第Ⅰ部分　数据分析基础

第 1 章　概率与统计基础

本章回顾一些概率知识和基本的统计概念。大多数结论只叙述而不证明，读者可以很容易找到相关书籍参考学习和理解。这些概念极为重要，是继续学习的基础、通往其他部分不可或缺的钥匙。

1.1　随 机 变 量

随机变量(random variable)是取值具有随机性的变量。随机变量按其取值情况可以分为离散型和连续型两种类型，离散型随机变量只能取有限或可数的多个数值，连续型随机变量的取值充满一个或若干有限或无限区间。

1.1.1　概率分布

1. 概率分布的含义

随机变量 X 取各个值 x_i 的概率称为 X 的概率分布。对一个离散型随机变量 X，可以给出如下概率分布：

$$P(X = x_i) = p_i, \quad i = 1,2,3,\cdots \tag{1.1.1}$$

例如，X 代表宏观经济所处的状态，假定只有经济增长率较高的繁荣和增长率较低的衰退两种状态，X 相应地取 1 和 2 两个值(图 1.1.1)，并假定概率分别为 p，q，即

$$P(X = 1) = p, \quad P(X = 2) = q$$

由概率的性质可知，概率分布满足以下两个条件：

$$p_i \geqslant 0, \quad i = 1,2,\cdots$$

$$\sum_{i=1}^{\infty} p_i = 1 \tag{1.1.2}$$

可以知道，对于上面例子中的 p 和 q，存在约束：$p \geqslant 0$，$q \geqslant 0$，$p+q=1$。

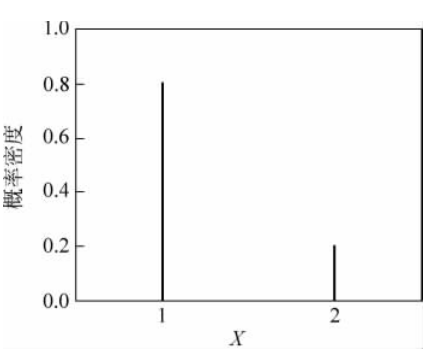

图 1.1.1　离散型概率分布

(经济状态概率分布：$p=0.8$，$q=0.2$)

2. 累积分布函数

对于随机变量 X(无论连续还是离散)可以确定实值函数 $F(x)$，称为累积分布函数(cumulative distribution function，CDF)，定义如下：

$$F(x) = P(X \leqslant x) \tag{1.1.3}$$

表示随机变量 X 小于或等于 x 的概率。显然，$F(-\infty)=0$，$F(+\infty)=1$。对于离散随机变量，累积分布函数的形式为

$$F(x)=\sum_{x_i\leqslant x}p_i \tag{1.1.4}$$

3. 连续型随机变量的分布函数及概率密度函数

对于连续型随机变量，取任何特定数值的概率都是0，因此度量该随机变量在某一特定范围或区间内的概率才有实际意义。设$F(x)$是随机变量X的分布函数，如果对任意实数x，存在非负函数$f(x)\geqslant 0$，使

$$F(x)=\int_{-\infty}^{x}f(t)\mathrm{d}t \tag{1.1.5}$$

就称$f(x)$为X的概率密度函数(PDF)，且$f(x)$具有性质：

$$f(x)\geqslant 0,\quad \int_{-\infty}^{\infty}f(x)\mathrm{d}x=1 \tag{1.1.6}$$

$$P(a<x<b)=\int_{a}^{b}f(x)\mathrm{d}x \tag{1.1.7}$$

令X代表身高，用厘米来度量，那么人的身高在某一区间内(比如160～170cm)的概率，由这两个值之间的密度函数之下的面积决定(图1.1.2)。

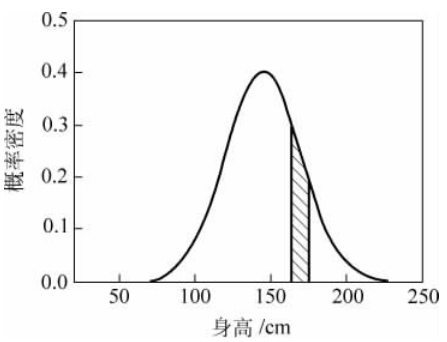

图1.1.2　连续型(身高)概率分布

例1.1　离散随机变量的CDF

抛币4次，求随机变量(正面朝上的次数)的概率密度函数和累积分布函数(图1.1.3和图1.1.4)。

正面朝上的次数	PDF		CDF	
$(X=x_i)$	X值	p_i	X值	$F(x)$
0	$0\leqslant X<1$	1/16	$X\leqslant 0$	1/16
1	$1\leqslant X<2$	4/16	$X\leqslant 1$	5/16
2	$2\leqslant X<3$	6/16	$X\leqslant 2$	11/16
3	$3\leqslant X<4$	4/16	$X\leqslant 3$	15/16
4	$4\leqslant X<5$	1/16	$X\leqslant 4$	1

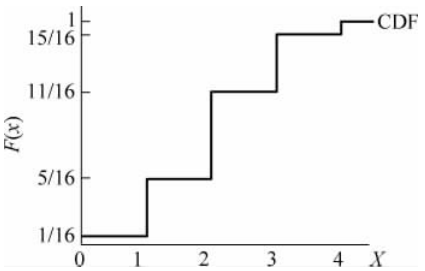

图 1.1.3　离散型随机变量的累积分布函数

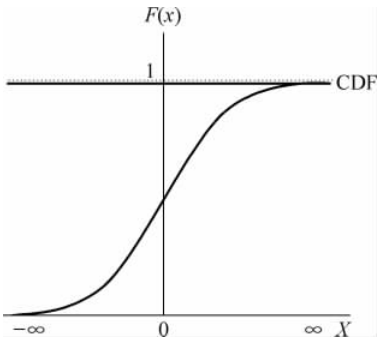

图 1.1.4　连续型随机变量的累积分布函数

1.1.2　随机变量的数字特征

有多种数值指标分别从不同角度描述随机变量分布的特征，其中最重要的是数学期望（也称均值）和方差。期望是随机变量的平均值，它度量了集中趋势；方差是随机变量偏离期望的离散程度的度量。

1. 数学期望

假设我们研究一个离散型随机变量 X，设 $x_1, x_2, \cdots, x_N$ 为该变量的 N 个取值，则均值或数学期望值是所有可能结果的加权平均值，权重为各个可能结果的发生概率，用 μ_X 代表 X 的数学期望，定义为

$$\mu_X = E(X) = p_1 x_1 + p_2 x_2 + \cdots + p_N x_N = \sum_{i=1}^{N} p_i x_i \tag{1.1.8}$$

式中：p_i 为 X_i 发生的概率，$\sum p_i = 1$。

如果 X 是连续型随机变量，则数学期望为

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x \tag{1.1.9}$$

数学期望有一个重要的性质：

$$E(a + bX) = a + bE(X) \tag{1.1.10}$$

式中：a, b 都是常数。

除了期望之外，用来描述随机变量集中趋势的还有中位数。中位数是满足 $P(X \leqslant m) \geqslant 0.5$ 和 $P(X \geqslant m) \leqslant 0.5$ 的 m 的值。粗略地说，中位数比均值更接近分布的中点，它不受极端值影响。

2. 方差

对于经济变量，我们经常关心其波动性，尤其证券市场中人们十分关心投资的风险大小，这可以通过变量的方差来描述。随机变量的方差刻画了随机变量偏离均值的程度，将方差记为 σ_X^2，对于离散的情形，方差为

$$\sigma_X^2 = \mathrm{var}(X) = E[X - E(X)]^2 = \sum_{i=1}^{N} p_i (x_i - \mu_X)^2 \tag{1.1.11}$$

对于连续情形，方差为

$$\sigma_X^2 = \mathrm{var}(X) = \int_{-\infty}^{\infty} (x-\mu_X)^2 f(x)\mathrm{d}x \tag{1.1.12}$$

方差不能为负值，如果 X 偏离均值幅度很大，则方差就较大；相反，则方差较小；如果 X 所有的值都等于 $E(X)$，则方差为 0。这意味着随机变量是常数。

方差有一个重要的性质：

$$\mathrm{var}(a+bX) = b^2\mathrm{var}(X) \tag{1.1.13}$$

经常用到的标准差 σ_X 是方差的正平方根。如果要我们猜测对一个随机变量进行一次抽样的结果，均值可能是不错的选择。但如果要给出一个区间，就可以根据希望正确的程度确定置信水平，在均值两侧延伸相应倍数的标准差产生一个区间（置信区间）。就是说虽然方差是衡量波动程度的指标，但与均值进行加减运算只能是标准差，因为标准差可以被认为和 μ 有相同的度量单位。对任意随机变量 X 和任意正常数 k，切比雪夫不等式表明：

$$P(\mu - k\sigma \leqslant X \leqslant \mu + k\sigma) \geqslant 1 - \frac{1}{k^2} \tag{1.1.14}$$

3. 偏度和峰度

除了最为常用的描述随机变量 X 集中趋势的期望和中位数、描述偏离均值程度的方差外，偏度 S（skewness）和峰度 K（kurtosis）也是描述随机变量 X 的数字特征。偏度 S 衡量了 X 围绕其均值的非对称性，峰度 K 度量凸起或平坦程度。

在定义偏度 S 和峰度 K 之前，首先需要了解 X 的高阶矩和高阶中心矩。一般 r 阶矩和 r 阶中心矩分别定义为

$$E(X)^r \quad 和 \quad E(X-\mu_X)^r$$

随机变量 X 的一阶矩即是数学期望值，方差是 X 的二阶中心矩，三阶中心矩表示为

$$E(X-\mu_X)^3$$

四阶中心矩表示为

$$E(X-\mu_X)^4$$

偏度 S 用三阶中心矩除以标准差的立方来计算：

$$S = \frac{E(X-\mu_X)^3}{\sigma_X^3} \tag{1.1.15}$$

如果概率密度函数是对称的，则 S 值为 0；正的 S 值意味着序列分布有长的右拖尾（右偏）；负的 S 值意味着序列分布有长的左拖尾（左偏）。

峰度 K 定义为

$$K = \frac{E(X-\mu_X)^4}{\sigma_X^4} \tag{1.1.16}$$

正态分布是最常见到的分布。对于正态分布，$K=3$，$S=0$。如果 K 值大于 3，分布的凸起程度大于标准正态分布；如果 K 值小于 3，分布相对于标准正态分布是平坦的。因此，了解标准正态分布峰度 K 和偏度 S 有助于比较其他概率分布函数。

1.1.3 随机变量的联合分布

对于两个或两个以上的随机变量，规律性由它们的联合分布所决定。联合分布有协

方差和相关系数等重要数字特征。

例 1.2　两个离散随机变量的联合分布

X 表示家庭收入，Y 表示是否受过大学教育，1、0 分别表示受过和没有受过大学教育，联合分布如下：

结　果	概率 $f(x,y)$	结　果	概率 $f(x,y)$
$X=600$ 元，$Y=1$	0	$X=600$ 元，$Y=0$	1/4
$X=1\,500$ 元，$Y=1$	1/8	$X=1\,500$ 元，$Y=0$	1/8
$X=3\,000$ 元，$Y=1$	1/3	$X=3\,000$ 元，$Y=0$	1/6

对于两个离散的随机变量 X,Y，它们的联合分布为

$$P(X=x_i,Y=y_j)=p_{ij},\quad i,j=1,2,\cdots \tag{1.1.17}$$

如例 1.2 中：

$$P(X=600,Y=0)=1/4$$

对于连续的随机变量 X,Y，它们的概率分布则由联合概率密度 $f(x,y)$ 决定：

$$P(a<X<b,c<Y<d)=\int_a^b \mathrm{d}x\int_c^d f(x,y)\mathrm{d}y \tag{1.1.18}$$

1. 边际概率

与联合概率函数 $f(x,y)$ 相对应，$f_X(x)$，$f_Y(y)$ 都称为边际概率函数。如例 1.2 中，$f_X(600)$ 应该是所有家庭收入为 600 元而无论是否受过大学教育的概率，即 $f_X(600)=P(X=600)=P(X=600,Y=0)+P(X=600,Y=1)=1/4$。因此，从联合分布得到某一个变量（比如 X）的边际密度，只需要将其对应的联合概率累加（离散）或积分（连续）起来：

$$f_X(x)=P(X=x_i)=\sum_{j=1}^{\infty}p_{ij},\quad Y\text{ 是离散的} \tag{1.1.19}$$

$$f_X(x)=\int_{-\infty}^{\infty}f(x,y)\mathrm{d}y,\quad Y\text{ 是连续的} \tag{1.1.20}$$

在计量经济和时间序列分析中经常假定两个随机变量之间独立且同分布（记为 i.i.d），当且仅当联合密度是边际密度的乘积时，两个随机变量才是独立的，即

$$f(x,y)=f_X(x)f_Y(y) \tag{1.1.21}$$

注意，这要求对所有取值都成立。对于例 1.2，显然式（1.1.21）是不成立的，如 $f(600,0)=1/4$，而 $f_X(600)=1/4$，$f_Y(0)=13/24$。因此，收入和是否受过大学教育这两个变量是不独立的。

2. 条件概率函数

在例 1.2 中，如果我们想要知道在受过大学教育的人中，收入为 3 000 元的比例，就是要得到在给定 $Y=1$ 的条件下，$X=3\,000$ 的概率为多少，这就归结为求条件概率的问题。这可以由下面的公式计算：

$$P(X=x_i\mid Y=y_j)=\frac{P(X=x_i,Y=y_j)}{P(Y=y_j)},\quad \text{离散情形} \tag{1.1.22}$$

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)}, \quad \text{连续情形} \tag{1.1.23}$$

例如：

$$P(X = 3\,000 \mid Y = 1) = \frac{P(X = 3\,000, Y = 1)}{P_Y(Y = 1)} = \frac{1/3}{1/3 + 1/8} = \frac{8}{11} \approx 0.727$$

这说明，在受过大学教育的条件下，X 取 3 000 的概率约为 0.727。如果没有这样的条件(无条件概率或边际概率)，X 取 3 000 的概率为 0.5(=1/3+1/6)。这也说明 X，Y 是不独立的变量，Y 的取值影响到 X 取值的概率分布。由式(1.1.21)可知，独立的两个随机变量的条件概率函数应该与无条件概率函数相同。对于连续型随机变量，也有类似性质，只要将概率函数换成概率密度即可。

3. 协方差和相关系数

两个随机变量 X、Y 的协方差定义为

$$\operatorname{cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] \tag{1.1.24}$$

式中：μ_X，μ_Y 分别表示 X，Y 的期望值。

协方差度量了两个变量的同时波动。如果两个变量同方向变动(比如一个变量增加，另一个变量也增加)，则协方差为正；如果两个变量反方向变动(比如一个变量增加，另一个变量却减少)，则协方差为负；如果两个随机变量是独立的，则协方差为 0。

如果两个变量不是独立的，即协方差不是 0，人们自然希望知道它们之间的相关程度有多大，相关系数刻画了这种特征。相关系数 ρ 定义如下：

$$\rho = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{1.1.25}$$

式中：σ_X，σ_Y 分别为 X，Y 的标准差。可以看出两个变量的相关系数等于它们的协方差与各自标准差之比。相关系数是两个随机变量线性相关程度的数字特征，其符号与协方差符号相同。但相关系数经过标准化处理，已经没有量纲，其值在 −1 和 1 之间。如果是 0，表明两变量不相关。

1.2 从总体到样本

1.2.1 基本统计量

统计中将所研究的对象称为总体(population)。总体的某种数量指标 X，作为随机变量称为总体随机变量，通常简称为总体 X，如中国人的年龄等。要想知道总体全部数据常常是困难的，甚至是做不到的。一般只能抽取一部分数据，$x_1, x_2, \cdots, x_N$，即所谓的样本(sample)。统计学的基本任务就是依据样本数据来推断总体，包括推断总体的分布及其数字特征等。

1. 样本均值(mean)

样本的算术平均值定义为

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.2.1}$$

它是总体均值 $E(X)$的一个好的估计量。在统计中，有时还使用加权平均，它是各个数据依照相对重要程度乘以相应的权重后再平均。例如，要计算一揽子商品的平均价格，就需要用每种商品的数量作为权重，价格指数的计算就是利用加权平均。除了算术平均，还有一种很重要的几何平均，即各个数据连乘积的 N 次方根，N 是样本观测值的个数。当根据一个国家一段时期内各期经济增长率数据，要得出这一段时期的平均增长率时，一定要用几何平均来计算。另外，样本中位数(median)是一个关于中心位置的度量，即样本按从小到大排列后的中间值。对于奇数个样本来说，中位数是位于中间的数据点；对于偶数个样本，中位数是两个中间数据的平均值。

2. 样本标准差(standard deviation)

样本标准差衡量了样本值对样本均值的偏离程度，记为 s_x，其计算公式如下：

$$s_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2} \tag{1.2.2}$$

式中：$\bar{x}$为样本均值。在式(1.2.2)中除以 $N-1$ 而不是除以 N，是因为这样得到的样本方差估计量才是无偏估计量。标准差的平方即样本方差 s_x^2 是样本二阶中心矩。类似地，样本三阶矩为

$$\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^3 \tag{1.2.3}$$

样本四阶矩定义为

$$\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^4 \tag{1.2.4}$$

由式(1.2.2)、式(1.2.3)和式(1.2.4)，可以类似总体偏度和峰度，计算样本偏度和峰度。

例 1.3　基本统计量

表 1.2.1 列出了我国 1992—2003 年的实际 GDP 增长率，求出我国这 12 年的平均增长率、标准差、偏度和峰度。

表 1.2.1　GDP 增长率(可比价格)　　%

年份	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
增长率	14.2	13.5	12.6	10.5	9.6	8.8	7.8	7.1	8.0	7.5	8.0	9.1

数据来源：中国统计年鉴. 北京：中国统计出版社，2004.

算术均值：9.725　　几何平均值：9.467　　标准差：2.45

偏度：0.78　　峰度：2.15

3. 样本协方差(covariance)

样本协方差记为 c_{xy}，计算公式如下：

$$c_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}) \tag{1.2.5}$$

式中：$y_1, y_2, \cdots, y_N$是随机变量 Y 的 N 个样本。进而可以计算样本相关系数：

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} = \frac{c_{xy}}{s_x s_y} \tag{1.2.6}$$

在进行经济分析时，经常考察两个变量之间的相关系数。如果相关系数较大，比如正相关接近 1，则说明这两个变量的波动性十分相似，很多个样本点上有这样的关系：一个变量大于其均值时，另一个变量也大于均值。波动的相似性为进一步建立模型等提供了依据。

相关系数计算的是两组样本的同期相关程度。在分析经济周期问题的时候，经常区分先行、一致和滞后经济指标，用来表明经济指标与整个经济景气的同步性。这时，往往需要计算交叉相关(cross correlation)系数。序列 X 与 Y 的交叉相关系数的计算公式如下：

$$r(l) = \frac{c_{xy}(l)}{s_x s_y}, \quad l = 0, \pm 1, \pm 2, \cdots \tag{1.2.7}$$

式中：

$$c_{xy}(l) = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{i=1}^{N-l}(x_i - \bar{x})(y_{i+l} - \bar{y}), & l = 0,1,2,3,\cdots \\ \dfrac{1}{N}\displaystyle\sum_{i=1}^{N+l}(y_i - \bar{y})(x_{i-l} - \bar{x}), & l = 0,-1,-2,\cdots \end{cases} \tag{1.2.8}$$

1.2.2 估计量性质

在许多实际运用中，仅有来自总体的一些样本，而且要用样本矩(如样本方差)来推断总体矩(方差)。如何用有限的样本对总体进行尽可能准确的推断呢？这无疑要求我们寻找到性质优良的统计量。

考虑总体任意的一个参数 β，我们选择一个样本统计量作为 β 的估计量。评判估计量优劣通常有以下几个标准。

1. 无偏性

由于估计量是随机变量，不同的样本值就有不同的估计值，自然希望这些估计值的平均值是待估参数的真值，即估计量的均值等于未知参数的真值。

当 $E(\hat{\beta})=\beta$ 时，称 $\hat{\beta}$ 为 β 的无偏估计量，否则称为有偏估计，并称

$$b = E(\hat{\beta}) - \beta \tag{1.2.9}$$

为估计量 $\hat{\beta}$ 的偏倚(bias)。

由于

$$E(\bar{x}) = \frac{1}{N}\sum_{i=1}^{N}E(x_i) = \mu_X \tag{1.2.10}$$

因此，$\bar{x}$ 是 μ_X 的无偏估计。同样，可以证明式(1.2.2)定义的样本标准差所确定的样本方差是总体方差的无偏估计。无偏估计和有偏估计如图 1.2.1 所示。