构简化方法** 主成分分析、因子分析、对应分析等
- **相关分析方法** 定性资料分析、典型相关分析、回归分析、主成分分析、因子分析、对应分析等
- **预测决策方法** 回归分析、判别分析、定性资料分析、聚类分析等

1.3 一些应用的实例

红学研究

韦博成教授以情景描写为基础，应用统计学方法研究了《红楼梦》前80回与后40回在若干情景描写上的差异。他选择了花卉、树木、饮食、医药与诗词这5个情景指标，统计出它们在前80回与后40回中出现的频数，并应用统计学中的“等价性检验”方法来检验二者的差异。例如，《红楼梦》在前80回中有34回涉及饮食方面的描写；后40回仅有8回涉及饮食方面的描写。结果表明，《红楼梦》前80回与后40回在饮食与花卉的描写上确实存在非常显著的差异；在树木的描写上也存在明显差异。这提供一个强有力的证据，说明《红楼梦》前80回与后40回在某些文风上确实存在非常显著的差异。

蝎子号事件

1968年5月，美国潜艇蝎子号（Scorpion）在完成北大西洋参观后，在返回纽波特纽斯（Newport News）途中消失了。克雷文组建了一个囊括各方面专家的团队。团队成员包括数学家、潜艇专家和救助人员等。有趣的是，他非但不是要求团队成员互相协商寻求一个答案，反而请每个成员提供自己对每个可能场景的发生概率的猜测。克雷文认为，如果他能把所有答案加在一起，构建一个蝎子号出事全景的复合图像，他应该会得到对潜艇最终位置的很好估计。事实证明这个集体的判断非常精彩。在蝎子号消失后的第五个月，海军发现了它。它和克雷文最后得到的位置只差约200米。

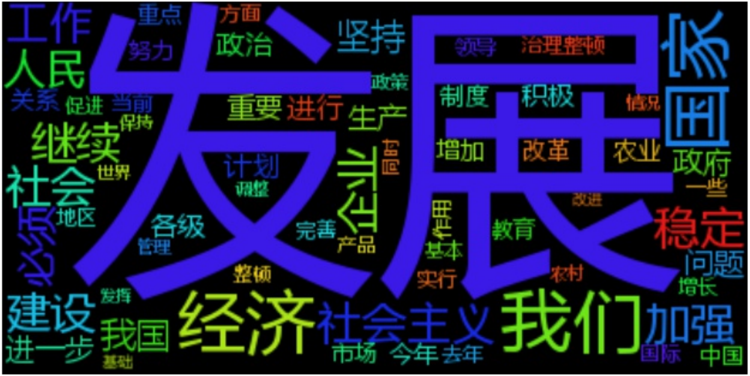

我们
经济
国家机关
已经
根据
一九五三年
他们
企业
技术
社会主义
这些
为了
应当
改善
生活
计划
对于
关系
许多
改造
国家
生产
需要
中国
并且
建立
可以
全国
一九五四年
资本主义
反对
台湾
和平
会议
集团
等于
建设
农业
现在
完成
方面
但是
没有
事业
几年
这个问题
重要
进行
增加
工业
正在
人民
一个
工作
发展
必须
组织
一九四九年
一九五

1.4 小项目——词云


人民
我们
教育
中国
国民经济
革命
资产阶级
发展
国家
建设
各国
生产
必须
工作
民族
运动
斗争
政策
一个
无产阶级
文化
工业
坚持
全世界
这个
一切
自己
力量
群众
资本主义
一定
支持
更加
提高
改造
思想
世界
周总理
社会主义
我国
进行
基础
反动派
干部
进一步
毛主席
问题
美帝国主义
帝国主义
阶级斗争
美国
自力更生
已经

1.4 小项目——词云

增加
万吨
改革
试点
亿元
提高
国民经济
工业
进行
安排
工作
计划
实行
可以
上半年
增长
发展
企业
调整
生产
方面
上年
机械
地区
部门
产量
加强
产品
农业
逐步
进一步
经济
基本建设
总产值
扩大
职工
国家
预计
取得
轻工业
重工业
水平
根据
联合
管理


发展
国家
经济
我们
社会主义
稳定
加强
建设
工作
人民
坚持
政治
继续
社会
必须
企业
改革
增加
制度
积极
农业
政府
重要
进行
生产
我国
进一步
问题
增长
中国
国际
市场
今年
去年
重点
方面
努力
领导
治理整顿
政策
关系
促进
当前
保持
情况
世界
地区
计划
调整
一些
改进
各级
管理
完善
产品
教育
作用
基本
实行
整顿
农村
发挥
基础

1.4 小项目——词云


发展
建设
企业
工作
管理
加强
继续
改革
调整
稳定
制度
各级
国家
增加
政府
教育
群众
社会
加快
进一步
积极
市场
实现
社会主义
进行
政策
坚持
增长
完善
基本
取得
我们
经济
中央
措施
国有企业
投资
促进
问题
扩大
地方
努力
干部
生活
维护
重点
支持
提高
推进
加大
同时
我国
重要
做好
部门

1.4 小项目——词云

发展
社会
政策
推进
继续
就业
工作
稳定
加快
扩大
进一步
经济
改革
建设
农村
提高
增加
人民
推动
鼓励
基本
实施
服务
管理
国家
应对
创新
加强
地区
完善
增长
国际
积极
文化
投资
我们
政府
亿元
合作
重点
农业
制度
企业
促进
市场
坚持
全面
支持
节能
教育
实现
民生
能力
改善
维护

1.4 小项目——词云

发展
建设
经济
改革
创新
推进
政府
实施
中国
合作
安全
投资
扩大
推动
积极
地方
服务
就业
我国
农村
完善
政策
人民
一批
支持
社会
工作
企业
国家
促进
我们
制度
国际
深化
实现
全面
加快
加强
取得
加大
农业
今年
继续
基本
落实
机制

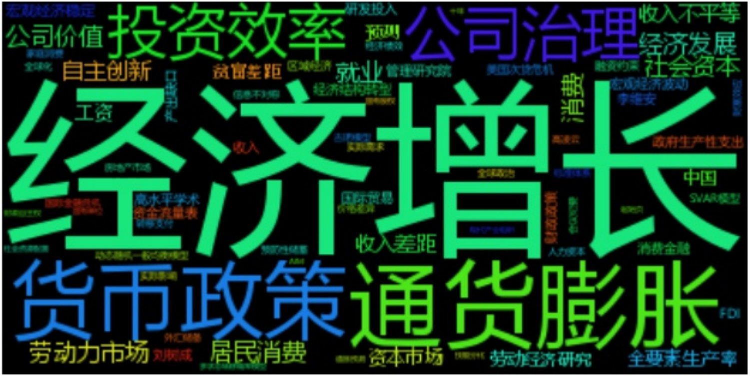
1.4 小项目——词云

货币政策
产出缺口
金融危机
经济发展方式
金融体制
资产价格
生产者价格指数
国有企业改革
国有企业
国际贸易
投资
城市化
面板数据
经济增长
信用卡
理论前沿
消费
中等收入陷阱
汇率
储蓄
利息收入
风险溢价
代理成本
分工
公司治理
技术进步
价格传递
危机时代
消费者价格指数
外包
薪酬激励
国际金融危机
经济可持续增长

1.4 小项目——词云


经济增长
体制性矛盾
资本市场
结构性问题
国际贸易
技术进步
医疗服务利用
经济发展
区域差异
汇率制度
福利
货币供给
金融加速器
信息不对称
实证分析
地区差异
公司价值
家庭资产
潜在增长率
技术溢出效应
经济波动
企业异质性
风险投资
负债
DEA分析
财政政策
农村经济
融资约束
社会保障
全要素生产率
经济运行
经济发展方式
货币政策
环境规制
中国经济

1.4 小项目——词云

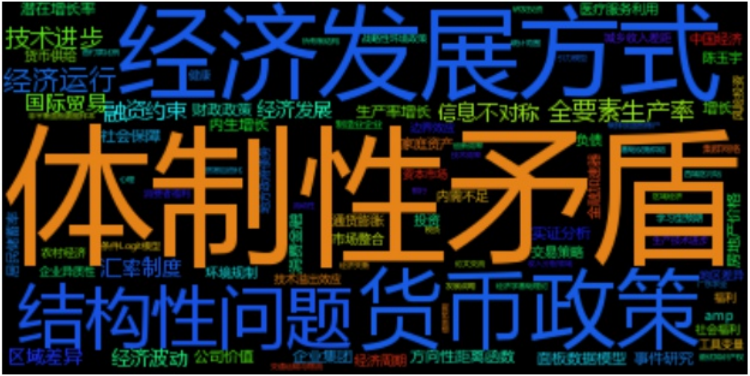


经济增长
货币政策
通货膨胀
投资效率
公司治理
公司价值
收入不平等
经济发展
社会资本
自主创新
贫富差距
就业
工资
消费
中国
收入差距
劳动力市场
居民消费
资本市场
劳动经济研究
全要素生产率
政府生产性支出
SVAR模型
FDI
研发投入

1.4 小项目——词云

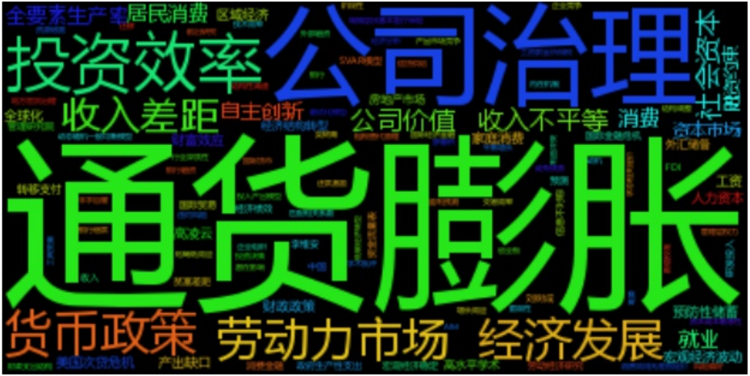

技术进步
公司治理
货币政策
金融危机
危机时代
产出缺口
面板数据
国有企业
代理成本
价格传递
生产者价格指数
城市化
薪酬激励
消费
资产价格
人力资本
保险
储蓄
经济发展方式
消费者价格指数
人民币汇率
生产率
财政分权
中等收入陷阱
地方政府竞争
二氧化碳排放
方向性距离函数
环境全要素生产率

1.4 小项目——词云



经济发展方式
体制性矛盾
结构性问题
货币政策
技术进步
经济运行
国际贸易
融资约束
财政政策
经济发展
信息不对称
全要素生产率
社会保障
汇率制度
区域差异
经济波动
公司价值

1.4 小项目——词云



通货膨胀
公司治理
投资效率
货币政策
劳动力市场
经济发展
收入差距
公司价值
收入不平等
居民消费
自主创新
消费
社会资本
资本市场
就业

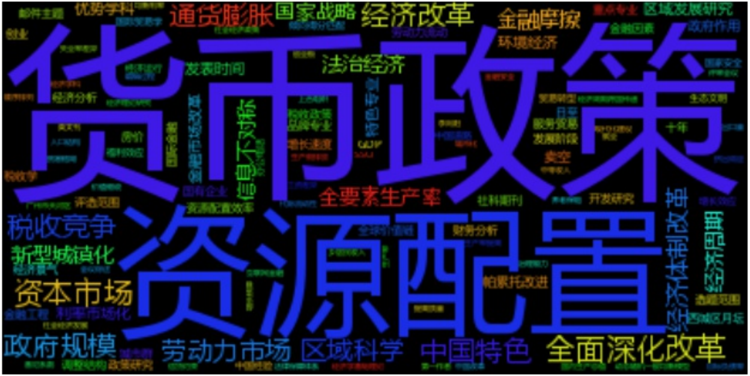
1.4 小项目——词云


收入分配
全要素生产率
经济体制改革
货币政策
顶层设计
异质性
二元经济
发展阶段
融资约束
碳排放
中等收入陷阱
市场分割
生产率
外汇管理体制
土地调整
系统性风险
居民消费
加工贸易
社会公平
税收制度
新型城镇化
收入差距
全面深化改革
劳动收入份额
经济发展方式
家族企业
计划经济体制
价格粘性
经典文献
腐败
医疗保险
养老金
出口
健康

1.4 小项目——词云


货币政策
全面深化改革
经济发展
经济体制改革
区域经济
融资约束
政府补助
中国经济
收入差距
资本市场
经济学科
税收激励
公司财务
金融市场
资产定价
证券市场
全球气候
收入差距
发起单位
金融工程
农民工
财务分析
人力资本
居民消费
新型城镇化
全要素生产率
资源配置
中等收入陷阱
中等收入

1.4 小项目——词云


货币政策
资源配置
通货膨胀
国家战略
经济改革
金融摩擦
区域发展研究
优势学科
环境经济
法治经济
发表时间
政府作用
经济分析
信息不对称
服务贸易
发展阶段
增长速度
生态文明
全要素生产率
国有企业
资源配置效率
评选范围
社科期刊
税收竞争
新型城镇化
资本市场
利率市场化
财务分析
帕累托改进
经济体制改革
经济周期
政府规模
劳动力市场
区域科学
中国特色
全面深化改革

1.5 概率统计简要回顾：集合论

定义 (样本空间)

某次试验全体可能的结果所构成的集合S为该试验的样本空间（*sample space*）。

定义 (事件)

一个事件（*event*）是一次试验若干可能的结果所构成的集合，即S的一个子集（可以是S本身）

集合的关系和运算

集合的关系 序关系，相等关系

集合的运算 并、交、补

1.5 概率统计简要回顾：集合论

定义 (不交)

如果两个事件$A \cap B = \varnothing$，称两个事件A和B不交（*disjoint*）

定义 (两两不交)

如果对于任意$i \neq j$，都有$A_i \cap A_j = \varnothing$，则称事件$A_1, A_2, \ldots$两两不交（*pairwise disjoint*）。

定义 (划分)

如果事件$A_1, A_2, \ldots$两两不交，并且$\bigcup_{i=1}^{\infty} A_i = S$，则称$A_1, A_2, \ldots$构成$S$的一个划分（*partition*）。

定义 (σ代数)

S的一族子集如果满足下列三个性质，就称为一个σ代数，记作Λ。

- $\varnothing \in \Lambda$
- 若$A \in \Lambda$,则$A^c \in \Lambda$
- 若$A_1, A_2, \ldots \in \Lambda$,则$\bigcup_{i=1}^{\infty} A_i \in \Lambda$

1.5 概率统计简要回顾：概率论的公理化基础

定义 (概率函数)

已知样本空间S和σ代数Λ，定义在Λ上且满足下列条件的函数P称为一个概率函数（*probability function*）

- 对任意$A \in \Lambda$，$P(A) \geq 0$
- $P(S) = 1$
- 若$A_1, A_2, \ldots \in \Lambda$且两两不交，则$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

1.5 概率统计简要回顾：条件概率和独立性

定义 (条件概率)

设A, B为S中的事件，且$P(B) > 0$，则在事件B发生的条件下事件A发生的条件概率记作$P(A \mid B)$，表示为

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

定义 (统计独立)

称事件A, B统计独立，如果

$$P(A \cap B) = P(A)P(B)$$

1.5 概率统计简要回顾：随机变量

定义 (随机变量)

从样本空间映射到实数的函数称为随机变量（*random variable*）

定义 (累积分布函数)

随机变量X的累积分布函数（*cumulative distribution function*，简记为*cdf*），记作$F_X(x)$，表示

$$F_X(x) = P_X(X \le x)$$

其中x任意。

1.5 概率统计简要回顾：随机变量

定理 (累积分布函数的性质)

函数$F_X(x)$是一个累积分布函数，当且仅当它同时满足下列三个条件：

- $\lim_{x\to-\infty} F(x) = 0$，且$\lim_{x\to+\infty} F(x) = 1$
- $F(x)$是x的单调递增函数
- $F(x)$右连续

定义 (连续和离散)

设X为一随机变量，如果$F_X(x)$是x的连续函数，则称X是连续的（*continuous*）；如果$F_X(x)$是x的阶梯函数，则称X是离散的（*discrete*）。

定义 (概率质量函数)

离散随机变量X的概率质量函数（*probability mass function*，简称*pmf*）为

$$f_X(x) = P_X(X = x)$$

定义 (概率密度函数)

连续随机变量X的概率密度函数$f_X(x)$（*probability density function*，简称*pdf*）是满足下式的函数

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$$

1.5 概率统计简要回顾：期望

定义 (期望)

连续随机变量$g(X)$的期望（*expected value*）或均值（*mean*），记作$Eg(x)$，定义为

$$Eg(X) = \int_{-\infty}^{+\infty} g(x) f_X(x)\, dx$$

1.5 概率统计简要回顾：期望

定义 (矩)

对任意整数n，X（或$F_X(x)$）的n阶矩，记作$\mu_n^{'}$，定义为

$$\mu_n^{'} = EX^n$$

X的n阶中心矩，记为μ_n，定义为

$$\mu_n = E(X-\mu)^n$$

1.5 概率统计简要回顾：多维随机变量

定义 (联合概率密度函数)

设(X, Y)为连续随机变量，称R^2到R的函数$f(x, y)$为(X, Y)的联合概率密度函数（*joint pdf*），如果对于任意$A \in R$，都有

$$P((X, Y) \in A) = \iint_A f(x, y)\, dx\, dy$$

X和Y的边缘概率密度函数分别为

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y)\, dy$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y)\, dx$$

1.5 概率统计简要回顾：多维随机变量

定义 (独立随机变量)

设(X, Y)是二维随机变量，其联合概率密度函数为$f(x, y)$，边缘概率密度函数分别为$f_X(x)$和$f_Y(y)$。称X和Y是独立随机变量，如果对于任意$x, y \in R$，都有

$$f(x, y) = f_X(x) f_Y(y)$$

定义 (协方差)

随机变量X和Y的协方差定义为：

$$Cov(X, Y) = E((X - \mu_X)(Y - \mu_Y))$$

1.5 概率统计简要回顾：随机样本的性质

定义 (随机样本)

如果随机变量$X_1, ..., X_n$相互独立且有相同的边缘概率密度函数$f(x)$，则称$X_1, ..., X_n$是总体$f(x)$的大小为n的随机样本，或称$X_1, ..., X_n$是概率密度函数为$f(x)$的独立同分布随机变量（简记为iid随机变量）。

定义 (统计量)

设$X_1, ..., X_n$是从总体中抽取的大小为n的随机样本，$T(x_1, ..., x_n)$是定义在$(X_1, ..., X_n)$的样本空间上的实值或向量值函数，则随机变量或随机向量$Y = T(X_1, ..., X_n)$称为一个统计量（*statistic*），Y的概率分布称为Y的抽样分布（*sampling distribution*）。

定义 (样本均值)

样本均值（*sample mean*）是随机样本值的算术平均，常记作

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{N} X_i$$