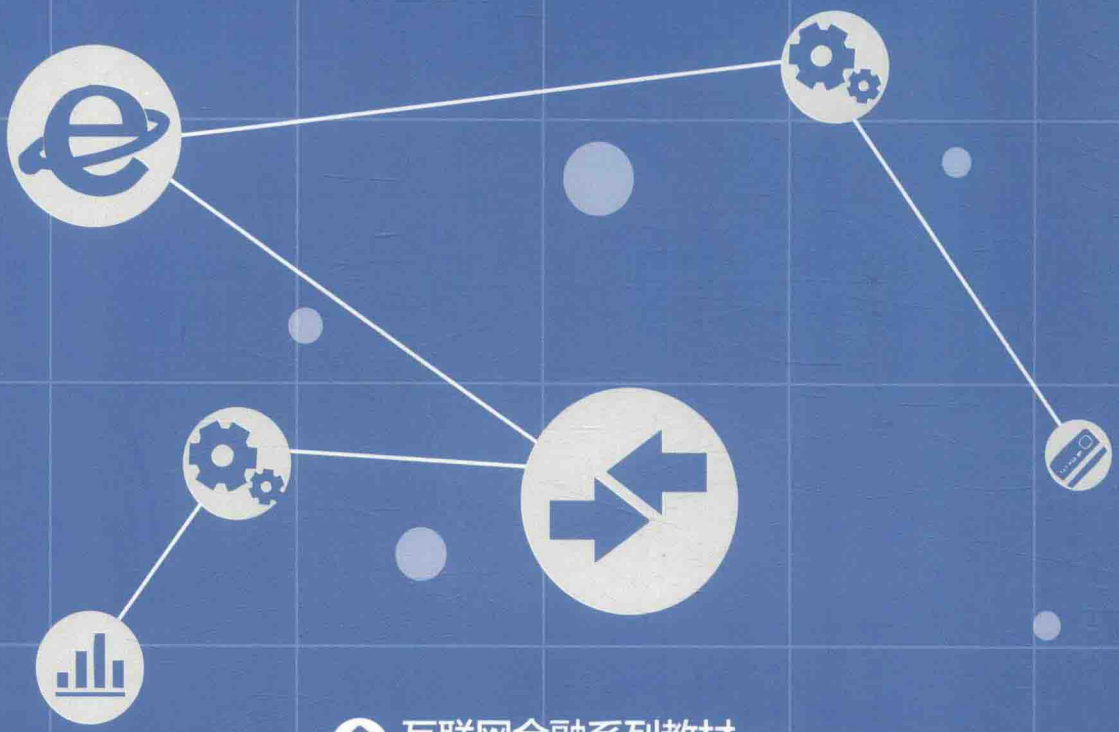

互联网金融系列教材
HULIANWANG JINRONG XILIE JIAOCAI

金融数据统计分析

JINRONG SHUJU TONGJI FENXI

主编◎何 岩

- 国内首套系列互联网金融专业教材
- 全景式解密互联网时代的金融变革
- 深度剖析互联网金融六大模式
- 生动展示互联网金融最新典型案例

中国金融出版社

互联网金融系列教材
HULIANWANG JINRONG XILIE JIAOCAI

金融数据统计分析

JINRONG SHUJU TONGJI FENXI

主编◎何　岩

中国金融出版社

责任编辑：王效端　张菊香
责任校对：张志文
责任印制：陈晓川

图书在版编目（CIP）数据

金融数据统计分析（Jinrong Shuju Tongji Fenxi）/何岩主编．—北京：中国金融出版社，2018.1
互联网金融系列教材
ISBN 978-7-5049-9306-9

Ⅰ.①金…　Ⅱ.①何…　Ⅲ.①金融统计—统计分析—教材　Ⅳ.①F830.2

中国版本图书馆 CIP 数据核字（2017）第 277865 号

出版
发行　中国金融出版社
社址　北京市丰台区益泽路 2 号
市场开发部　(010)63266347，63805472，63439533（传真）
网上书店　http://www.chinafph.com
　　(010)63286832，63365686（传真）
读者服务部　(010)66070833，62568380
邮编　100071
经销　新华书店
印刷　北京市松源印刷有限公司
尺寸　185 毫米×260 毫米
印张　12.5
字数　271 千
版次　2018 年 1 月第 1 版
印次　2018 年 1 月第 1 次印刷
定价　35.00 元
ISBN 978-7-5049-9306-9
如出现印装错误本社负责调换　联系电话（010）63263947

前言

互联网的快速发展使得金融、电信、零售等行业从传统的线下销售转为线上推广。互联网金融的蓬勃发展，使得人们越来越重视数据及其价值。从业人员和研究者将目光越来越多地聚集在已有的海量数据和互联网金融行业中每天产生的大量金融数据上，这些数据如同一座含有丰富信息和知识宝藏的矿山，等待人们去发掘。目前，金融行业和互联网金融行业都急需大量的能够综合运用数学理论、信息技术并理解金融业务的金融数据分析人才。

本书主要面向高职院校学生和非数学专业人士，对读者的数学基础要求不高，力图使更多的人能够运用统计学知识和方法，借由 Excel 工具进行基本的数据分析。各章内容以应用为导向，从基本的金融数据出发，利用统计学的概念和方法来处理金融、商务和经济领域的各种问题。鉴于笔者行业经验不够丰富，部分章节选取的实例不能与互联网金融行业背景完全契合，深感歉意。建议读者在掌握相应章节的知识和方法后，以项目背景为切入点，收集更多的数据进行分析和讨论，在实践中理解金融数据分析的概念和方法。

全书分为七个项目，前六个项目是基本知识的学习和实践，每个项目由三个以上的实验环节构成。第七个项目是综合实训项目，融合所学的知识结合实际任务进行实践演练。项目以互联网金融发展的主题为背景，明确项目任务，使得读者能够带着问题从知识点中寻找答案，再通过操作示范掌握 Excel 软件环境下的问题解决过程，最后通过能力拓展开拓相关知识视野。项目一主要介绍量化投资和概率论的基本知识；项目二主要介绍 P2P 网贷的关键问题以及获取数据、清洗数据和描述数据的基本方法；项目三主要介绍移动互联网营销方式和营销数据的参数描述方法；项目四主要介绍第三方支付和平行数据的对比方法；项目五主要介绍互联网众筹和数据关联的描述；项目六主要介绍互联网金融的发展和趋势分析方法；项目七为综合实践。

作者

2017 年 11 月

互联网金融系列教材
HULIANWANG JINRONG XILIE JIAOCAI

目录

项目一

基本概率

一、学习目标

◇ 理解投资与数据分析之间的关系
◇ 掌握投资分析中使用的概率知识
◇ 掌握资产收益率分布函数的使用方法
◇ 能够利用 Excel 完成概率的计算

二、项目背景　量化投资和概率论

量化投资之父詹姆斯·西蒙斯曾说："粒子的行动看似杂乱无章，实际上却存在着内在的规律。"量化投资研究就是用方程式来描述看似混乱的证券市场中隐藏的数学规律。

对于一般投资者，甚至是部分金融从业者来说，量化投资都是一门高深的技术，充斥着模型代码和算法假设，门槛非常高。其实，生活中的量化思想无处不在。以最常见的赌掷硬币为例，按照传统理论，正反面的概率各为50%，赌的次数少的时候凭运气，次数多的时候基本就是输赢各半。但美国斯坦福大学的数学教授佩尔西·戴康尼斯发现，如果在掷硬币前把硬币的正面朝上，掷硬币后依然是正面朝上的概率不是50%，而是51%；同样，如果反面朝上，结果反面继续朝上的概率也不是50%，而是51%。所以只要参与人看到哪一面在掷硬币前朝上，就赌哪一面，短期不一定赢，但长期看肯定赢率在51%左右，这就是量化模型。

量化投资是将人们总结出的投资思想，利用现代统计学、数学方法，形成若干可运算测量的数学模型，并借助现代计算机技术在海量历史数据中对模型进行验证，以寻找到能够带来超额收益的多种"大概率"模型，然后再严格地按照这些策略所构建的数量化模型运算结果来指导投资。量化投资克服了投资者情绪波动的影响，使投资的稳定性大为增加，避免因市场极度狂热或悲观的情况而导致作出非理性的投资决策，以保证在控制风险的前提下实现收益最大化。

量化投资的原理一：将每次赚钱概率提高到50%以上。也许从每次投资来看，成功的概率略微超过50%并不是很出彩，但是很多次加起来，投资所靠的"运气"可能

被变成风险有限的高额投资回报。

量化投资的原理二：如果每次交易赔钱的概率超过50%，但是每次赔的数量都很小，相对而言如果每次赚钱的概率虽然小于50%，但是赚的数目都很大的话，成功的概率也有可能超过50%。经过多次交易之后，只要盈利交易多于亏损交易，总体交易结果就是盈利的。对于这种情况，将交易进行分组，如果最大连续亏损次数为3次，则可以将6次交易分为一组，这样就可以看到每组赚钱概率提高到50%以上了。

下面对量化投资中常用的概念和方法进行介绍。

（一）资产收益率

资产收益率，也叫资产回报率（ROA），它是用来衡量每单位资产创造多少净利润的指标，也可以解释为企业利润额与企业平均资产的比率。其计算公式为

$$\text{资产收益率} = \text{净利润} / \text{平均资产总额} \times 100\%$$

$$\begin{aligned}\text{单期资产的收益率} &= \text{资产价值(价格)的增值} / \text{期初资产价值(价格)} \\ &= [\text{利息(股息)收益} + \text{资本利得}] / \text{期初资产价值(价格)} \\ &= \text{利息(股息)收益率} + \text{资本利得收益率}\end{aligned}$$

资产收益率是反映企业资产综合利用效果的指标，也是衡量企业利用债权人和所有者权益总额所取得盈利的重要指标，资产收益率越高，说明企业资产的利用效率越高，利用资产创造的利润越多，整个企业的获利能力也就越强，企业经营管理水平越高；反之，资产收益越低，说明企业资产的利用效率不高，利用资产创造的利润越少，整个企业的获利能力也就越差，企业经营管理水平越低。

资产收益率是财务分析的一个重要比率，其主要意义表现在：

第一，资产收益率集中地体现了资金运动速度与资产利用效果之间的关系。从计算公式不难看出，资金运动速度快，必然资金占用额少而业务量大，表现为较少的资产投资能够获得较多的利润。通过资产收益率的分析，能使企业资产运用与利润实现很好的挂钩，使投资者对企业“所得”与“所费”间的比例妥当与否有清晰认识。

第二，在资产一定的情况下，利润的波动必然引起资产收益率的波动，因此利用资产收益率这一指标，可以分析企业盈利的稳定性和持久性，从而确定企业的经营风险。盈利的稳定性表明企业盈利水平变动的基本态势。有时，尽管企业的盈利水平很高，但是缺乏稳定性，这很可能就是经营状况欠佳的一种反映。

第三，资产收益率的高低反映了企业经营管理水平的高低和经济责任制的落实情况。企业经营管理水平的高低，经济责任落实的情况如何，直接反映在利润的高低和资产的运用状况上。通过资产收益率将利润与资产相比较，可以更好地反映企业经营情况。

对于投资者来说，资产收益率总是越高越好，资产收益率越高，说明企业运用资产获取利润的能力越强，反之则越弱。在实际评价某一特定企业的资产收益率时，首先，应与本企业前期水平相比，以确定该年度的盈利水平；其次，应将连续几年的资产收益率水平进行比较，以观察企业资产收益率的变动趋势；最后，还应将该企业的资产收益率水平同其他企业的资产收益率水平以及同行业的平均水平相比较，才能对

该企业的获利能力作出正确评价。

在实际工作中，由于工作角度和出发点不同，收益率可以有以下一些类型：

1. 实际收益率。实际收益率表示已经实现或者确定可以实现的资产收益率，表述为已实现或确定可以实现的利息（股息）收益率与资本利得收益率之和。当存在通货膨胀时，应扣除通货膨胀率的影响。

2. 预期收益率。预期收益率也称为期望收益率，是指在不确定的条件下，预测的某资产未来可能实现的收益率。计算公式为

$$E(R) = \sum_i P_i R_i$$

式中，P_i 是情况 i 可能发生的概率，R_i 是情况 i 发生时的收益率。

证券资产组合的预期收益率就是组成证券资产组合的各种资产收益率的加权平均数，其权数为各种资产在组合中的价值比例。即

$$E(R_p) = \sum_i W_i \cdot E(R_i)$$

式中，$E[R_i]$ 表示组合内第 i 项资产的预期收益率，W_i 表示第 i 项资产在整个组合中所占的价值比例。

3. 必要收益率。必要收益率也称为最低必要报酬率或最低要求的收益率，表示投资者对某资产合理要求的最低收益率。必要收益率由两部分构成，即无风险收益率和风险收益率

$$R = R_f + \beta(R_m - R_f)$$

式中，R_f 表示无风险收益率，通常以短期国债的利率近似代替；R_m 表示市场投资组合收益率，通常用股票价格收益指数收益率的平均值或所有股票的平均收益率来代替；β 表示该资产的系统性风险系数。

（1）无风险收益率。无风险收益率也称无风险利率，它是指无风险资产的收益率，它的大小由纯粹利率（资金的时间价值）和通货膨胀补贴两部分组成。无风险资产一般满足两个条件：一是不存在违约风险；二是不存在再投资收益率的不确定性。实际上，满足这两个条件的资产几乎是不存在的，一般用与所分析的资产的现金流量期限相同的国债来表示。因此，一般用国债的利率表示无风险利率，该国债应与所分析的资产的现金流量有相同的期限。

（2）风险收益率。风险收益率是指某资产持有者因承担该资产的风险而要求的超过无风险利率的额外收益。风险收益率衡量了投资者将资金从无风险资产转移到风险资产而要求得到的“额外补偿”，它的大小取决于以下两个因素：一是风险的大小；二是投资者对风险的偏好。

（二）货币的时间价值

金融专业里有一句话叫做“今天的一块钱不等于明天的一块钱”。一般来讲，投资的最主要特点是时间价值和风险。投资是对资本的利用，而资本是那些能够带来未来收益的价值，所以投资就是为了获得未来收益而对资本的利用过程。作为一种与未来收益相联系的价值，随着时间的推移，资本具有不断增值的能力，时间越长，从最终

结果看，增值的价值量越大。

资本的时间价值又被称为货币的时间价值，这是因为资本最初总是以货币的预付为起点、货币的回流为终点。

货币的时间价值最直接的表现就是利率，利率是货币价值增值能力的体现，利率越高，则货币的增值能力越大，反之则越小。利率的存在，使得不同时间点上的货币量之间可以进行价值量大小的比较。由于投资总是以某个时间点的货币预付为起点，而以其后某个时期或在某个期间内的货币回流为终点，因此要反映投资是否达到预期的效益，就必须将它们转化为某个特定时间的货币价值来进行比较。这种转化一般是以市场上货币的机会成本或平均利率为标准，或者是以投资者的预期毛利率为标准的。

要求回报率是投资人对于风险投资所要求的收益率，它是反映未来现金流风险的报酬，也称为人们愿意进行投资所必须赚得的必须收益。大部分投资者的目标是获得大的投资回报和承担小的投资风险。然而要求回报率并不是漫天要价的，它的决定基础是资金的供给和需求的水平，即价值规律。与经济学中需求和供给决定了商品价格一样，资金的需求和供给决定了资金的价格，而资金的价格就是利率。

要求回报率由无风险利率和风险溢价组成。无风险利率分为名义无风险利率和实际风险利率。假设银行存款利率为7%，这是否意味着资金可以每年增值7%呢？答案是否定的。因为如果每年的通货膨胀率为6%，则实际只能获得1%的收益率，所以7%称为名义无风险利率，1%称为实际无风险利率。

风险溢价是一个人在面对高低不同的风险，且清楚高风险高报酬、低风险低报酬的情况下，会如何因个人对风险的承受度而影响其是否要冒风险获得较高的报酬，或是只接受已经确定的收入，放弃冒风险可能得到的较高报酬。确定的收入与较高的报酬之间的差，即为风险溢价。从投资学的角度而言，风险溢价可以视为投资者对于高风险所要求的较高报酬。衡量风险时，通常的方法就是使用无风险利率即政府公债之利率来与其他高风险的投资比较。高于无风险利率的报酬，这部分即称为风险溢价。风险溢价有多种，以违约风险溢价为例，违约风险溢价是指债券发行者在规定时间内不能支付利息和本金的风险。债券信用等级越高，违约风险越小；债券信用等级越低，违约风险越大。违约风险越大，债券的到期收益率越高。违约风险溢价一般会被添加进无风险真实利率里，以补偿投资者对违约风险的承受。

当利息周期与计息周期不一致时，出现了名义利率和有效年利率的概念。利息周期指的是利率以多长时间为一个周期计算，例如，“年利率12%”的利息周期为一年。所有给出的利率都成为名义利率，例如“年利率12%”中名义利率为12%。如果名义利率中利息周期为一年，则名义利率也是有效年利率，但如果利息周期不是以一年为单位，则名义利率不是有效年利率。名义利率按不同的计息期调整后算得的利率为有效年利率（EAR）。

如果一年内计算复利次数为 n，名义利率为 r，则有效年利率为

$EAR = (1 + r/n)^n - 1$。如果是连续复利，则 $EAR = e^r - 1$。

（三）投资与随机变量

几乎在所有的投资决策中，都会用到随机变量，例如股票收益率和每股收益都是常见的随机变量的例子。下面分别介绍两个常见的概率应用。

二项期权定价模型是由考克斯（J. C. Cox）、罗斯（S. A. Ross）、鲁宾斯坦（M. Rubinstein）和夏普（Sharpe）等人提出的一种期权定价模型，主要用于计算美式期权的价值。二项期权定价模型假设股价波动只有向上和向下两个方向，且假设在整个考察期内，股价每次向上（或向下）波动的概率和幅度不变。模型将考察的存续期分为若干阶段，根据股价的历史波动率模拟出正股在整个存续期内所有可能的发展路径，并对每一路径上的每一节点计算权证行权收益和用贴现法计算出权证价格。以连续三天的股价变动为例，三天中，每天股价上升下降都是一个独立的事件，股价以常数概率 p 向上变动。如果股价向上变动，U 等于 1 加上向上变动带来的收益率。股票以常数概率 $1-p$ 向下变动，如果股价向下变动，D 等于 1 加上向下变动所带来的收益率。图 1-1 表示了连续三天股票变动的情况。从图 1-1 可以看到，在 $t=3$ 时股票价格有 4 个可能值：UUUS、UUDS、UDDS、DDDS。导致 UUDS 结果的序列有三个，分别是 UUD、UDU 和 DUU，这三个序列中都有两次上升和一次下降，所以概率都是 $p^2(1-p)$，所以得到 UUDS 结果的概率为 $3p^2(1-p)$。股票价格等于最后任一一个结果的概率都可以用二项分布给出。股票价格是二项随机变量的一个函数。

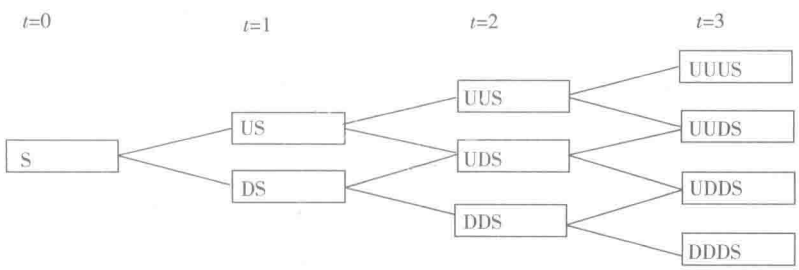

图 1-1　股价变动二叉树

在投资组合中，超亏风险是指资产组合的价值低于某个可以接受的最低值的风险。如果一个投资者能够忍受的最低收益率是 6%，则当投资的收益率低于 6% 时，他正面临超亏风险。所以所有投资组合将尽可能降低投资收益率低于可承担的最低收益率，即罗伊第一安全比率准则。用符号表示，投资者的目标是选择一个投资组合使得 $P(R_P < R_L)$，其中 R_P 是实际的投资收益率，R_L 是临界收益率。当投资组合收益率服从正态分布时，R_P 的期望是 $E(R_P)$，标准差为 σ_P，则安全第一准则下，最优投资组合将最大化安全比率：

$$SFR = \frac{E(R_P) - R_L}{\sigma_P}$$

例如，假设一个投资者的临界收益率为 2%，现在有两种投资组合的选择。投资组合 1 的预期收益率为 12%，标准差为 15%。投资组合 2 有 14% 的收益率和 16% 的标准

差，则投资组合1的 $SFR = (12-2)/15 = 0.667$，投资组合2的 $SFR = (14-2)/16 = 0.75$。投资组合2的安全比率优于投资组合1。根据正态分布的计算，投资组合2收益率低于临界收益率2%的概率为23%。

三、知识要点

在日常生活中，我们常常会遇到一些涉及可能性或发生机会等概念的事件。一个事件的可能性或一个事件的发生机会是与数学有关的。例如："从一个班40名学生中随意选出一人，这人会是男生吗?"

概率常用来量化对于某些不确定命题的想法，命题一般会是以下的形式："某个特定事件会发生吗?"，对应的想法则是："这个事件会发生的可能性是多少?"。确定的程度可以用0到1之间的数值来表示（0表示不可能发生，1表示一定会发生），这个数值就是概率。

有些事件是确定性事件。确定性事件包含必然事件和不可能事件。如太阳从东方升起，或者在标准大气压下，水在100℃时会沸腾。这些事件称为必然事件。一个普通的骰子，掷出向上一面的数字是7，这是不可能发生的，这种事件被称为不可能事件。

大量事件在一定条件下是否发生，是无法确定的。如明天的气温比今天低、掷骰子得到的点数为2，又比如中东冲突可能会对油价有某种程度的影响，而油价对世界经济可能会有涟漪效应的影响。这些可能发生也可能不会发生的事件称为随机事件。

（一）概率

一个事件的概率值通常以一个介于0到1的实数来表示。一个不可能事件其概率值为0，而确定事件其概率值则为1。

设随机事件的样本空间为 Ω，Ω 的一个子集称为事件。对于 Ω 中的每一个事件 A，都有实函数 $P(A)$，$P(A)$ 为 Ω 中事件 A 的概率。概率满足如下条件：

非负性：$P(A) \geqslant 0$；

规范性：$P(\Omega) = 1$。

1. 古典概率和主观概率。概率有三种主要的含义，第一种是古典先验概率，即以有关过程的先验知识为基础的事件发生的可能性。例如掷骰子，每次投掷得到点数是不确定的，在10次投掷中，可能得到2点的次数是7次，但不能给出"掷骰子得到2点的概率是0.7"的结论。根据先验知识，我们知道，掷骰子得到1点到6点的概率都是相同的，即1/6。

第二种含义是古典经验概率，这类概率可以表示调查得知的支持某政策的居民比例，线上购买电子产品的比例等。它是利用过去历史资料计算得到的先验概率。模型要求满足两个条件：（1）试验的所有可能结果是有限的；（2）每一种可能结果出现的概率相等。若所有可能结果的总数为 n，随机事件 A 包括 m 个可能结果，则事件 A 发生的概率为 m/n。

当历史资料无从取得或资料不完全时，凭人们的主观经验来判断而得到的概率，称为主观概率，这是概率的第三种含义。例如，对于一个新产品，该产品的开发人员

和市场营销人员对该产品赢得市场的可能性会有不同的判断。个人的过往经验、观点和立场影响人们赋予各种事件的概率值。主观概率对于决策尤为有用。

2. 多个事件的概率运算

（1）互斥事件。有些事件是互斥事件，即不可能同时发生的事件，例如掷骰子时，得到1个点和得到2个点是不可能同时发生的，它们是互斥事件。直观描述互斥事件在样本空间中的位置关系如图1-2所示。如果要计算投掷一次骰子得到1点或2点的概率，以A表示得到1点，以B表示得到2点，则得到1点或2点可表示为$P(A \cup B)$。

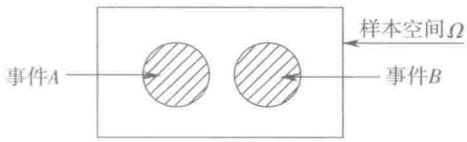

图1-2　互斥事件

对于数个两两互斥事件$\{A_i\}_{i\in N}$，有$\sum_{i=1}^{\infty} P(A_i) = P(\bigcup_{i=1}^{\infty} A_i)$。在掷骰子中，想计算得到1点或2点的概率，因为$A$、$B$是互斥事件，所以$P(A \cup B) = P(A) + P(B) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$。

（2）广义加法公式。对于任意两个事件A和事件B，计算A或B发生的概率如图1-3所示。计算公式为

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{1.1}$$

式中，$P(A \cap B)$表示事件A和事件B同时发生的概率，称为联合概率。

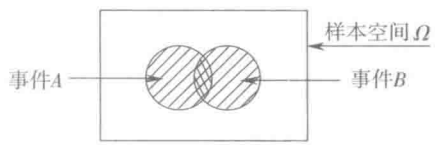

图1-3　任意事件的合并

例如，一家科技公司的人事部门对近两年离职人员的离职原因进行调查，被调查的人中有40%对工资不满意，30%是因为工作强度太大，有15%是对工资和工作强度都不满意。现在想知道离职员工中因为对工资不满意或者对工作强度不满意而离职的员工比例。

以A表示员工离职是因为对工资不满意，以B表示员工离职时因为对工作强度不满意，则要计算的对象为$P(A \cup B)$。根据公式（1.1），有$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.4 + 0.3 - 0.15 = 0.55$。

3. 条件概率（Conditional Probability）。条件概率就是事件A在另外一个事件B已经发生条件下的发生概率。条件概率表示为$P(A|B)$，读作“在B条件下A的概率”。A与B之间不一定有因果或者时间顺序关系。A可能会先于B发生，也可能相反，也可能二者同时发生。A可能会导致B的发生，也可能相反，也可能二者之间根本就没有因果关系。

（1）条件概率的计算。设A与B为样本空间Ω中的两个事件，其中$P(B) > 0$。那么在事件B发生的条件下，事件A发生的条件概率为

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1.2}$$

条件概率有时候也称为后验概率。条件概率的图示见图 1-4。

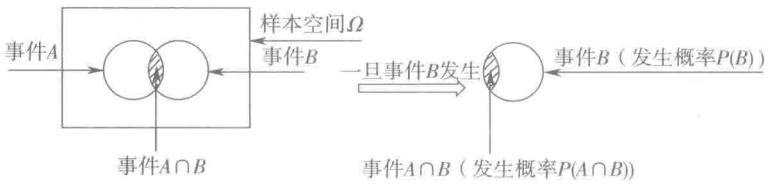

图 1-4　条件概率

（2）概率的乘法公式。要计算两个事件同时发生的概率，即 $P(A \cap B)$，根据条件概率公式（1.2），有

$$P(AB) = P(B)P(A \mid B) \tag{1.3}$$

$P(A \cap B)$ 通常表示为 $P(AB)$。若事件 A 与 B 相互独立，则有

$$P(AB) = P(A)P(B) \tag{1.4}$$

将公式（1.3）推广到 n 个事件，有

$$P(A_1A_2\cdots A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1A_2)\cdots P(A_n \mid A_1A_2\cdots A_{n-1}) \tag{1.5}$$

将公式（1.4）推广到 n 个相互独立的事件，有

$$P(A_1A_2\cdots A_n) = P(A_1)P(A_2)\cdots P(A_n) \tag{1.6}$$

4. 概率密度和累积分布。随机变量根据其取值特点分为离散型随机变量和连续刑随机变量。如果随机变量 X 只可能取有限个或者是可数无穷尽的值 $X = x_1, x_2, x_3, \cdots$，则称 X 为离散型随机变量。例如掷骰子中每掷一次，只能得到 1 点到 6 点中的某一个数值。如果随机变量 X 由全部实数或者由一部分区间组成，$X = \{x \mid a \leqslant x \leqslant b\}$，$-\infty < a < b < \infty$，则称 X 为连续型随机变量。连续随机变量的值是不可数及无穷尽的。

在数学中，连续型随机变量的概率密度函数是一个描述这个随机变量在某个确定的取值点附近的可能性的函数。当概率密度函数存在的时候，累积分布函数是随机变量的取值落在某个区域之内的概率，是概率密度函数的积分，一般以大写“PDF”（Probability Density Function）标记。对一维随机变量 X，如果它的概率密度函数为 $f_X(x)$，则它的累积分布函数是

$$F_X(a) = \int_{-\infty}^{a} f_X(x)\,\mathrm{d}x, \quad -\infty < a < +\infty \tag{1.7}$$

连续型随机变量的概率密度函数有如下性质：

$$\forall\ -\infty < x < \infty,\ f_X(x) \geqslant 0$$

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$$

$$\forall\ -\infty < a < b < \infty,\ P[a < X \leqslant b] = F_X(b) - F_X(a) = \int_{a}^{b} f_X(x)\,\mathrm{d}x$$

如果概率密度函数 $f_X(x)$ 在一点 x 上连续，那么累积分布函数可导，并且它的导数：$F'_X(x) = f_X(x)$。

由于随机变量 X 的取值 $P[a < X \leqslant b]$ 只取决于概率密度函数的积分，所以概率密度函数在个别点上的取值并不会影响随机变量的表现。

连续型的随机变量取值在任意一点的概率都是0。作为推论，连续型随机变量在区间上取值的概率与这个区间是开区间还是闭区间无关。要注意的是，概率 $P[X = a] = 0$，但 $\{X = a\}$ 并不是不可能事件。

最简单的概率密度函数是均匀分布的密度函数，如图1-5所示。对于一个取值在区间 $[a,b]$ 上的均匀分布函数 $I_{[a,b]}$，它的概率密度函数 $f_{I_{[a,b]}}(x) = \dfrac{1}{b-a} I_{[a,b]}$。

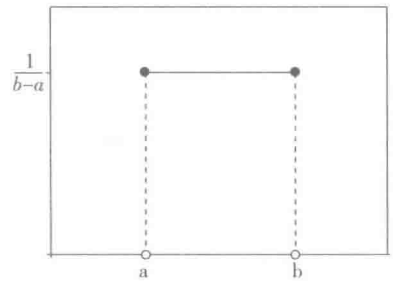

图1-5　均匀分布的概率密度

（二）几种概率分布

结合本书后续章节关注的概率知识，本节重点介绍二项分布和正态分布。

1. 二项分布。二项分布又叫贝努里分布，是一种具有广泛用途的离散型随机变量的概率分布。

二项分布是指统计变量中只有性质不同的两项群体的概率分布。所谓两项群体，是按两种不同性质划分的统计变量，即各个变量都可归为两个不同性质中的一个，两个观测值是对立的。例如，对病人治疗结果的有效与无效，进入商店的客户购买还是不购买商品。

考虑只有两种可能结果的随机试验，当成功的概率是恒定的，且各次试验相互独立，这种试验在统计学上称为伯努利试验（Bernoulli Trial）。二项分布即重复 n 次的伯努利试验。

如果事件发生的概率是 p，则不发生的概率 $q = 1 - p$，n 次独立重复试验中发生 k 次的概率是

$$p(X = k) = C_n^k p^k (1-p)^{n-k} \tag{1.8}$$

例如，有10道判断题，由于答题者完全不懂，只能靠猜测答题，那么他答对6题的概率是多少？

靠猜测回答问题，答对和答错的概率各占一半，回答正确概率 $p = 0.5$，回答错误的概率 $q = 1 - p = 0.5$，则答对6题的概率 $C_{10}^6 \times 0.5^6 \times (1-0.5)^4 = 0.205$，即有20.5%的可能性猜对6题。如果答对8题，则概率为4.4%。

2. 正态分布（Normal Distribution）。正态分布是一个在数学、物理、工程及金融等领域都非常重要的概率分布。正态分布在自然界中随处可见，比如说人的身高和智力都服从正态分布。

若随机变量 X 服从一个位置参数为 μ、标准差为 σ 的概率分布，记为

$$X \sim N(\mu, \sigma^2) \tag{1.9}$$

则其概率密度函数为

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e - \frac{(x-\mu)^2}{2\sigma^2} \tag{1.10}$$

标准差 σ 的平方 σ^2 称为方差。

正态分布的概率密度图如图 1－6 所示。正态分布的概率密度函数曲线呈钟形，因此人们又经常称之为钟形曲线。

累积分布函数是随机变数 X 小于或等于 x 的概率，连续随机变量的累积分布函数是概率密度函数从 $-\infty$ 到 x 的积分函数。正态分布的累积分布函数表示为

$$F(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)}\mathrm{d}t \tag{1.11}$$

正态分布的累积分布函数图如图 1－7 所示。

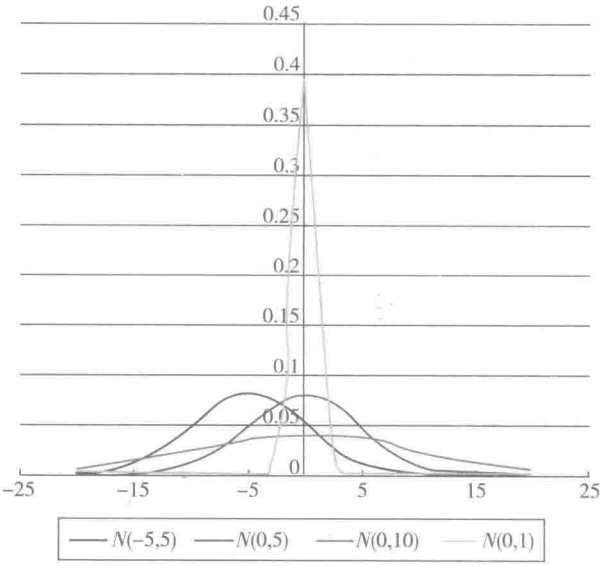

图 1－6　正态分布概率密度图

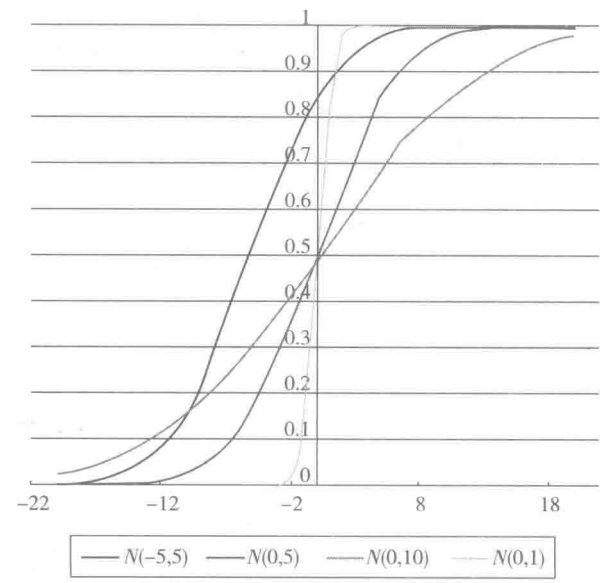

图 1－7　正态分布的累计分布函数图

正态分布有一个被称为“经验法则”的“68－95－99.7 法则”，即约 68.3% 的数值分布在距离平均值有 1 个标准差之内的范围，约 95.4% 的数值分布在距离平均值有 2 个标准差之内的范围，以及约 99.7% 的数值分布在距离平均值有 3 个标准差之内的范围，如图 1－8 所示。因此也称正态分布的均值 μ 是“位置参数”，它决定了分布的中心位置，标准差 σ 是“尺度参数”，它决定了分布的幅度。

根据经验法则，在给出正态分布均值和标准差的条件下，可以快速作出估计。例如，在某次高中的数学考试中（满分为 150 分），考生成绩服从均值为 100，标准差为 10 的正态分布。任意选取一位学生，他的考试成绩在 80～120 分的概率为多少？由于 80～120 在均值的正负两个标准差范围，所以可以马上给出“成绩在 80～120 分的概率为 95.4%”这样的结论。

正态分布中，均值为 0，标准差为 1 的分布称为标准正态分布，通常用 $Z \sim N(0,1)$ 表示标准正态分布。

3. 正态分布与其他分布之间的关系。正态分布有一个非常重要的性质：在特定条件下，

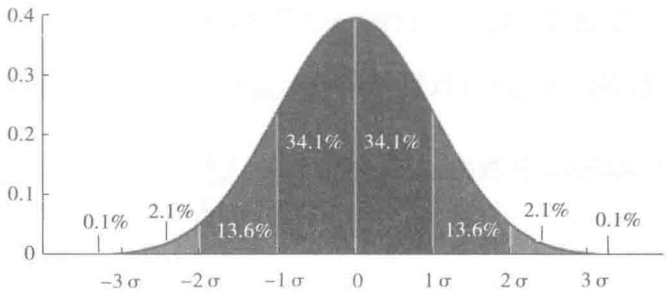

图 1－8　正态分布“经验法则”

大量统计独立的随机变量的平均值的分布趋于正态分布，这就是中心极限定理。中心极限定理的重要意义在于，根据这一定理的结论，其他概率分布可以用正态分布作为近似。

以人的智力都服从正态分布为例，除了由许多不同的基因调控以外，后天的营养、环境、健康，甚至偶然的意外，都有着各自的影响。在这种情况下，如果将每个因素看成一个基本事件，并且假定这些因素各自的影响都差不多，将这些因素综合考虑，根据中心极限定理，得到的结果就非常接近正态分布。

参数为 n 和 p 的二项分布，在 n 相当大而且 p 接近 0.5 时近似于正态分布。近似正态分布平均数为 $\mu = np$，且方差为 $\sigma^2 = np(1-p)$。例如，样本数为 $n = 48$，$p = 0.25$ 的二项分布，趋近于均值为 12，标准差为 3 的正态分布，如图 1－9 所示。

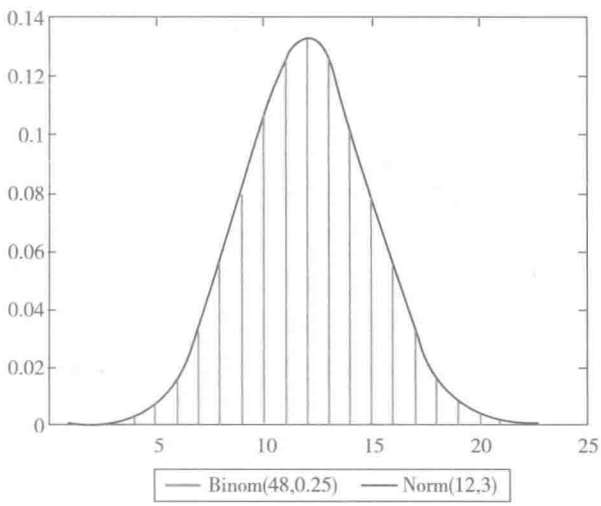

图 1－9　二项分布与正态分布

四、项目任务

【任务概览】

本项目通过完成几种应用情景中的数据表格，掌握二项分布、正态分布在 Excel 中的计算方法，理解正态分布的特点。

1. 利用已有的统计数据计算事件发生的概率。
2. 多事件概率的计算方法。
3. 选择合适的概率分布对数据进行概率计算。
4. 能够灵活运用 NORM. DIST( ) 函数和 NORM. INV( ) 函数计算正态分布。