第 1 章　概率与统计基础[①]

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. 离散型随机变量概率分布的含义

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：$p\geqslant 0$，$q\geqslant 0$，$p+q=1$。

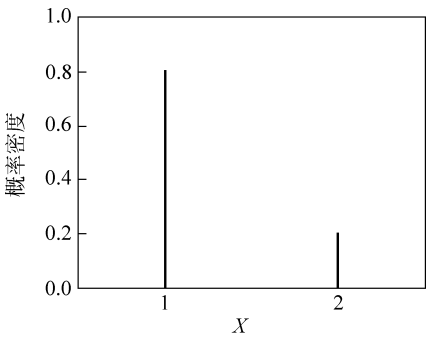


图 1.1.1　离散型概率分布
(经济状态概率分布：$p=0.8$，$q=0.2$)

2. 累积分布函数

对于随机变量 X(无论是连续还是离散)可

① 古扎拉蒂，波特. 经济计量学精要[M]. 张涛，译. 4 版. 北京：机械工业出版社，2010.
伍德里奇. 计量经济学导论[M]. 费剑平，译. 4 版. 北京：中国人民大学出版社，2010.

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,$F(-\infty)=0$,$F(+\infty)=1$。对于离散型随机变量,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(probability density function,PDF),且 $f(x)$具有以下性质:

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

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

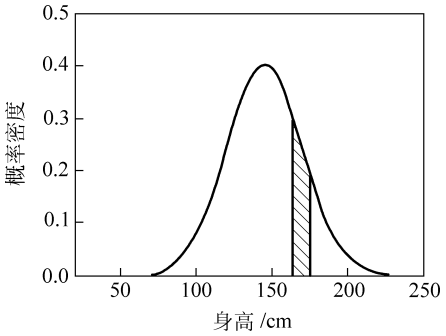


图 1.1.2　连续型(身高)概率分布

令 X 代表身高,用厘米来度量,那么人的身高在某一区间内(如 160～170 cm)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)。

例 1.1　离散型随机变量的 CDF

抛币 4 次,求随机变量(正面朝上的次数)的概率密度函数(PDF)和累积分布函数(CDF)(表 1.1.1)。

表 1.1.1　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

正面朝上的次数 ($X=x_i$)	PDF		CDF	
	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 是例 1.1 的离散型随机变量累积分布函数(CDF)的示意图,图 1.1.4 是连续型随机变量累积分布函数(CDF)的示意图。

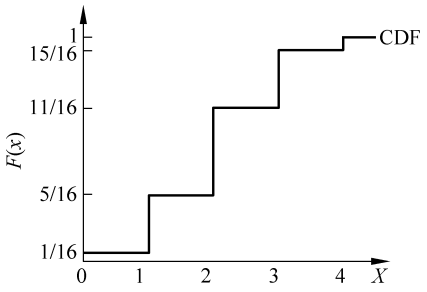


图 1.1.3　离散型随机变量的累积分布函数

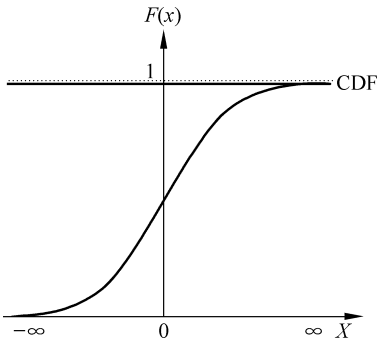


图 1.1.4　连续型随机变量的累积分布函数

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,或简称期望)和方差。期望是随机变量的平均值,它度量了集中趋势;方差是对随机变量偏离期望的离散程度的度量。

1. 数学期望和中位数

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

$$\mu_X = E(X) = p_1x_1 + p_2x_2 + \cdots + p_Nx_N = \sum_{i=1}^{N} p_ix_i \tag{1.1.8}$$

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

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

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

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

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;反之,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_X 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　两个离散型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 分别表示受过大学教育和没有受过大学教育，联合分布如表 1.1.2 所示。

表 1.1.2　联 合 分 布

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（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(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(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相乘的积之比。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。如果值接近于 0,表明变量相关性比较弱;如果绝对值接近 1,说明两个变量的相关性比较强;符号代表相关的方向,即是正相关还是负相关。

1.1.4　从总体到样本

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. 样本均值和中位数

样本的算术平均值(mean)定义为

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.1.26}$$

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。除了算术平均外，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。

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

样本标准差(standard deviation)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：

$$s_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{1.1.27}$$

式中：$\bar{x}$ 为样本均值。在式(1.1.27)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。样本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

$$\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^3 \tag{1.1.28}$$

样本四阶矩为

$$\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^4 \tag{1.1.29}$$

由式(1.1.27)～式(1.1.29)，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 1.1.3 列出了我国 1992—2003 年的实际 GDP(国内生产总值)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 1.1.3　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

资料来源：国家统计局.中国统计年鉴[M].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. 样本协方差和样本相关系数

样本协方差(covariance)记为 c_{xy},计算公式如下：

$$c_{xy}=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}) \tag{1.1.30}$$

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：

$$r=\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2}}=\frac{c_{xy}}{s_x s_y} \tag{1.1.31}$$

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,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：

$$r(l)=\frac{c_{xy}(l)}{s_x s_y},\quad l=0,\pm1,\pm2,\cdots \tag{1.1.32}$$

式中：

$$c_{xy}(l)=\begin{cases}\dfrac{1}{N}\displaystyle\sum_{i=1}^{N-l}(x_i-\bar{x})(y_{i+l}-\bar{y}), & l=0,1,2,3,\cdots\\ \dfrac{1}{N}\displaystyle\sum_{i=1}^{N+l}(y_i-\bar{y})(x_{i-l}-\bar{x}), & l=0,-1,-2,\cdots\end{cases} \tag{1.1.33}$$

1.2　一些重要的概率分布

在计量经济分析中,有几类分布尤为重要,其中最常用的有以下几种：正态分布、χ^2 分布、t 分布和 F 分布。在进行统计推断时,经常假定样本来自正态分布的总体。

1.2.1　正态分布

正态分布(normal distribution)是一个连续的、形状为钟形的概率分布。如图 1.2.1 所示,正态分布可以由它的均值 μ 和方差 σ^2 完全描述出来。如果 X 服从正态分布,可以记为 $X\sim N(\mu,\sigma^2)$,密度函数为

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-(x-\mu)^2/(2\sigma^2)} \tag{1.2.1}$$

由图 1.2.2 可以看出,方差小的正态分布更尖,这是因为均值处的密度值为 $1/(\sqrt{2\pi}\sigma)$。

我们将均值为 0、方差为 1 的正态分布称为标准正态分布(形如图 1.2.2 中方差为 1 的图形),这时的密度函数将更加简洁:

$$\phi(z)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-z^2/2} \tag{1.2.2}$$

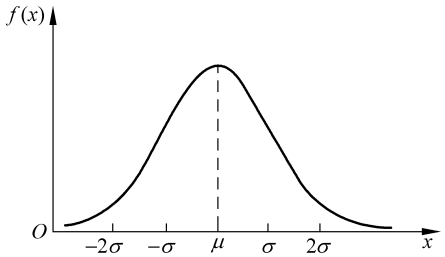


图 1.2.1　正态分布的密度函数

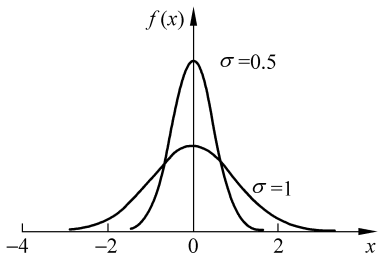


图 1.2.2　同均值不同方差的密度函数

正态分布的几个重要性质是极其有用的。

① 正态分布密度函数以其均值为中心对称。

② 正态分布的概率密度函数的图形中间高、两边低,在均值处达到最高,而两端概率很小。

③ 正态分布随机变量的线性组合仍服从正态分布。令

$$X\sim N(\mu_X,\sigma_X^2)$$

$$Y\sim N(\mu_Y,\sigma_Y^2)$$

并且假定 X、Y 相互独立,则它们的线性组合 $W=aX+bY$ 也服从正态分布,且

$$W\sim N(a\mu_X+b\mu_Y,a^2\sigma_X^2+b^2\sigma_Y^2) \tag{1.2.3}$$

④ 正态分布的线性变换仍然服从正态分布,即若 X 服从均值为 μ、方差为 σ^2 的正态分布,则 $a+bX$ 也服从正态分布,均值为 $a+b\mu$,方差为 $b^2\sigma^2$。这个性质很重要,因为任何一个正态分布都能够根据这个性质变换成标准正态分布,即

$$Z=\frac{X-\mu}{\sigma}\sim N(0,1) \tag{1.2.4}$$

再根据

$$P(a<X<b)=P\left(\frac{a-\mu}{\sigma}<\frac{X-\mu}{\sigma}<\frac{b-\mu}{\sigma}\right) \tag{1.2.5}$$

这样就可以通过标准正态分布表查到正态分布变量在某个区间取值的概率。

在使用标准正态分布 $Z\sim N(0,1)$时,常用记号 z_α 表示满足条件

$$P(Z>z_\alpha)=\alpha,\quad 0<\alpha<1 \tag{1.2.6}$$

的点,称点 z_α 为标准正态分布的上 α 分位数。例如,由查表(附录 A,表 A1)可知 $P(Z>1.96)=0.025$,即标准正态分布的上 0.025 分位数 $z_{0.025}$ 是 1.96。或者可以定义下 α 分位数 z_α,即满足 $P(X<z_\alpha)=\alpha$。

由于正态分布是对称的,根据分位数的定义,有下式成立:

$$P(-z_{\alpha/2}<Z<z_{\alpha/2})=1-\alpha \tag{1.2.7}$$

对于 $\alpha=0.05$,式(1.2.7)就成为

$$P(-1.96<Z<1.96)=95\%$$

因此,对于任意的正态分布 X,再由式(1.2.5)可以知道

$$P\left(-1.96<\frac{X-\mu}{\sigma}<1.96\right)=P(\mu-1.96\sigma<X<\mu+1.96\sigma)=95\% \tag{1.2.8}$$

即正态随机变量 X 的观测值落在距均值距离为两倍标准差范围内的概率约为 0.95。因此,一般经验地认为,正态分布曲线下的面积约有 68%处于$(\mu-\sigma)$和$(\mu+\sigma)$之间;约有 95%的面积位于$(\mu-2\sigma)$和$(\mu+2\sigma)$之间;约有 99.7%的面积位于$(\mu-3\sigma)$和$(\mu+3\sigma)$之间。这些结果在统计推断中将有重要的作用。

除此之外,有时还可以借助分位数来判断某一个序列与哪一种概率分布最为接近,从而对这个序列所来自的总体作出符合实际情况的假定。

实际中,很多随机变量服从正态分布。人类的身高、体重和考试得分等的分布大体都类似于正态图形状,并且中心极限定理告诉我们:如果一个随机变量 X 具有均值 μ 和方差 σ^2,则随着样本容量 N 增加,$\bar{x}$ 越来越接近于均值为 μ、方差为 σ^2/N 的正态分布。还有一些变量可以通过变换具有正态性,最常见的是取自然对数。当所研究的变量 X 是正值,如收入、价格,如果取自然对数后服从正态分布,就说 X 服从对数正态分布。

1.2.2　χ^2 分布

统计学中另一个常用的概率分布是 χ^2 分布[chisquare(χ^2)distribution]。我们已经知道如果随机变量 X 服从均值为 μ、方差为 σ^2 的正态分布,则 $Z=(X-\mu)/\sigma\sim N(0,1)$。统计理论证明:标准正态分布的平方服从自由度(degree of freedom,d.f.)为 1 的 χ^2 分布,即

$$Z^2\sim\chi^2(1) \tag{1.2.9}$$

上式括号中 1 表示自由度,正如均值、方差是正态分布的参数一样,自由度是 χ^2 分布的参数。在这里,自由度是平方和中独立变量的个数。如果令 $Z_1,Z_2,\cdots,Z_k$ 为 k 个独立的服从标准正态分布的随机变量,则它们的平方和服从自由度为 k 的 χ^2 分布,即

$$X=\sum Z_i^2=Z_1^2+Z_2^2+\cdots+Z_k^2\sim\chi^2(k) \tag{1.2.10}$$

上式括号中 k 为 χ^2 分布的自由度,因为在此平方和中有 k 个独立的变量自由取值。如

果这些变量存在约束，自由度将降低。不同 k 值的 χ^2 分布的密度函数如图 1.2.3 所示。

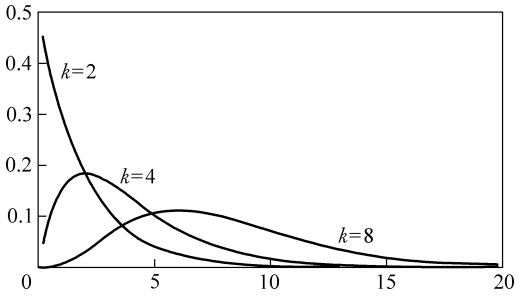


图 1.2.3　χ^2 分布的密度函数

χ^2 分布具有如下重要性质。

① 与正态分布不同，χ^2 分布只取正值，并且是偏斜分布。其偏度取决于自由度的大小，自由度越小越右偏，随着自由度增大，χ^2 分布逐渐对称，接近正态分布。R. A. Fisher 曾证明，当 N 充分大时，近似地有

$$Z=\sqrt{2\chi^2}-\sqrt{2N+1}\sim N(0,1) \tag{1.2.11}$$

② χ^2 分布具有期望为其自由度 k、方差为 $2k$ 的特殊性质。

例 1.4　正态分布的一种检验方法——Jarque-Bera 统计量

Jarque-Bera 统计量是用来检验一组样本是否能够认为来自正态总体的一种方法，其计算公式如下：

$$JB=\frac{T-k}{6}\left[S^2+\frac{1}{4}(K-3)^2\right] \tag{1.2.12}$$

式中：S、K 分别表示偏度和峰度。在正态分布的假设下，Jarque-Bera 统计量服从自由度为 2 的 χ^2 分布。若为原始序列，$k=0$；若序列是通过模型估计得到的，k 为估计的参数个数。本例中，$k=0$。例 1.3 中计算了我国 1992—2003 年的实际 GDP 增长率的一些重要的样本统计量。本例计算出实际 GDP 增长率的 JB 统计量值为 1.57，自由度为 2 的 χ^2 分布，查表（附录 A，表 A2）得，$\chi^2(2)$ 值大于等于 1.386 的概率是 0.5，大于等于 2.773 的概率是 0.25，因此 $\chi^2(2)$ 值大于等于 1.57 的概率应为 0.25～0.5。事实上，EViews 软件能直接计算出这个概率，本例 $\chi^2(2)$ 值大于等于 1.57 的概率约为 0.455。如果认为这个概率够大，可以认为样本的确来自正态分布总体。相反，如果 JB 统计量值较大，如为 11，则可以计算出 $\chi^2(2)$ 值大于 11 的概率约为 0.004，这个概率过小，因此不能认为样本来自正态分布。

1.2.3　t 分布

计量经济学中另一个广泛使用的概率分布是 t 分布（t distribution），又称学生氏的 t 分布（student's t distribution）。它与正态分布密切相关，可以从一个标准正态分布和

一个 χ^2 分布得到。

设 Z 服从标准正态分布,X 服从自由度为 k 的 χ^2 分布,并且两者相互独立,于是随机变量

$$t=\frac{Z}{\sqrt{X/k}} \tag{1.2.13}$$

服从自由度为 k 的 t 分布。

对于来自正态总体的样本,对样本均值 $\bar{x}$ 进行标准化可以得到 $(\bar{x}-\mu)/(\sigma/\sqrt{N})$。它是一个均值为 0、方差为 1 的标准正态分布,并且,可以证明,如果来自方差为 σ^2 的一个正态分布的 N 个观测值的样本方差为 s^2,则 $(N-1)s^2/\sigma^2\sim\chi^2(N-1)$,因此有

$$\frac{(\bar{x}-\mu)/(\sigma/\sqrt{N})}{\sqrt{(N-1)s^2/\sigma^2(N-1)}}=\frac{(\bar{x}-\mu)}{s/\sqrt{N}}\sim t(N-1) \tag{1.2.14}$$

式(1.2.14)告诉我们,当总体方差 σ^2 已知时,可以利用减去均值除以标准差的方式对正态分布进行标准化;而当总体方差 σ^2 未知时,可以用样本标准差代替总体标准差,只是这时得到的分布不再是标准正态分布,而是自由度为 $N-1$ 的 t 分布。

和正态分布一样,t 分布是对称的。t 分布的随机变量期望值为 0,方差为 $k/(k-2)$。这可以看出,其方差大于标准正态分布的方差 1,因此 t 分布的尾部比正态分布更厚。但随着自由度 k 的增加,方差收敛于 1,即当自由度很大时,它趋近于正态分布。这些特征通过图 1.2.4 都可以表现出来。在图中,将自由度分别为 2 和 60 的 t 分布与标准正态分布密度函数画在一起,但由于标准正态分布和自由度为 60 的 t 分布几乎重合,因此只看到两条曲线。事实上,对于样本容量较大的 t 分布,可以用正态分布来近似。

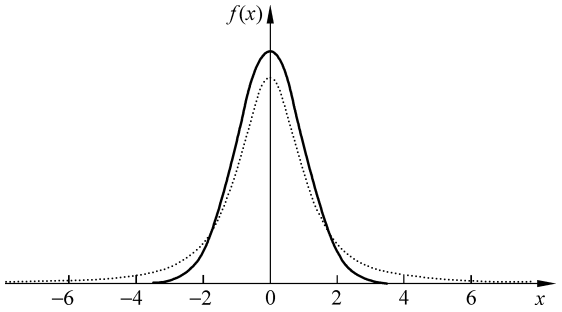


图 1.2.4　t 分布(虚线)和标准正态分布(实线)密度函数

由于当自由度较大时,t 分布趋近于服从标准正态分布,因此有

$$p(-1.96<t<1.96)\approx 95\% \tag{1.2.15}$$

这对假设检验最有用。当人们对回归分析得到的结果进行分析时,首先看各个变量的系数是否显著异于 0。这可以通过 t 值的绝对值是否大于 2 来判断。再次提醒注意的是,自由度必须足够大,否则将是危险的,因为这对于自由度较小的 t 分布不成立,如自由度为 5 的时候,$t_{0.025}=2.571$。

1.2.4　*F* 分布

在多元回归分析中，常用到 F 分布检验模型的显著性。F 分布是计量经济学中又一种重要的概率分布。如果两个服从 χ^2 分布的随机变量相互独立，其自由度分别为 k_1 和 k_2，则

$$F(k_1,k_2)=\frac{\chi^2(k_1)/k_1}{\chi^2(k_2)/k_2} \tag{1.2.16}$$

服从自由度为(k_1,k_2)的 F 分布，其中 k_1 和 k_2 分别为分子自由度和分母自由度。

为说明 F 分布的作用，假设有两组样本容量分别为 N_1 和 N_2 的样本，分别来自两个正态分布 X,Z，并且不妨假定方差相同：$\sigma_X^2=\sigma_Z^2=\sigma^2$。由于对于正态分布的总体而言，$(N-1)s^2/\sigma^2$ 服从自由度为 $N-1$ 的 χ^2 分布，因此可以依据 F 分布的定义构造如下 F 分布：

$$\frac{(N_1-1)s_X^2/\sigma_X^2}{N_1-1}\Big/\frac{(N_2-1)s_Z^2/\sigma_Z^2}{N_2-1}=\frac{s_X^2}{s_Z^2}\sim F(N_1-1,N_2-1) \tag{1.2.17}$$

这样，如果要检验方差相同的假设是否成立就很容易了。

F 分布与 χ^2 分布类似，只取非负值并且是斜分布，随自由度逐渐增大，F 分布逐渐对称，接近正态分布(图 1.2.5)。

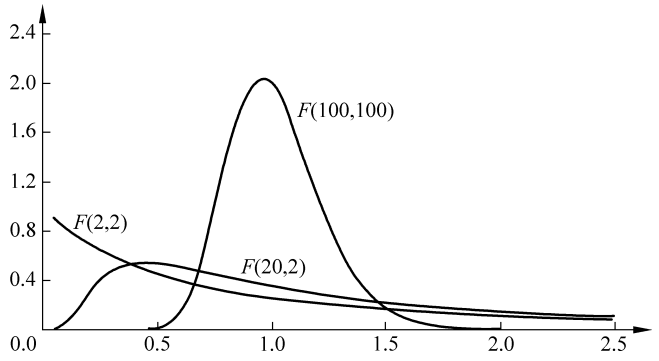


图 1.2.5　不同自由度下的 F 分布密度函数

由 t 分布和 F 分布的定义可以看出，t 分布变量的平方服从分子自由度为 1、分母自由度为 k 的 F 分布，即

$$t^2(k)=F(1,k) \tag{1.2.18}$$

在计量经济学中，具有大分母自由度的 F 分布很普遍，当 k_2 无限大时，F 的分母收敛为 1，这时 F 分布与 χ^2 分布存在如下关系：

$$F(k_1,k_2)=\chi^2(k_1)/k_1 \tag{1.2.19}$$

即 $\chi^2(k_1)$变量与其自由度之比近似为分子自由度为 k_1、分母自由度很大的 F 分布。

例 1.5　消费和 GDP 波动相等性检验(方差相等性检验)

表 1.2.1 列出了我国 1992—2000 年的实际 GDP 增长率和居民消费增长率。在经济周期分析中,人们十分关心相对于宏观经济总的波动,经济系统中哪些经济行为的波动更剧烈。本例用 GDP 的波动表示宏观经济总的波动,简单地研究我国居民消费行为和 GDP 波动的对比。

表 1.2.1　*GDP_R* 和 *CS_R* 数据　%

年份	*GDP_R*	*CS_R*
1992	14.2	12.9
1993	13.5	8.1
1994	12.6	4.3
1995	10.5	7.5
1996	9.6	9.1
1997	8.8	4.2
1998	7.8	5.5
1999	7.1	7.9
2000	8.0	9.1

资料来源:国家统计局.中国统计年鉴[M].北京:中国统计出版社,2003.

首先计算得到实际 GDP 增长率(GDP_R)和居民实际消费增长率(CS_R)的样本标准差分别为 2.628 和 2.733。根据样本方差的数值,可以看出在 1992—2000 年,消费增长率的波动程度与 GDP 增长率波动程度很接近。

下面假设这两个样本都来自正态总体(JB 统计量分别为 0.91 和 0.32,其 p 值分别为 0.63 和 0.85,表明正态性假设是合理的),那么能否认为这两个总体是同方差的呢?

利用式(1.2.16)计算 F 值:

$$F=\frac{2.733^2}{2.628^2}\approx 1.082$$

它服从分子、分母自由度分别为 $k_1=8, k_2=8$ 的 F 分布,F 值大于 1.082 的概率为 $p\approx 0.91$。如果我们认为这个概率相当大,则可以得出结论,两总体同方差,但由假设检验的知识可以知道,作出接受假设的判断需要谨慎。

1.3　统 计 推 断

统计推断就是根据来自总体的样本对总体的种种统计特征作出判断,参数估计和假设检验是统计推断的两个孪生分支问题。对于前者我们并不陌生,在前面的介绍中已经利用样本均值和方差作为总体的均值和方差的估计量,这样的估计是参数的点估计。这部分将介绍参数的区间估计和各种假设检验问题。

1.3.1　参数估计

假设一个服从正态分布的随机变量 X 均值 μ_X 未知,现在通过抽样得到一组样本,样本容量为 N,显然可以选择样本均值作为总体均值 μ_X 的估计值。由于每次抽取的样本可能不同,可以计算得到不同的数值,因此从这个意义上说它是随机变量,通常称为点估计量。这是对参数进行的点估计(point estimation)。通常进行点估计的方法有

矩估计和极大似然估计方法，如用样本均值（样本一阶矩）作为总体均值（总体一阶矩）的估计就属于矩估计。本书后面各章节将用到这些方法对各种不同的模型进行参数估计。

由于总体均值是某一个特定的数值，然而点估计值随着抽取的样本不同可能取不同的值，无法断定总体均值确切的数值，因此自然的想法是给出一个包含了总体均值的区间，这就是区间估计（interval estimation）的基本思想。可以想象，给出的区间越大，包括均值的概率将越大，然而区间过大，则几乎不能提供任何有用的信息。因此，既希望包含总体均值的区间小些，同时还希望这个区间包含总体均值的概率大一些，这是一对矛盾，应该尽量给出一个能以较大的概率将未知均值包括进来的较小区间。

我们知道，如果随机变量 $X\sim N(\mu_X,\sigma^2)$，则

$$\bar{x}\sim N(\mu_X,\sigma^2/N) \tag{1.3.1}$$

将其标准化得到

$$Z=\frac{(\bar{x}-\mu_X)}{\sigma/\sqrt{N}}\sim N(0,1) \tag{1.3.2}$$

通常来讲，方差 σ^2 也是未知的，但可以用其样本估计量 $s^2=\sum(x_i-\bar{x})^2/(N-1)$ 代替，则有

$$t=\frac{\bar{x}-\mu_X}{s/\sqrt{N}} \tag{1.3.3}$$

服从自由度为$(N-1)$的 t 分布。这样就有

$$P\left(-t_{\alpha/2}<\frac{\bar{x}-\mu_X}{s/\sqrt{N}}<t_{\alpha/2}\right)=1-\alpha \tag{1.3.4}$$

整理得到

$$P(\bar{x}-t_{\alpha/2}s/\sqrt{N}<\mu_X<\bar{x}+t_{\alpha/2}s/\sqrt{N})=1-\alpha \tag{1.3.5}$$

此时，称这个区间为置信度是 $1-\alpha$ 的置信区间。如果取 $\alpha=0.05$，此区间即为 μ_X 的置信度为 95%的置信区间。

需要注意，区间是随机的，根据不同样本观测值会得到不同的区间。而总体均值 μ_X 虽然未知，却是一个固定的值，是非随机的。所以，式(1.3.5)应理解为区间包括真实 μ_X 的概率是 0.95（这意味着抽样 100 次得到 100 个置信区间，将有 95 次的区间包括真实值），而不能读作 μ_X 落在区间中的概率是 0.95，因为 μ_X 不是随机变量。

更一般地，假定随机变量 X 服从某一概率分布，若要对其参数 θ 进行估计，选取容量为 N 的随机样本，再由选取的置信度$(1-\alpha)$找出两个统计量 L 和 U：

$$P(L\leqslant\theta\leqslant U)=1-\alpha,\quad 0<\alpha<1 \tag{1.3.6}$$

即从 L 到 U 的随机区间包括真实 θ 的概率是 $1-\alpha$，称随机区间$[L,U]$为 θ 的置信区间，α 称为显著水平（level of significance）或犯第一类错误的概率，显著性水平一般为 5%或 1%。由式(1.3.5)可以看出，不同的显著水平对应不同宽度的置信区间，显著水平越小即

置信度越大,置信区间越宽。例如,选择 1%的显著性水平就会得到一个比选择 5%的显著水平宽的置信区间。置信区间的建立使得统计推断的另一分支——假设检验的理解更加容易。

例 1.6　参 数 估 计

在例 1.3 中,如果认为 GDP 增长率是来自正态分布的样本,则根据样本值求正态总体均值的矩估计和置信度为 95%的置信区间。

由于样本均值 $\bar{x}$ 为 9.725%,因此总体均值 μ 的矩估计值是 9.725%。

查表(见附录 A,表 A3)得出自由度为 11 时,$t_{0.025}=2.201$,由式(1.3.5)得

$$P(9.725-2.2\times 2.45/\sqrt{12}<\mu<9.725+2.2\times 2.45/\sqrt{12})=95\%$$

即置信度为 95%的置信区间为(8.169%,11.281%)。

1.3.2　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估计量。

在例 1.6 中,用样本均值 $\bar{x}$ 作为总体均值 μ_X 的点估计量,并得到了 μ_X 的区间估计量。在实践中,样本均值是度量总体均值最广泛使用的统计量,下面以样本均值为例,介绍评判估计量优劣的几个标准。

1. 无偏性

如果总体参数有若干个估计量,且其中的一个或几个估计量的均值等于未知参数的真实值,就称这些估计量是参数的无偏估计量(unbiased estimator)。设未知参数为 θ,$\hat{\theta}$ 为参数估计量,如果估计量 $\hat{\theta}$ 满足

$$E(\hat{\theta})=\theta \tag{1.3.7}$$

则 $\hat{\theta}$ 称为 θ 的无偏估计量,否则称为有偏估计,并称

$$\delta=E(\hat{\theta})-\theta \tag{1.3.8}$$

为估计量 $\hat{\theta}$ 的偏倚(bias)。

由于

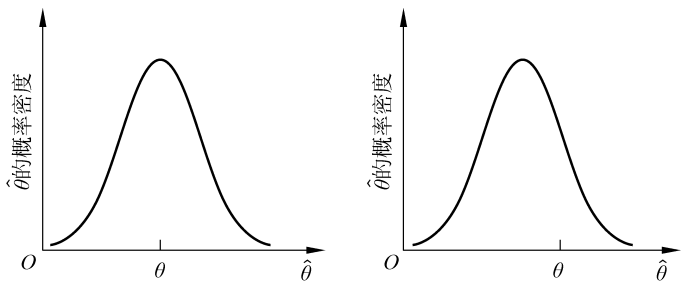
$$E(\bar{x})=\frac{1}{N}\sum_{i=1}^{N}E(x_i)=\mu_X \tag{1.3.9}$$

因此,$\bar{x}$ 是 μ_X 的无偏估计。同样,可以证明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 1.3.1 所示。

2. 最小方差性

对于 θ 的两个估计量,如果其中一个更密集在 θ 附近,则更加适合作为它的估计。方

差可用来衡量密集程度，方差越小越密集。如果估计量 $\hat{\theta}$ 的方差比其他任何估计量的方差都小，则 $\hat{\theta}$ 称为最小方差估计量。可以证明，$\bar{x}$ 是 μ_X 的最小方差估计量。



(a) 无偏估计　　(b) 有偏估计

图 1.3.1　无偏估计和有偏估计

3. 有效性

虽然无偏性是一个理想的性质，但它本身并不充分。如果参数 θ 有两个或更多个无偏估计量，该如何选择呢？

在参数 θ 的所有无偏估计量中，方差最小的估计量被称为最优或有效估计量(efficient estimator)。因此，$\bar{x}$ 是 μ_X 的最小方差无偏估计量，即最优或有效估计量。

4. 一致性(相容性)

一致性是指样本容量增加时，估计量 $\hat{\theta}$ 越来越接近真值 θ。更准确地说，如果 $\hat{\theta}$ 依概率收敛于 θ，则称 $\hat{\theta}$ 为 θ 的一致估计量，即

$$\lim_{N\to\infty} P(|\theta-\hat{\theta}|<\delta)=1 \tag{1.3.10}$$

也就是说，当样本数趋向无穷时，对于任何 $\delta>0$，$|\theta-\hat{\theta}|$ 小于一个任意小的正数 δ 的概率趋于 1。

1.3.3　假设检验

假设检验与置信区间关系密切。如例 1.6，假如认为我国 GDP 增长率的均值为 10%，如何根据样本检验这个假设的正确性呢？用假设检验的语言，类似均值为 10%这样的假设称为原假设或零假设，记为 H_0：$\mu_X=10\%$，它通常和备选假设 H_1 成对出现。备选假设有单边和双边两种形式，如 H_1：$\mu_X>10\%$或 H_1：$\mu_X<10\%$都属于单边检验，而 H_1：$\mu_X\neq10\%$是双边检验。

为了检验原假设，我们根据样本数据以及统计理论建立判定规则来判断样本信息是否支持原假设。如果支持，我们就不拒绝 H_0；否则，拒绝 H_0 而接受备择假设 H_1。有两个互补的办法来判定原假设：置信区间法和显著性检验。

1. 置信区间法

例 1.6 求出了总体均值置信度为 95%的置信区间是[8.169%,11.281%],对于总体均值为 10%的原假设的真伪判断是清楚的。既然假设的值被根据样本求出来的置信度较高的区间包含,那么我们无法拒绝这样的假设。因此,置信区间可以用来判断原假设的真伪,如果置信区间包含了原假设,则不能拒绝原假设。相反,如果原假设没有包含在置信区间内,如原假设是 $H_0:\mu_X=8\%$,则拒绝原假设,而接受备选假设。用假设检验的语言,置信区间可称为接受区域(acceptance region),接受区域以外的称为临界区域(critical region)或拒绝区域(region of rejection),接受区域的上界和下界称为**临界值**(critical values)。这样,可以表述为:如果参数值在原假设下位于接受区域内,则不拒绝原假设;如果落在接受区域外(落在拒绝区域内),则拒绝原假设。显然,临界值是判断接受或拒绝原假设的分界线。

对同一组样本而言,临界值与显著性水平的选择一一对应。在进行假设检验时,如何选择显著性水平呢?这涉及下面的两类错误,即第一类错误和第二类错误。第一类错误也称为弃真错误。如果检验的原假设是 $H_0:\mu_X=10\%$,计算出来的置信区间没有包括 10%,经过检验(显著性水平为 5%),则将拒绝原假设,但是如果总体均值的确是 10%,显然这个判断是错误的,这种错误称为第一类错误。由于置信区间包含真值的概率是 95%,显然犯第一类错误的概率(没有包含真值的概率)为 5%。如果经检验置信区间包含了原假设的值,正像我们列举的这种情况,则不能拒绝原假设,然而置信区间内包含大量的数值,任何一个值作为原假设都无法拒绝,因此,如果事实上真值为 9%而不是 10%,10%在置信区间内,因而只能接受原假设,这种错误称为第二类错误,也叫存伪错误。我们自然想减少这两类错误,但是对于一个给定样本,我们无法做到犯两类错误的概率都很小,通常的做法是选择相当小的显著性水平减少犯第一类错误的概率。因此,拒绝原假设通常是有把握的,然而接受原假设犯错误的风险可能较大。

2. 显著性检验

显著性检验(test of significance)是另一种较为简洁但完备的假设检验方法。

根据式(1.2.13):

$$t=\frac{\bar{x}-\mu_X}{s/\sqrt{N}} \tag{1.3.11}$$

在利用置信区间进行假设检验时,要根据式(1.3.5)计算得到真值 μ_X 的置信区间,然后判断这个区间是否包含原假设设定值。现在我们另辟蹊径,直接将原假设的设定值代入式(1.3.11),可以计算出唯一的 t 的数值来,同时也很容易得到大于等于该 t 值的概率。如果 $\bar{x}$ 和 μ_X 差别不大,则 $|t|$ 可能会很小,在此情况下,接受原假设;如果 $\bar{x}$ 和 μ_X 差别大,则 $|t|$ 可能会很大。根据 t 分布表,对于给定的自由度,$|t|$ 越大,则获得此 $|t|$ 的概率越小。由于在原假设成立的条件下,根据式(1.3.11)得到的 t 值是服从 t 分布的,所以如果 $|t|$ 超出了给定显著性水平对应的临界值,则拒绝原假设,这就是显著性检验的基

本思想。

以上假设检验是通过构造出服从 t 分布的统计量实现的，因此称为 t 检验。在对回归模型的系数进行显著性检验时，用到的就是 t 检验。所谓统计显著，一般是指能够拒绝原假设，接受了原假设则认为统计不显著。

例 1.7　t 检　验

例 1.6 求出了我国 GDP 增长率均值的置信度为 95%的置信区间，也介绍了由置信区间法判断能否拒绝某一个原假设的方法。下面利用 t 检验再次进行检验 $H_0:\mu_X=10\%$，$H_1:\mu_X\neq10\%$。

直接将原假设的设定值代入式(1.3.11)，可以计算出 t 的数值：

$$t=\frac{\bar{x}-\mu_X}{s/\sqrt{N}}=\frac{9.725-10}{2.45/\sqrt{12}}\approx-0.389$$

给定显著性水平为 5%，自由度为 12，查表(附录 A，表 A3)得出 $t_{0.025}=2.179$，由于 $|t|$ 为 0.389，小于临界值 2.179，因此不能拒绝原假设。

这种检验与置信区间法是完全对应的，只要原假设的值处于置信区间中，计算出来的 t 值也一定处于临界值内，这是显而易见的。这也说明，本例中虽然不能拒绝原假设，但是并不能认为 10%就是总体均值，因为只要原假设的值在[8.169%，11.281%]内，计算出来的 t 值一定会在 5%的显著性水平所对应的临界值内。

在例 1.7 中，由于备选假设是双边的，所以，可以将犯第一类错误的风险均分在 t 分布的两侧，即 t 值大于 $t_{0.025}$ 或小于 $-t_{0.025}$ 都将拒绝原假设。如果检验的备选假设是单边的，比如是 $H_1:\mu_X>10\%$，则只需要一个临界值 $t_{0.05}$，如果 t 值大于 $t_{0.05}$ 就拒绝原假设。

3. 显著性水平的选择与 p 值

除了用计算出的 t 值与根据设定的显著性水平查表得到的临界值比较进行检验的经典方法，实践中更好的方法是利用 p(probability)值来判断。例如在自由度为 20 时，我们计算出了 t 值为 3.552，根据 t 分布表(附录 A，表 A3)，得出 $p(t>3.552)=0.001$(单边检验)，即此 t 值对应的 p 值为 0.001。如果此时设定的显著性水平是 1%，其所对应的临界值(2.528)一定在 t 值的左侧，根据显著性检验知道，这时将拒绝原假设。可见，t 值越大，p 值越小，就越能拒绝原假设。给出了 p 值就避免在选择显著水平时的随意性，如我们可以根据计算的 p 值(如 0.001)，得出在 0.001 的显著性水平下拒绝原假设的结论。双边检验的情况下，为了可以和显著性水平直接对比，p 值是单边情形的两倍。在例 1.4 中，检验了单边检验；在例 1.5 中，如果是单边检验，结论不变，如果是双边检验，p 值应是原来的两倍，即 91.39%。

4. χ^2 显著性检验和 F 显著性检验

除了 t 检验外,在后面的章节中还将遇到建立在 χ^2 分布和 F 分布基础上的显著性检验,它们的检验原理和机制都是相同的。我们可以通过例 1.4 和例 1.5 看到这两种检验方法,这里不再说明。

1.4　EViews 软件的相关操作[①]

1.4.1　单序列的统计量

打开工作文件,双击一个序列名(series),进入序列的对话框。单击序列对象菜单的视图(view)可看到此菜单分为四个部分:第一部分为序列显示形式,第二部分和第三部分提供数据统计方法,第四部分是标签。

第二部分的第一项是描述统计量(Descriptive Statistics & Test),其中的第一项为 Histogram and Stats,即以直方图显示序列的频率分布,同直方图一起显示的还有一些标准的描述统计量。这些统计量都是由样本中的观测值计算出来的,包括均值(mean)、中位数(median)、最大和最小值(max and min)、标准差(standard deviation)、偏度(skewness)、峰度(kurtosis)、Jarque-Bera 统计量及概率(probability),即 1.3.3 节中介绍的 p 值,例 1.4 详细介绍了 Jarque-Bera 统计量。

1.4.2　多序列的显示和统计量

对几个感兴趣的序列之间的关系的直观描述可以通过"组"中提供的方便功能来实现。通过组,可以计算几种统计量、描述不同序列之间的关系,并以表格、数据表和图等各种方式显示出来。组窗口内的 View 下拉菜单分为 4 个部分:第一部分包括组中数据的各种显示形式;第二部分包括各种基本统计量;第三部分为时间序列的特殊统计量;第四部分为标签项,提供组对象的相关信息。以下假定组中包含 G 个序列。

1. N 步列表、联合概率和独立性检验

例 1.2 给出的联合分布表可以通过 N 步列表以更加清晰的方式来表现。

打开由 X,Y 组成的组,在组对象菜单中选择 View/N-way Tabulation...选项,然后在出现的对话框的复选框上选择 Overall%(每个区间观测值占总个数的百分比),在出现的结果窗口的最下方的表格中显示了联合密度和边际密度。

对于一组样本,它计算出满足 X,Y 的各种分类条件的样本数的比例。例 1.2 中 X 分为 3 类,Y 分为两类;最后一列的每一个值是对应行的总和,即 X 的边际密度;最后一行的每一个值是对应列的总和,即 Y 的边际密度。

---

① EViews 10 User's Guide Ⅰ,IHS Global Inc.,2017,Chapter 11:401-545;Chapter 12:547-616.

如果这是根据总体得到的，可以严格地检验这两个总体的独立性，即是否总是有联合密度等于两个边际密度的乘积。如果是由样本得到的结果，这时需要通过假设检验来判断 X，Y 是否独立。默认情况下，EViews 软件会显示 χ^2 统计量检验组中序列的独立性。

$$\text{Pearson}\chi^2=\sum_{ij}\frac{(\hat{n}_{ij}-n_{ij})^2}{\hat{n}_{ij}} \tag{1.4.1}$$

$$\text{Likelihood Ratio}=2\sum_{ij}n_{ij}\log\left(\frac{n_{ij}}{\hat{n}_{ij}}\right) \tag{1.4.2}$$

式中：n_{ij} 为满足 X 第 i 类($i=1,2,3$)和 Y 第 j 类($j=1,2$)的条件的样本个数，也就是样本总数乘以相应的联合密度；$\hat{n}_{ij}$ 为 X 和 Y 相应的边际密度的乘积再乘以样本总数。理论上，如果两个总体 X 和 Y 独立，联合密度应该等于边际密度乘积，因而将有 n_{ij} 等于 $\hat{n}_{ij}$。因此，对于样本序列 X 和 Y 独立的检验统计量的构造基于 n_{ij} 和 $\hat{n}_{ij}$ 之间的差别，如果独立的原假设成立，它们的差别应该很小。两个统计量都渐近服从自由度为$(I-1)(J-1)$的 χ^2 分布，其中，I，J 为每个序列类的个数。

2. 协方差和交叉相关系数

EViews 软件可以计算两个序列的协方差、相关系数和交叉相关系数等统计量。在 View 菜单中选择 Covariances Analysis 选项，可以根据式(1.1.30)、式(1.1.31)计算协方差矩阵、相关矩阵，也可以计算其他统计量。选择 Cross Correlation 选项，可以根据式(1.1.32)计算交叉相关系数。唯一不同的是，EViews 软件在这里计算协方差和方差时，自由度是样本个数 N 而不是 $N-1$。

交叉相关系数显示组中前两个序列的交叉相关。交叉相关图的每栏中两侧虚线对应着正负 2 倍标准差，近似计算为 $\pm 2/\sqrt{N}$。图 1.4.1 计算了例 1.5 中消费 CS_R 和 GDP_R 两个序列的交叉相关系数。第 1 列与式(1.1.33)中第 2 个公式相对应，第 2 列与式(1.1.33)中第 1 个公式相对应，i 是延迟数，lag 和 lead 是相应的交叉相关系数。可

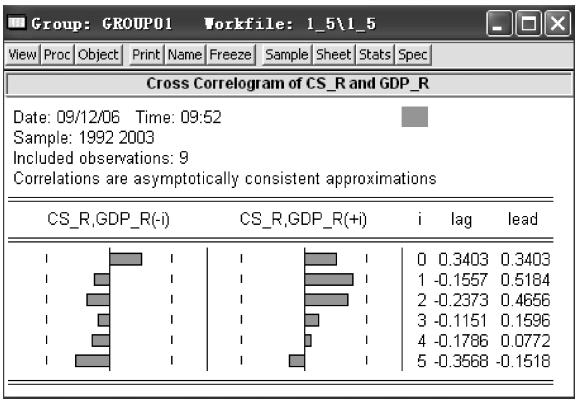


图 1.4.1　消费增长率和 GDP 增长率交叉相关系数的输出结果

以看出,CS_R 和 GDP_R 的同期相关系数、CS_R 与 $GDP\_R(+i)$ 和 $GDP\_R(-i)$ 的相关系数都较高,因此可以认为消费增长率与 GDP 增长率波动基本一致,略微先行(lead)。同时要注意,最大的相关系数虽然接近,但也没有超过 2 倍的标准差,因此还不是非常显著的相关。

1.4.3　分布函数

对一个序列,例 1.4 介绍了利用 JB 统计量来判断其是否服从正态分布。EViews 软件中为我们提供了判断序列与哪种理论上的分布函数最为接近的几种方法。当选择 View/Graphs 命令时,会出现图 1.4.2 所示界面。

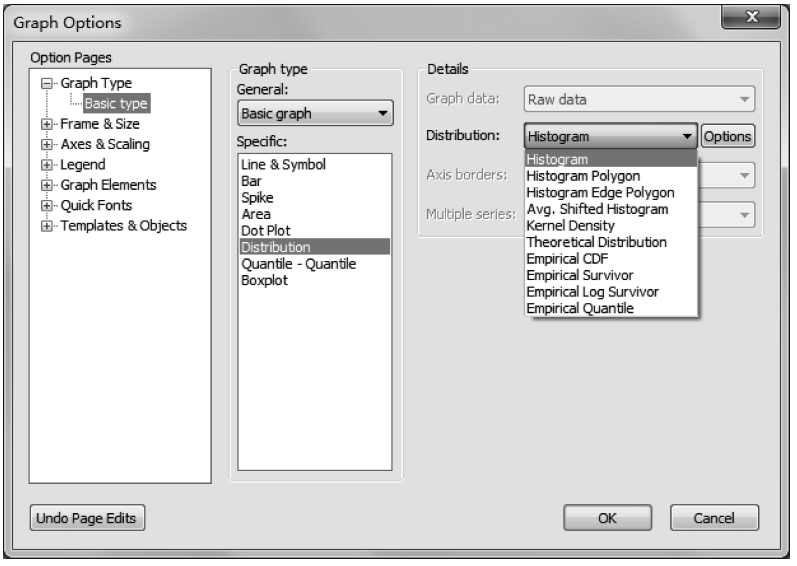


图 1.4.2　图形功能选项窗口

1. CDF-Surivor-Quantile 图

EViews 软件可以描绘出序列的带有加、减 2 倍标准差的经验累积分布函数(CDF)和分位数分布。

Options(选项)提供了几种计算经验 CDF 的方法,它们的不同之处在于对 CDF 进行非连续性的调整,随着样本数的增加,这些差别将很小。

通过观察分布函数或分位数分布的图形,与几种理论分布对比,可以大致判断应该服从哪种分布。EViews 软件还提供了基于分位数的对比判断序列分布的功能——QQ 图。

2. Quantile-Quantile 图

Quantile-Quantile(QQ 图)是一种简单但重要的比较两个分布的工具。这个图绘出序列分位数分布相对于某一种理论分布的异同。如果这两个分布是相同的,QQ 图将在一条直线上;如果 QQ 图不在一条直线上,说明这两个分布是不同的。选择 View/

Graphs 命令，在图 1.4.2 选择 Quantile-Quantile，将出现 Q-Q graph 对话框，单击右边的 Options 按钮可以选择理论分布：Normal（正态）分布、Uniform（均匀）分布、Exponential（指数）分布、Logistic（逻辑）分布和 Extreme Value（极值）分布等。

1.4.4　假设检验

1. 均值、方差和中位数的假设检验

选择 View/Descriptive Statistics &Test /Simple Hypothesis Tests，可以进行均值、方差、中位数检验。在对话框 Mean 后面的文本框中输入 μ 值，如例 1.7，输入 10。如果已知总体标准差，在对话框右边的框内输入标准差值，可以计算 Z 统计量，否则，用样本标准差代替总体标准差，运用 t 检验。在双边假设下，如果 p 值小于检验的显著水平（如 0.05），则拒绝原假设，否则无法拒绝。

2. 相等性检验

相等性检验（tests of equality）的原假设是组内所有的序列具有相同的均值、中位数或方差。例 1.5 中的方差相等性检验可以通过这个功能实现。选择 View/Tests of Equality 命令后，在出现的对话框中选择 variance，出现方差相等性检验结果，列出了各种不同的检验方法，第 2 列为自由度，第 3 列是统计量值，第 4 列为 p 值。

与例 1.5 对比可知，EViews 软件里方差相等的 F 检验是双边检验。除了 F 检验外，EViews 软件还提供了其他几种检验的结果。Siegel-tukey 统计量近似服从正态分布（Sheskin，1997）；Bartlett 检验近似服从自由度是 $G-1$ 的 χ^2 分布（Sokal&Rohlf，1995；Judge et al.，1985）；Levene 检验以方差分析为基础，近似服从分子自由度为 $G-1$、分母自由度为 $2N-G$ 的 F 分布（Levene，1960）；Brown-Forsythe 检验是修正的 Levene 检验（Conover et al.，1981；Brown & Forsythe，1974a，1974b；Neter et al.，1996）。

1.5　习　　题