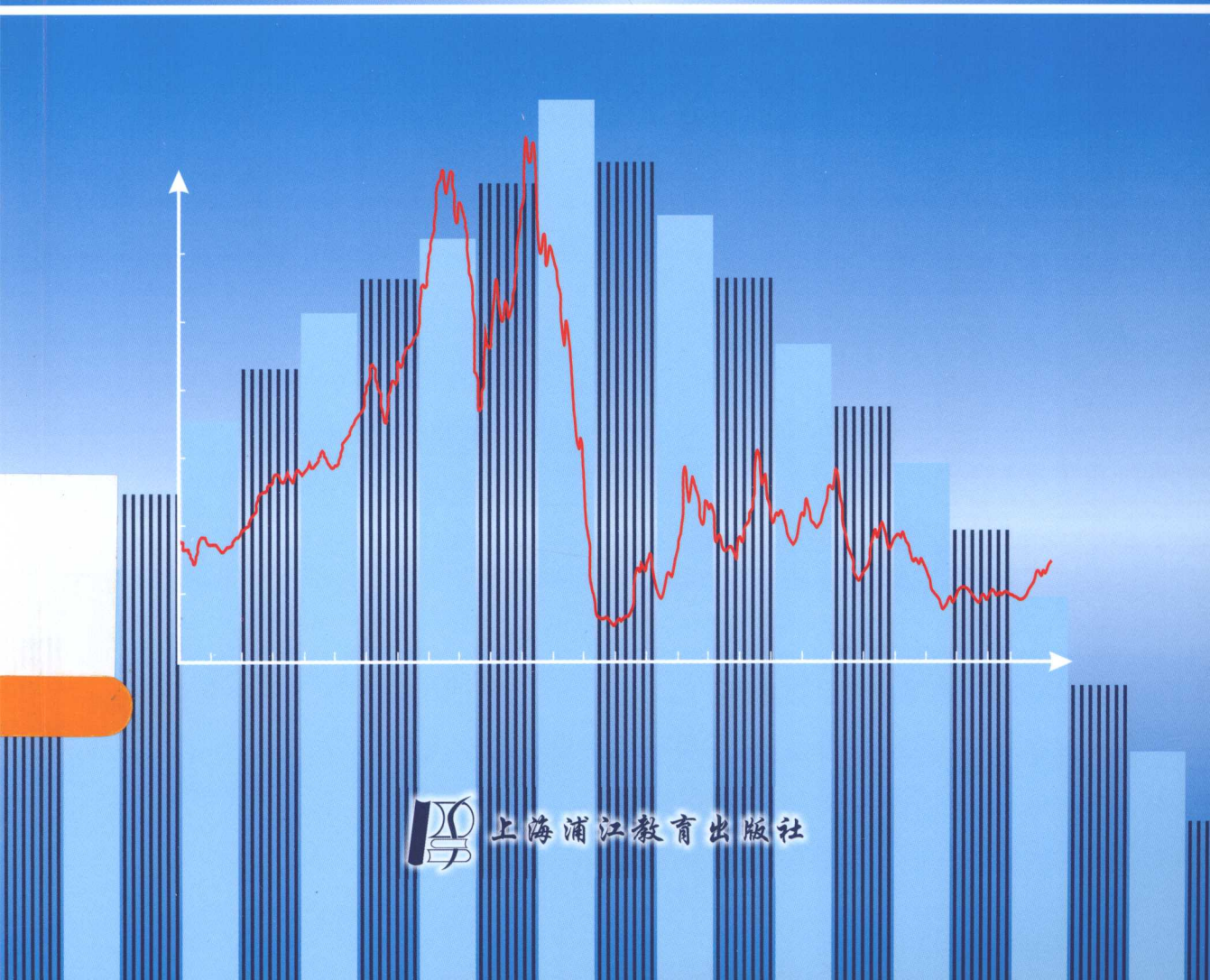
应用统计分析方法

李序颖 编著



上海浦江教育出版社

应用统计分析方法

李序颖　编著

上海浦江教育出版社

内 容 提 要

本书根据编者多年来积累的研究生教学经验编写而成。本书共分 11 章：第 1 章介绍统计学的任务、性质；第 2 章回顾概率统计的基础知识；第 3 章介绍主要的抽样方法；第 4 章介绍回归分析方法；第 5 章介绍定性变量的若干统计分析方法；第 6 和第 7 章介绍聚类分析、判别分析、主成分分析、因子分析和典型相关分析等多元统计分析方法；第 8 章介绍时间序列分析方法；第 9 章介绍空间截面数据的分析方法；第 10 章介绍面板数据的基本模型；第 11 章应用实例介绍经济管理问题中的统计方法应用。

本书可作为经济管理类研究生学习统计分析方法课程的教材，也可作为相关领域实际工作者的参考书。

图书在版编目（CIP）数据

应用统计分析方法 / 李序颖编著. — 上海：上海浦江教育出版社，2012.3
ISBN 978-7-81121-215-0

Ⅰ. ①应… Ⅱ. ①李… Ⅲ. ①统计方法－研究生－教材 Ⅳ. ①C81

中国版本图书馆 CIP 数据核字(2012)第 032408 号

上海浦江教育出版社出版

社址：上海海港大道 1550 号上海海事大学校内 邮政编码：201306
电话：(021)51322547(发行) 38284923(总编室) 38284916(传真)
E-mail: cbs@shmtu.edu.cn URL: http://www.pujiangpress.cn
上海图宇印刷有限公司印装 上海浦江教育出版社发行
幅面尺寸：185 mm×260 mm 印张：15.25 字数：300 千字
2012 年 3 月第 1 版 2012 年 4 月第 1 次印刷
责任编辑：谢 尘 封面设计：赵宏义
定价：35.00 元

前　言

统计学方法是研究生开展研究工作必备的工具，统计学类课程作为最重要的方法类课程之一，是经济管理类专业研究生的必修课程。根据作者多年的教学体验，愈来愈多的学生对统计学方法的学习充满热情，但对统计学方法有时又感觉内容太多，无从下手，为此，本书希望围绕数据处理问题来安排，在章节的编排上，从搜集数据的抽样调查方法，再到常用的数据分析方法。

第 1 章对统计学进行概要介绍。第 2 章对概率统计基础知识的主要内容进行回顾，这一章对统计学基础知识进行简要概括，并加入了本科阶段可能没有接触但本书以后章节需要掌握的若干基础知识，例如极大似然估计、广义矩估计、大样本分布理论等内容。对于熟悉本科阶段统计学知识的学生，本章大部分内容可以跳过。

第 3 章对抽样的主要方法进行介绍，抽样调查在我国的应用愈来愈广泛，它是科学地获取数据最重要的方法。本章介绍的刀切法和自助法等再抽样技术，可以用于估计复杂样本的方差，也可以用于研究估计量性质。本章有些符号表示与其他各章不同，学习时应加以注意，对于只关心数据分析方法的学生，可以跳过本章。

第 4 章介绍回归分析方法。回归模型是经济管理问题研究中使用最广泛的模型，这一章介绍的内容是第 5 章部分内容、第 8 章、第 9 章、第 10 章的基础。本章涉及回归模型、模型参数估计以及估计量的性质、统计检验、模型诊断等问题，本章在介绍相关结论时尽可能只列出必要的假设条件。

第 5 章介绍属性变量的分析方法，本书只在这一章介绍属性变量的常用分析方法，对于空间数据、时间序列数据、面板数据中的属性变量问题，则属于比较专门化的问题，本书没有介绍。

第 6 章介绍聚类与判别分析；第 7 章主要介绍主成分分析与因子分析。第 6 章和第 7 章的内容都是多元统计分析方法，它们是多维数据的分析方法。

第 8 章介绍线性时间序列数据的分析模型。 20 世纪 70 年代我国时间序列分析研究在理论和应用方面取得了若干成果， 20 世纪 90 年代以来，时间序列分析方法在我国金融领域得以广泛应用， ENGLE 和 GRANGER 在经济时间序列分析领域取得的成果获得 2003 年诺贝尔经济学奖，吸引了更多学者参与时间序列分析方面的应用研究。

第 9 章介绍用于研究空间截面数据的空间回归模型和空间差异性模型。经济问题尤其是区域经济问题中，数据中存在空间方向相关的问题，近几年来愈来愈受到学界的关注，但在教材中包括此内容的还较少。

第 10 章简要介绍面板数据 (也称截面时间序列数据) 的基本模型及参数估计问题。随着我国统计工作的积累，获得面板形式的数据变得容易，针对这类实际数据的研究工作将愈来愈多。

统计分析方法很多，本书主要介绍常用的统计分析方法，力求内容广泛、方法实用，同时有所侧重。本书将一些结论的推导过程作为附录放到每章的章后，供有兴趣的学生参考。

本书每一章内容都可以形成专门的教材，因此，本书希望重点放在统计思想和方法的介绍上，建议学生在学习相关内容的同时，阅读每章之后列出的参考文献，并利用统

计软件进行计算机模拟操作。本书部分章节附录附有模拟操作的 R 软件代码，供 R 软件初学者参考。

对于本书介绍的每一种方法，要找到恰好能够使用相应方法的实际问题并不容易，因此，本书大多数方法介绍之后未安排相关案例。本书在第 11 章介绍了一些实际研究工作中数据处理的实例，仅供读者参考。

本书在编写过程中，参考了国内外相关研究工作以及教材论著的成果，未能一一列出，在此谨向有关资料的原作者表示深深的谢意！

本书可以作为经济管理，尤其是航运经济管理类研究生的教材，需要学生已经掌握高等数学、线性代数的基本知识，同时修过本科的统计学课程，教师可以根据学生的情况，对教学内容加以取舍。本书也可以作为相关研究领域科研工作人员的参考资料。

本书的出版获得上海海事大学研究生教材建设项目资助，在此，对资助单位表示感谢！本书编写过程中，得到上海海事大学经济管理学院曲林迟教授、周溪召教授和上海浦江教育出版社袁林新编审的热情鼓励和支持。周溪召教授仔细阅读了本书的初稿，并提出许多建设性的意见。研究生顾贤斌等参与了本书文字校对工作。在此对大家的帮助表示感谢！

限于作者水平有限，书中肯定存在错误、谬误之处，恳请读者批评、指正。

李序颖

2012 年 2 月，上海

目　　录

第1章 概 述

1.1 统计学的任务

尽管不同的书中对统计学定义的表述有些许不同，但收集、分析、推断是其共同的关键词。如陈希孺 (1989) 引述《中国大百科全书 · 数学卷》对数理统计学的定义：统计学是一门科学，它研究怎样以有效的方式收集、整理、分析带随机性的数据，并在此基础上，对所研究的问题作出统计性的推断，直至对可能作出的决策提供依据或建议。《大英百科全书》将统计学定义为收集数据、分析数据，并根据数据进行推断的艺术与科学。下面以调查某一年全国大学毕业生的起薪水平为例，明确统计学的任务。

以全国某一年大学毕业生作为研究对象 (以下称其为“总体”)，通常先调查当年部分大学毕业生 (以下称这部分人为“样本”) 的起薪。这个调查工作需要设计抽样方案，也就是如何抽出这部分人群。

由于所记录样本的起薪只是一堆数据，还不能表达样本的起薪水平信息，因此进一步的工作是对数据进行整理，即将数据分门别类，用醒目且便于使用的形式表达。如，计算样本的平均起薪、最高起薪、最低起薪、起薪水平的差异程度 (如标准差、全距) 等，或按毕业于“985”学校、“211”学校、“其他”学校等对毕业生进行分类，用图表表示不同类型学校毕业生的起薪水平。对数据的整理过程实际上也是一个分析的过程，如何有效地用图表反映所收集数据中的信息，要用到描述统计学的方法。

由于最终想得到的是总体的起薪水平，而不是样本的起薪水平，还需要根据样本数据对总体特征进行统计推断，这主要包括对总体特征的估计和检验。如，如何根据样本数据对总体平均起薪进行估计 (用样本的平均起薪去估计总体的平均起薪似乎是一个比较自然的想法)，不同类型学校毕业生的起薪水平是否有差异等，这些都是统计推断的工作。

1.2 统计学的性质

因为统计学用到许多数学工具，许多初学者认为统计学就是数学，但实际上，国际上普遍认为统计学与数学是相互独立的学科。

从研究方法来说，统计学是归纳式的。例如，研究吸烟与某些呼吸道疾病的关系，统计学通过大量观察，以确认吸烟者患病的可能性是否远远大于不吸烟者，如果是，则吸烟与某些呼吸道疾病有关系。

数学则是演绎式的。例如，研究等腰三角形的底角是否相等：数学的研究方法是从“等腰”的条件和“几何公理”出发，证明出等腰三角形底角相等；而统计学的研究方法是寻找许多大小、形状不一的等腰三角形，量测其底角，根据所获得的资料推断等腰三角形的底角是否相等。

从研究对象来说，统计学研究的是数据，希望从数据中获得信息，而数学则是研究

数与形本身，如歌德巴赫猜想、几何问题等。

用统计方法处理的数据，通常具有随机性 (random)。如 1+1=2，数学上通常是

“绝对的 1” + “绝对的 1” = “绝对的 2”

而在统计学上则是

“大概的 1” + “大概的 1” = “大概的 2”

下面以测量人体身高为例来说明数据中随机性的来源。

第一种情形，量测误差。由于量测有误差，且误差是随机的，因此得不到确切的身高。为了得到准确的人体身高，可以多量测几次，用平均值估计身高。这时推断的随机性来自于量测误差。

第二种情形，抽样误差。假设测量全国成年男子的平均身高，且量测是无误差的，则通过量测全体人员的身高并求平均值就可以得到平均身高，但这样做，量测与计算工作量太大。统计学采取的方法是从全体人员中随机抽出一部分量测其身高，并用这部分人员的平均身高估计全国成年男子的平均身高。这时推断的随机性来自于被抽出人员的随机性，称其为抽样误差。

由此看出，即使量测无误差，但用部分个体推断总体时，由于部分个体的抽出是随机的，因此对总体的推断存在随机误差。对第一种情形，量测有误差，若设想无穷次量测的平均可以得到精确值，则有限次量测只是无限次量测的一部分，从这个意义上说，它也可以与第二种情形统一起来。但下面还是将其区分开来，即实际工作中，随机性来源于量测误差，或者来源于抽样误差，或者既包括量测误差又包括抽样误差。

1.3 统计学的历史

统计学与人类文明共生，统计思想方法是人类文明的一个组成部分。有学者认为，统计学的历史可以追溯到五千年前中国先人的“结绳记事”。

早期的统计活动主要是收集、整理数据，如中国古代关于钱粮、户口的记载，皇帝的起居注等。这一阶段只是对有关事情的记录，还没有超出数据范围的推断工作，因此不是现代意义上的统计学。

陈希孺 (1989) 将现代统计学的发展划分为三个阶段：

第一阶段始于 18 世纪末 19 世纪初，以高斯 (C. F. GAUSS) 引进正态分布、使用最小二乘估计方法为标志。这一时期，虽然统计学理论发展不快，但也出现了一些奠基性的工作。如贝叶斯提出的关于统计推断的一种方法论，发展成为现在的贝叶斯学派，贝叶斯估计成为统计推断最重要的方法之一；又如高尔登 (F. GALTON) 提出“回归”一词，而回归分析是研究带随机性的变量之间关系的最重要方法。

第二阶段始于 19 世纪末，至第二次世界大战结束。这一时期是统计学发展最快的时期，统计学的主要分支形成于这一时期，CRAMER(1946) 的《统计学数学方法》是统计学成为一门成熟学科的标志。这一阶段的重要人物有 R. A. FISHER，K. PEARSON，J. NEYMAN 和 E. PEARSON 等。我们将在本书学习 FISHER 的极大似然估计、方差分析、多元分析、相关回归等；K. PEARSON 的矩估计法、χ^2 检验；NYEMAN 的区间估计；NEYMAN 和 E. PEARSON 的假设检验等内容。

第三阶段始于第二次世界大战结束，直到现在。这一阶段统计学向深度和广度方向

发展。这一阶段统计学的应用和发展得益于计算机的普及。计算机解决了数据计算的问题，这使得大量数据尤其是高维数据的分析成为可能，而计算机的普及使这些方法得以广泛应用。另一方面，计算机的应用又为统计学的研究提供了“随机模拟”的新途径。

1.4 关于本书的学习

本书要求学生已学过本科的统计学课程，对统计学有一定了解。没有学习过本科统计学课程的学生，可以通过第 2 章及所附参考文献进行自学。

统计学类课程是经济管理类研究生最重要的方法课之一。学习统计学，不应停留在记住一些公式并加以套用的水平上，而应正确理解所使用的统计方法并在实际中加以合理使用。

学习了本书介绍的各类统计方法后，建议利用统计软件做一些模拟练习，这对加深理解统计方法、实际感受数据处理有好处。目前统计软件有很多，比较有名的如 SAS，S+ 和 R 等，R 软件作为免费软件，与 S+ 的代码大部分相同，目前受到国内统计教育界的重视，相应的教材也已经出版若干，如薛毅等 (2007)、汤银才 (2008)。R 软件可以通过 R 官方网站 www.r-project.org 下载。

在学习统计软件的同时，要注意避免陷入误区，即认为有了电脑软件，用不着学习统计理论与方法。要知道，统计分析方法的使用很多时候是有条件的，许多统计软件尤其是“傻瓜”软件是依据标准教科书编写的，在处理实际问题时，软件的确带来了方便，但数据能否满足所用方法的假设条件，需要实际工作者加以研究，而这就要求实际工作者熟悉所用的理论、方法。而对于 SAS，S+ 和 R 等需要编程的软件，在调用各种命令、函数、模块时涉及许多选项，这些选项如何设定、为什么这样设定，也需要实际工作者熟悉相关的理论和方法。

实际上，在获得数据后，如何对数据进行“清洗”，使其符合统计方法的使用条件，这是统计学作为“艺术”之所在，是比本书要学习的具体方法难得多的事情，也正是“研究”工作的价值之所在。

参考文献

[1] 陈希孺. 统计学概貌[M]. 北京：科学技术文献出版社， 1989.
[2] 吴喜之. 统计学到底是什么？一个本不应成为问题的问题[J]. 中国统计，1997(12): 30-31.
[3] 何晓群. 现代统计分析方法与应用[M]. 北京：中国人民大学出版社， 1998.

第 2 章　概率统计基础知识

2.1　随机变量及其数字特征

2.1.1　随机变量及其概率分布

2.1.1.1　随机变量

在概率论中，一般将对自然现象的观察或科学试验统称为试验。如果一个试验可以在相同条件下重复进行、每次试验的可能结果不止一个且能事先明确试验的所有可能结果、进行一次试验之前不能确定哪一个结果会出现，则称这个试验为随机试验。

例如，体育比赛开始前，裁判抛一枚硬币决定谁先发球，这个试验的结果可能是抛出硬币后出现 4 种情况：落下出现正面、背面、站立，一直向上未落下。又如，测量一个人的身高、体重，测量结果有无数种可能，但均在一个可预知的范围内。

将随机试验的结果对应于单一的数值 X，且对应是一对一的，则结果变量 X 是一个随机变量，在试验完成之前，不能确定 X 的取值。如抛硬币的例子中，将可能的试验结果用数字 1，2，3，4 来对应，则随机试验的结果就是 1，2，3，4 中的一种。在观测人体身高、体重的例子中，记 X_1 和 X_2 分别代表人体的身高和体重，则其取值不妨设定在 (0 m,3 m) 和 (0 kg,1 000 kg) 的范围内。

若随机试验的结果集合是有限的或可数的，则随机变量是离散的。若随机试验的结果集合是无限可分因而不可数的，则随机变量是连续的。在抛硬币的例子中，随机变量 X 是离散的；测量人体身高、体重的例子中，随机变量 X_1 和 X_2 都是连续的。

随机变量又分为一维的和多维的。如抛硬币的例子中，随机变量 X 是一维的；在测量人体身高、体重的例子中，随机变量 $X = (X_1, X_2)'$ 是二维的。

虽然在试验完成之前，不能确切地知道随机变量的取值，但对随机变量取某个值的可能性大小，可以用概率描述。如硬币若是均匀的，则人们相信出现正面和出现背面的概率应该各为 1/2 。对于某些事件，它可能发生，但其概率为 0，如抛硬币时，试验结果为“站立”这一事件是可能发生的，但通常人们认为其概率为 0。而有些事件是不可能事件，如“一直向上未落下”是不可能事件，其概率为 0。

2.1.1.2　随机变量的概率分布

对于离散型随机变量 X 所取的一系列值 x 及其相应的概率称为概率分布 $p(x)$，即

$$p(x) = P(X = x) \tag{2.1.1}$$

由概率的性质，要求：

(1) $0 \leqslant p(x) \leqslant 1$ (2.1.2)

(2) $\sum_{x} p(x) = 1$ (2.1.3)

对于连续型随机变量 X，与任一特定值 x 相对应的概率为 0。定义概率密度函数 (PDF) 为 $f(x)$，它满足：

(1) $f(x) \geqslant 0$ (2.1.4)

(2) $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$ (2.1.5)

(3) $P(a < X \leqslant b) = \int_{a}^{b} f(x)\mathrm{d}x$ (2.1.6)

对任意随机变量 X，称 $F(x)$ 为累积分布函数 (CDF)：

$$F(x) = P(X \leqslant x) = \begin{cases} \sum_{X \leqslant x} p(X)\ , & \text{若}X\text{是离散的} \\ \int_{-\infty}^{x} f(t)\mathrm{d}t\ , & \text{若}X\text{是连续的} \end{cases} \tag{2.1.7}$$

$F(x)$ 应满足下列性质：

(1) $0 \leqslant F(x) \leqslant 1$

(2) 若 $a < b$，则 $F(a) \leqslant F(b)$，且有 $P(a < x \leqslant b) = F(b) - F(a)$

(3) $F(-\infty) = 0$

(4) $F(+\infty) = 1$

2.1.2　随机变量的数字特征

2.1.2.1　期望值

定义：一个随机变量 X，若

$$\mu = \begin{cases} \sum_{x} xp(x)\ , & \text{若}X\text{是离散的} \\ \int_{x} xf(x)\mathrm{d}x\ , & \text{若}X\text{是连续的} \end{cases}$$

存在，则 X 的期望值或均值 (Mean) 为

$$E(X) = \mu \tag{2.1.8}$$

性质：若 a 和 b 为常数，且随机变量 X 的期望值为 μ，则

$$E(a + bX) = a + b\,E(X) = a + b\,\mu \tag{2.1.9}$$

特别地，常数 a 的期望值就是 a。

2.1.2.2　方差与标准差

定义：一个随机变量 X 的方差 (Variance) 为

$$\mathrm{Var}(X) = E(X - \mu)^2 = \sigma^2 \tag{2.1.10}$$

若 $E(X-\mu)^2$ 存在。有时也用 $D(X)$ 表示随机变量 X 的方差。方差的一个简化计算式为

$$\mathrm{Var}(X) = E(X^2) - \mu^2 \tag{2.1.11}$$

定义：一个随机变量 X 的标准差 (Standard Deviation) 为

$$\sigma = \sqrt{\sigma^2} \tag{2.1.12}$$

性质：若 a 和 b 为常数，X 为随机变量，其方差为 σ^2，则

$$\mathrm{Var}(a+bX) = b^2\,\mathrm{Var}(X) = b^2\,\sigma^2 \tag{2.1.13}$$

注意，对于常数 a，其方差 $\mathrm{Var}(a) = 0$。

定义：若随机变量 X 有 $E(X) = \mu$，$\mathrm{Var}(X) = \sigma^2$，则称 Cv 为 X 的变异系数

$$Cv = \frac{\sigma}{\mu} \tag{2.1.14}$$

2.1.2.3 协方差

定义：两个随机变量 X 与 Y 之间的协方差 (Covariance) 为

$$\mathrm{Cov}(X,Y) = E(X-\mu_x)(Y-\mu_y) = E(XY) - \mu_x\,\mu_y \tag{2.1.15}$$

两个随机变量的协方差度量了它们相互变化的趋势。协方差大于 0，表示两个随机变量的取值有同方向的趋势，换句话说，就是两个随机变量的取值有“同大同小的趋势”；如果协方差小于 0，则表示它们的取值有相反的趋势；如果协方差等于 0，则称两个随机变量不相关，这时两个随机变量之间没有线性关系。

若两个随机变量 X 与 Y 不相关，则由

$$\mathrm{Cov}(X,Y) = E(XY) - \mu_x\,\mu_y = 0$$

有

$$E(XY) = \mu_x\,\mu_y = E(X)\,E(Y) \tag{2.1.16}$$

2.1.2.4 相关系数

定义：随机变量 X 与 Y 的相关系数为

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_x\,\sigma_y} \tag{2.1.17}$$

两个随机变量之间的协方差受量纲的影响，但相关系数不受量纲的影响，因此，相关系数可以很好地度量两个随机变量线性关系的密切程度。

ρ 的取值在 -1 和 $+1$ 之间。为对这一点加以说明，考虑以 X 的线性函数 $A+BX$ 近似表示 Y，以均方误差 (Mean Square Error, MSE)

$$\begin{aligned}\mathrm{MSE} &= E[Y-(A+BX)]^2\\ &= E(Y^2) + A^2 + B^2E(X^2) - 2AE(Y) - 2BE(XY) + 2ABE(X)\end{aligned}$$

衡量 $A+BX$ 近似 Y 的优劣程度，显然，MSE 愈小，说明近似程度愈好。注意到 MSE 是参数 A 和 B 的函数，MSE 要达到极小值的必要条件是它关于参数 A 和 B 的偏导为 0，即

$$\begin{cases}\dfrac{\partial \mathrm{MSE}}{\partial A} &= 2A - 2E(Y) + 2BE(X) = 0\\ \dfrac{\partial \mathrm{MSE}}{\partial B} &= 2BE(X^2) - 2E(XY) + 2AE(X) = 0\end{cases}$$

解方程，得

$$\begin{aligned} b &= \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \\ a &= E(Y) - bE(X) \end{aligned}$$

将 a, b 代入 MSE，得

$$\begin{aligned} \min_{A,B} \mathrm{MSE} &= E\left\{[Y-(a+bX)]\right\}^2 \\ &= (1-\rho^2)\mathrm{Var}(Y) \end{aligned}$$

由 MSE 和 Var(Y) 的非负性，得

$$1-\rho^2 \geqslant 0\text{，也就是 } -1 \leqslant \rho \leqslant 1$$

若两个随机变量之间具有正的相关关系，则相关系数大于 0；若两个随机变量具有负的相关关系，则相关系数小于 0；若两个随机变量不相关，则相关系数为 0。特别地，如果相关系数为 1，则两随机变量之间具有完全正的线性相关关系；若相关系数为 −1，则两随机变量之间具有完全负的线性相关关系。

2.1.2.5　偏度与峰度

定义：随机变量 X 的偏度 (skewness) 系数为

$$b = \frac{E[(X-\mu)^3]}{\sigma^3} \tag{2.1.18}$$

峰度 (kurtosis) 系数为

$$k = \frac{E[(X-\mu)^4]}{\sigma^4} \tag{2.1.19}$$

偏度和峰度是描述分布形状的指标。偏度是用来度量分布对称性的指标。偏度系数为 0 的分布是对称分布，即

$$f(x-\mu) = f(x+\mu) \tag{2.1.20}$$

偏度系数为正，分布的形状是“长尾”在正方向，反之，则在负方向。

峰度是度量分布尾部厚度的指标。正态分布的峰度系数为 3。峰度系数大于 3 的分布，具有比正态分布尾部更厚的特征。

2.1.2.6　矩

定义：一个随机变量 X，若

$$\mu_k = E(X^k), \quad k = 1, 2, \cdots \tag{2.1.21}$$

存在，称它为 X 的 k 阶原点矩，简称 k 阶矩 (moment)。若

$$E\left\{[X-E(X)]^k\right\}, \quad k = 1, 2, \cdots \tag{2.1.22}$$

存在，称它为 X 的 k 阶中心矩。

随机变量的期望值、方差、偏度和峰度分别是随机变量一阶矩、二阶矩、三阶矩和四阶矩的函数。

2.1.3 随机向量的数字特征

记 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)'$，$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_q)'$

(1) 向量 $\boldsymbol{X}$ 的均值为

$$\boldsymbol{\mu}=E(\boldsymbol{X})=(E(X_1),E(X_2),\cdots,E(X_p))' \tag{2.1.23}$$

(2) 向量 $\boldsymbol{X}$ 的自协 (方) 差阵为

$$\mathrm{Var}(\boldsymbol{X})=\boldsymbol{\Sigma}=E(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})'=\left(\mathrm{Cov}(X_i,X_j)\right)=\left(\sigma_{ij}\right) \tag{2.1.24}$$

$\boldsymbol{\Sigma}$ 是对称、非负定阵，大多数情况下是正定阵。

(3) 向量 $\boldsymbol{X}$ 和向量 $\boldsymbol{Y}$ 的协 (方) 差阵为

$$\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})=E(\boldsymbol{X}-\boldsymbol{\mu}_X)(\boldsymbol{Y}-\boldsymbol{\mu}_Y)'=\left(\mathrm{Cov}(X_i,Y_j)\right) \tag{2.1.25}$$

若 $\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})=0\Leftrightarrow\boldsymbol{X},\boldsymbol{Y}$ 不相关。

设 $\boldsymbol{A},\boldsymbol{B}$ 为常数阵，则

$$\mathrm{Var}(\boldsymbol{AX})=\boldsymbol{A}\mathrm{Var}(\boldsymbol{X})\boldsymbol{A}'=\boldsymbol{A\Sigma A}' \tag{2.1.26}$$

$$\mathrm{Cov}(\boldsymbol{AX},\boldsymbol{BY})=\boldsymbol{A}\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{B}' \tag{2.1.27}$$

$$E(\boldsymbol{X}'\boldsymbol{AX})=\mathrm{tr}(\boldsymbol{A\Sigma})+\boldsymbol{\mu}'\boldsymbol{A\mu} \tag{2.1.28}$$

(4) 相关阵为

$$\boldsymbol{R}=\left(r_{ij}\right),\quad i=1,\cdots,p,\ j=1,\cdots,q \tag{2.1.29}$$

式中：

$$r_{ij}=\frac{\mathrm{Cov}(X_i,Y_j)}{\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(Y_j)}},\quad |r_{ij}|\leqslant 1 \tag{2.1.30}$$

2.2 正态分布及其有关的分布

常用的概率分布有很多，本节介绍正态分布，以及与其有关的 χ^2 分布、t 分布、F 分布。

2.2.1 正态分布

定义：若 X 服从均值为 μ，方差为 σ^2 的正态 (Normal) 分布，则记 $X\sim N(\mu,\sigma^2)$。其概率密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{2.2.1}$$

正态分布也称为高斯 (GAUSS) 分布。正态分布具有对称性；还具有线性性，即正态分布的线性组合仍服从正态分布。

若 $X\sim N(\mu,\sigma^2)$，则经过标准化处理

$$Z=\frac{X-\mu}{\sigma} \tag{2.2.2}$$

服从标准正态分布，记为 $Z \sim N(0,1)$。

总体方差已知情况下，对均值的统计推断通常用服从正态分布的统计量进行。

定义：若 $\ln X \sim N(\mu, \sigma^2)$，则称 X 服从对数正态分布。

对数正态分布通常用于对规模数据的分布进行描述。

2.2.2　χ^2 分布

定义：若 $X_1, X_2, \cdots, X_p \overset{\text{iid}}{\sim} N(0,1)$(符号“iid”表示独立同分布，是 identical independent distribution 的缩写)，则

$$Y = \sum_{i=1}^{p} X_i^2 \sim \chi^2(p) \tag{2.2.3}$$

性质：

(1) 若 $Y \sim \chi^2(p)$，则

$$E(Y) = p \tag{2.2.4}$$

$$D(Y) = 2p \tag{2.2.5}$$

(2) 若 $Y_1 \sim \chi^2(p)$，$Y_2 \sim \chi^2(q)$，且 Y_1, Y_2 相互独立，则

$$Y_1 + Y_2 \sim \chi^2(p+q) \tag{2.2.6}$$

对总体方差的统计推断通常用服从 χ^2 分布的统计量进行。

2.2.3　t 分布

定义：若 $X \sim N(0,1)$，$Y \sim \chi^2(p)$，X, Y 相互独立，则

$$t = \frac{X}{\sqrt{Y/p}} \sim t(p) \tag{2.2.7}$$

t 分布也称学生氏 (student) 分布，它是对称分布，但其分布的形状与正态分布相比，具有尾部更厚的特征。当 $p \to \infty$，有 $t(p) \to N(0,1)$。实际工作中，通常 $p > 30$，就可以用 $N(0,1)$ 近似 $t(p)$。

正态总体方差未知的情况下，对均值的统计推断通常用服从 t 分布的统计量进行。

2.2.4　F 分布

定义：若 $X \sim \chi^2(p)$，$Y \sim \chi^2(q)$，X, Y 相互独立，则

$$F = \frac{X/p}{Y/q} \sim F(p,q) \tag{2.2.8}$$

显然，若 $X \sim t(p)$，则

$$X^2 \sim F(1,p) \tag{2.2.9}$$

两个总体方差是否相同的统计推断通常用服从 F 分布的统计量进行。