第一章

基本概念

统计学是研究如何有效地收集、整理、分析和解释数据的科学, 其目的在于探索数据内在的某些规律性, 以达到对研究对象的科学认识. 可以说, “由数据探索事物内在规律”的思想始终贯穿于统计学. 随着智能时代的到来, 无论在科学研究还是社会生活中都产生了大量数据, 数据无处不在. 另外, 万物数字化已是必然, 数据已成为生产要素、一种资源, 甚至是战略资源, 如何让这些数据产生更大的价值, 更好地服务人类则成为当前多个学科的前沿课题之一. 而统计学作为研究数据的科学, 在自然科学、工程技术、航空航天、人工智能、制造业、军事等领域, 以及生物医药、经济金融、社会科学、人文和管理科学等许多学科都有着广泛的应用, 并且推动着这些学科的发展. 统计学在民生保障、经济生产、可持续发展、现代化建设等方面也发挥着关键的促进作用.

本章将首先通过几个例子简述统计学的几个研究方向, 之后给出统计学的定义及研究内容, 最后给出样本、参数、抽样分布、统计量, 以及充分、完全统计量等一些基本概念.

1.1　引言

我们先通过几个例子尝试说明统计学有许多研究方向与内容, 本书所讲内容只是统计学知识海洋中的沧海一粟; 之后再简述什么是统计学.

1.1.1　几个例子

例 1.1.1 (抽样调查 (sampling survey))　随着高校扩招和自主招生的推广, 每所院校都越来越重视各自的生源情况, 也都想出了各种各样的招生方法以吸引各地优秀中学生报考. 导致这种现象的一个重要原因在于各高校间竞争的本质就是人才的竞争. 大家普遍认为, 好的生源是培养高质量人才的前提. 我们自然要问: 一个人是否成才, 与其大学成绩相关吗? 大学成绩与高中成绩有关吗?

为回答第一个问题, 就要给出衡量一个人是否成才的标准, 或制定一套用来衡量一个人成才与否的可量化的指标体系, 之后再研究与大学成绩的关系. 而指标体系的建立, 以及相关性度量就是统计中的一个研究内容.

为回答第二个问题, 即大学成绩是否与高中成绩有关, 有必要考虑如下问题:

(1) 由于全国大学众多, 普查很难实现, 故在全国选取多少所大学以及哪几所大学进行调查比较合适?

(2) 由于学生数量也非常多, 故在选定的大学中选取多少以及哪些大学生参与调查, 调查结果才可靠?

(3) 由于学生在校期间, 还参加多种多样的课外竞赛或活动, 故如何衡量一名学生的高中及大学成绩?

(4) 不同科系之间、不同院校之间的成绩具有可比性吗?

(5) 数据收集之后, 对这些数据如何进行科学的分析?

在解决上述问题过程中, 我们感兴趣的问题可能还包括: 重点大学与一般大学的区别有多大? 男生与女生有区别吗? 东西部大学及生源有区别吗? 在进行数据分析过程中可能会遇到如下问题: 在收回的问卷中, 如仅回答了部分问题, 则这份问卷如何处理? 等等.

为得到一个科学合理的答案, 在问卷设计之初, 就有必要考虑上述问题, 并且在收集到数据后, 也有必要利用统计方法对其进行合理的数据分析. 问卷设计及数据分析就是统计学中抽样调查这一研究方向的研究内容. 历史上最早的抽样调查是法国 1802 年进行的人口抽样调查. 1800 年左右, 法国数学家、天文学家、概率学家 P.-S. Laplace (拉普拉斯, 1749—1827) 受政府委托估计全国人口. 此时 Laplace 就采用了抽样调查, 且提出了等概率估计.

当然, 我们日常生活中的许多实际问题都与抽样调查有关, 如某品牌的市场占有率、我国农产品产量、全国总人口, 以及消费物价指数等.

在设计抽样调查方案或进行抽样调查时, 应公平地对待每一群体及个体, 否则会导致错误的结论. 比如 1936 年美国总统大选, 候选人为民主党的 Roosevelt(罗斯福) 与共和党的 Landon(兰登). 当时最有名的民意调查机构《文学摘要》根据其民意调查, 预测 Landon 的得票率为 $57\% \pm 19\%$, 而盖洛普根据 5 万人的民意调查, 预测 Landon 的得票率为 $44\% \pm 6\%$. 最终, Landon 以实际得票率 48% 输了大选. 《文学摘要》调查的偏差主要来自其选取的数据代表性不足:《文学摘要》根据电话簿或其俱乐部会员名单上的地址将问卷寄给 1 000 万个选民, 收回问卷 238 万份. 但我们注意到, 1936 年的美国大概只有四分之一的家庭安装了电话, 且由于经济好转, 穷人普遍倾向 Roosevelt, 富人倾向 Landon. 于是,《文学摘要》的调查结果更多地代表了富人的意愿, 从而导致预测不准.

另外, 一些敏感问题, 比如艾滋病患者的比例、官员受贿比例等, 也是抽样调查要研究的内容. 对于这样的敏感问题, 通过随机问卷的方式很难得到真实的答案. Warner (华纳, 1965) 提出了随机应答技术 (randomized response technique, 简记为 RRT) 来处理上述敏感问题调查.

假如有 240 人参与是否受贿的敏感性问题调查, 此时为得到一个较客观的结论, 我们可以让接受问卷者通过扔硬币方式回答问题 A 或 B:

问题 A: 您是否受过贿?

问题 B: 您是否 5 月出生?

假设回答“是”和“不是”的人分别为 30 人和 210 人, 请问最终的受贿比例如何?

假设硬币均匀, 故有 120 人分别回答了问题 A 与 B, 而在回答问题 B 的 120 人中回

答“是”的人数应该是 1/12, 即在回答问题 B 的 120 人中有 10 人回答了“是”, 于是我们知道共有 30−10=20 人在回答问题 A 时回答了“是”, 于是受贿比例为 20/120=16.67%.

在抽样调查中, 有多种抽样方法可供选择, 如简单随机抽样、分层抽样、系统抽样、整群抽样、小域抽样等, 请有兴趣的读者关注统计学的抽样调查这一研究方向.

例 1.1.2 (随机对照双盲试验 (randomized controlled double blind trial))　如何衡量一种新药是否安全有效?

在药物研制过程中, 为验证其有效性、毒副作用等, 制药公司首先需要做离体实验和动物实验. 一旦通过, 就可以向药监部门申请做临床试验 (clinical trial). 临床试验共分四期: I、II、III 及 IV 期. 按照国家药品监督管理局颁布的《药物临床试验质量管理规范》, 四期试验为:

I 期试验: 时间短 (数月)、规模小 (20∼30 人), 健康志愿者或患者均可参加, 目的在于检查新药是否有急性毒副作用、合适的安全给药剂量及药物动力学试验. 如果没有严重问题, 则转入 II 期.

II 期试验: 时间及规模适当 (几个月到两年, 100∼300 人), 试验对象为患者, 目的在于观察新药是否有疗效, 也对短期的安全性做进一步观察. 通过后, 转入 III 期.

III 期试验: 长期大规模 (一到四年, 不少于 300 人), 试验对象为患者, 目的在于确认新药疗效及安全性, 确定给药剂量. 通过后, 制药公司可向药监部门提出上市申请, 待通过药监部门组织的专家鉴定后, 经药监部门批准后就可上市.

IV 期试验: 在新药批准上市后, 进一步观察药物在大范围长时间内的疗效和安全性, 并与其他药物进行比较, 且观察在 I、II、III 期试验中被排除在外的儿童、孕妇、老人患者群体的疗效、安全性和给药剂量.

在 II、III 期临床试验中, 常采用随机对照双盲试验: 将患者随机分两组, 一组吃新药 A, 另一组吃安慰剂 B(placebo, 与新药形状、颜色、味道完全相同, 但没有任何药效); 患者及试验者均不知 A 药与 B 药哪个为试验药 (在 II 期试验中也可不设盲). 双盲目的在于消除心理因素的影响. 吃新药组称为处理组 (treatment, 用 T 表示), 吃安慰剂组称为对照组 (contrast, 用 C 表示). 在 IV 期试验中, 可不设对照组.

随机对照双盲试验最早应用于 20 世纪 60 年代进行的脊髓灰质炎疫苗检测. 20 世纪 50 年代初, 美国小儿麻痹症防治基金会 (NFIP) 认为由匹兹堡大学 Jonas Salk (乔纳斯・索尔克) 研制的疫苗有抗体, 但是否推广, 要进行一次大规模试验检测. 1954 年美国公共卫生总署决定组织此次试验. 试验对象是那些容易感染的人群: 小学一、二、三年级的学生, 试验方案有如下五种:

方案一: 与过去比较;

方案二: 不同地区间比较;

方案三: 给取得父母同意的儿童接种;

方案四: NFIP 的试验方案;

方案五: 随机对照双盲试验.

对于方案一, 由于脊髓灰质炎是一种流行病, 而流行病每年的发病率变化很大, 故被否决. 对于方案二, 由于脊髓灰质炎很可能在某些地区流行, 而在另一地区不流行, 故也被否决. 对于方案三, 人们发现脊髓灰质炎似乎偏爱那些卫生保健条件较好的人. 由于教育程度高的父母更愿意接种, 且其家境比较富裕, 居住条件也较好, 故将导致当疫苗有效时, 而试验结果不利于人们认为疫苗有效的结果. 这是一个条件不对等的对照比较, 也被否决. 对于方案四, 由于仅给所有二年级并取得父母同意的儿童接种疫苗, 而将一、三年级的儿童作为对照组, 故此方案除有方案三的不足之外, 还需考虑到: 脊髓灰质炎是通过接触传染, 二年级的发病率可能比一、三年级的发病率低或高; 另外, 医生明确知道一、三年级没有接种, 故在诊断时容易产生偏差. 方案五按如下安排试验:

- 双盲: 医生不知道谁接种与不接种, 试验者也不知道自己是在处理组还是对照组.
- 对等: 由于试验有风险, 但也可能有效, 故是否接种都要事先征得父母同意.
- 随机: 处理组与对照组随机分配.

表 1.1.1 汇总了方案五的试验结果, 结果显示该疫苗有效.

表 1.1.1 脊髓灰质炎试验检测结果

		人数	患病数	10 万人患病比例/%
父母同意	接种	200 745	57	28.4
参与试验	不接种	201 229	142	70.6
父母不同意参与试验		338 778	157	46.3

在有些临床试验中, 观察到的数据可能不全. 比如在检测一个治疗癌症药物的五年生存率时, 需要患者定期到医院取药, 但某些患者在试验过程中, 由于搬家、出国等原因而失去联系, 这就导致部分数据删失 (censor). 虽然这部分删失数据不全, 但仍包含某些有用信息, 不能丢弃不用. 对于上述临床试验方案的安排、数据分析, 以及删失数据的统计分析, 就是生物统计的研究内容.

例 1.1.3 (试验设计 (design of experiment)) 在生产日光灯之前, 假设知道影响其亮度的因素包括: 灯管的长度、灯管直径、管内气体、灯丝种类、灯管材质、镇流器、灯管涂层 7 个因素 (假设电压稳定). 如果每个因素均有 2 个取值, 则共有 $2^7 = 128$ 种组合, 如何选取一种组合生产, 以使日光灯的亮度最大?

在许多工业产品的设计阶段都会面临上述问题, 在许多科学实验的方案设计阶段, 也有类似的问题需要解决. 对于上述问题, 我们当然可以把 128 种组合都做一次, 之后选一个亮度最大的组合以指导生产. 是否有更科学的方法能让实验者减少实验次数, 而得到全局最优的组合? 这就是试验设计的研究范畴.

试验设计的产生与发展与 R. A. Fisher (费希尔, 1890—1962) 密不可分. 他 1909 年

入剑桥大学攻读数学和物理, 其间受 K. Pearson (皮尔逊, 1857—1936) 的影响, 主要兴趣集中在生物学和统计学上. 1914 年第一次世界大战爆发后, Fisher 也打算投笔从戎, 但因视力不好未果. 此后 5 年, 他曾做过中学教师, 也曾在短时间内经营过一个小型农场. 1919 年由 Darwin(达尔文) 一位亲戚介绍进入洛桑实验站 (Rothamsted Experimental Station) 工作, 且一直工作到 1933 年, 才到高校任职. 在此实验站工作期间, 他对此实验站积累了近百年的农作物产量、施肥情况、田间管理, 以及气象等数据进行了深入研究与分析, 提出了许多著名的统计概念与方法, 如充分统计量、自由度、方差分析等, 并开创了试验设计这一研究方向.

在统计史中, 36 名军官排队问题是一个著名的试验设计问题: 18 世纪欧洲一位皇帝想在阅兵典礼上, 由来自 6 个军种各 6 个军衔的 36 名军官排成一个 6×6 方阵, 要求此方阵的每行每列每个军种每个军衔各有一名军官. 多人冥思苦想, 终不得解, 于是求助于著名数学家 Leonhard Euler (欧拉, 1707—1783). 欧拉也无解, 但提出一个猜想: 当 N 是 4 的倍数加 2 时, 即 $N = 4k + 2$ 时, 上述 $N \times N$ 方阵不存在. 直至 20 世纪 50 年代才被证明: 上述猜想仅当 $N = 2, 6$ 时成立. 表 1.1.2则为一个 5×5 方阵, 这即试验设计中的正交拉丁方 (Latin square) 或希腊拉丁方 (Graeco-Latin square).

表 1.1.2　5 阶正交拉丁方

4B	2C	5D	3E	1A
3C	1D	4E	2A	5B
2D	5E	3A	1B	4C
1E	4A	2B	5C	3D
5A	3B	1C	4D	2E

《周易》是一部中国哲学古籍, 八卦在其中占有极重要的地位. 如果在八卦图 1.1.1 中, 以 1 和 0 分别表示其中的阳爻和阴爻, 则八卦就是如下的 8 个分量为 0, 1 的三维向量:

$$(0,0,0),(0,0,1),(0,1,0),(0,1,1),(1,0,0),(1,0,1),(1,1,0),(1,1,1).$$

如果我们把上述 8 个分量为 0,1 的三维向量看成 3 个分量为 0,1 的 8 维向量, 且对其中任何两个作模 2 加法形成另外 3 个, 再把这 3 个作模 2 加法形成第 7 个, 就构成如表 1.1.3 所示的表格, 这就是试验设计中的 $L_8(2^7)$ 正交表 (orthogonal array), 也是部分析因设计 (fractional factorial design) 中的 2^{7-4}_{III} 设计.

针对上述日光灯生产问题, 如果不考虑 7 个因素间的相互作用, 则我们可以按照表 1.1.3 所示的行组合进行生产. 虽然此时仅有 8 次试验, 但可以利用统计知识, 找到其最亮组合. 至于如何求取及数据分析, 请参看试验设计方面的内容.

图 1.1.1 《周易》中的八卦图

表 1.1.3 $L_8(2^7)$ 正交表

行号	1	2	3	4	5	6	7
1	0	0	0	0	0	0	0
2	0	0	0	1	1	1	1
3	0	1	1	0	0	1	1
4	0	1	1	1	1	0	0
5	1	0	1	0	1	0	1
6	1	0	1	1	0	1	0
7	1	1	0	0	1	1	0
8	1	1	0	1	0	0	1

采集数据不是统计的目的, 而如何让数据产生价值才是统计的核心. 故在统计学中, 估计与检验始终是统计学的两大任务.

例 1.1.4 (估计 (estimation)) 如何估计一个人的体重?

针对此问题, 一个最简单的方法就是找一个可以测量体重的秤, 称量一下即可. 但一次称量无法确定此估计的误差, 如果多称几次, 我们可以用平均值、中位数等估计其体重, 也可以得到估计的误差, 但我们应采用哪个估计呢? 或者说哪个估计比较好? 有没有最优的估计? 这就是统计估计部分的研究内容.

类似上述的问题还有许多. 估计是统计中应用最早、最广的内容之一. 比如在《管子·问》中: "问死事之孤其未有田宅者有乎? 问少壮而未胜甲兵者几何人? ……问一民有几年之食也?" 这些问题的解决都需要估计与抽样调查. 再比如如何估算一个池塘中鱼的总量、野生东北虎的数量等问题也都是抽样调查与估计的综合.

统计估计方法在第二次世界大战中也得到了很好的应用. 比如, 二战时, 盟军想估计德军某种型号坦克的产量, 但情报很难收集, 于是转而求助统计学家. 现收集到的数据仅有在战场上被缴获或摧毁的德军坦克数量 N, 以及这 N 台坦克的最大编号 M, 于

是统计研究人员基于矩估计, 得到了德军坦克产量的估计值为

$$M\left(1+\frac{1}{N}\right).$$

据此公式, 他们得到的估计值见表 1.1.4, 与战后得知的实际产量还是非常接近的. 我们将在第二章讲述什么是矩估计.

表 1.1.4　德军坦克产量

时间	统计估计	情报估计	实际产量
1940 年 6 月	169	1 000	122
1941 年 6 月	244	1 550	271
1942 年 8 月	327	1 550	342

例 1.1.5(假设检验 (hypothesis testing))　在许多体育比赛中, 裁判多通过掷硬币的方法让一方先进行选择. 大家之所以接受这种方法是由于相信硬币是均匀的, 即掷一次硬币出现正面与反面的概率是相同的. 为验证某硬币的均匀性, 将这枚硬币投掷 495 次后, 正反面分别出现 220 次和 275 次, 请问: 这枚硬币均匀吗?

类似的检验问题还有许多. 我们在高中生物学课程中都学过, 豌豆杂交二代四个性状的比例为 9:3:3:1, 此结果由遗传学奠基人、来自奥地利的 Gregor Johann Mendel (孟德尔, 1822—1884) 于 1866 年发表的一篇论文中提出. Mendel 自幼就爱好园艺, 中学毕业后曾考入奥尔谬茨大学哲学院学习, 但因家境贫寒, 被迫中途辍学, 进入奥地利布隆城的一所修道院当修道士. 从 1851 年到 1853 年, Mendel 在维也纳大学学习了 4 个学期, 系统学习了植物学、动物学、物理学和化学等课程. 1854 年 Mendel 回到家乡, 继续在修道院任职, 并利用业余时间开始了长达 12 年的植物杂交实验. 经过整整 8 年 (1856—1864) 的不懈努力, 于 1866 年提出了上述豌豆遗传规律, 并提出了基因的论点. 但他的遗传学理论直至 1900 年被欧洲三位植物学家分别证实后才被大家所认可. 在此期间, 现代统计学奠基人之一的 K. Pearson 于 1900 年提出的 χ^2 拟合优度检验说明豌豆杂交实验数据符合上述遗传规律.

关于假设检验的系统理论基础则是由 R. A. Fisher 于 20 世纪 20 年代基于女士品茶实验提出的. 20 世纪 20 年代后期, 在英国剑桥的一个夏日的午后, 一群大学的绅士和他们的夫人们, 还有来访者, 正围坐在户外的桌旁享用着下午茶. 在品茶过程中, 一位女士坚称: 把茶加进奶里, 或把奶加进茶里, 不同的做法, 会使茶的味道品起来不同. 在场绝大多数对这位女士的“胡言乱语”嗤之以鼻. 然而, 在座的一个身材矮小、戴着厚眼镜、下巴上蓄着的短尖髯开始变灰的先生——R. A. Fisher, 却不这么看, 他对这个问题很有兴趣, 并据此提出了统计学的一大研究内容——显著性假设检验的框架.

对于硬币均匀性检验问题, 如果假设硬币均匀, 则连续投掷的结果应该是正反面出

现次数相差不多. 但如何确定此枚硬币的 495 次投掷中, 正反面相差 55 次是大还是小呢? 也就是说, 我们能否找到一个正反面出现次数相差大与小的显著性界限呢?

关于硬币均匀性问题, 以及女士是否有品茶能力问题, 将在本书第三章之假设检验内容给出解决方案, 但关于检验豌豆杂交数据的 χ^2 拟合优度检验则将在第六章讲述.

从遗传学角度看, 遗传会把一种性状 (如身高) 的优势传递给下一代. 如果真是这样的话, 将会看到一代代人中, 个子很高和很矮的人的比例会日渐升高, 而中间部分的比例会日渐下降, 但实际上, 一代代人的身高却稳定地服从正态分布, 请问如何解释这种情况?

例 1.1.6 (回归 (regression)) 身高遗传吗?

为解决此问题, F. Galton (高尔顿, 1822—1911) 花费十几年时间收集了 1 078 个家庭中父母及成年子女的身高数据. 为平衡性别的差异, 他把女性身高乘上 1.08 倍. 之后通过研究父代身高 (父母身高平均值) 与子代身高 (成年子女身高的平均值) 的关系, 他发现

- 父代身高高的子代身高也倾向于高;
- 在父代身高高的家庭, 其子代身高超过父代身高的比例小; 在父代身高矮的家庭, 其子代身高低于父代身高的比例也小.

也就是说, 在上述亲子身高问题中, Galton 发现了亲子代间性状遗传中, 性状有向中心回归的现象, 即高个子的后代平均来说也高些, 但不如其父代那么高, 要向平均身高的方向“回归”一些. 另外, 他研究发现二者间的关系为 (单位: cm)

$$\text{成年儿子身高} = 85.67 + 0.516 \times \text{父亲身高} \pm 9.51. \tag{1.1.1}$$

上述例子就产生了统计学中一个研究方向——回归分析, 上述关系的直线就称为回归直线.

另外, 湖北体育科学研究所根据他们所观测的数据, 得到中国成年孩子与父母身高的关系为 (单位: cm):

$$\text{成年儿子身高} = 56.699 + 0.419 \times \text{父亲身高} + 0.265 \times \text{母亲身高} \pm 3,$$

$$\text{成年女儿身高} = 40.089 + 0.306 \times \text{父亲身高} + 0.431 \times \text{母亲身高} \pm 3.$$

关于回归分析, 本书在第五章将加以简单讲述, 详细的内容可参见相关参考文献.

例 1.1.7 (时间序列 (time series)) 在利用股票进行投资理财时, 会关注大盘及所关注股票每天的收盘价. 如以 $\{X_t\}_{t=1}^n$ 记过去 n 天的收盘价, 则投资者关心的问题就是如何利用这些数据预测未来的收盘价.

如何利用依时间或位置观测到的数据 $\{X_t\}_{t=1}^n$ 来预测 $X_{n+1}, X_{n+2}, \cdots$, 就是统计学中时间序列这一方向的研究内容. 关于时间序列数据模型, 最常用的三类平稳模型有:

(1) 自回归 (autoregressive, 简记为 AR) 模型:

$$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_p X_{t-p} = \varepsilon_t;$$

(2) 移动平均 (moving average, 简记为 MA) 模型:

$$X_t = \varepsilon_t - \beta_1 \varepsilon_{t-1} - \cdots - \beta_q \varepsilon_{t-q};$$

(3) 自回归移动平均 (autoregressive and moving average, 简记为 ARMA) 模型:

$$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_p X_{t-p} = \varepsilon_t - \beta_1 \varepsilon_{t-1} - \cdots - \beta_q \varepsilon_{t-q},$$

上述模型中 ε_t 是白噪声, α_i, β_j 是未知的常量, p, q 为模型阶数.

自回归模型是由英国统计学家 G. U. Yule(尤尔, 1871—1951) 于 1927 年提出的, 后两种是由英国统计学家 G. T. Walker(沃克, 1868—1958) 于 1931 年提出. 如何利用这三类模型进行预测等, 请参见时间序列方向的相关参考文献.

例 1.1.8 (多元统计分析 (multivariate statistical analysis))　由于衡量一个人健康与否的指标有许多, 故体检时需要检查多个指标. 如何利用这些指标衡量一个地区或一个国家人民群众或一个人的整体健康水平?

对于此问题, 假设共收集到 n 个人的体检数据, 也假设这 n 个人的体检指标都有 p 个, 如以 X_{ij} 记第 i 个人的第 j 项检验结果, 则此时收集到的数据即为

$$\begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1p} \end{pmatrix}, \begin{pmatrix} X_{21} \\ X_{22} \\ \vdots \\ X_{2p} \end{pmatrix}, \cdots, \begin{pmatrix} X_{n1} \\ X_{n2} \\ \vdots \\ X_{np} \end{pmatrix}. \tag{1.1.2}$$

要根据这些数据回答整体健康水平, 就要利用统计学一个分支——多元统计分析方法. 另外, 近 20 年在统计学文献中经常提到的高维数据统计推断, 就是指当 $p > n$ 或 p 是 n 的函数时的研究内容. 本书第四章仅考虑某些简单多元情况下的统计方法, 但不涉及高维情形, 详细内容请参见多元统计分析的相关文献.

上面仅通过几个例子, 简单提出了统计学中的几个研究方向, 但统计学的研究方向并不仅这些, 还包括非参数统计、生物信息、统计机器学习、统计计算、因果推断、函数型数据分析等, 也包括如统计质量过程、教育统计、心理统计、体育统计、空间统计等许多研究分支.

1.1.2　什么是统计

关于中西方统计一词的含义, 陈希孺在《中国大百科全书·数学》中指出: “按《不列颠百科全书》上的说法, 统计学 (Statistics) 是‘收集和分析数据的科学与艺术’, 而没

有标出 Mathematical Statistics 一词, 这是因为在‘Statistics’一词的使用上, 我们与西方不同. 我们所说的数理统计 (Mathematical Statistics) 即为西方所说的‘Statistics’, 其原因是与在我国被视为社会科学的经济统计学加以区别. 在西方, 也有 Mathematical Statistics 的提法, 但那是特指统计方法的理论基础那一部分.” 基于此, 我们罗列三个关于统计学的定义, 但都是陈先生所指的数理统计, 而不是国内的经济统计. 本书所涉及的统计也均指数理统计.

《不列颠百科全书》给出的定义是:

定义 1.1.1 (不列颠百科) 统计学是收集和分析数据的科学与艺术.

陈希孺给出的定义是:

定义 1.1.2 (陈希孺) 统计是数学的一个分支, 它是一门用有效的方法收集和分析带有随机影响的数据的学科, 且其目的是解决特定的问题.

华东师范大学茆诗松在华东师范大学出版社出版的“数理统计丛书”总序中给出的定义是:

定义 1.1.3 (茆诗松) 统计是一门应用性很强的学科, 它是研究如何有效地收集、整理和分析受随机影响的数据, 并对所考虑的问题作出推断或预测, 直至为采取决策和行动提供依据和建议的一门学科.

三个定义大同小异, 只是《不列颠百科全书》强调统计也是艺术, 陈希孺强调统计是数学的一个分支, 茆诗松强调统计是应用性很强的学科. 不论如何, 他们都认为统计是研究数据的科学. 那什么是数据? 数据是大自然与人类活动的记录, 是信息的载体. 统计学所研究的数据是以编码形式存在的信息载体, 因此, 无论是阿拉伯数字形式的数据, 还是文本、图像, 甚至是音频、视频等形式的数据都是统计学的研究对象.

在统计学的上述定义中, 有如下几个共性:

(1) 必须是受到随机影响的数据, 才能成为数理统计学的研究对象.

数理统计与其他学科的一个很重要区别在于“随机性”. 随机性的主要来源是试验误差 (不是系统误差), 其次是由于影响研究问题结果的因素非常多, 故我们只能随机地抽取部分来进行研究所造成的.

(2) 如何“有效”地收集数据.

“有效” 有两个方面的含义: 一是可以建立一个模型来描述所得数据; 二是数据中要尽可能多地包含与研究问题有关的信息. 例如想调查某地区共 1 万农户的经济状况, 由于条件的限制, 我们不可能逐户去调查, 现决定从中随机地抽取 100 户作实际调查, 那问题是:

- 100 户是否合适? (如何权衡精度与费用)
- 如果调查 100 户合适, 则这 100 户如何去选? (如何抽样以得到更有代表性的数据)
- 针对这 100 户都调查什么指标? (如何选取合适的指标, 以反映农户的经济状况)

(3) 如何“有效”地利用数据.

获取数据的目的在于提供所研究问题的相关信息, 这种信息有时并不是一目了然的, 而需要用“有效”的方法去提取或提炼, 之后再对所研究的问题作出一个结论, 这种“结论”在统计上被称为推断 (inference). 为有效地使用数据进行统计推断, 就要涉及统计中的一些准则, 以评价推断的优劣.

另外, 统计学还有频率学派与 Bayes 学派之别. 所谓频率学派是指基于频率估计概率的统计理论与方法, 此时没有先验信息; 而 Bayes 学派则指基于先验信息的统计理论与方法, 其理论基础为英国数学家 Thomas Bayes (贝叶斯, 1702—1761) 发展的 Bayes 定理. 本书所讲的统计理论与方法, 除 2.4 节与 2.5 节之外, 均是从频率学派的角度进行阐述的.

经典统计学的研究内容主要包括以下几个部分:

(1) 抽样分布 (sampling distribution).

(2) 估计 (estimation).

(3) 假设检验 (hypothesis test).

(4) 随机模拟 (simulation).

上述内容, 我们将在后面分别讲述, 但本书的重点在于估计与假设检验. 而在衡量估计与检验的优良性质时所用到的大样本理论, 将在附录中加以介绍. 另外, 在目前智能时代, 统计学习是统计学现在的研究热点, 也可以把它列为统计学的第五大研究内容或重要的研究分支.

1.2　样本、参数、统计量及抽样分布

本节将介绍一些后面经常用到的基本概念, 如样本、总体、参数、统计量、抽样分布等.

1.2.1　样本、总体与参数

从统计的定义可以看出, 统计学是研究数据的科学, 数据是统计学的研究对象, 而数据就被称为**样本** (sample).

样本或数据来自哪里? 或者说, 样本是怎么产生的? 在实际生活或生产活动中, 样本的来源大概有两个渠道: 一是通过实验, 如通过化学或物理实验, 抽样调查或试验设计等方法收集而得到; 二是自动产生, 如网络日志、商品买卖、实时监控等. 在进行统计研究时, 也经常采用随机模拟以比较方法间的优劣, 此时的数据则是人为随机产生的.

收集样本的目的是什么呢? 比如在评估某批产品的次品率时, 抽检是常用方法. 检测的目的是用来估计这整批产品的次品率. 为此, 称这整批产品为总体, 而从中抽出来的产品就是样本. 再比如买橘子时, 我们经常会随机拿一个橘子, 尝尝其甜不甜. 此时关心的有两个问题: 所选出橘子的代表性以及其甜度的代表性. 在关心橘子的代表性时, 我们称这批橘子为总体, 而单个橘子称为个体, 所尝橘子称为样本; 在关心橘子的甜度时, 我们称所有橘子的甜度为总体, 而单个橘子的甜度是个体, 所尝橘子的甜度称为样本.

在统计学中, 称根据一定目的确定的所要研究对象的全体为**总体** (population), 以 X 记. 而从总体 X 中选出的个体为样本, 记之为 $X_1, X_2, \cdots, X_n \sim X$, 其中 n 称为**样本容量** (sample size), X_i 称为第 i 个样本, 并称从总体中抽取样本的工作为抽样 (sampling).

样本有一维、多维之分, 有连续、离散之别, 也可能是文本、图像, 或音频及视频数据. 如果仅从应用角度考虑, 样本就是一组已知的数或给定的数字化信息, 比如身高及体重例子中的 n 次测量值. 但是, 也要注意到, 样本是一组受到随机因素影响的数, 因此它们也是随机的. 这就是样本的二重性.

总体中个体的数量有时是确定的, 有时较难确定, 但并不影响问题的确定与解决. 比如上述橘子的代表性问题, 其总体中个体的数量就是确定的, 而在甜度代表性问题中, 我们没有必要知道一共有多少个橘子, 而此时的总体则被认为是甜度的所有可能值. 再比如, 在测量身高及体重的例子中, 我们称所有可能的测量结果就是总体, 而每次测量结果就是一个样本.

由于每次抽样产生的样本都不同, 故我们称所有样本构成的集合为**样本空间** (sample space), 记为 $\mathcal{X}$, 即

$$\mathcal{X} = \{(X_1, X_2, \cdots, X_n) : X_i \sim X, i = 1, 2, \cdots, n\}.$$

如果假设在抽样过程中, 每次均独立进行, 则称 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的简单随机样本, 或**独立同分布** (independent identical distribution, 简记为 iid) 的样本, 记为 $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} X$. 此时, 我们也把样本看成是总体的一组实现或观测值.

从总体中抽取样本的目的在于对总体的某些特征进行统计推断, 我们称这些特征为**参数** (parameter). 基于样本的随机性, 我们也把总体 X 看成一个随机变量, 而参数就是总体分布中的未知常量, 常以 θ 记.

为了以后叙述方便, 以及避免严密的数学语言, 我们给出本书所说的概率分布之含义如下: 如果分布是离散的, 则概率分布表示累积概率分布函数 $F(x;\theta)$; 如果分布是连续的, 则概率分布表示概率密度函数 $f(x;\theta)$.

虽然参数是未知的, 但在许多实际问题中, 我们对参数都有一定的认知与了解. 如在身高测量例子中, 我们知道人的身高基本都在一个合理的范围内. 于是, 称参数可能取值的范围为**参数空间** (parameter space), 并记之为 Θ.

于是, 当样本 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的独立同分布样本时, 记之为

$$X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} X \sim F(x;\theta) \text{ 或 } f(x;\theta),\ \theta \in \Theta.$$

如果总体概率分布可以用有限个参数来刻画, 则用 n 个随机样本对参数 θ 进行的统计推断, 就称为参数统计. 当总体概率分布不能用有限个参数刻画时的统计推断, 则称之为**非参数统计** (non-parametric statistics).

由于样本具有随机性, 而在概率论中, 概率分布被认为是刻画随机性的一个很好度量, 于是我们就用样本联合分布 (简称为样本分布) 来刻画样本的随机性. 如果样本 X_1, $X_2, \cdots, X_n$ 为来自总体 $X \sim F(x;\theta)$ 的独立同分布样本, 则样本分布为

$$F(x_1, x_2, \cdots, x_n;\theta) = P(X_1 \leqslant x_1, X_2 \leqslant x_2, \cdots, X_n \leqslant x_n) = \prod_{i=1}^{n} F(x_i;\theta).$$

如果总体是连续的, 且概率密度函数为 $f(x;\theta)$, 则样本分布就指如下的联合概率密度函数:

$$f(x_1, x_2, \cdots, x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

例 1.2.1　某企业为检测某批产品的次品率, 现从这批共 N 件的产品中不返回随机抽出 n 件, 求样本分布.

解　如以 M 记这批产品中的次品个数, 它是我们感兴趣的未知常量. 再记

$$X_i = \begin{cases} 1, & \text{如果第 } i \text{ 次抽到的样本为次品,} \\ 0, & \text{否则.} \end{cases}$$

由于 n 个样本 $X_1, X_2, \cdots, X_n$ 是不返回随机抽取的, 且 $X_i \sim B(1, M/N)$, 则样本分布为

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\ &= \frac{M(M-1)\cdots(M-t+1)}{N(N-1)\cdots(N-t+1)} \frac{(N-M)\cdots(N-M-n+t+1)}{(N-t)\cdots(N-n+1)}, \end{aligned} \tag{1.2.1}$$

其中 $t = \sum\limits_{i=1}^{n} x_i$, M 即为参数. □

我们回忆一下概率论所学的超几何分布 (hypergeometric distribution), 知 $\sum\limits_{i=1}^{n} X_i$ 的分布为超几何分布, 即

$$P\left(\sum_{i=1}^{n} X_i = t\right) = \frac{\begin{pmatrix} M \\ t \end{pmatrix} \begin{pmatrix} N-M \\ n-t \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}. \tag{1.2.2}$$

如果在例 1.2.1中采用的抽样方案为有返回的, 则由于样本是独立的, 故样本分布为

$$P\left(X_1=x_1,X_2=x_2,\cdots,X_n=x_n\right)=\left(\frac{M}{N}\right)^{\sum\limits_{i=1}^{n}x_i}\left(1-\frac{M}{N}\right)^{n-\sum\limits_{i=1}^{n}x_i},\tag{1.2.3}$$

而此时样本中次品数 $\sum\limits_{i=1}^{n}X_i$ 的分布为二项分布:

$$P\left(\sum_{i=1}^{n}X_i=t\right)=\binom{N}{n}\left(\frac{M}{N}\right)^{t}\left(1-\frac{M}{N}\right)^{n-t}.\tag{1.2.4}$$

例 1.2.2 现对某人身高独立测量 n 次, 且测量是在“相同条件下”进行的, 记得到的样本为 $X_1,X_2,\cdots,X_n$, 求样本分布.

解 此时, 我们把总体理解为所有观测可能值的集合. 由于每次测量都围绕身高真值波动, 且影响其测量结果的因素可归结为随机噪声, 故我们可以假设总体服从正态分布 $N(\mu,\sigma^2)$, 其中 (μ,σ^2) 为参数, 而 μ 则反映此人身高, σ^2 反映测量误差. 由于 n 次测量是独立的, 故样本分布为

$$f(x_1,x_2,\cdots,x_n;\mu,\sigma^2)=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right\}.$$

此时参数空间可取为

$$\Theta=\{(\mu,\sigma^2):\mu>0,\sigma^2>0\}.$$

根据经验, 成年人的身高基本在 140∼200 cm 之间, 故我们也可以取

$$\Theta=\{(\mu,\sigma^2):\mu\in[140,200],\sigma^2>0\}.$$ □

在统计中, 称那些不感兴趣的参数为**讨厌参数** (或称冗余参数, nuisance parameter). 如在例 1.2.2中, 如果仅对身高 μ 有兴趣, 则 σ^2 就被称为讨厌参数.

1.2.2 统计量

统计学的目的在于利用从总体中抽取的样本对总体某些特征, 即参数或参数的函数进行统计推断. 而利用样本进行统计推断时, 由于样本是一堆杂乱无章的数据, 故我们有必要根据参数特点, 对这些数据进行加工整理后再进行统计推断, 并且在许多实际问题中, 加工整理后的数据更能反映出参数本身的性质. 比如在例 1.2.1中, 我们多利用汇总后的次品数 $\sum\limits_{i=1}^{n}X_i$ 来对次品率作统计推断.

在统计中, 这些由样本加工整理后得到的量, 就被称为**统计量** (statistic). 用数学语言来讲, 统计量就是样本的可测函数, 通常记为 $T(X_1, X_2, \cdots, X_n)$, 在不引起混淆时, 简记为 $T(X)$ 或 $T_n(X)$. 在某些情况下, 为避免样本与总体混淆, 也记样本为 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$, 样本值为 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$. 此时统计量也简记为 $T(\boldsymbol{X})$ 或 $T_n(\boldsymbol{X})$.

从上述定义可以看出, 统计量仅是样本的函数, 与参数无关, 即给定样本值后, 统计量就是一个已知的数. 下面给出几个常用的统计量.

样本均值与样本方差 (sample mean and sample variance): 对于 n 个样本 X_1, $X_2, \cdots, X_n$, 分别称

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{1.2.5}$$

为样本均值与样本方差.

从上述定义可以看出, 样本均值与方差分别度量样本平均值与样本的分散程度或波动性, 也反映总体期望 $E(X) = \mu$ 与方差 $\mathrm{Var}(X) = \sigma^2$. 另外, 注意到在上述样本方差的定义中, 分母没有用 n 而用 $n-1$, 其原因有二:

(1) 如定义 $Y_i = X_i - \bar{X}$, 则 $(n-1)S_n^2 = \sum\limits_{i=1}^{n} Y_i^2$. 虽然 S_n^2 是 n 个数的平方和, 但由于 $\sum\limits_{i=1}^{n} Y_i = 0$, 即它有一个约束, 故在 n 个加和中, 仅有 $n-1$ 个可自由变化.

(2) 由于 S_n^2 是一个二次型, 故可以把它改写成 $(n-1)S_n^2 = \boldsymbol{X}^{\mathrm{T}} A \boldsymbol{X}$, 其中 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^{\mathrm{T}}$, $A = I_n - \dfrac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}$, 其中 I_n 为 n 阶单位矩阵, $\mathbf{1}$ 表示分量均为 1 的向量. 而矩阵 A 的秩为 $n-1$.

基于以上考虑, $n-1$ 被称为样本方差 S_n^2 或 $(n-1)S_n^2$ 的**自由度** (degree of freedom). 自由度是统计学中一个非常重要的基本概念, 它在后续的抽样分布、估计和假设检验中都发挥着重要的作用.

由于在后面某些场合, 我们也会用到分母为 n 的情形, 故称统计量

$$S_{n*}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{1.2.6}$$

为修正的样本方差.

样本矩 (sample moment): 对于样本 $X_1, X_2, \cdots, X_n$, 以及给定的正整数 k, 分别称

$$a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad m_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$$

为样本 k 阶原点矩与中心矩.

样本变异系数 (coefficient of variability): 对于样本 $X_1, X_2, \cdots, X_n$, 称 $S_n/\bar{X}$ 为样本变异系数.

样本变异系数是消除量纲后的样本波动性的度量, 它反映着总体变异系数 σ/μ.

次序统计量 (或称顺序统计量, order statistic): 对于样本 $X_1, X_2, \cdots, X_n$, 把它由小到大排列成 $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$, 则称 $(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$ 为次序统计量, $X_{(i)}$ 称为第 i 个次序统计量.

通过次序统计量, 我们有如下一些很实用的统计量, 如

(1) 样本 p 分位数 (quantile): 对于给定的 $p \in (0,1)$, 称

$$m_{n,p} = X_{([np])} + (n+1)\left(p - \frac{[np]}{n+1}\right)(X_{([np]+1)} - X_{([np])}) \tag{1.2.7}$$

为此样本的 p 分位数, 其中 $[x]$ 表示不超过 x 的最大整数. 特别地, 样本中位数 (median) 定义为

$$X_{\text{med}} = \begin{cases} X_{((n+1)/2)}, & n \text{ 是奇数}, \\ (X_{(n/2)} + X_{(n/2+1)})/2, & n \text{ 是偶数}. \end{cases}$$

(2) 极值 (extreme value): 称 $X_{(1)}$ 与 $X_{(n)}$ 为极小值与极大值统计量.

(3) 极差 (range): $R = X_{(n)} - X_{(1)}$.

注 1.2.1 上述定义的第 i 个次序统计量的随机性可如下理解: 对于给定的一组样本值 $x_1, x_2, \cdots, x_n$, $X_{(i)}$ 取值 $x_{(i)}$.

注 1.2.2 关于样本 p 分位数的定义还有许多种, 如 $X_{([np])}$, $X_{([np]+1)}$, 和 $\inf\{x : F_n(x) \geqslant p\}$ 等, 其中 $F_n(x)$ 为下面给出的经验分布函数. 这些定义虽然有所不同, 但它们均是一个 (或两个) 次序统计量, 且随着样本容量的加大, 它们之间的差别并不大.

样本相关系数 (sample coefficient of correlation)): 设 $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, \cdots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ 为一个二维样本, 称

$$r(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{1.2.8}$$

为 X, Y 间的样本相关系数.

上述样本相关系数 $r(X,Y)$ 是两组观测: $X_1, X_2, \cdots, X_n$ 与 $Y_1, Y_2, \cdots, Y_n$ 间的相关性的度量. 如以 $\begin{pmatrix} X \\ Y \end{pmatrix}$ 记上述样本的总体, 则样本相关系数也反映着 X, Y 间的 Pearson 矩相关系数

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

的大小.

经验分布函数 (empirical distribution function): 设 $X_1, X_2, \cdots, X_n$ 为取自总体分布函数是 $F(x)$ 的样本, $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$ 为其次序统计量, 称

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i) = \begin{cases} 0, & x < X_{(1)}, \\ \dfrac{k}{n}, & X_{(k)} \leqslant x < X_{(k+1)}, k = 1, 2, \cdots, n-1, \\ 1, & x \geqslant X_{(n)} \end{cases} \tag{1.2.9}$$

为样本 $X_1, X_2, \cdots, X_n$ 的经验分布函数.

经验分布函数与总体分布间的关系很密切. 在 6.2.1 小节将证明: 如果 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的独立同分布样本, 则对于任意给定的 $x \in \mathbb{R}$, 由大数定律及中心极限定理有

$$nF_n(x) \sim B(n, F(x)),$$

$$F_n(x) \xrightarrow{\text{P}} F(x),$$

$$\frac{\sqrt{n}[F_n(x) - F(x)]}{\sqrt{F(x)(1-F(x))}} \xrightarrow{\text{d}} N(0,1).$$

从上述两个极限性质可以看到, 经验分布函数是总体分布的一个很好的估计. 关于经验分布函数的详细讨论见 6.2.1 小节.

1.2.3　抽样分布

由于样本具有二重性, 故作为样本函数的统计量也具有二重性, 即给定样本值后, 统计量就是一个给定的值; 但在样本给定前, 统计量也是随机变量. 因此, 作为随机变量的统计量之分布被称为**抽样分布** (sampling distribution). 抽样分布概念是由 R. A. Fisher 于 1922 年提出的.

如在例 1.2.1中, 当抽样是不返回的时, 由 (1.2.2) 式给出的超几何分布就是统计量 $T(X) = \sum\limits_{i=1}^{n} X_i$ 的抽样分布; 当抽样有返回时, 由 (1.2.4) 式给出的分布就是此时加和统计量的抽样分布. 在例 1.2.2 中, 如果 n 次测量是独立进行的, 则由概率论知识知道样本均值的抽样分布为

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

抽样分布在统计学中占有非常重要的地位, 那抽样分布有什么用呢? 比如在例 1.2.2 中, 由于此时我们感兴趣的是估计参数 μ, 且一个显然的估计就是样本均值 $\bar{X}$. 那一个自然的问题就是: 这个估计准确吗? 或者说, 二者之差在一个允许范围内的概率有多大?

此时, 我们根据样本均值的抽样分布可知,

$$P(|\bar{X}-\mu|\leqslant c)=P\left(\frac{\sqrt{n}|\bar{X}-\mu|}{\sigma}\leqslant\frac{\sqrt{n}c}{\sigma}\right)=1-2\Phi\left(-\frac{\sqrt{n}c}{\sigma}\right),$$

其中 $\Phi(x)$ 为标准正态分布的累积分布函数, 由此则回答了上述问题.

另外, 对于上述身高例子, 如果我们的目的在于判断此人身高是否为某个给定的值 μ_0, 即要检验命题

$$H:\mu=\mu_0$$

是否可接受, 则要从统计角度回答此问题, 仍然要利用上述样本均值的抽样分布, 即可利用样本均值 $\bar{X}$ 离 μ_0 的远近来判断上述命题, 请见第三章的假设检验.

例 1.2.3 设 $X_1,X_2,\cdots,X_n$ 为来自总体 $X\sim F(x)$ 的独立同分布样本, 求第 k 个次序统计量 $X_{(k)}$ 的抽样分布.

解 实际上, 我们有

$$\begin{aligned}G_k(x)&=P(X_{(k)}\leqslant x)=P(\text{在}X_1,X_2,\cdots,X_n\text{中至少有}k\text{个不大于}x)\\&=\sum_{m=k}^{n}P(\text{在}X_1,X_2,\cdots,X_n\text{中恰有}m\text{个不大于}x)\\&=\sum_{m=k}^{n}\left(\begin{array}{c}n\\m\end{array}\right)F^m(x)[1-F(x)]^{n-m},\end{aligned}$$

即

$$G_k(x)=\sum_{m=k}^{n}\left(\begin{array}{c}n\\m\end{array}\right)F^m(x)[1-F(x)]^{n-m}.$$

另外, 我们还可以把上式改写成如下积分形式 (留作习题):

$$G_k(x)=k\left(\begin{array}{c}n\\k\end{array}\right)\int_0^{F(x)}t^{k-1}(1-t)^{n-k}\mathrm{d}t. \tag{1.2.10}$$

□

如总体分布 $F(x)$ 有概率密度函数 $f(x)$, 则 $X_{(k)}$ 的概率密度函数为

$$g_k(x)=k\left(\begin{array}{c}n\\k\end{array}\right)F^{k-1}(x)[1-F(x)]^{n-k}f(x).$$

特别地, 当 $k=1$ 和 $k=n$ 时, 极小与极大次序统计量的累积分布函数与概率密度函数如下:

$$\begin{aligned}&G_1(x)=1-[1-F(x)]^n,\quad &&g_1(x)=n[1-F(x)]^{n-1}f(x),\\&G_n(x)=F^n(x),\quad &&g_n(x)=nF^{n-1}(x)f(x).\end{aligned}$$

注 1.2.3　关于极小次序统计量的抽样分布, 我们也可以如下求得:

$$\begin{aligned} G_1(x) &= P(X_{(1)} \leqslant x) = 1 - P(X_{(1)} > x) \\ &= 1 - P(X_1, X_2, \cdots, X_n \text{均大于} x) \\ &= 1 - \prod_{i=1}^{n} P(X_i > x) = 1 - [1 - F(x)]^n. \end{aligned}$$

注 1.2.4　利用类似于例 1.2.3 的方法, 可求得任意两个次序统计量的联合概率密度函数如下: 设 $1 \leqslant r < s \leqslant n$, 则 $X_{(r)}$ 与 $X_{(s)}$ 的联合概率密度函数为 $(x \leqslant y)$

$$\begin{aligned} g_{r,s}(x,y) =& \frac{n!}{(r-1)!(s-r-1)!(n-s)!} \cdot \\ & F^{r-1}(x)[F(y) - F(x)]^{s-r-1}[1 - F(y)]^{n-s} f(x) f(y), \end{aligned}$$

其中 $f(x)$ 为总体 X 的概率密度函数.

注 1.2.5　独立同分布样本的 n 个次序统计量的联合概率密度函数为

$$g(x_1, x_2, \cdots, x_n) = n! \prod_{i=1}^{n} f(x_i), \quad x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n, \tag{1.2.11}$$

其中 $f(x)$ 为总体 X 的概率密度函数.

由 (1.2.11) 式给出的联合概率密度, 可求得任意几个次序统计量的联合概率密度.

1.3　一些常用的抽样分布

本节将讲述统计中经常用到的几个抽样分布: χ^2 分布、t 分布、F 分布, 以及指数型分布、Γ 分布等. 另外, 我们也把几个经常遇到的分布, 如次 Gauss(高斯) 分布、β 分布、Laplace 分布、Cauchy(柯西) 分布、Pareto(帕雷托) 分布、logistic 分布、对数正态分布、Weibull(韦布尔) 分布、双指数分布等的概率密度函数, 以及期望、方差列出, 供大家参考使用.

1.3.1 χ^2 分布

定义 1.3.1 (χ^2 分布) 设 $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} N(0,1)$, 则称随机变量

$$\xi = \sum_{i=1}^{n} X_i^2$$

所服从的分布为自由度 n 的 χ^2 分布, 也称 ξ 为自由度 n 的 χ^2 随机变量, 并记为 $\xi \sim \chi^2(n)$.

定理 1.3.1 设 $X \sim \chi^2(n)$, 则其概率密度为

$$f(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} \mathrm{e}^{-\frac{x}{2}} x^{\frac{n}{2}-1}, & x \geqslant 0, \\ 0, & x < 0. \end{cases} \tag{1.3.1}$$

证明 为求 X 的概率密度, 我们先求其累积分布函数 $F(x) = P(X \leqslant x)$. 显然, 当 $x < 0$ 时, $F(x) = 0$; 当 $x \geqslant 0$ 时,

$$F(x) = P\left(\sum_{i=1}^{n} X_i^2 \leqslant x\right).$$

因为 $X_1, X_2, \cdots, X_n$ 相互独立且同分布于 $N(0,1)$, 所以 $(X_1, X_2, \cdots, X_n)$ 的联合概率密度为

$$\frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right\}.$$

于是,

$$F(x) = \frac{1}{(2\pi)^{n/2}} \int \cdots \iint_{\sum_{i=1}^{n} x_i^2 \leqslant x} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right\} \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_n.$$

为求此积分, 做球面坐标变换

$$\begin{cases} x_1 = \rho \cos\theta_1 \cos\theta_2 \cdots \cos\theta_{n-1}, \\ x_2 = \rho \cos\theta_1 \cos\theta_2 \cdots \sin\theta_{n-1}, \\ \cdots\cdots\cdots\cdots \\ x_n = \rho \sin\theta_1, \end{cases}$$

此变换的 Jacobi(雅可比) 行列式的绝对值为

$$J = \left|\frac{\partial(x_1, x_2, \cdots, x_n)}{\partial(\rho, \theta_1, \theta_2, \cdots, \theta_{n-1})}\right| = \rho^{n-1} D(\theta_1, \theta_2, \cdots, \theta_{n-1}),$$

其中 $D(\theta_1,\theta_2,\cdots,\theta_{n-1})$ 与 ρ 无关. 于是,

$$\begin{aligned}F(x)&=\frac{1}{(2\pi)^{n/2}}\int_0^{\sqrt{x}}\mathrm{d}\rho\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\cdots\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\int_{-\pi}^{\pi}\mathrm{e}^{-\rho^2/2}\rho^{n-1}D(\theta_1,\theta_2,\cdots,\theta_{n-1})\mathrm{d}\theta_1\mathrm{d}\theta_2\cdots\mathrm{d}\theta_{n-1}\\&=C_n\int_0^{\sqrt{x}}\mathrm{e}^{-\rho^2/2}\rho^{n-1}\mathrm{d}\rho,\end{aligned}$$

其中 $C_n=\dfrac{1}{(2\pi)^{n/2}}\displaystyle\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\cdots\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\int_{-\pi}^{\pi}D(\theta_1,\theta_2,\cdots,\theta_{n-1})\mathrm{d}\theta_1\mathrm{d}\theta_2\cdots\mathrm{d}\theta_{n-1}$.

令 $\rho=\sqrt{y}$, 则 $\mathrm{d}\rho=\dfrac{\mathrm{d}y}{2\sqrt{y}}$, 于是,

$$\begin{aligned}F(x)&=C_n\int_0^x\mathrm{e}^{-y/2}y^{(n-1)/2}\frac{1}{2y^{1/2}}\mathrm{d}y\\&=\frac{C_n}{2}\int_0^x\mathrm{e}^{-y/2}y^{n/2-1}\mathrm{d}y.\end{aligned}$$

又因为

$$1=F(\infty)=\frac{C_n}{2}\int_0^\infty\mathrm{e}^{-y/2}y^{n/2-1}\mathrm{d}y,$$

而

$$\int_0^\infty\mathrm{e}^{-y}y^{n/2-1}\mathrm{d}y=\Gamma(n/2),$$

故

$$C_n=\frac{1}{2^{\frac{n}{2}-1}\Gamma\left(\dfrac{n}{2}\right)}.$$

这样, $\chi^2(n)$ 分布的累积分布函数为

$$F(x)=\frac{1}{2^{\frac{n}{2}}\Gamma\left(\dfrac{n}{2}\right)}\int_0^x\mathrm{e}^{-\frac{y}{2}}y^{\frac{n}{2}-1}\mathrm{d}y,$$

其概率密度为

$$f(x)=\frac{1}{2^{\frac{n}{2}}\Gamma\left(\dfrac{n}{2}\right)}\mathrm{e}^{-\frac{x}{2}}x^{\frac{n}{2}-1},\ \forall\ x\geqslant 0.\qquad\square$$

图 1.3.1给出了自由度分别为 5, 10 和 20 的 χ^2 分布的概率密度函数 (记为 pdf). 从中可以看出, 随着自由度增大, 其概率密度函数右移, 对称性越来越好, 且越来越“像”正态分布的概率密度函数.

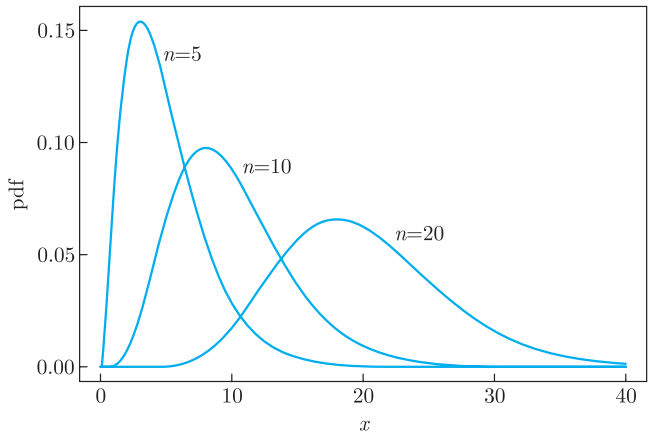

图 1.3.1 χ^2 分布的概率密度函数

注 1.3.1 由于 $\chi^2(n)$ 分布是 n 个独立的标准正态分布的平方和, 故由中心极限定理可知, 把其中心标准化后的极限分布就是正态分布. 此结论也可由 (1.3.2) 式给出的其特征函数证明.

利用定理 1.3.1的结论, 可以求得 $\chi^2(n)$ 分布的特征函数为

$$\psi(t) = (1 - 2\mathrm{i}t)^{-n/2}, \tag{1.3.2}$$

并由此可求得其期望与方差分别为 n 与 $2n$. 另外, 利用上述特征函数还可以证明: 如果 $X_i \sim \chi^2(n_i)$, $i = 1, 2, \cdots, k$, 且 $X_1, X_2, \cdots, X_k$ 相互独立, 则 $\sum\limits_{i=1}^{k} X_i \sim \chi^2\left(\sum\limits_{i=1}^{k} n_i\right)$.

由于抽样分布是某统计量的分布, 自然要问: 哪个统计量的分布是 χ^2 分布? 为此, 我们先给一个引理.

引理 1.3.1 设 $X_1, X_2, \cdots, X_n$ 相互独立, 且 $X_i \sim N(\mu_i, \sigma^2)$, $A = (a_{ij})$ 为 n 阶正交矩阵, 记

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = A \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}.$$

则 $Y_1, Y_2, \cdots, Y_n$ 相互独立, 且 $Y_i \sim N(\beta_i, \sigma^2)$, 其中 $\beta_i = \sum\limits_{j=1}^{n} a_{ij}\mu_j$.

证明 因为 A 为正交矩阵, 所以由 X 到 Y 的变换的 Jacobi 行列式的绝对值为 1, 且

$$\sum_{i=1}^{n}(Y_i - \beta_i)^2 = \sum_{i=1}^{n}(X_i - \mu_i)^2.$$

又因为 $(X_1, X_2, \cdots, X_n)$ 的联合概率密度为

$$(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_i)^2\right\},$$

则知道 $(Y_1, Y_2, \cdots, Y_n)$ 的联合概率密度为

$$(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_i)^2\right\},$$

故结论得证. □

定理 1.3.2　设 $X_1, X_2, \cdots, X_n$ 为来自 $N(\mu, \sigma^2)$ 的独立同分布样本, $\bar{X}, S_n^2$ 为样本均值与样本方差, 则

(1) $\bar{X} \sim N(\mu, \sigma^2/n)$.

(2) $(n-1)S_n^2/\sigma^2 \sim \chi^2(n-1)$.

(3) $\bar{X}$ 与 S_n^2 独立.

证明　设 $A = (a_{ij})$ 为 n 阶正交矩阵, 且其第一行的元素均为 $1/\sqrt{n}$. 作变换

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = A \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}.$$

由引理 1.3.1知, $Y_1, Y_2, \cdots, Y_n$ 相互独立, 且 $Y_i \sim N(\beta_i, \sigma^2)$, 其中 $\beta_i = \mu\sum_{j=1}^{n} a_{ij}$. 由于 A 为正交矩阵, 且第一行为 $(1, 1, \cdots, 1)/\sqrt{n}$, 则

- $\beta_1 = \sqrt{n}\mu$, $\beta_i = 0$ $(i = 2, 3, \cdots, n)$.
- $\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} X_i^2$.
- $Y_1 = \sqrt{n}\bar{X}$.

而

$$(n-1)S_n^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = \sum_{i=1}^{n} Y_i^2 - Y_1^2 = \sum_{i=2}^{n} Y_i^2.$$

综上可知

$$\bar{X} = Y_1/\sqrt{n} \sim N(\mu, \sigma^2/n), (n-1)S_n^2/\sigma^2 \sim \chi^2(n-1).$$

另外, 由于 $\bar{X}$ 仅依赖于 Y_1, S_n^2 仅依赖于 $Y_2, Y_3, \cdots, Y_n$, 而 $Y_1, Y_2, \cdots, Y_n$ 相互独立, 则知 $\bar{X}$ 与 S_n^2 独立. □

注 1.3.2 设 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的独立同分布样本, 如果其样本均值 $\bar{X}$ 与样本方差 S_n^2 独立, 则总体 X 必为正态分布 (见 Stuart et al.(1987)). 实际上, 样本均值与样本方差的协方差为 $\mathrm{Cov}(\bar{X}, S_n^2) = \dfrac{\nu_3}{n}$, 其中 $\nu_k = E[X - E(X)]^k$ 为总体 k 阶中心矩.

从 χ^2 分布的定义不难看出, χ^2 分布是若干个独立的标准正态随机变量的平方和, 而平方和是二次型的一个特殊形式, 那一个正态随机向量的二次型的分布如何呢? 这即是下面的 Cochran(科克伦) 定理.

定理 1.3.3 (Cochran 定理) 设 $X_1, X_2, \cdots, X_n$ 相互独立, 且 $X_i \sim N(\mu_i, \sigma^2)$, $i = 1, 2, \cdots, n$, 令 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^{\mathrm{T}}$. 又设 $A_1, A_2, \cdots, A_m$ 是 m 个 n 阶非负定矩阵, 且 $A_1 + A_2 + \cdots + A_m = I_n$, $\sum\limits_{i=1}^{m} \mathrm{rank}(A_i) = n$ (rank 表示矩阵的秩). 记 $\xi_i = \boldsymbol{X}^{\mathrm{T}} A_i \boldsymbol{X}$, $i = 1, 2, \cdots, m$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)^{\mathrm{T}}$. 则

(1) $\xi_1, \xi_2, \cdots, \xi_m$ 相互独立;

(2) 如果 $\boldsymbol{\mu}^{\mathrm{T}} A_1 \boldsymbol{\mu} = 0$, 则 $\xi_1/\sigma^2 \sim \chi^2(\mathrm{rank}(A_1))$.

证明 记 $n_i = \mathrm{rank}(A_i)$, $i = 1, 2, \cdots, m$. 因为 A_i 是 n 阶非负定矩阵, 则存在一个 $n \times n_i$ 矩阵 B_i, 使得 $A_i = B_i B_i^{\mathrm{T}}$. 记 $B = (B_1 \vdots B_2 \vdots \cdots \vdots B_m)$, 则由已知条件可知 B 是一个 n 阶正交矩阵.

作如下正交变换:

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = B^{\mathrm{T}} \boldsymbol{X} = \begin{pmatrix} B_1^{\mathrm{T}} \boldsymbol{X} \\ B_2^{\mathrm{T}} \boldsymbol{X} \\ \vdots \\ B_m^{\mathrm{T}} \boldsymbol{X} \end{pmatrix},$$

即

$$B_i^{\mathrm{T}} \boldsymbol{X} = \begin{pmatrix} Y_{n_1+\cdots+n_{i-1}+1} \\ Y_{n_1+\cdots+n_{i-1}+2} \\ \vdots \\ Y_{n_1+\cdots+n_{i-1}+n_i} \end{pmatrix}, \ i = 1, 2, \cdots, m,$$

其中 $n_0 = 0$.

由于 B 是一正交矩阵, 故由引理 1.3.1 知道: $Y_1, Y_2, \cdots, Y_n$ 独立, 且 $Y_i \sim N(\beta_i, \sigma^2)$, $i = 1, 2, \cdots, n$, 其中 $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_n)^{\mathrm{T}} = B^{\mathrm{T}} \boldsymbol{\mu}$.

又因为

$$\xi_i = \boldsymbol{X}^{\mathrm{T}} A_i \boldsymbol{X} = \boldsymbol{X}^{\mathrm{T}} B_i B_i^{\mathrm{T}} \boldsymbol{X} = (B_i^{\mathrm{T}} \boldsymbol{X})^{\mathrm{T}} B_i^{\mathrm{T}} \boldsymbol{X} = \sum_{j=n_1+\cdots+n_{i-1}+1}^{n_1+\cdots+n_{i-1}+n_i} Y_j^2, \ i = 1, 2, \cdots, m,$$

所以 $\xi_1,\xi_2,\cdots,\xi_m$ 相互独立, 结论 (1) 得证.

下证结论 (2). 由已知条件知, $\boldsymbol{\mu}^{\mathrm{T}}A_1\boldsymbol{\mu}=0$, 即 $\boldsymbol{\mu}^{\mathrm{T}}B_1B_1^{\mathrm{T}}\boldsymbol{\mu}=0$, 于是, $B_1^{\mathrm{T}}\boldsymbol{\mu}=\mathbf{0}$, 即 $\beta_i=0,\ i=1,2,\cdots,n_1$. 又由于 $\xi_1/\sigma^2=\sum\limits_{i=1}^{n_1}Y_i^2/\sigma^2$, 而 $Y_1,Y_2,\cdots,Y_{n_1}\overset{\text{iid}}{\sim}N(0,\sigma^2)$, 故 $\xi_1/\sigma^2\sim\chi^2(n_1)$. □

当在定义 1.3.1 中的正态分布的均值不全为零时, 我们有如下的非中心 χ^2 分布的定义.

定义 1.3.2 (非中心 χ^2 分布)　设 $X_1,X_2,\cdots,X_n$ 相互独立, 且 $X_i\sim N(\mu_i,1)$, μ_i 不全为零, 则称

$$\xi=\sum_{i=1}^{n}X_i^2$$

为非中心的 χ^2 随机变量, 称它的分布为非中心 χ^2 分布, 自由度为 n, 非中心参数为 $\delta=\sum\limits_{i=1}^{n}\mu_i^2$, 记为 $\xi\sim\chi^2(n,\delta)$.

相较于上述的非中心 χ^2 分布, 定义 1.3.1中的 χ^2 也称为中心 χ^2 分布.

用类似于定理 1.3.1 的方法, 经过一些复杂计算可求得非中心 χ^2 分布的概率密度为

$$\mathrm{e}^{-\delta^2/2}\sum_{m=0}^{\infty}\frac{(\delta^2/2)^m}{m!}\chi^2(x;2m+n),$$

其中 $\chi^2(x;l)$ 为 $\chi^2(l)$ 分布的概率密度. 另外, 非中心 χ^2 分布的特征函数为

$$\psi(t)=(1-2\mathrm{i}t)^{-n/2}\exp\{\mathrm{i}\delta^2t/(1-2\mathrm{i}t)\},$$

其期望与方差分别为 $n+\delta$ 和 $2n+4\delta$.

1.3.2　t 分布

在 20 世纪初以前, 或更具体地说, 在 1908 年以前, 统计学的主要用武之地是社会统计, 尤其是人口统计, 后来加入生物统计. 在这些领域中, 数据一般都是大量的、自然采集的, 所用方法多以中心极限定理为依据. 但到了 20 世纪, 由人工试验所得数据的统计分析问题, 日渐引人注意. 这个方向的先驱就是 W. S. Gosset (戈塞特, 1876—1937) 和 R. A. Fisher.

1899 年, Gosset 作为一名酿酒师进入爱尔兰都柏林一家啤酒厂工作, 那里涉及有关酿造过程中的数据处理问题. 1906—1907 年, 他到 K. Pearson 那里学习统计学, 并着重研究少量数据的统计分析问题. 1908 年, 他在 *Biometrika* 上以笔名 Student 发表了论文《均值的或然误差》(*The probable error of a mean*). 在这篇文章中, 他提出了如下结果: 设 $X_1,X_2,\cdots,X_n$ 是来自 $N(\mu,\sigma^2)$ 的独立同分布样本, μ,σ^2 均未知, 则当 n 比较

小时, $\dfrac{\sqrt{n}(\bar{X}-\mu)}{S_n}$ 并不服从标准正态分布, 而是一个未知分布, 且给出了这个未知分布的一些概率取值. 最早注意到这个小样本均值分布问题的是 R. A. Fisher, 并于 1922 年给出了此未知分布的严格数学定义, 即本小节的 t 分布. 然而, 直到 Gosset 于 1937 年去世, 尚有许多统计学家并不知道他就是 Student. 有兴趣的读者可参见陈希孺 (2000).

定义 1.3.3 (t 分布) 设 $\xi\sim N(0,1)$, $\eta\sim\chi^2(n)$, 且 ξ,η 相互独立, 则称随机变量

$$T=\frac{\xi}{\sqrt{\eta/n}}$$

服从 t 分布, n 为其自由度, 记 $T\sim t(n)$.

由于 t 分布是由 Student 提出的, 故有时也称之为 Student t 分布, 或学生 t 分布, 其概率密度由下述定理给出.

定理 1.3.4 设 $T\sim t(n)$, 其概率密度为

$$f(x)=\frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{n\pi}}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

证明 因为 ξ,η 相互独立, 所以可知 ξ,η 的联合概率密度为

$$C\mathrm{e}^{-x^2/2}\mathrm{e}^{-y/2}y^{n/2-1},$$

其中 C 为不依赖于 x,y 的常数, 但与 n 有关.

为求 $\xi/\sqrt{\eta/n}$ 的概率密度, 作如下变换:

$$\begin{cases}\xi=R\sin\Theta, & 0<R<\infty,\\ \eta=R^2\cos^2\Theta, & -\dfrac{\pi}{2}<\Theta<\dfrac{\pi}{2},\end{cases}$$

利用概率密度变换公式, 求得 R,Θ 的联合概率密度为

$$2C\mathrm{e}^{-r^2/2}r^n(\cos\theta)^{n-1}.$$

利用概率论知识, 由上式可知, R 与 Θ 是相互独立的, 且 Θ 的概率密度为

$$C_1(\cos\theta)^{n-1}, \tag{$*$}$$

其中 $C_1^{-1}=\displaystyle\int_{-\pi/2}^{\pi/2}(\cos\theta)^{n-1}\mathrm{d}\theta=\frac{\Gamma(1/2)\Gamma(n/2)}{\Gamma((n+1)/2)}$.

又由于 $T=\dfrac{\sin\Theta}{\cos\Theta}\sqrt{n}=\sqrt{n}\tan\Theta$, 如令 $t=\sqrt{n}\tan\theta$, 则 $\cos\theta=(1+t^2/n)^{-1/2}$, $\mathrm{d}t=\sqrt{n}\sec^2\theta\mathrm{d}\theta=\sqrt{n}(1+t^2/n)\mathrm{d}\theta$, 于是, 由 $(*)$ 式可得到 t 分布的概率密度为

$$\begin{aligned}f(t)&=C_1\left[(1+t^2/n)^{-1/2}\right]^{n-1}\frac{1}{\sqrt{n}(1+t^2/n)}\\&=\frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{n\pi}}(1+t^2/n)^{-\frac{n+1}{2}}.\end{aligned}$$

□

图 1.3.2 给出了 $n=2,5,50$ 时 t 分布的概率密度函数曲线. 此图显示

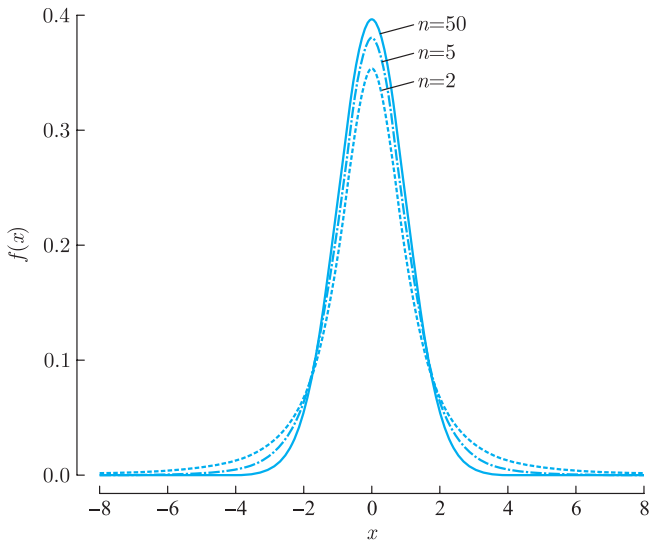

图 1.3.2　t 分布的概率密度函数

- $t(n)$ 的概率密度关于 y 轴对称, 且 $\lim\limits_{|x|\to\infty} f(x)=0$.
- 随着 n 的增大, 其峰度越来越高, 尾部越来越薄.
- t 分布的概率密度很像标准正态分布的概率密度.

为了区分它与正态分布的差异, 表 1.3.1 列出了标准正态分布、$t(4)$、$t(10)$ 与 $t(50)$ 的尾部概率. 从中可以看出, t 分布的尾部概率大于正态分布的相应概率. 但是, 对于固定的 x, 由于

$$\lim_{n\to\infty}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}=\mathrm{e}^{-x^2/2},$$

故当 n 很大时, t 分布的概率密度接近于标准正态分布的概率密度, 其常数可用 Stirling(斯特林) 公式:

$$\Gamma(x)\approx\sqrt{2\pi}x^{x-1/2}\mathrm{e}^{-x}$$

求得.

表 1.3.1　标准正态分布与 t 分布的尾部概率: $100\times P(T>x)$

x	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
$N(0,1)$	30.853 7	15.865 5	6.680 7	2.275 0	0.620 9	0.135 0	0.023 3	0.003 2
$t(4)$	32.166 5	18.695 0	10.400 0	5.805 8	3.338 3	1.997 1	1.244 8	0.806 5
$t(10)$	31.394 7	17.044 6	8.225 3	3.669 4	1.572 3	0.667 2	0.286 3	0.125 9
$t(50)$	30.963 4	16.106 2	6.995 1	2.547 3	0.787 2	0.210 0	0.049 4	0.010 4

注 1.3.3 基于 t 分布与标准正态分布概率密度函数的比较, 在统计学中, 称尾部概率大于正态分布相应尾部概率的分布为厚尾分布 (heavy-tailed distribution). 经济金融等许多行业中的数据多服从厚尾分布.

由上述概率密度函数, 可求得 $t(n)$ 分布的 $r(<n)$ 阶中心矩为 $(n>1)$:

$$E(T^r)=\begin{cases}0, & r\text{是奇数},\\ n^{r/2}\dfrac{\Gamma((r+1)/2)\Gamma((n-r)/2)}{\Gamma(1/2)\Gamma(n/2)}, & r\text{是偶数}.\end{cases}$$

并由此可知: 当 $n>2$ 时, 其方差为 $n/(n-2)$.

基于定理 1.3.2 与上述 t 分布定义, 可以证明: 对于来自 $N(\mu,\sigma^2)$ 的独立同分布样本 $X_1,X_2,\cdots,X_n$, 有

$$\frac{\sqrt{n}(\bar{X}-\mu)}{S_n}\sim t(n-1). \tag{1.3.3}$$

这即是 Gosset 于 1908 年论文的结论.

定义 1.3.4 (非中心 t 分布) 设 $\xi\sim N(\mu,\sigma^2)$, $\eta/\sigma^2\sim\chi^2(n)$, 且二者独立, $\mu\neq 0$, 则称

$$T=\frac{\xi}{\sqrt{\eta/n}}$$

服从非中心 t 分布, 其非中心参数为 $\delta=\mu/\sigma$, 自由度为 n.

经过一些复杂计算可求得非中心 $t(n)$ 分布的概率密度为

$$\frac{n^{n/2}\mathrm{e}^{-\delta^2/2}}{\sqrt{\pi}\Gamma(n/2)(n+x^2)^{(n+1)/2}}\sum_{m=0}^{\infty}\Gamma\left(\frac{n+m+1}{2}\right)\left(\frac{\delta^m}{m!}\right)\left(\frac{\sqrt{2}x}{\sqrt{n+x^2}}\right)^m,$$

其期望与方差分别为

$$E(T)=\delta\frac{\Gamma((n-1)/2)}{\Gamma(n/2)}\sqrt{\frac{n}{2}},\ n>1,$$

$$\mathrm{Var}(T)=\frac{n(1+\delta^2)}{n-2}-\frac{\delta^2 n}{2}\left(\frac{\Gamma((n-1)/2)}{\Gamma(n/2)}\right)^2,\ n>2.$$

1.3.3 F 分布

定义 1.3.5 (F 分布) 设 ξ,η 相互独立, 且服从自由度分别为 m,n 的 χ^2 分布. 称

$$F=\frac{\xi/m}{\eta/n}$$

服从的分布为 F 分布, 自由度为 (m,n), 记为 $F\sim F(m,n)$.

由上述定义知道,

- 如果 $X\sim F(1,n)$, $Y\sim t(n)$, 则 X 与 Y^2 依分布相等.
- 如果 $X\sim F(m,n)$, 且 $m\neq n$, 则 $X^{-1}\sim F(n,m)$.

定理 1.3.5　*$F(m,n)$ 分布的概率密度为*

$$f(x;m,n)=\begin{cases}0, & x<0,\\ \dfrac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)}\left(\dfrac{m}{n}\right)\left(\dfrac{mx}{n}\right)^{m/2-1}\left(1+\dfrac{mx}{n}\right)^{-(m+n)/2}, & x\geqslant 0.\end{cases}$$

证明　由于 $\xi\sim\chi^2(m),\eta\sim\chi^2(n)$, 且二者独立, 故它们的联合概率密度为

$$f(x,y)=\frac{1}{2^{(m+n)/2}\Gamma(m/2)\Gamma(n/2)}\mathrm{e}^{-(x+y)/2}x^{m/2-1}y^{n/2-1},\ x\geqslant 0,\ y\geqslant 0.$$

作变换

$$\begin{cases}u=x+y,\\ v=\dfrac{x}{y}\dfrac{n}{m},\end{cases}$$

则 (U,V) 的联合概率密度为

$$\begin{aligned}g(u,v)=&\frac{1}{2^{(m+n)/2}\Gamma(m/2)\Gamma(n/2)}\mathrm{e}^{-u/2}u^{\frac{m+n}{2}-2}\cdot\\ &\left(\frac{m}{n}\right)^{m/2-1}\frac{v^{\frac{m}{2}-1}}{(1+mv/n)^{\frac{m+n}{2}-1}}\frac{m}{n}\frac{u}{(1+mv/n)^2}\\ =&\frac{1}{2^{\frac{m+n}{2}}\Gamma((m+n)/2)}\mathrm{e}^{-u/2}u^{\frac{m+n}{2}-1}\cdot\\ &\frac{\Gamma((m+n)/2)(m/n)^{m/2}}{\Gamma(m/2)\Gamma(n/2)}\frac{v^{\frac{m}{2}-1}}{(1+mv/n)^{\frac{m+n}{2}}},\end{aligned}$$

由此可知 U,V 相互独立, 且 $U\sim\chi^2(m+n)$, 而第二部分即为 V 的概率密度, 即为所求. □

由上面定理的证明, 知如下结论:

推论 1.3.1　*设 $\xi\sim\chi^2(m),\eta\sim\chi^2(n)$, 且 ξ 与 η 独立, 则 $Y=\xi+\eta$ 与 $Z=\xi/\eta$ 独立.*

图 1.3.3 给出了自由度分别为 $(m,n)=(5,10),(10,10),(10,5)$ 的 F 分布的概率密度函数曲线.

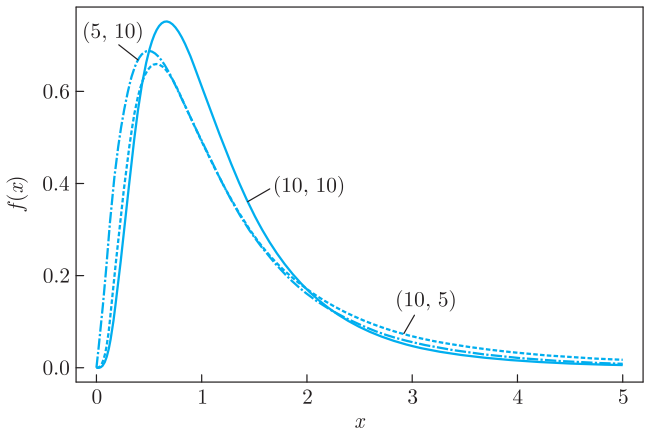

图 1.3.3 F 分布的概率密度函数

由上述概率密度函数可求得 F 分布的数字特征如下: 设 $F \sim F(m,n)$, 对 $0 < r < n/2$, 有

$$E(F^r) = \left(\frac{n}{m}\right)^r \frac{\Gamma\left(r+\frac{m}{2}\right)\Gamma\left(\frac{n}{2}-r\right)}{\Gamma(m/2)\Gamma(n/2)}.$$

由此求得其期望与方差分别为

$$E(F) = \frac{n}{n-2}, \quad n > 2,$$

$$\mathrm{Var}(F) = \frac{n^2(2m+2n-4)}{m(n-2)^2(n-4)}, \quad n > 4.$$

定义 1.3.6 (非中心 F 分布) 设 $\xi \sim \chi^2(m,\delta)$, $\eta \sim \chi^2(n)$, $\delta \neq 0$, 且 ξ, η 独立, 则称 $F = \dfrac{\xi/m}{\eta/n}$ 服从非中心 F 分布, 其自由度为 (m,n), 非中心参数为 δ.

由于非中心 F 分布的概率密度非常复杂, 且不常用, 故这里就不再给出. 如有兴趣, 请参见复旦大学编著的《概率论》第二册. 这里仅给出其期望与方差:

$$E(F) = \frac{n(m+\delta)}{m(n-2)}, \quad n > 2,$$

$$\mathrm{Var}(F) = \frac{2n^2}{m^2(n-2)^2(n-4)}\left[(m+\delta)^2 + (n-2)(m+2\delta)\right], \quad n > 4.$$

为考察 F 分布是哪个统计量的抽样分布, 假设有两组独立同分布样本: $X_1, X_2, \cdots, X_m$ 来自正态总体 $N(\mu_1, \sigma^2)$, $Y_1, Y_2, \cdots, Y_n$ 来自正态总体 $N(\mu_2, \sigma^2)$, 且两组样本独立. 如记

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

$$S_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2, \quad S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

则由定理 1.3.2 知道,

$$(m-1)S_x^2/\sigma^2 \sim \chi^2(m-1), \quad (n-1)S_y^2/\sigma^2 \sim \chi^2(n-1),$$

由于两组样本独立, 故两组样本方差独立, 于是, 由 F 分布定义可知

$$\frac{S_x^2}{S_y^2} \sim F(m-1, n-1). \tag{1.3.4}$$

1.3.4　一些常用分布

在本小节, 我们介绍几个统计经常用到的分布. 为了简单, 我们仅以总体分布的形式给出. 在有的参考文献中, 考虑到一些分布还包含若干常用分布, 故也称之为分布族.

定义 1.3.7 (Γ 分布)　*称概率密度为*

$$f(x;\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\mathrm{e}^{-\lambda x}, \ x \geqslant 0 \tag{1.3.5}$$

的分布为 Γ 分布, 记为 $\Gamma(\alpha,\lambda)$, 其中 $\alpha>0$ 为形状参数 (shape parameter), *$\lambda>0$ 为尺度参数* (scale parameter).

通过比较式 (1.3.1) 的 $\chi^2(n)$ 与上述 $\Gamma(\alpha,\lambda)$ 分布的概率密度函数, 可以发现: $\Gamma(n/2, 1/2)$ 分布就是 $\chi^2(n)$. 另外, 通过与指数分布 (指数分布的定义见本小节最后内容) 的比较, 我们发现, $\Gamma(1,\lambda)$ 就是指数分布 $Exp(\lambda)$. 再者, 之所以称 α 为形状参数, 是由于

- 当 $\alpha \leqslant 1$ 时, 其概率密度严格单调下降;
- 当 $1<\alpha\leqslant 2$ 时, 其概率密度先上凸后下凸, 且拐点为 $\dfrac{\alpha-1}{\lambda}$;
- 当 $\alpha>2$ 时, 其概率密度先下凸, 再上凸, 最后又下凸, 且峰值在 $\dfrac{\alpha-1}{\lambda}$ 处达到.

由 Γ 分布的概率密度, 可求得 $\xi\sim\Gamma(\alpha,\lambda)$ 分布的特征函数为

$$\psi(t) = \left(1-\frac{\mathrm{i}t}{\lambda}\right)^{-\alpha},$$

并由此可以求得其 k 阶矩为

$$E(\xi^k) = \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}\frac{1}{\lambda^k},$$

特别地, 其期望、方差分别为

$$E(\xi) = \frac{\alpha}{\lambda}, \quad \mathrm{Var}(\xi) = \frac{\alpha}{\lambda^2}.$$

利用其特征函数, 可以证明:

定理 1.3.6 设 $X_i \sim \Gamma(\alpha_i, \lambda)$, $i = 1, 2$.

(1) 如 X_1, X_2 独立, 则 $X_1 + X_2 \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$.

(2) 对于给定的正数 c, $cX_1 \sim \Gamma(\alpha_1, \lambda/c)$.

定义 1.3.8 (指数型分布) 称概率分布为

$$f(x;\boldsymbol{\theta}) = \exp\left\{a_0(\boldsymbol{\theta}) + b_0(x) + \sum_{i=1}^{q} c_i(\boldsymbol{\theta}) b_i(x)\right\} \tag{1.3.6}$$

的分布为指数型分布, 其中 $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$ 为参数, 且要求其支撑集 $S = \{x : f(x;\boldsymbol{\theta}) > 0\}$ 与 $\boldsymbol{\theta}$ 无关. 如定义 $\eta_i = c_i(\boldsymbol{\theta})(i = 1, 2, \cdots, q)$, 则称 $(\eta_1, \eta_2, \cdots, \eta_q)$ 为自然参数, $\mathcal{G} = \{(\eta_1, \eta_2, \cdots, \eta_q) : \eta_i = c_i(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ 为自然参数空间.

例 1.3.1 设 $X \sim B(1, p)$, 其中 $p \in (0, 1)$ 为参数, 验证它服从指数型分布.

证明 此时 X 的概率分布函数为

$$f(x;p) = P(X = x) = p^x (1-p)^{1-x}, \quad x = 0, 1.$$

我们可以改写上式为

$$f(x;p) = \left(\frac{p}{1-p}\right)^x (1-p) = \exp\left\{\ln(1-p) + x \ln \frac{p}{1-p}\right\},$$

故可知, 两点分布是指数型分布. 如取 $\theta = \ln[p/(1-p)]$, 则上式为

$$f(x;\theta) = \exp\left\{\ln \frac{1}{1+\mathrm{e}^{\theta}} + \theta x\right\},$$

其自然参数空间为 $\mathcal{G} = \mathbb{R}$. □

例 1.3.2 设 $X \sim N(\mu, \sigma^2)$, 验证其为指数型分布.

证明 此时 X 的概率密度为

$$\begin{aligned} f(x;\mu,\sigma^2) &= (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right\} \\ &= \exp\left\{-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{\mu^2}{2\sigma^2} - \frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2}\right\}, \end{aligned}$$

由此可知, 正态分布是指数型分布. 如取 $\eta_1 = -1/(2\sigma^2), \eta_2 = \mu/\sigma^2$, 则其自然参数空间为 $\mathcal{G} = \{(\eta_1, \eta_2) : \eta_1 < 0, \eta_2 \in \mathbb{R}\}$. □

定义 1.3.9 (次 Gauss 分布) 称 X 服从参数为 $\sigma > 0$ 的次 Gauss 分布 (sub-Gauss), 是指满足

$$E[\mathrm{e}^{t[X - E(X)]}] \leqslant \mathrm{e}^{t^2\sigma^2/2}, \quad \forall\, t \in \mathbb{R}.$$

如果 X 服从均值为 0、参数为 σ 的次 Gauss 分布, 则

$$P(X>t)\leqslant \exp\left\{-\frac{t^2}{2\sigma^2}\right\},\ \forall t>0.$$

实际上,

$$P(X>t)=P\left(\mathrm{e}^{tX/\sigma^2}>\mathrm{e}^{t^2/\sigma^2}\right)\leqslant \frac{E[\exp\{tX/\sigma^2\}]}{\exp\{t^2/\sigma^2\}}\leqslant \exp\left\{-\frac{t^2}{2\sigma^2}\right\}.$$

为什么称上述分布为次 Gauss 分布? 原因如下: 设 $X\sim N(0,\sigma^2)$, 则对于任意的 $t>0$, 有

$$P(X>t)=\frac{1}{\sqrt{2\pi\sigma^2}}\int_t^{\infty}\mathrm{e}^{-\frac{x^2}{2\sigma^2}}\mathrm{d}x\leqslant \exp\left\{-\frac{t^2}{2\sigma^2}\right\}.$$

关于次 Gauss 分布, 我们有如下结论:

定理 1.3.7　设 $X_1,X_2,\cdots,X_n$ 为来自均值为 0、参数为 σ 的次 Gauss 分布, 则

$$E\left(\max_{1\leqslant i\leqslant n}X_i\right)\leqslant \sigma\sqrt{2\ln n},\quad P\left(\max_{1\leqslant i\leqslant n}X_i>t\right)\leqslant n\exp\left\{-\frac{t^2}{2\sigma^2}\right\}.$$

证明

$$\begin{aligned}E\left(\max_{1\leqslant i\leqslant n}X_i\right)&=\frac{1}{t}E\left(\ln \mathrm{e}^{t\max\limits_i X_i}\right)\leqslant \frac{1}{t}\ln E\left(\mathrm{e}^{t\max\limits_i X_i}\right)\quad \text{(Jensen(延森) 不等式)}\\&=\frac{1}{t}\ln E\left(\max_i \mathrm{e}^{tX_i}\right)\leqslant \frac{1}{t}\ln\sum_{i=1}^{n}E(\mathrm{e}^{tX_i})\\&\leqslant \frac{1}{t}\ln\sum_{i=1}^{n}\mathrm{e}^{\sigma^2t^2/2}=\frac{\ln n}{t}+\frac{\sigma^2t}{2},\end{aligned}$$

取 $t=\sqrt{(2\ln n)/\sigma^2}$ 即证明了第一个结论.

另外,

$$P\left(\max_i X_i>t\right)=P\left(\bigcup_{i=1}^{n}\{X_i>t\}\right)\leqslant \sum_{i=1}^{n}P(X_i>t)\leqslant n\mathrm{e}^{-\frac{t^2}{2\sigma^2}}.$$

第二个结论得证. □

在风险预警及管理中, 极值统计量应用最广, 并由此产生了统计学的一个研究方向——极值统计. 极值统计最早是由 L. H. Tippett (蒂珀特, 1902—1985) 在 Shirley Institute 做统计分析时提出的. 当时, 他发现一根棉线的强度取决于最弱一根纤维的强度. 为解决此强度的建模问题, 他于 1924 年到伦敦大学学院的 Galton 生物统计实验室跟随 K. Pearson 学习, 之后发现了一个能联系极值分布与数据分布很简单的方程式, 但他无法求解, 而猜出了一个答案. 后来他求助 R. A. Fisher, Fisher 推导出了 Tippett 的解, 还给出了另外两个解. 这就是下面的三个极值分布.

为了给出极值分布的一般形式, 先给出分布等价的定义: 两个累积分布函数 F_1, F_2, 如存在常数 $a>0, b$, 使

$$F_1(ax+b)=F_2(x),$$

则称两者等价. 另外, 对于极大次序统计量 $X_{(n)}$, 如果存在两个常数 $c_n>0, b_n$, 使得

$$c_nX_{(n)}+b_n \xrightarrow{\mathrm{d}} G(x),$$

则称 $G(X)$ 为极值分布.

定义 1.3.10 (极值分布) 如果 $c_nX_{(n)}+b_n$ 有连续的极限分布 $G(X)$, 则 $G(x)$ 必属于以下三种分布所代表的分布类之一:

(1) I 型: $G(x)=\exp\{-\mathrm{e}^{-x}\}$, $x\in\mathbb{R}$.

(2) II 型: $G(x)=\exp\{-x^{-\alpha}\}I_{(0,\infty)}(x)$, $\alpha>0$.

(3) III 型: $G(x)=\begin{cases}\exp\{-(-x)^{-\alpha}\}, & x<0,\\ 1, & x\geqslant 0,\end{cases}$ $\alpha>0$.

例 1.3.3 设总体 $X\sim Exp(1)$, 求 $X_{(n)}$ 的极限分布.

解 由前面讲述的极大次序统计量的分布可知, $X_{(n)}$ 的累积分布为

$$F(x)=(1-\mathrm{e}^{-x})^n.$$

为求其极限分布, 我们先求其期望. 为此, 记 $Y_i=X_{(i)}-X_{(i-1)}, i=1,2,\cdots,n$, 其中 $X_{(0)}=0$.

由于 Y_1 的概率密度为 $n\mathrm{e}^{-nx}$, 故可求得 $E(Y_1)=1/n$.

由例 1.2.3 知, $X_{(k)}$ 的概率密度为

$$k\binom{n}{k}(1-\mathrm{e}^{-x})^{k-1}\mathrm{e}^{-(n-k+1)x}.$$

由此可求得

$$E(Y_k)=E[X_{(k)}]-E[X_{(k-1)}]=\frac{1}{n-k+1}.$$

于是, $E[X_{(n)}]=\sum\limits_{k=1}^{n}E(Y_k)=\sum\limits_{k=1}^{n}\dfrac{1}{k}$, 与 $\ln n$ 同阶.

由于 $X_{(n)}-\ln n$ 的累积分布函数满足

$$P(X_{(n)}-\ln n\leqslant t)=(1-\mathrm{e}^{-t-\ln n})^n=(1-\mathrm{e}^{-t}/n)^n\to\exp\{-\mathrm{e}^{-t}\},$$

故知 $X_{(n)}$ 是 I 型极值分布. □

注 1.3.4 陈希孺 (2009) 给出了 $X_{(n)}$ 服从三种极值分布而要求总体分布所满足的充要条件, 请有兴趣的读者参阅.

统计中还有多种常用分布, 现罗列出连续总体的概率密度、期望与方差如下:

(1) 指数分布 $Exp(\mu,\lambda)$, 其概率密度为

$$\lambda \mathrm{e}^{-\lambda(x-\mu)} I_{[\mu,\infty)}(x),\ \lambda>0, \mu\in\mathbb{R},$$

期望与方差分别为 $\mu+\lambda^{-1}, \lambda^{-2}$. 为了简单, 记 $\mu=0$ 的指数分布为 $Exp(\lambda)$.

(2) 双指数分布 $DE(\mu,\lambda)$, 其概率密度为

$$\frac{\lambda}{2}\mathrm{e}^{-\lambda|x-\mu|},\ \lambda>0, \mu\in\mathbb{R},$$

期望与方差分别为 $\mu, 2/\lambda^2$.

(3) Laplace 分布 $L(\lambda)$, 其概率密度为

$$\lambda \mathrm{e}^{-\lambda|x|},\ \lambda>0,$$

期望与方差分别为 $0, 1/\lambda$.

(4) β 分布 $Beta(\alpha,\beta)$, 其概率密度为

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{(0,1)}(x),\ \alpha>0, \beta>0,$$

期望与方差分别为 $\alpha/(\alpha+\beta), \alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]$.

(5) Cauchy 分布 $C(\mu,\sigma)$, 其概率密度为

$$\frac{1}{\pi\sigma\left[1+\left(\dfrac{x-\mu}{\sigma}\right)^2\right]},\ \mu\in\mathbb{R}, \sigma>0,$$

期望与方差均不存在. 称 $C(0,1)$ 为标准 Cauchy 分布.

(6) 对数正态分布 $LN(\mu,\sigma^2)$, 其概率密度为

$$\frac{1}{\sqrt{2\pi}\sigma} x^{-1}\mathrm{e}^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} I_{(0,\infty)}(x),\ \mu\in\mathbb{R}, \sigma>0,$$

期望与方差分别为 $\mathrm{e}^{\mu+\sigma^2/2}, \mathrm{e}^{2\mu+\sigma^2}\left(\mathrm{e}^{\sigma^2}-1\right)$.

(7) Weibull 分布 $W(\alpha,\beta)$, 其概率密度为

$$\frac{\alpha}{\beta} x^{\alpha-1}\mathrm{e}^{-\frac{x^\alpha}{\beta}} I_{(0,\infty)}(x),\ \alpha>0, \beta>0,$$

期望与方差分别为 $\beta^{\alpha^{-1}}\Gamma(\alpha^{-1}+1)$ 和 $\beta^{2/\alpha}\left[\Gamma(2\alpha^{-1}+1)-[\Gamma(\alpha^{-1}+1)]^2\right]$.

(8) Pareto 分布 $Pa(a,\theta)$, 其概率密度为

$$\theta a^\theta x^{-(\theta+1)} I_{[a,\infty)}(x), \theta>0, a>0,$$

期望与方差分别为 $\theta a/(\theta-1)$(要求 $\theta>1$) 和 $\theta a^2/[(\theta-1)^2(\theta-2)]$(要求 $\theta>2$).

(9) logistic 分布 $LG(\mu,\sigma)$, 其概率密度为

$$\frac{1}{\sigma}\frac{\mathrm{e}^{-\frac{x-\mu}{\sigma}}}{1+\mathrm{e}^{-\frac{x-\mu}{\sigma}}},$$

期望与方差分别为 $\mu,\pi^2\sigma^2/3$.

其他一些常用离散总体的概率分布、期望及方差如下:

(1) 几何分布 $G(p)$, 其概率分布为

$$(1-p)^{x-1}p,\quad x=1,2,\cdots,\quad p\in[0,1],$$

其期望与方差分别为 $1/p$ 和 $(1-p)/p^2$.

(2) 超几何分布 $HG(r,n,m)$, 其概率分布为

$$\frac{\binom{n}{x}\binom{m}{r-x}}{\binom{n+m}{r}},\quad x=0,1,\cdots,\min\{r,n\},r-x\leqslant m,$$

其期望与方差分别为 $rn/(n+m)$ 和 $rnm(n+m-r)/[(n+m)^2(n+m-1)]$.

(3) 负二项分布 $NB(p,r)$, 其概率分布为

$$\binom{x-1}{r-1}p^r(1-p)^{x-r},\quad x=r,r+1,\cdots,\quad p\in[0,1],$$

其期望与方差分别为 r/p 和 $r(1-p)/p^2$.

1.4 充分与完全统计量

在结束本章之前, 我们引进两个在估计中非常有用的统计量: 充分统计量、完全统计量.

20 世纪 20 年代初, 在 R. A. Fisher 与天文学家 A. S. Eddington (爱丁顿, 1882—1944) 之间发生了如下争论: 对于 n 个来自正态总体 $N(\mu,\sigma^2)$ 的独立同分布样本 X_1, $X_2,\cdots,X_n$, 如何估计总体标准差 σ. Fisher 主张用样本标准差 S_n, 而 Eddington 主张用如下的平均绝对偏差:

$$D_n=\sqrt{\frac{\pi}{2}}\frac{1}{n}\sum_{i=1}^{n}|X_i-\bar{X}|. \tag{1.4.1}$$

Fisher 给出的原因是: 样本标准差 S_n 包含了样本中有关 σ 的全部信息, 而 D_n 则没有. 由此, Fisher 于 1922 年提出了**充分统计量** (sufficient statistic).

对于一个统计量 $T(X)$, 我们知道 $E[T(X)]$ 是由样本空间 $\mathcal{X}$ 到参数空间 Θ 的一个映射. 请问, 这个映射是一一对应的吗? 当 $T(X)$ 是**完全统计量** (complete statistic) 时, 答案是肯定的.

1.4.1　充分统计量

要衡量一个统计量是否把样本中关于参数的所有信息都包含进来, 就要回答两个问题: 一、如何度量样本中关于参数的信息, 二、如何度量统计量是否包含了样本中关于参数的全部信息.

实际上, 关于参数的信息有两部分, 一是总体分布所提供的, 二是样本所提供的. 综合这两部分我们知道, 关于参数的所有信息就包含在样本分布之中. 如果统计量 $T(X)$ 包含了参数的全部信息, 则意味着扣除统计量所包含的信息外, 样本分布中就不再有参数的信息了. 于是, Fisher 给出充分统计量的定义如下:

定义 1.4.1　*设 $X_1, X_2, \cdots, X_n$ 为来自总体概率分布为 $f(x;\theta)(\theta \in \Theta)$ 的样本, 称统计量 $T(X)$ 为参数 θ 的充分统计量, 是指在给定 $T(X)$ 下, 样本的条件联合分布与参数 θ 无关.*

从上述定义可以看出, 充分统计量并不唯一, 且充分统计量的可逆函数, 只要它仍是统计量, 则它也是充分统计量.

例 1.4.1　在例 1.2.1 中, 假设抽样是有返回的, 证明 $\sum\limits_{i=1}^{n} X_i$ 是次品率参数的充分统计量.

证明　如记 $p = M/N$, 则 n 次抽样所得样本 $X_1, X_2, \cdots, X_n$ 就是来自两点分布总体 $B(1,p)$ 的独立同分布样本. 样本分布及加和统计量 $T(X) = \sum\limits_{i=1}^{n} X_i$ 的分布分别为

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = p^{\sum\limits_{i=1}^{n} x_i}(1-p)^{n-\sum\limits_{i=1}^{n} x_i},$$

$$P(T(X) = t) = \binom{n}{t} p^t (1-p)^{n-t}.$$

于是, 在给定统计量 $T = t$ 下, 样本的条件概率分布为

$$\begin{aligned} & P(X_1 = x_1, \cdots, X_n = x_n | T = t) \\ = & \frac{P\left(X_1 = x_1, \cdots, X_{n-1} = x_{n-1}, X_n = t - \sum\limits_{i=1}^{n-1} x_i\right)}{P(T = t)} \\ = & \frac{p^t (1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} = \binom{n}{t}^{-1}, \end{aligned}$$

它与参数 p 无关. □

在例 1.4.1 中, 当 $n>2$ 时, 如果采用统计量 $T_1(X)=X_1+X_2$, 则由于

$$\begin{aligned}
&P(X_1=x_1,\cdots,X_n=x_n|T_1=t)\\
=&\frac{P(X_1=x_1,X_2=t-x_1,X_3=x_3,\cdots,X_n=x_n)}{P(T_1=t)}\\
=&\frac{p^{t+\sum\limits_{i=3}^{n}x_i}(1-p)^{n-t-\sum\limits_{i=3}^{n}x_i}}{\begin{pmatrix}2\\t\end{pmatrix}p^t(1-p)^{2-t}}
\end{aligned}$$

与 p 有关, 则说明 $T_1(X)$ 不是充分统计量.

例 1.4.2 设 $X_1,X_2,\cdots,X_n$ 为来自 Poisson(泊松) 分布 $P(\lambda)$ 的独立同分布样本, 证明 $T=\sum\limits_{i=1}^{n}X_i$ 是 λ 的充分统计量.

证明 由 Poisson 分布的可加性知, $T=\sum\limits_{i=1}^{n}X_i\sim P(n\lambda)$, 故当 $T=t$ 时, 样本的条件概率分布为

$$\begin{aligned}
&P(X_1=x_1,\cdots,X_n=x_n|T=t)\\
=&\frac{P(X_1=x_1)\cdots P(X_{n-1}=x_{n-1})P\left(X_n=t-\sum\limits_{i=1}^{n-1}x_i\right)}{P(T=t)}\\
=&\frac{\left(\prod\limits_{j=1}^{n-1}\frac{\lambda^{x_j}\mathrm{e}^{-\lambda}}{x_j!}\right)\frac{\lambda^{t-\sum\limits_{i=1}^{n-1}x_i}\mathrm{e}^{-\lambda}}{\left(t-\sum\limits_{i=1}^{n-1}x_i\right)!}}{\frac{(n\lambda)^t}{t!}\mathrm{e}^{-n\lambda}}\\
=&\frac{t!}{\prod\limits_{j=1}^{n-1}x_j!\left(t-\sum\limits_{i=1}^{n-1}x_i\right)!}\frac{1}{n^t},
\end{aligned}$$

与 λ 无关, 故 T 是充分统计量. □

例 1.4.3 设 $X_1,X_2,\cdots,X_n$ 为来自正态总体 $N(\mu,1)$ 的独立同分布样本, 证明 $T(X)=\sum\limits_{i=1}^{n}X_i$ 是 μ 的充分统计量.

证明 由于此时总体分布是连续的, 故不易直接计算条件分布. 为此, 作正交变换 $\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)^{\mathrm{T}}=\boldsymbol{C}\boldsymbol{X}$, 其中 $\boldsymbol{C}$ 是第一行为 $(1/\sqrt{n},1/\sqrt{n},\cdots,1/\sqrt{n})$ 的正交矩阵, $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$. 此时, $Y_1=T(X)/\sqrt{n}$, 且 $Y_1,Y_2,\cdots,Y_n$ 独立, $Y_i\sim N(0,1),2\leqslant i\leqslant n$. 由此可知, 在给定 $T(X)$ 即 Y_1 的条件下, $(Y_2,Y_3,\cdots,Y_n)$ 的条件分布与 μ 无关. 由于正交矩阵 $\boldsymbol{C}$ 与 μ 无关, 故知 $X_1,X_2,\cdots,X_n$ 在给定 $T(X)$ 下的条件分布也与 μ 无关, 即得证. □

从上述三个例子可以看出, 相较于离散总体, 关于连续总体的充分统计量的验证就要麻烦许多, 即使是离散总体, 其计算也较复杂. 为了更加容易验证一个统计量是否为充分的, 下面的因子分解定理就非常有用.

定理 1.4.1 (因子分解定理 (factorization theorem))　设 $X_1, X_2, \cdots, X_n$ 为来自总体概率分布为 $f(x;\theta)$ 的样本, 则统计量 $T(X)$ 为充分统计量的充要条件是: 样本分布 $f(x_1, x_2, \cdots, x_n;\theta)$ 可如下分解:

$$f(x_1, x_2, \cdots, x_n;\theta) = g_\theta(T(x_1, x_2, \cdots, x_n)) \cdot h(x_1, x_2, \cdots, x_n), \tag{1.4.2}$$

其中 $h(x_1, x_2, \cdots, x_n)$ 不依赖于参数 θ.

证明　为简单, 这里仅考虑总体分布连续的证明, 且在不引起混淆的情况下, 以 $\boldsymbol{X}$ 记样本 $X_1, X_2, \cdots, X_n$. 假设统计量 $T(\boldsymbol{X}) = (T_1(\boldsymbol{X}), T_2(\boldsymbol{X}), \cdots, T_k(\boldsymbol{X}))$, $W(\boldsymbol{X}) = (W_1(\boldsymbol{X}), W_2(\boldsymbol{X}), \cdots, W_{n-k}(\boldsymbol{X}))$, 其中 $k < n$, 且要求变换

$$(X_1, X_2, \cdots, X_n) \to (T_1(\boldsymbol{X}), T_2(\boldsymbol{X}), \cdots, T_k(\boldsymbol{X}), W_1(\boldsymbol{X}), W_2(\boldsymbol{X}), \cdots, W_{n-k}(\boldsymbol{X})) \quad (*1)$$

是一对一的. 如以 J 记上述变换的 Jacobi 行列式的绝对值, 以 $H(t,w)$ 表示 (1.4.2) 式中的 $h(x_1, x_2, \cdots, x_n)/J$, 则 (T, W) 的概率密度函数为 $g_\theta(t)H(t,w)$. 由此可求得给定 T 时, W 的条件概率密度为

$$f_{W|T}(w) = \frac{g_\theta(t)H(t,w)}{\displaystyle\int_{\mathbb{R}^{n-k}} g_\theta(t)H(t,w)\mathrm{d}w} = \frac{H(t,w)}{\displaystyle\int_{\mathbb{R}^{n-k}} H(t,w)\mathrm{d}w}. \tag{*2}$$

下证充分性:

由前述可知, $\forall\, t_1, t_2, \cdots, t_k$, 如记

$$B(t) = \{\boldsymbol{X} \in \mathcal{X} : T(\boldsymbol{X}) = (t_1, t_2, \cdots, t_k)\},$$

则在集合 $B(t)$ 内, $\boldsymbol{X}$ 与 $W(\boldsymbol{X})$ 也一一对应. 因此, 要确定给定 $T(\boldsymbol{X})$ 下, $\boldsymbol{X}$ 的条件分布与参数 θ 无关, 只需确定在给定 $T(\boldsymbol{X})$ 下, $W(\boldsymbol{X})$ 的分布与参数 θ 无关即可.

由 (*2) 式可知, 当样本分布满足 (1.4.2) 式的形式时, 给定 T 下, W 的条件分布与参数 θ 无关, 于是, $T(X)$ 即为充分统计量.

下证必要性:

设 $T(X)$ 为充分统计量, 且记 $G(t,w)$ 为给定 T 下 W 的条件概率密度. 由充分统计量定义知, $G(t,w)$ 与 θ 无关. 如记 $g_\theta(t)$ 为 T 的概率密度, 则知 (T, W) 的联合概率密度为

$$f_{(T,W)}(t,w) = g_\theta(t)G(t,w).$$

如通过变换 (*1) 式, 把 $G(t,w)$ 表示为 $h_1(x_1, x_2, \cdots, x_n)$, 且利用密度变换公式, 则可求得样本分布为

$$f(x_1, x_2, \cdots, x_n;\theta) = f_{(T,W)}(t,w)J = g_\theta(x_1, x_2, \cdots, x_n)h_1(x_1, x_2, \cdots, x_n)J.$$

如记 $h_1(x_1,x_2,\cdots,x_n)J=h(x_1,x_2,\cdots,x_n)$, 由于它与 θ 无关, 则结论得证. □

注 1.4.1 Fisher (1922) 首先给出此定理, Halmos et al. (1949) 给出了其一般形式及严格数学证明. 上述定理的证明来自陈希孺等 (2009), 其一般形式的严格证明参见陈希孺 (2009).

例 1.4.4 设 $X_1,X_2,\cdots,X_n$ 为来自正态总体 $N(\mu,\sigma^2)$ 的独立同分布样本, 求其充分统计量.

解 此时, 样本分布为

$$f(\boldsymbol{x};\mu,\sigma^2)=(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right\}$$

$$=(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{\sum\limits_{i=1}^{n}x_i^2}{2\sigma^2}+\frac{\mu\sum\limits_{i=1}^{n}x_i}{\sigma^2}-\frac{n\mu^2}{2\sigma^2}\right\}.$$

由因子分解定理可知, 当 σ^2 已知时, $\sum\limits_{i=1}^{n}X_i$ 为充分统计量, 这与例 1.4.3 的结论一致; 当 μ,σ^2 均未知时, $\left(\sum\limits_{i=1}^{n}X_i,\sum\limits_{i=1}^{n}X_i^2\right)$ 为充分统计量. □

由上例可知, $(n-1)S_n^2=\sum\limits_{i=1}^{n}X_i^2-n\bar{X}^2$ 为充分统计量的函数, 这即是本节开始时提到的 Fisher 的观点.

例 1.4.5 设 $X_1,X_2,\cdots,X_n$ 为来自均匀分布 $U(0,\theta)$ 的独立同分布样本, 求 θ 的充分统计量.

解 此时样本分布为

$$f_\theta(\boldsymbol{x})=\begin{cases}\theta^{-n}, & \max\{x_1,x_2,\cdots,x_n\}<\theta,\\ 0, & \text{否则}.\end{cases}$$

如记 $t=\max\{x_1,x_2,\cdots,x_n\}$, 则上式可改写为

$$f_\theta(\boldsymbol{x})=\frac{1}{\theta^n}I_{(0,\theta)}(t),$$

则由因子分解定理可知, $T=X_{(n)}$ 即为 θ 的充分统计量. □

例 1.4.6 设 $X_1,X_2,\cdots,X_n$ 为来自均匀分布 $U\left(-\frac{1}{2}+\theta,\frac{1}{2}+\theta\right)$ 的独立同分布样本, 求 θ 的充分统计量.

解 此时样本分布为

$$f_\theta(\boldsymbol{x})=\prod_{i=1}^{n}I_{(-\frac{1}{2}+\theta,\frac{1}{2}+\theta)}(x_i)=I_{(x_{(n)}-\frac{1}{2},x_{(1)}+\frac{1}{2})}(\theta),$$

故由因子分解定理可知, $(X_{(1)}, X_{(n)})$ 是 θ 的充分统计量. □

此例说明, 一维参数的充分统计量并不一定是一维的. 另外, 从充分统计量定义或因子分解定理可知, 充分统计量不唯一. 为此, 我们引入最小充分统计量的定义.

定义 1.4.2(最小充分统计量)　设 T 为充分统计量, 如果对于任何一个充分统计量 S, 都存在一一映射 ψ, 使得 $T = \psi(S)$, 则称统计量 T 为最小充分统计量.

从上述定义可以看出, 在一一映射相等的条件下, 最小充分统计量是唯一的.

例 1.4.7　设 $X_1, X_2, \cdots, X_n$ 为来自均匀分布 $U(\theta, \theta+1)$ 的独立同分布样本, $\theta \in \mathbb{R}$ 为参数, 且 $n > 1$. 证明 $(X_{(1)}, X_{(n)})$ 是最小充分统计量.

证明　此时样本分布为

$$f_\theta(\boldsymbol{x}) = \prod_{i=1}^{n} I_{(\theta,\theta+1)}(x_i) = I_{(x_{(n)}-1, x_{(1)})}(\theta),$$

则由因子分解定理知, $T = (X_{(1)}, X_{(n)})$ 为充分统计量, 且

$$x_{(1)} = \sup\{\theta : f_\theta(\boldsymbol{x}) > 0\}, \quad x_{(n)} = 1 + \inf\{\theta : f_\theta(\boldsymbol{x}) > 0\}.$$

设 $S(X)$ 为 θ 的任一充分统计量, 则由因子分解定理知, 存在两个函数 h, g_θ, 使得

$$f_\theta(\boldsymbol{x}) = g_\theta(S(\boldsymbol{x}))h(\boldsymbol{x}).$$

由于对于任一满足 $h(\boldsymbol{x}) > 0$ 的 $\boldsymbol{x}$, 有

$$x_{(1)} = \sup\{\theta : g_\theta(S(\boldsymbol{x})) > 0\}, \quad x_{(n)} = 1 + \inf\{\theta : g_\theta(S(\boldsymbol{x})) > 0\}.$$

故当 $h(\boldsymbol{x}) > 0$ 时, 存在可测函数 ϕ, 使得 $T(\boldsymbol{x}) = \phi(S(\boldsymbol{x}))$. 由于 $P(h(\boldsymbol{X}) > 0) = 1$, 故 T 是最小充分统计量. □

定理 1.4.2　设总体 X 的概率分布为 $f(x;\theta) = g_\theta(T(x))h(x)$, 且满足: 如果由存在不依赖于 θ 的常数 $c(x,y)$ 使得 $f(x;\theta) = c(x,y)f(y;\theta)$, 就可推出 $T(x) = T(y)$, 则 $T(\boldsymbol{X})$ 是 θ 的最小充分统计量.

证明　设 T' 为一充分统计量, 则由因子分解定理知,

$$f(\boldsymbol{x};\theta) = \tilde{g}_\theta(T'(\boldsymbol{x}))\tilde{h}(\boldsymbol{x}).$$

如果 T 不是最小充分统计量, 即 T 不是 T' 的函数, 则存在两组样本点 $\boldsymbol{x}_0, \boldsymbol{y}_0$, 使得 $T'(\boldsymbol{x}_0) = T'(\boldsymbol{y}_0)$, 但是 $T(\boldsymbol{x}_0) \neq T(\boldsymbol{y}_0)$.

由于 $T'(\boldsymbol{x}_0) = T'(\boldsymbol{y}_0)$, 故存在不依赖于 θ 的常数 $c(\boldsymbol{x}_0, \boldsymbol{y}_0)$, 使得

$$f(\boldsymbol{x}_0;\theta) = \tilde{g}_\theta(T'(\boldsymbol{x}_0))\tilde{h}(\boldsymbol{x}_0) = c(\boldsymbol{x}_0, \boldsymbol{y}_0)\tilde{g}_\theta(T'(\boldsymbol{y}_0))\tilde{h}(\boldsymbol{y}_0) = c(\boldsymbol{x}_0, \boldsymbol{y}_0)f(\boldsymbol{y}_0;\theta),$$

于是, 由已知条件有 $T(\boldsymbol{x}_0) = T(\boldsymbol{y}_0)$. 矛盾. □

例 1.4.8 设 $X_1, X_2, \cdots, X_n$ 为来自双指数分布 $DE(\theta, 1)$ 的独立同分布样本, 其中 $\theta \in \mathbb{R}$. 证明次序统计量 $(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$ 是最小充分统计量.

证明 此时, 样本分布为

$$f(\boldsymbol{x}; \theta) = \frac{1}{2^n} \exp\left\{ -\sum_{i=1}^{n} |x_i - \theta| \right\}.$$

由因子分解定理可知, 次序统计量 $(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$ 是充分统计量, 下证它是最小的. 假设存在常数 $c(\boldsymbol{x}, \boldsymbol{y})$, 使得

$$f(\boldsymbol{x}; \theta) = c(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{y}; \theta), \quad \text{即} \quad \sum_{i=1}^{n} |x_i - \theta| = c_0(\boldsymbol{x}, \boldsymbol{y}) + \sum_{i=1}^{n} |y_i - \theta|,$$

其中 $c_0(\boldsymbol{x}, \boldsymbol{y}) = -\ln(c(\boldsymbol{x}, \boldsymbol{y}))$. 由于上述函数都是 θ 的分段线性函数, 而斜率仅依赖于每个次序统计量中的两个, 且递增. 注意到如果 $\boldsymbol{x}, \boldsymbol{y}$ 有相同的次序统计量, 则二者的差仅是一个常数 $c_0(\boldsymbol{x}, \boldsymbol{y})$. 因此, 由上面定理知, 次序统计量是最小充分统计量. □

1.4.2 完全统计量

完全统计量是 E. L. Lehmann (莱曼, 1917—2009) 和 H. Scheffé (谢费, 1907—1977) 提出的, 参见 Lehmann et al. (1950, 1955). 有文献也称之为完备统计量, 它类似于泛函分析中函数系的完全性概念.

定义 1.4.3 设 $X_1, X_2, \cdots, X_n$ 是来自某参数总体的样本, 参数 $\theta \in \Theta$. 统计量 $T(X)$ 如满足: 对任何满足条件

$$E_\theta[g(T(X))] = 0, \ \forall\, \theta \in \Theta$$

的统计量 $g(T)$, 都有

$$P_\theta(g(T(X)) = 0) = 1, \ \forall\, \theta \in \Theta,$$

则称之为完全统计量.

在统计中, 还有如下的完全分布族的概念:

定义 1.4.4 对于总体 X, 参数 $\theta \in \Theta$, 如果对于任一函数 $\psi(x)$, 由

$$E_\theta[\psi(X)] = 0, \ \forall\, \theta \in \Theta,$$

总可推出 $P_\theta(\psi(X) = 0) = 1$, 则称此总体分布族是完全的.

从上述两个定义可以看出, 统计量 $T(X)$ 的完全性, 与此统计量的抽样分布族的完全性等价. 另外, 如果上述定义中的函数 g 是有界的, 那么称统计量 T 是有界完全的或其分布族是有界完全的. 显然, 一个完全统计量肯定是有界完全的, 但反之不一定成立.

例 1.4.9　考虑二项分布族 $\{B(n,p):\ 0<p<1\}$ 的完全性.

解　设函数 $\psi(x)$ 满足

$$E_p[\psi(X)]=\sum_{x=0}^{n}\psi(x)\left(\begin{array}{c}n\\x\end{array}\right)p^x(1-p)^{n-x}=0,\ \forall\ 0<p<1.$$

令 $\theta=p/(1-p)$, 则 $\theta>0$, 且上式可以写成

$$\sum_{x=0}^{n}\psi(x)\left(\begin{array}{c}n\\x\end{array}\right)\theta^x=0,\ \forall\ \theta>0.$$

由于上式左边是 θ 的一个 n 次多项式, 故有 $\psi(x)=0,\ x=0,1,\cdots,n$, 故二项分布族是完全的. □

例 1.4.10　设 $X_1,X_2,\cdots,X_n$ 为来自均匀分布 $U(0,\theta)$ 的独立同分布样本, 证明其充分统计量 $T_n=X_{(n)}$ 的完全性.

证明　由于统计量 T_n 的概率密度为

$$f(t;\theta)=\begin{cases}\dfrac{nt^{n-1}}{\theta^n}, & 0<t<\theta,\\ 0, & \text{否则},\end{cases}$$

又设 $\psi(T)$ 满足 $E_\theta[\psi(T)]=0,\ \forall\ \theta>0$, 即

$$\int_0^\theta\psi(t)f(t;\theta)\mathrm{d}t=0,\ \forall\ \theta>0,$$

即

$$\int_0^\theta\psi(t)t^{n-1}\mathrm{d}t=0,\ \forall\ \theta>0,$$

两边关于 θ 求导, 有

$$\psi(\theta)\theta^{n-1}=0,\ \forall\ \theta>0,$$

即

$$\psi(\theta)=0,\ \forall\ \theta>0,$$

所以, 均匀分布族 $U(0,\theta)$ 的充分统计量 $T_n=X_{(n)}$ 是完全的. □

定理 1.4.3　设 $X_1,X_2,\cdots,X_n$ 为来自指数型分布 (1.3.6) 的独立同分布样本, 则

(1) $T(X)=\left(\sum\limits_{i=1}^{n}b_1(X_i),\sum\limits_{i=1}^{n}b_2(X_i),\cdots,\sum\limits_{i=1}^{n}b_q(X_i)\right)$ 是充分统计量.

(2) 如果其自然参数空间 $\mathcal{G}$ 包含一个开集, 则 $T(X)$ 也是完全的.

证明请参见陈希孺 (1997).

例 1.4.11 设 $X_1, X_2, \cdots, X_n$ 为来自 Poisson 分布 $P(\lambda)$ 的独立同分布样本, 证明 $\sum_{i=1}^{n} X_i$ 是充分完全统计量.

证明 由 Poisson 分布的可加性知, 统计量 $T = \sum_{i=1}^{n} X_i \sim P(n\lambda)$, 且其概率分布为

$$P(T=t) = \frac{(n\lambda)^t}{t!}\mathrm{e}^{-n\lambda} = \mathrm{e}^{-n\lambda}\exp\{t\ln(n\lambda)\}/t!,\ t=0,1,\cdots.$$

如令 $\eta = \ln(n\lambda)$, 则知 $T(X)$ 的分布是指数型分布, 且由于其自然参数空间

$$\mathcal{G} = \{\eta : \eta = \ln(n\lambda), \lambda > 0\} = \mathbb{R},$$

故由定理 1.4.3 知, 统计量 $T = \sum_{i=1}^{n} X_i$ 是充分完全的. □

充分完全统计量在参数估计中, 以及假设检验中均有很好的应用. 另外, 关于充分完全统计量的应用还有一个非常著名的 Basu(巴苏) 定理 (Basu, 1955). 为此, 我们先引入如下定义.

定义 1.4.5(辅助统计量) 设 $X_1, X_2, \cdots, X_n$ 为来自某参数分布的样本, 参数为 θ. 如果统计量 $T(X)$ 的分布与参数 θ 无关, 则称之为辅助统计量.

定理 1.4.4 (Basu 定理) 设 $X_1, X_2, \cdots, X_n$ 为来自某参数分布的样本, 参数为 θ. 设 $T(X)$ 为充分且有界完全统计量, $S(X)$ 为辅助统计量, 则统计量 $T(X)$ 与 $S(X)$ 独立.

证明 为证 T 与 S 独立, 仅需证明

$$P(S(X) \in B | T(X) = t) = P(S(X) \in B), \quad \forall B. \tag{$*1$}$$

因为 T 是充分统计量, S 是辅助统计量, 所以上式中左右端的概率均与 θ 无关. 如记 $C = \{X : S(X) \in B\}$, 则上式可以写为

$$P(X \in C | T = t) = P(X \in C),$$

它等价于下式

$$E[I_C(X)|T(X)=t] = P(X \in C) \xlongequal{\text{记为}} \alpha. \tag{$*2$}$$

其中 α 仅与 C 有关, 与 θ 无关. 为此, 我们令

$$h(T) = E[I_C(X)|T(X)=t] - \alpha.$$

显然, 它是一个统计量, 且可以求得其期望为零, 即

$$E_\theta(h(T)) = 0, \quad \forall \theta \in \Theta,$$

则因为 T 是完全统计量, 可知 $h(T) = 0$, 即 $(*2)$ 式成立. □

例 1.4.12　设 $X_1, X_2, \cdots, X_n$ 为来自 $N(\mu, \sigma^2)$ 的样本, 如果 $g(x_1, x_2, \cdots, x_n)$ 满足平移不变性, 即任给 c, $g(x_1 + c, x_2 + c, \cdots, x_n + c) = g(x_1, x_2, \cdots, x_n)$, 证明 $g(x_1, x_2, \cdots, x_n)$ 与 $\bar{X}$ 独立.

证明　当 σ 已知时, 易证 $\bar{X}$ 是 μ 的充分完全统计量. 由于 g 的平移不变性, 如取 $c = -\mu$, 则知 $g(X)$ 为辅助统计量, 故由 Basu 定理知二者独立. 由于上述对任给的 σ 都成立, 故结论得证. □

习题一

1. 证明例 1.2.3中的 (1.2.10) 式.

2. 设 X 的累积分布为 $F(x)$, 且 F 连续, 证明: $F(X) \sim U(0, 1)$.

3. 设 $X \sim U(0, 1)$, 且 F 为一累积分布函数. 定义 $Y = F^{-1}(X)$, 证明: Y 的累积分布为 F, 其中 $F^{-1}(t) = \inf\{x : F(x) \geqslant t\}$.

4. 设 X, Y 独立同分布, 证明: $E|X + Y| \geqslant E|X|$.

5. 设 X, Y 为两随机变量, 证明: (1)$E[E(X|Y)] = E(X)$. (2) 对于任一可测函数 g, 如 $E[X^2] < \infty, E[g^2(Y)] < \infty$, 则 $E[X - E(X|Y)]^2 = \min\limits_{g} E[X - g(Y)]^2$.

6. 设 X 为一随机变量, $g(x)$ 为给定函数, 且 $E[g(X) - \theta]^2$ 在 $\theta \in [a, b]$ 上存在. 定义

$$S(x) = \begin{cases} a, & g(x) < a, \\ g(x), & a \leqslant g(x) \leqslant b, \\ b, & g(x) > b. \end{cases}$$

证明: $E[S(X) - \theta]^2 \leqslant E[g(X) - \theta]^2, \quad \forall \theta \in [a, b]$.

7. 设 $E(X) = E(Y) = 0, E(X^2) = E(Y^2) = 1, E(XY) = \rho$. 证明如下的二元 Chebyshev(切比雪夫) 不等式: 对任一 $a > 0$,

$$P(\max\{|X|, |Y|\} \geqslant a) \leqslant \frac{1 + \sqrt{1 - \rho^2}}{a^2}.$$

8. 设 X, Y 独立, 且均服从标准正态分布, 求 $Z = Y/X$ 的概率密度.

9. 设 X_1, X_2, X_3 相互独立, 且服从 $N(0, \sigma^2)$. 证明: $(1)(X_1 + X_2X_3)/\sqrt{1 + X_2^2} \sim N(0, \sigma^2)$; (2) X_1/X_2 与 $X_1^2 + X_2^2 + X_3^2$ 独立.

10. 对于 n 个样本 $X_1, X_2, \cdots, X_n$, 记 $\bar{X}_m = \dfrac{1}{m}\sum\limits_{i=1}^{m} X_i, T_m = \sum\limits_{i=1}^{m}(X_i - \bar{X}_m)^2$. 证明:

$$T_m = T_{m-1} + \frac{m-1}{m}(X_m - \bar{X}_{m-1})^2.$$

11. 设 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的独立同分布样本, 对于 $m \leqslant n$, 记 $T_m = \sum_{i=1}^{m} X_i$. 证明: (X_i, T_m) 的分布与 i 无关.

12. 设 $X_1, X_2, \cdots, X_n$ 为来自两点分布 $B(1, \theta)$ 的独立同分布样本, 其中 $\theta \in (0, 1)$. 求在给定 $\sum_{i=1}^{n} X_i$ 下, $X_1(1 - X_2)$ 的条件分布.

13. 设 (X_1, X_2) 的概率分布为 $f_X(x_1, x_2)$. 记 $g(x) = (x_1, x_1+x_2), h(x) = (x_1/x_2, x_2)$. 证明: (1)$g(X)$ 的概率分布为 $f_g(x_1, y) = f_X(x_1, y - x_1)$; (2) $h(X)$ 的概率分布为 $f_h(z, x_2) = |x_2| f_X(zx_2, x_2)$; (3) 如 X_1, X_2 独立, 且均服从标准正态分布, 则 $Z = X_1/X_2 \sim C(0, 1)$, 其中 $C(0, 1)$ 为标准 Cauchy 分布.

14. 设 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的独立同分布样本, 且总体 X 关于常数 c 对称, 即 $X - c$ 与 $-X + c$ 同分布. 证明: (1) $E(X) = c$; (2) $\mathrm{Cov}(\bar{X}; S_n^2) = 0$; (3) 如设总体 X 连续, 且记 $f_i(x)$ 为第 i 个次序统计量的概率密度, 则 $f_i(x+c) = f_{n-i+1}(c - x)$, $\forall x \in \mathbb{R}, i = 1, 2, \cdots, n$.

15. 设 $X_1, X_2, \cdots, X_n$ 为来自总体 $N(\mu, 1)$ 的独立同分布样本, 以 $\Phi(x)$ 记标准正态分布的累积分布, 求 a, 使得 $E[a\Phi(\bar{X})] = \Phi(\mu)$.

16. 设 $X_1, X_2, \cdots, X_n$ 为来自指数分布 $Exp(\lambda)$ 的独立同分布样本, 记 $T_n = \sum_{i=1}^{n} X_i$, 以 $f_n(x; \lambda)$ 记 T_n 的概率密度, 证明:

$$f_n(x; \lambda) = \int_{\infty}^{\lambda} \lambda \mathrm{e}^{-(x-y)} f_{n-1}(y; \lambda) \mathrm{d}y,$$

并由此求 $\bar{X}$ 的概率密度.

17. 设 $X_1, X_2, \cdots, X_n$ 为来自总体累积分布为 F 的独立同分布样本, 对于 $1 \leqslant i < j \leqslant n$, 求 $X_{(j)} - X_{(i)}$ 的累积分布, 并由此给出极差 $R = X_{(n)} - X_{(1)}$ 的累积分布. 如总体 $X \sim U(0, 1)$, 则给出具体表达式. (提示: 令 $U = X_{(i)}, V = X_{(j)} - X_{(i)}$.)

18. 设 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的独立同分布样本, 且总体累积分布 $F(x)$ 连续. 记其次序统计量为 $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$, 定义

$$\bar{F}(x) = [F(x) - F(x_{(1)})]/[1 - F(x_{(1)})] I_{(x_{(1)}, \infty)}(x).$$

证明: 在给定 $X_{(1)} = x_{(1)}$ 的条件下, $X_{(2)}, X_{(3)}, \cdots, X_{(n)}$ 的条件分布即为从 $\bar{F}$ 中抽取的 $n - 1$ 个独立同分布样本的次序统计量的分布.

19. 设 $X_1, X_2, \cdots, X_n$ 为来自指数分布 $Exp(1)$ 的独立同分布样本, 证明: $nX_{(1)}$, $(n-1)[X_{(2)} - X_{(1)}], \cdots, (n - r + 1)[X_{(r)} - X_{(r-1)}], \cdots, X_{(n)} - X_{(n-1)}$ 仍为来自 $Exp(1)$ 的独立同分布样本.

20. 设 $X_1, X_2, \cdots, X_n$ 为来自总体累积分布为 F 的独立同分布样本, 且 F 连续,

证明:

$$E[F(X_{(i)})]=\frac{i}{n+1},\ \mathrm{Var}[F(X_{(i)})]=\frac{i(n+1-i)}{(n+1)^2(n+2)}.$$

21. 设 $X_1,X_2,\cdots,X_n$ 为来自 Weibull 分布 $W(\alpha,\beta)$ 的独立同分布样本, 证明: $X_{(1)}$ 仍服从 Weibull 分布.

22. 设 $X_1,X_2,\cdots,X_n$ 为来自总体 X 的独立同分布样本, 总体 X 服从指数分布 $Exp(a,b)$. 考虑定数截尾试验, 即对于事先给定的正整数 r, 当事件 $X_{(r)}$ 发生时试验停止. 记 $T_1=nX_{(1)},T_2=\sum\limits_{i=1}^{r}[X_{(i)}-X_{(1)}]+(n-r)X_{(r)}$, 证明: T_1,T_2 独立, 且 $T_1\sim Exp(na,b),2bT_2\sim\chi^2(2r-2)$.

23. 设 $X\sim\chi^2(n)$, 证明: 当 $n\to\infty$ 时, $(X-n)/\sqrt{2n}\xrightarrow{\mathrm{d}}N(0,1)$. 利用此结果, 给出 $\chi^2(n)$ 的 p 分位数 $\chi_p^2(n)$ 与标准正态分布的 p 分位数 u_p 间的近似关系.

24. 设 $X_1,X_2,\cdots,X_n$ 为来自某总体 $X\sim F$ 的独立同分布样本, 且 $E(X)=\mu$, $\mathrm{Var}(X)=\sigma^2$. 记 $T_n=\dfrac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$, 则在某些连续性假设下, 证明: T_n 的分布函数可展开成如下形式:

$$\Phi(x)-\phi(x)\left\{\frac{\gamma_1}{6\sqrt{n}}H_2(x)+\frac{1}{n}\left[\frac{\gamma_2}{24}H_3(x)+\frac{\gamma_1^2}{72}H_5(x)\right]+r_n\right\},$$

其中 γ_1,γ_2 分别为总体 X 的偏度与峰度, $r_n=o(n^{-1})$, $H_r(\cdot)$ 为 r 阶 Hermite(埃尔米特) 多项式, 其定义如下: $\dfrac{\mathrm{d}^r(\mathrm{e}^{-x^2/2})}{\mathrm{d}x^r}=(-1)^rH_r(x)\mathrm{e}^{-x^2/2}$. 上述展开式被称为 Edgeworth(埃奇沃思) 展开. 由上述 Edgeworth 展开, 请验证: T_n 的 α 分位数可近似为

$$u_\alpha+\frac{\gamma_1}{6\sqrt{n}}(u_\alpha^2-1)+\frac{\gamma_2}{24n}(u_\alpha^3-3u_\alpha)-\frac{\gamma_1^2}{36n}(2u_\alpha^3-5u_\alpha)+r_n,$$

其中 u_α 为标准正态分布的 α 分位数, 且上述展开式被称为 Cornish-Fisher(科尼什–费希尔) 展开.

25. 设 $X_1,X_2,\cdots,X_m$ 和 $Y_1,Y_2,\cdots,Y_n$ 为分别来自正态总体 $N(\mu,\sigma_1^2)$ 和 $N(\mu,\sigma_2^2)$ 的独立同分布样本, 且全样本独立, 记 S_{1x}^2,S_{2y}^2 分别为两组样本方差, $\alpha_1=S_{1x}^2/(S_{1x}^2+S_{2y}^2),\alpha_2=S_{2y}^2/(S_{1x}^2+S_{2y}^2)$. 求 $\alpha_1\bar{X}+\alpha_2\bar{Y}$ 的期望.

26. 设 $U_1,U_2,\cdots,U_n$ 为来自 $U(0,1)$ 的独立同分布样本的次序统计量, $X_1,X_2,\cdots,X_n$ 为来自指数分布 $Exp(1)$ 的独立同分布样本, $Y_r=\sum\limits_{i=1}^{r}X_i/(n-i+1)$. 证明: $(-\ln U_1,-\ln U_2,\cdots,-\ln U_n)$ 与 $(Y_1,Y_2,\cdots,Y_n)$ 依分布相等.

27. 设 X_1,X_2 独立同分布于 $N(0,1)$, 证明: $X_1/|X_2|$ 服从标准 Cauchy 分布.

28. 设 $X_1,X_2,\cdots,X_n$ 为来自标准 Cauchy 分布的独立同分布样本, 证明: $\bar{X}$ 也服从标准 Cauchy 分布.

29. 设 $X_1, X_2, \cdots, X_n$ 为来自 $N(\mu, \sigma^2)$ 的样本, 且 $\mathrm{Cov}(X_i, X_j) = \rho\sigma^2$, 其中 $|\rho| < 1$. 证明: $\bar{X}$ 与 S_n^2 独立, 且

$$\bar{X} \sim N(\mu, (1 + (n-1)\rho)\sigma^2/n), (n-1)S_n^2/[(1-\rho)\sigma^2] \sim \chi^2(n-1).$$

30. 设 $X_1, X_2, \cdots, X_n$ 相互独立, 且 $X_i \sim \chi^2(n_i)$. 记

$$U_i = \frac{\sum\limits_{k=1}^{i} X_k}{\sum\limits_{k=1}^{i+1} X_k}, \ i = 1, 2, \cdots, n-1.$$

证明: $U_1, U_2, \cdots, U_{n-1}$ 相互独立, 且 $U_i \sim B\left(\sum\limits_{k=1}^{i} n_k/2, n_{i+1}/2\right)$.

31. 设 $X_1, X_2, \cdots, X_n$ 为来自 β 分布 $Beta(\theta, 1)$ 的独立同分布样本, 证明:

$$-2\theta \sum_{i=1}^{n} \ln X_i \sim \chi^2(2n).$$

32. 设 $X_1, X_2, \cdots, X_n$ 是 n 个相互独立的随机变量, 且 $X_i \sim N(0, \sigma_i^2)$. 记

$$Y = \frac{\sum\limits_{i=1}^{n} X_i/\sigma_i^2}{\sum\limits_{i=1}^{n} 1/\sigma_i^2}, \quad Z = \sum_{i=1}^{n} \frac{(X_i - Y)^2}{\sigma_i^2},$$

试证明 $Z \sim \chi^2(n-1)$.

33. 设 $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\boldsymbol{\mu}, \Sigma)$, 其中 $\boldsymbol{\mu} = (\mu_1, \mu_2)^{\mathrm{T}}$, $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. 证明: (1) $X + Y$ 与 $X - Y$ 独立, 当且仅当 $\sigma_1^2 = \sigma_2^2$; (2) 如果 $\sigma_1\sigma_2 > 0, |\rho| < 1$, 则

$$\frac{1}{1-\rho^2}\left[\frac{(X-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(X-\mu_1)(Y-\mu_2)}{\sigma_1\sigma_2} + \frac{(Y-\mu_2)^2}{\sigma_2^2}\right] \sim \chi^2(2).$$

34. 设 $X \sim N(0,1), T \sim t(n)$, 证明: 存在一个正数 t_0, 使得

$$P(|T| \geqslant t_0) \geqslant P(|X| \geqslant t_0).$$

35. 设 $X_1, X_2, \cdots, X_n$ 为来自正态总体 $N(\mu, \sigma^2)$ 的独立同分布样本, 记 $\tau = \dfrac{X_1 - \bar{X}}{S_n}\sqrt{\dfrac{n}{n-1}}$, 证明:

$$T = \frac{\sqrt{n-2}\tau}{\sqrt{n-1-\tau^2}} \sim t(n-2).$$

36. 设 $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, \cdots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ 为来自二元正态分布 $N(\boldsymbol{\mu}, \Sigma)$ 的独立同分布样本, 其中 $\boldsymbol{\mu} = (\mu_1, \mu_2)^{\mathrm{T}}$, $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. 定义两个分量间的相关系数为

$$r(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right]^{1/2}},$$

证明: 如果 $\rho = 0$, 则统计量

$$T = \sqrt{n-2}\frac{r(X,Y)}{\sqrt{1 - r^2(X,Y)}} \sim t(n-2).$$

37. 设 $X \sim F(n,m)$, 证明: $nX/(m+nX) \sim Beta(n/2, m/2)$.

38. 设 $X_1, X_2, \cdots, X_n$ 为来自 Weibull 分布的独立同分布样本, X_1 的概率密度为

$$f(x;\lambda) = \alpha\lambda x^{\alpha-1}\mathrm{e}^{-\lambda x^{\alpha}} I_{[0,\infty)}(x).$$

证明: $-2\lambda\sum_{i=1}^{n} X_i^{\alpha} \sim \chi^2(2n)$.

39. 设 $X_1, X_2, \cdots, X_n$ 为来自正态总体的独立同分布样本, 证明: $(X_1 - \bar{X})/S_n$ 与 $(n-1)Y/\sqrt{n(Y^2+n-2)}$ 分布相等, 其中 $Y \sim t(n-2)$, $\bar{X}, S_n^2$ 分别为样本均值和方差.

40. 设 $X_1, X_2, \cdots, X_n$ 和 $Y_1, Y_2, \cdots, Y_n$ 为分别来自 Γ 分布 $\Gamma(\alpha, \lambda_1)$ 和 $\Gamma(\alpha, \lambda_2)$ 的独立同分布样本, 且独立, 求 $\bar{X}/\bar{Y}$ 的分布.

41. 设 $X \sim \Gamma(\alpha, 1), Y \sim \Gamma(\alpha + 1/2, 1)$, 且二者独立, 证明: $2\sqrt{XY} \sim \Gamma(2\alpha, 1)$.

42. 设 $X \sim N(\mu, 1)$, 证明: X^2 的累积分布函数是 μ^2 的递减函数.

43. 设 T, S 是两个统计量, 且 $S = \phi(T)$, 其中 ϕ 是一可测函数. 证明: (1) 如果 T 是完全的, 则 S 也是完全的; (2) 如果 T 是充分完全的, 且 ϕ 是一一映射, 则 S 也是充分完全的.

44. 设 $X_1, X_2, \cdots, X_n$ 为来自 Γ 分布 $\Gamma(\alpha, \lambda)$ 的独立同分布样本. 求 (1) 当 α 已知时, λ 的充分统计量; (2) 当 λ 已知时, α 的充分统计量.

45. 设 $X_1, X_2, \cdots, X_n$ 为来自 $N(\theta, \theta^2)(\theta > 0)$ 的独立同分布样本. 问 $\bar{X}$ 是否仍为充分统计量?

46. 设样本 $Y_1, Y_2, \cdots, Y_n$ 相互独立, 且 Y_i 服从 Poisson 分布 $P(\lambda_i)$. 另外, 假设

$$\ln\lambda_i = \alpha + \beta x_i, \quad i = 1, 2, \cdots, n,$$

其中 x_i 为已知常数. 求 α, β 的最小充分统计量.

47. 设 $X_1, X_2, \cdots, X_n$ 为来自 $N(0, \sigma^2)$ 的独立同分布样本, 其中 $\sigma > 0$ 为参数. 证明: 统计量 $\left(\sum\limits_{i=1}^{n} X_i, \sum\limits_{i=1}^{n} X_i^2\right)$ 是充分的但不是有界完全的.

48. 设 $X_1, X_2, \cdots, X_n$ 为来自均匀分布 $U(\theta - 1/2, \theta + 1/2)$ 的独立同分布样本, 其中 $\theta \in \mathbb{R}$ 为参数. 证明: $(X_{(1)}, X_{(n)})$ 不是完全统计量.

49. 设 $X_1, X_2, \cdots, X_n$ 为来自均匀分布 $U(-\theta, \theta)$ 的独立同分布样本, 其中 $\theta > 0$ 为参数. 记 $T_1 = X_{(n)} - X_{(1)}, T_2 = \max\{X_{(n)}, -X_{(1)}\}$. (1) 证明 T_2 是充分统计量, 且 T_1/T_2 与 T_2 独立; (2) 求 $T_1/(2T_2)$ 的分布.

50. 设 $Y_1, Y_2, \cdots, Y_n$ 相互独立, 且 $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$, 其中 $x_1, x_2, \cdots, x_n$ 为给定的常数, $\alpha, \beta, \sigma > 0$ 为参数. 求充分与完全统计量.

51. 设 $X_1, X_2, \cdots, X_m$ 和 $Y_1, Y_2, \cdots, Y_n$ 为分别来自总体 $N(\mu, \sigma_1^2)$ 和 $N(\mu, \sigma_2^2)$ 的独立同分布样本, 且全体独立. 当 $\sigma_1^2/\sigma_2^2 = c$ 为常数时, 求充分完全统计量.

52. 设 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的独立同分布样本, 总体 X 的概率密度为 $\theta^{-1}\mathrm{e}^{-(x-\theta)/\theta} I_{[\theta,\infty)}(x)$, 其中 $\theta > 0$ 为参数. (1) 求 θ 的最小充分统计量; (2) 证明上述最小充分统计量是完全的.

53. 设 $X_1, X_2, \cdots, X_n$ 为来自均匀分布总体 $U(\theta_1, \theta_2)$ 的独立同分布样本, 证明: $(X_{(1)}, X_{(n)})$ 是完全统计量.

54. 设 $X_1, X_2, \cdots, X_n$ 相互独立, g, h 为两个可测函数, $1 \leqslant k < n$, 证明: $g(X_1, X_2, \cdots, X_k)$ 与 $h(X_{k+1}, X_{k+2}, \cdots, X_n)$ 独立.

55. 设 $X_1, X_2, \cdots, X_n$ 为来自 Γ 分布 $\Gamma(\alpha, \lambda)$ 的独立同分布样本, 证明: $\sum\limits_{i=1}^{n} X_i$ 与 $\sum\limits_{i=1}^{n} [\ln X_i - \ln X_{(1)}]$ 独立.

56. 设 $X_1, X_2, \cdots, X_n$ 为来自总体 $N(0, \sigma^2)$ 的独立同分布样本, 证明: $\bar{X}/S_n$ 与 $\sum\limits_{i=1}^{n} X_i^2$ 独立.

57. 设 X, Y 独立且均服从 $N(0, \sigma^2)$, 证明: $X^2 + Y^2$ 与 $X/\sqrt{X^2 + Y^2}$ 独立.

58. 设 $X_1, X_2, \cdots, X_n$ 为来自均匀分布 $U(a, b)$ 的独立同分布样本, $-\infty < a < b < \infty$. 证明: 对于任何 a 和 b, $(X_{(i)} - X_{(1)})/(X_{(n)} - X_{(1)}), i = 2, 3, \cdots, n-1$ 独立于 $(X_{(1)}, X_{(n)})$.

59. 设 $X_1, X_2, \cdots, X_n$ 为来自均匀分布总体 $U(0, \theta)$ 的独立同分布样本, 证明: $X_1/X_{(n)}$ 与 $X_{(n)}$ 独立.

60. 设 $X_i, i = 1, 2, 3$ 独立且均服从指数分布 $Exp(\lambda)$. 记 $Y_1 = X_1 + X_2 + X_3$, $Y_2 = X_1/(X_1 + X_2), Y_3 = (X_1 + X_2)/(X_1 + X_2 + X_3)$. Y_1, Y_2, Y_3 独立吗?

61. 设 $X_1, X_2, \cdots, X_n$ 为来自指数分布 $Exp(\lambda)$ 的独立同分布样本, 其中 $\lambda > 0$ 为参数. 证明: $\bar{X}$ 与 $X_{(1)}/X_{(n)}$ 独立.

62. 设样本 $X_1, X_2, \cdots, X_n$ 相互独立, 且 $X_i \sim N(t_i\theta, 1)$, 其中 $t_1, t_2, \cdots, t_n$ 是非零的已知常数, θ 为参数. (1) 求 θ 的充分完全统计量; (2) 记 $\hat{\theta} = \sum\limits_{i=1}^{n} t_i X_i / \sum\limits_{i=1}^{n} t_i^2$, 证明: $\hat{\theta}$ 与 $\sum\limits_{i=1}^{n}(X_i - t_i\hat{\theta})^2$ 独立.

63. 设 $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, \cdots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ 为来自二元正态分布 $N(\boldsymbol{\mu}, \Sigma)$ 的独立同分布样本, 其中 $\boldsymbol{\mu} = (\mu_1, \mu_2)^{\mathrm{T}}$, $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. 分别记两组样本的均值与方差为 $\bar{X}, S_x^2$ 和 $\bar{Y}, S_y^2$, 再记 $S_{xy}^2 = \sum\limits_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})/n$. 证明: $(\bar{X}, \bar{Y})$ 与 (S_x^2, S_y^2, S_{xy}^2) 独立.