

高等学校教材

概率论与数理统计

东华大学概率统计教研组 编

教育出版社

高等学校教材

概率论与数理统计

东华大学概率统计教研组 编

高等教育出版社·北京

内容提要

本书强调概率论与数理统计的应用性，主要包括概率与统计简介、描述统计学、概率论的基础、随机变量的概率分布与数字特征、几种常见的分布、统计量的分布、参数估计、假设检验和线性回归等内容。书中主要统计计算使用 Excel 软件完成。学习本书的主要理论仅需要读者具有一元微积分的数学基础。

本书可作为高等学校非数学类专业“概率论与数理统计”课程的教材，也适合科技工作者自学与参考。

图书在版编目（C I P）数据

概率论与数理统计 / 东华大学概率统计教研组编

— 北京：高等教育出版社，2017. 1

ISBN 978-7-04-046861-8

I. ①概… II. ①东… III. ①概率论－高等学校－教材②数理统计－高等学校－教材 IV. ①O21

中国版本图书馆 CIP 数据核字（2016）第 280842 号

概率论与数理统计
Gailülun yu Shuli Tongji

策划编辑	张彦云	责任编辑	张彦云	封面设计	姜磊	版式设计	马敬茹
插图绘制	黄云燕	责任校对	李大鹏	责任印制	韩刚		

出版发行 高等教育出版社
社 址 北京市西城区德外大街4号
邮政编码 100120
印 刷 保定市中画美凯印刷有限公司

开 本 787mm×960mm 1/16
印 张 14.75
字 数 220 千字
购书热线 010-58581118

咨询电话 400-810-0598
网 址 <http://www.hep.edu.cn>
<http://www.hep.com.cn>
网上订购 <http://www.hepmall.com.cn>
<http://www.hepmall.com>
<http://www.hepmall.cn>

版 次 2017年1月第1版
印 次 2017年1月第1次印刷
定 价 28.40 元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换
版权所有 侵权必究
物 料 号 46861-00

前言

自从 20 世纪末我国高等教育大规模扩招以来,学生的学习能力和发展方向的差异性增大。对于一部分有志考研的学生,大学数学教学必须遵从考研大纲的要求,着重数学理论的系统性和严谨性培养,而对于大部分其他学生,学习数学的主要目的是为后续专业课程的学习打下基础,从而对教学内容选择的灵活性就更大一些。本教材是我们近几年来在东华大学实施“分流培养、分层教学、分类成才”的教学改革实践中编写而成的,主要面向不考研的学生,强调概率论与数理统计课程的应用性和统计软件的使用。

与目前国内概率论与数理统计的主流教材相比,本教材具有如下特色:

(1) 内容包含了描述统计学和线性回归这两个应用统计学的重要内容。这两部分内容在多数教材中往往不出现或者是在教学实践中来不及安排的。

(2) 删除了部分数学理论推导较复杂的内容。例如,对于多维随机变量,主要介绍二维离散型随机变量的联合分布律,在此基础上引入独立性和协方差等概念;而对于多维连续型随机变量,仅介绍密度函数和独立性的基本概念,但没有布置这方面的习题。在区间估计和假设检验中仅介绍单个总体均值的估计和检验,略去了比较复杂的两个正态总体的估计和检验,我们相信这些简化不会影响学生对于概率论和数理统计基本思想的理解。这样处理也使得本教材只要求学生预先学过一元微积分,不需要多元微积分和线性代数的知识,因而可在大学一年级的第二学期安排课程教学。

(3) 尽管本教材不强调数学上的严密性,但对所有理论的由来都给出了推导过程,而不是直接给出定理的结果。例如,本书对正态分布的独立可加性给出了一个简单却又完整的证明。我们还利用蒙特卡罗方法,画出独立均匀分布和的直方图,从而较自然地呈现出中心极限定理的正态分布曲线。又如,关于样本方差的分布以及回归模型残差平方和的分布,我们是从与卡方分布定义作类比的方

法得到的。这个方法很自然,学生容易理解,尽管它不是数学上的严格证明(该证明需要利用正交矩阵)。

(4) 将 Excel 软件与教学内容有机地结合起来。Excel 是大部分学生都熟悉的办公自动化软件,易于上手。首先,我们介绍了它的一些描述统计函数,如均值和标准差等,这使学生更容易处理较大规模数据样本的统计计算及其可视化表达。其次,我们利用它的概率分布计算函数,可避免查询很多统计分位数表。同时,在假设检验中,我们主要采用 p 值检验法而不是临界值检验法,前者是绝大多数统计软件采用的方法。另外,我们还利用 Excel 随机数生成函数来介绍蒙特卡罗模拟,并由此引出中心极限定理。与此同时,我们并不追求对 Excel 软件统计功能的完整介绍,没有介绍怎样利用 Excel 软件直接得出一些假设检验和回归分析的计算结果,因为这毕竟不是本课程的主要目的。教材中采用的是 Excel 2007 版,其中的函数在更高的版本中是被兼容的。在附录中,我们还给出了 2007 版与 2010 版函数的对照表。

对于部分可选的内容,我们用 ★ 号作标记,略过这些内容基本不会影响后续内容的学习。如果不包括打 ★ 号的部分,本教材的全部内容可以用 40 学时(每学时 45 分钟)讲完,大致分配如下表。学时较充裕时,建议教师也讲解一些打 ★ 号的内容,这可使得课程内容的体系更完整。

章	1	2	3	4	5	6	7	8	9
学时	1	2	5	8	7	5	4	4	4

此外,除了纸质教材内容,我们还通过扫描二维码链接的网址介绍了一些补充内容供学生自学。补充内容主要包含两部分:第一部分是课本上介绍的 Excel 函数的操作演示视频,这是教师不方便在课堂上介绍的;另一部分是一些较深入内容的讲解,主要提供给部分希望在理论上要求更高的读者(比如考研学生)选学。

本书第 1、6 章由胡良剑编写,第 2 章由姜荣编写,第 3 章由谢臻贇编写,第 4 章由闫理坦编写,第 5 章由何坤、葛勇编写,第 7 章由陆允生编写,第 8 章由赵伟国、刘欣编写,第 9 章由童金英、张振中编写。全书由胡良剑、陈珊敏、孙

晓君全面修改和统稿,并制作网络补充内容。

由于编者水平有限,书中不妥之处在所难免,请各位专家、读者不吝指正。

编者

2016 年 9 月于东华大学

目录

第 1 章 概率与统计简介	001
习题	008
第 2 章 描述统计学	010
2.1 数据的描述	010
2.1.1 频数表和频数图	010
2.1.2 相对频数表和相对频数图	011
2.1.3 分组数据、直方图和累积频数曲线	012
2.2 数据的汇总	015
2.2.1 样本均值、样本中位数和样本众数	015
2.2.2 样本方差和样本标准差	018
2.2.3 样本分位数和盒形图	021
2.2.4 成对数据集和样本相关系数	024
习题	028
第 3 章 概率论的基础	033
3.1 概率的基本概念	033
3.1.1 样本空间和随机事件	033
3.1.2 概率的定义与性质	036
3.2 等可能概型	039
3.2.1 古典概型	039

3.2.2 几何概型	044
3.3 条件概率	045
3.3.1 条件概率与乘法公式	045
3.3.2 全概率公式与贝叶斯公式	047
3.4 事件的独立性	051
习题	055
第 4 章 随机变量的概率分布与数字特征	060
4.1 随机变量及其概率分布	060
4.1.1 离散型随机变量及其分布律	060
4.1.2 分布函数	062
4.1.3 连续型随机变量及其密度函数	064
4.1.4 随机变量函数的分布	066
4.2 随机变量的联合概率分布及独立性	068
4.2.1 离散型随机变量的联合分布律	068
4.2.2 连续型随机变量的联合密度函数	070
4.2.3 随机变量的独立性	071
4.3 数学期望	074
4.3.1 数学期望的定义	075
4.3.2 随机变量函数的数学期望	076
4.3.3 数学期望的性质	079
4.4 方差和协方差	081
4.4.1 方差和标准差	082
4.4.2 协方差和相关系数	086
习题	089

第 5 章 几种常见的分布	096
5.1 伯努利分布和二项分布	096
5.1.1 伯努利试验与伯努利分布	096
5.1.2 二项分布的定义与计算	097
5.1.3 二项分布的性质及应用	100
5.2 泊松分布	104
5.2.1 泊松分布的定义与计算	104
5.2.2 泊松分布的性质与应用	105
5.3 均匀分布	109
5.4 指数分布	112
5.5 正态分布	114
5.5.1 标准正态分布	115
5.5.2 正态分布及其性质	118
5.6 正态分布生成的分布	124
5.6.1 χ^2 分布	124
5.6.2 t 分布	126
5.6.3 F 分布	128
习题	130
第 6 章 统计量的分布	135
6.1 随机样本	135
6.2 正态总体统计量的分布	137
6.3 大数定律与随机模拟	141
6.3.1 辛钦大数定律和伯努利大数定律	141
6.3.2 蒙特卡罗模拟	142
6.4 中心极限定理	144
6.4.1 独立随机变量和的蒙特卡罗模拟	144
6.4.2 两个中心极限定理	145

6.5 大样本均值的近似分布	151
习题	153
第 7 章 参数估计	157
7.1 点估计	157
7.1.1 点估计及其评价	157
7.1.2 最大似然估计	160
7.2 总体均值的区间估计	167
7.2.1 方差已知时正态总体均值的区间估计	167
7.2.2 方差未知时正态总体均值的区间估计	169
7.2.3 大样本情形总体均值的区间估计	170
习题	171
第 8 章 假设检验	174
8.1 假设检验的基本概念	174
8.2 总体均值的假设检验	177
8.2.1 方差已知时正态总体均值的假设检验	177
8.2.2 方差未知时正态总体均值的假设检验	186
8.2.3 大样本情形总体均值的假设检验	188
习题	191
第 9 章 线性回归	194
9.1 回归模型的参数估计	194
9.1.1 一元线性回归模型	194
9.1.2 回归系数的估计	195
9.2 回归模型的假设检验	201
9.2.1 方差分析	201
9.2.2 估计量的分布	202

9.3 ★ 回归预测	205
9.3.1 回归系数的区间估计	206
9.3.2 因变量的预测区间	208
习题	212
附表	217

第 1 章

概率与统计简介

统计学是研究如何收集、整理、分析和解释数据,并从数据中得出结论的科学. 它的历史可追溯到公元前 3 世纪的古希腊,起源于研究社会经济问题. 统计学的英文单词 statistics 源于现代拉丁文 statisticum collegium (国会) 以及意大利文 statista (政治家), 代表对国家的资料进行分析的学问. 在统计中, 我们把感兴趣的研究对象称为总体. 例如, 某企业近三年来每个月的总产值, 某连锁宾馆各分店某一天入住的顾客数等. 在古典统计学中, 总体就是所有的观测数据. 我们对这些数据进行汇总, 得到简明扼要的结果, 从而获取总体的本质特征. 这类对数据的整理, 描述和表达的方法统称为描述统计学 (descriptive statistics).

例 1.0.1 某寄宿制中学有 100 名男生, 测得其身高 (单位: cm) 和体重 (单位: kg) 数据如表 1.1, 求这 100 名男生的平均身高和平均体重, 并描述其分布特征.

表 1.1 100 名男生的身高和体重数据

编号	身高	体重	编号	身高	体重	编号	身高	体重	编号	身高	体重
1	167	50	9	162	53	17		62	25	173	66
2	179	63	10	177	67	18	171	63	26	179	69
3	168	54	11	179	68	19	169	56	27	172	59
4	187	79	12	172	61	20	167	64	28	169	59
5	173	62	13	170	58	21	169	64	29	167	56
6	176	70	14	177	67	22	166	53	30	163	51
7	170	57	15	172	62	23	163	50	31	158	44
8	170	57	16	166	53	24	175	66	32	175	69

续表

编号	身高	体重	编号	身高	体重	编号	身高	体重	编号	身高	体重
33	174	61	50	166	50	67	169	51	84	175	66
34	159	53	51	174	61	68	171	54	85	168	55
35	160	47	52	166	48	69	175	63	86	163	57
36	168	63	53	167	55	70	173	65	87	170	63
37	169	53	54	177	65	71	175	72	88	177	
38	171	63	55	162	62	72	162	47	89	172	58
39	160	51	56	182	70	73	172	62	90	167	58
40	165	53	57	171	60	74	170	63	91	169	59
41	174	63	58	175	58	75	171	68	92	171	63
42	164	53	59	179	66	76	174	61	93	179	65
43	180	73	60	172	61	77	173	65	94	168	58
44	170	58	61	174	61	78	159	48	95	160	45
45	175	68	62	160	45	79	170	68	96	165	52
46	1580	55	63	164	57	80	179	66	97	167	53
47	174	67	64	176	60	81	167	51	98	164	56
48	162	50	65	182	73	82	174	65	99	156	45
49	172	63	66	173	59	83	170	52	100	166	50

解 首先对原始数据作清理, 删除或修复一些残缺或错误的数 据, 以免影响后续的分析. 编号 17 的身高数据和编号 88 的体重数据残缺, 我们将这两条学生记录删除, 这样还剩余 98 条记录. 我们还注意到, 编号 46 的身高数据 1580 cm 出现了单位错误, 所以将它修正为 158 cm. 由于原始数据相当琐碎, 不易看出数据的分布特征, 我们需要用适当方法进行汇总.

数据的描述分为图形可视化描述和统计量数字化描述两类方法. 将清理后的 98 组数据依次写入 Excel 表格的某两列, 用 Excel 函数 average 计算得身高均值 170 cm 和体重均值 59.17 kg. 用函数 min 计算得身高最小值 156 cm 和体重最小值 44 kg. 用函数 max 计算得身高最大值 187 cm 和体重最大值 79 kg. 进一步, 我们将数据按大小适当分为若干组, 利用 Excel 数据分析工具计算各组的频数并画出直方图. 身高数据的分组和直方图



例 1.0.1
Excel 演示

见图 1.1, 体重数据的分组和直方图见图 1.2. 直方图直观地反映了这批数据的分布特征, 在均值附近数据有较高的频数, 而靠近最小值或最大值附近频数较低.

□

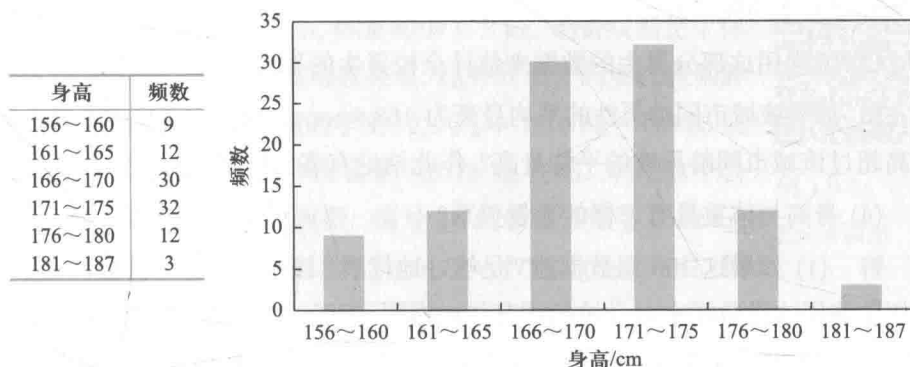


图 1.1 身高数据分组和直方图

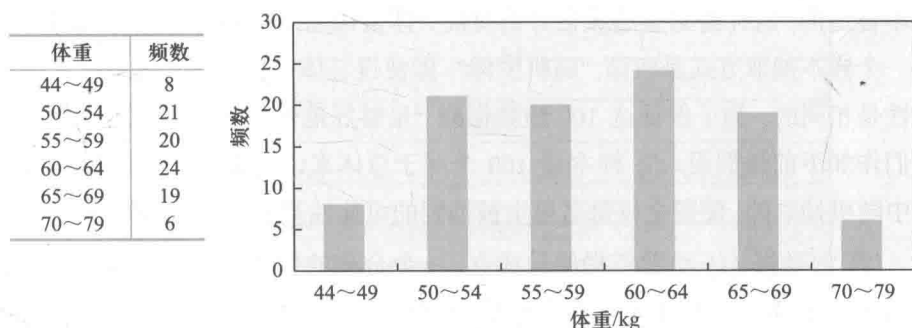


图 1.2 体重数据分组和直方图

随着时代的发展, 统计学已延伸到商业、医疗、自然以及社会科学等领域. 总体的数据量变得越来越大, 收集总体中所有个体的数据费时费力, 成本过高, 甚至成为不可能完成的任务. 例如, 总体是上海市的所有居民, 或东风汽车集团生产的所有汽车, 或中央电视台春节联欢晚会的所有观众等. 这时, 我们可以抽取一部分个体的数据来进行研究, 这部分个体称为样本. 如果统计方法使用得当, 利用这一小部分样本数据仍然可以得到关于总体的正确结论. 利用少量样本数据来对未知总体作出推断的方法, 称为推断统计学 (inferential statistics).

例 1.0.2 为了调查某寄宿制中学 900 多名男生的身体素质情况, 抽取了 100 名男生, 测得身高和体重数据如表 1.1, 现要求研究下列问题:

(1) 是否可以用这 100 名男生的数据来推测全校 900 多名男生的身高和体重? 需要怎样的前提假设?

(2) 怎样用这部分男生的数据来估计全校男生的平均身高? 精确度如何?

(3) 如果该城市同龄男性的平均身高为 168.8 cm, 能否认为该校男生的平均身高超过该城市同龄男性的平均身高? 作此结论有多大把握?

(4) 身高与体重具有怎样的数量关系?

解 (1) 如果这 100 组数据能“足够好地代表”该中学的全体男生, 我们可以相信能用这批数据来对全校的情况作出推测. 但是, 怎样判断样本是否“足够好”呢? 这与样本的数量和抽取方式有关. 一方面, 因为这所中学有 900 多名男生, 只抽取 10 名男生就显得太少了, 而 100 名男生从数量上看应该是够了. 另一方面, 如果这 100 名男生都是从高年级选取或者很多是从篮球队员中选取也是不合适的, 这样势必会造成估计的身高、体重偏高. 保证样本质量“足够好”的一个样本抽取方式是所谓“随机抽样”, 即使得总体中的每个个体被抽到的可能性是相同的. 为了保证这 100 组数据能“足够好地代表”该中学的全体男生, 我们作如下前提假设: (i) 样本量 100 个对于总体来说足够大; (ii) 样本是从总体中随机抽取的, 使得全校每名男生被抽到的可能性基本上是一样的.

(2) 如果第 (1) 小题前提假设成立, 一个合理的想法就是用样本的平均值 170 cm 来作为全校男生平均身高的估计. 可以理解的是, 这样估计必定会存在误差, 那么这个误差有多大呢? 这是一个难以回答的问题.

(3) 这 100 名男生身高的样本平均值为 170 cm, 超过了该城市同龄男性的平均身高 168.8 cm. 如果第 (1) 小题前提假设成立, 我们似乎应该认定该校男生的平均身高超过 168.8 cm. 但由于样本只是全校男生的一小部分, 我们并不能断定这一结论是正确的. 即使全校男生的平均身高实际上不超过 168.8 cm, 也有可能由于样本抽取的偶然性造成平均值偏大. 那么, 有多大把握肯定这一结论呢? 这也是一个难以回答的问题.

(4) 为了探讨身高与体重的关系, 我们用 Excel 软件作出散点图, 加上线性趋势线, 并显示公式, 得到图 1.3. 如果第 (1) 小题前提假设成立, 我们有理由

相信这批数据显示的身高与体重的关系也适用于这所学校的全体男生. 该图显示, 总体上说身材较高的人体重也是偏重的, 但身高与体重之间并没有很规则的数量关系. 有的人与比他矮的人相比还要轻一些, 如编号 19 的人比编号 20 的人高了 2 cm, 体重却少了 8 kg. 趋势线给出了体重 y 与身高 x 之间的大致函数关系. 趋势线上方的点代表偏胖的学生, 下方的点代表偏瘦的学生. 公式 $y = 1.0178x - 113.85$ 还告诉我们, 身高长 1 cm, 体重会重 1.0178 kg 左右. 利用这个公式我们还可以修复残缺的数据. 编号 17 的学生体重 $y = 62$ kg, 可推测其身高在 $x = (y + 113.85)/1.0178 = 172.77$ (cm) 左右. 编号 88 的学生身高 $x = 177$ cm, 有理由相信其体重在 $y = 1.0178x - 113.85 = 66.3$ (kg) 左右. □



例 1.0.2

Excel 演示

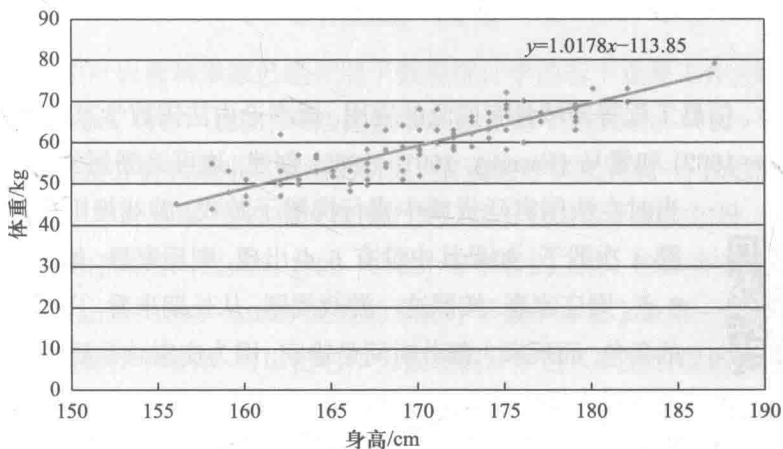


图 1.3 身高 - 体重散点图及趋势线

由于样本的选取带有不确定性, 评估样本推断总体的效果不是一件容易的事情, 需要较深刻的数学理论来支撑. 数理统计学 (mathematical statistics) 就是推断统计学的理论依据. 与古典统计学不同的是, 数理统计学认为总体是一个数学模型 (代表一大批未知的数据), 样本则是从总体抽取出来的一部分观测数据. 古典统计学与数理统计学的区别见图 1.4 和图 1.5.

数理统计学主要建立在研究不确定现象的数学理论——概率论 (probability theory) 的基础上. 概率论不仅是数理统计学的理论基础, 还在金融、保险、计

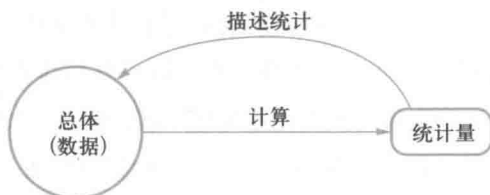


图 1.4 古典统计学

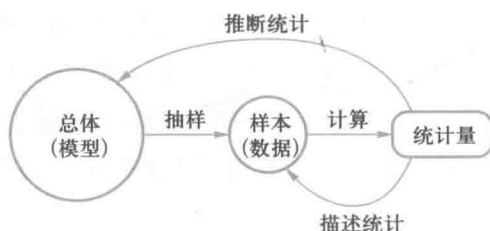


图 1.5 数理统计学

计算机科学、信息工程等领域具有广泛的应用。概率论由法国数学家帕斯卡 (Pascal, 1623—1662) 和费马 (Fermat, 1601—1665) 创建, 这可追溯到公元 17 世纪, 当时在法国宫廷贵族中盛行掷骰子游戏, 游戏规则是玩家连续掷 4 次骰子, 如果其中没有 6 点出现, 则玩家赢, 如果出现一次 6 点, 则庄家赢。按照这一游戏规则, 从长期来看, 庄家扮演赢家的角色, 而玩家大部分时间是输家, 因为庄家总是要靠此为生的, 而当时人们也接受了这种现象。后来为了使游戏更刺激, 游戏规则发生了些许变化, 玩家用 2 个骰子连续掷 24 次, 如果不同时出现 2 个 6 点, 玩家赢, 否则庄家赢。当时人们普遍认为, 出现 2 个 6 点的概率是出现一个 6 点的概率的 $\frac{1}{6}$, 因此 6 倍于前一种规则的次数, 也就是掷 24 次时赢或输的概率与以前是相等的。然而事实却并非如此, 从长期来看, 此时庄家处于输家的状态, 于是他们去请教当时的数学家帕斯卡, 希望他对这种现象作出解释, 由此, 帕斯卡在与费马的相互通信中建立了古典概率论。

概率论的起源
与发展

瑞士数学家伯努利 (Bernoulli, 1654—1705) 建立了概率论的第一个极限定理 (即伯努利大数定律), 阐明了事件的频率稳定于它的概率。随后法国数学家棣莫弗 (De Moivre, 1667—1754) 和拉普拉斯 (Laplace, 1749—1827) 又导出了