# 计量经济学

JILIANG JINGJIXUE

云俊 编著

人民交通出版社

# 计量经济学

Jiliang Jingjixue

云　　俊　编著



人民交通出版社

## 内 容 简 介

本书系统地阐述了计量经济学的基本理论与方法。全书共分六章,第一章介绍了计量经济学的基本概念及所需的数理统计基础知识;第二章至第五章介绍了单方程和联立方程计量经济学模型及模型的估计和检验;第六章介绍了计量经济学的应用。本书重点强调了学科的基本体系、基本理论的解释及基本方法的应用,注重理论与实际相结合,提供了大量的实例,既可作为经济与管理类研究生和本科生的教材,也可作为从事经济与管理研究的科技工作者的参考书。

图书在版编目(CIP)数据

计量经济学/云俊编著. —北京:人民交通出版社,2000.8

ISBN 7-114-03691-4

Ⅰ.计... Ⅱ.云... Ⅲ.计量经济学 Ⅳ.F224.0

中国版本图书馆 CIP 数据核字(2000)第 62972 号

计量经济学

云 俊 编著

责任印制:杨柏力 正文设计:王秋红 责任校对:尹 静

人民交通出版社出版

(100013 北京和平里东街 10 号)

新华书店北京发行所发行

各地新华书店经销

北京凯通印刷厂印刷

开本:$787 \times 1092 \frac{1}{16}$ 印张:13 字数:320 千

2000 年 9 月 第 1 版

2000 年 9 月 第 1 版 第 1 次印刷

印数:0001—1500 册 定价:20.00 元

ISBN 7-114-03691-4

F·00356

# 前　　言

计量经济学是在国民经济发展中应用最为广泛的当代新学科之一。从国民经济宏观分析到工矿企业的微观控制，处处可以看到其研究与应用的成果，目前已经成为各大学经济与管理专业研究生和本科生学习的重要课程。对于从事经济与管理的人员，掌握计量经济学的思想与方法，将其应用于解决生产实际问题，具有十分重要的意义。

本书是作者在从事多年研究生的《计量经济学》课程教学的基础上写成的。书中介绍了计量经济学的特点、研究内容及应用步骤。详细、系统地介绍了单方程计量经济学模型和联立方程计量经济学模型的建立过程、参数的估计和模型的检验。其中每种方法都给出了严格的理论证明和推导，有较强的理论、严密而系统的阐述。对证明过程中用到的较深的数学知识进行了解释，对一些较难理解的部分，用适当的例子进行了说明和描述。使读者易于理解，这正是本书的特点之一。本书还用了一定的篇幅介绍了计量经济学方法在经济领域中的应用，能帮助读者将所学的计量经济学方法应用于解决实际问题。

本书在编著过程中，广泛征求计量经济工作者以及所教过的研究生的意见与看法，并参考了若干计量经济学书籍和研究论文，有些内容为本书所引用，在此向有关作者及给予巨大支持与帮助的人员表示衷心感谢。初稿完成后，陈庆虎博士进行了审阅，并提出了许多宝贵意见；研究生马开平、王威等承担了该书的资料整理及校对工作，在此特别致谢。

希望通过本书能引起更多同行的共鸣，大家一起共同探讨与研究《计量经济学》方面的问题，不断推陈出新，使《计量经济学》的应用更加广泛，使理论研究更加深入。书中不当之处恳请广大读者批评指正。

作　　者

2000 年 5 月

# 第一章 绪 论

计量经济学成为一门独立的学科,至今约60余年,它产生于资本主义社会,起初并不被人们所重视,随着一些发达国家计量经济学的成功应用以及计算机的广泛普及使用,大量复杂的计量经济模型得以建立和应用,使这门学科得到了迅速的发展,正如美国著名经济学家萨缪尔森曾说过的:"二次世界大战后的经济学是计量经济学时代。"

## 第一节 计量经济学的基本概念

### 一、什么是计量经济学

英文"Econometrics"最早是由挪威经济学家费里希(R. Frish)在1926年仿照生物计量经济学(Biometrics)一词提出的。其中文译名有两种:计量经济学与经济计量学。前者试图强调它是一门经济学科;而后者是英文的直译,且强调该学科的主要内容是经济计量方法。这两种不同的强调正反应了计量经济学有两个主要的研究内容:一是计量经济学方法的理论研究,称为理论计量经济学;二是将这些理论广泛应用于实际的经济活动中,称为应用计量经济学。本书介于二者之间,偏重于应用计量经济学。大多学者认为:计量经济学是用经济理论、数学与统计推断作为工具来分析经济现象的一门学科。

### 二、计量经济学与其它学科的关系及区别

1. 数理经济学与计量经济学的区别

数理经济学与经济理论之间没有本质的区别。简单地说,数理经济学是将经济理论数学公式化;而计量经济学是研究经济现象各因素之间的数量描述,即将经济行为理论的定量化。二者的关系可以通过下述例子来说明。

例如:由经济理论知某一特定商品的需要量 $Q$ 主要与它的价格 $P$,其它商品的价格 $P_0$,消费者的收入 $Y$ 以及消费者的爱好 $X$ 有关,且假设是线性关系。将这一理论假设公式化(即用数理经济学描述),得到:

$$Q = b_0 + b_1P + b_2P_0 + b_3Y + b_4X$$

这样将需要量 $Q$ 与该商品的价格 $P$,其它商品的价格 $P_0$,消费者收入 $Y$ 及爱好 $X$ 之间的关系直观地用数学形式加以描述,称为数理经济学方程。它揭示了变量之间的数学关系,但未揭示其定量关系。

然而,在现实的经济活动中,除了上述四个因素对该商品的需要量 $Q$ 有影响外,可能存在更多的因素影响需要量。比如新产品的发明,战争、经济政策的改变,人口大量增加或迁移等等。而且人们的行为往往是不稳定的,会受到宣传、梦想、偏见、传统的和其它因素的影响,即使个人的收入和市场条件保持不变,人们的经济行为也会由于这些随机因素而有所不同。在计量经济学中引入了随机项,即将这些随机因素用一个随机项来表示加入到模型中。即:

$$Q = b_0 + b_1P + b_2P_0 + b_3Y + b_4X + u$$

式中:$u$——影响需要量的随机因素(从这点上说,计量经济学是较精确的数理经济学)。

通过对样本数据进行适当的数学处理,进而得到诸如

$$\hat{Q} = 0.5 + 0.7P + 0.88P_0 + 1.2Y + 0.8X$$

之类的定量方程称为计量经济学方程。

式中: $\hat{Q}$——$Q$ 的估计值。

2. 经济统计学与计量经济学的区别

经济统计学实际上是数理统计学在经济学领域的应用,它是针对于经济问题中已发生的实际现象,进行数据的整理分析(列表或图示)来发现问题。

而计量经济学同样也是以数理统计学作为基本工具,它侧重于对所有的经济规律首先用数理统计模型进行描述,然后再借助经济模型来对现实现象进行分析、推理,研究其变化趋势。同时将随机因素考虑到所要建立的模型中并对变数的系数作参数估计。

## 第二节　计量经济学的应用步骤

计量经济学研究经济问题,可分为六个阶段:理论分析、样本数据的收集、理论模型的设定、参数估计、模型检验及模型应用。

### 一、理论分析

依据经济理论或经验,对所研究的经济问题,按照客观事物的本来面目分析影响它的因素,揭示各因素之间的内在联系和这些因素对问题的影响程度。

理论分析是计量经济工作的重要阶段,它不仅要找出影响所研究问题的各种因素,而且要说明哪些是主要因素、次要因素、直接因素、间接因素,即确定将要建立的模型中的所有变量:解释变量和被解释变量。并要分析随着时间的流逝,因素会发生什么变化等等。

### 二、样本数据的收集

确定了影响所研究问题包含的各主要因素后,然后根据其因素的含义,收集并整理各因素的样本数据。这是一个非常繁杂的工作,然而样本数据的质量直接影响将要建立的模型的质量。

1. 几类常用的样本数据

1)时间序列数据:是一批按时间先后排列的统计数据,一般由统计部门提供。在利用时间序列数据作样本时,有两点要特别注意。一是数据的口径问题,如果出现在不同的样本点上口径不一致的情况,要认真调整。二是用时间序列数据作样本容易产生模型中随机误差项的序列相关。

2)截面数据:这是一批发生在同一时间截面上的调查数据。例如:某年某市各个企业的总产值,或者某月某市各个家庭的收入。利用截面数据作样本时,容易产生异方差性。

3)虚变量数据:也叫二进制变量数据,一般情况下取值为 0 和 1。例如:在农业生产函数研究中,若设置虚变量表示气候环境对农业生产的影响,那么相对于灾年,该变量取值为 1,相对于正常年份该变量取值为 0(近年来,在某些研究工作中,虚变量选用了 *0* 和 *1* 以外的数字以表示它所反映的非经济因素的不同程度,取得了较好的效果)。

2. 如何选择样本数据

在选择样本数据时，除了要考虑数据的可得性以外，还必须考虑数据的可用性。研究问题的目的不同，对样本数据的要求也不相同。如果研究问题的目的是为了预测，对参数估计值的最小方差性要求较高；如果研究的目的是进行结构分析或政策评价，那么参数估计值的无偏性更为重要。在存在多组样本数据可供选择时，应比较参数估计值的统计性质以选择较好的样本数据。

3. 样本数据的质量

样本数据的质量问题可概括为：完整性、准确性、可比性和一致性。

(1)完整性：是系统对数据的要求。人们所研究的经济现象，无论是国家经济、地区经济还是部门经济，都是一个系统。样本数据不完整，无法正确地建立模型。百分之百的完整是难以达到的，但对于少数“遗失数据”，必须采用科学的方法人为地补充以达到完整。

(2)准确性：要求统计数据或调查数据本身是准确的，而这往往容易为人们忽视；要求研究人员准确地选择、应用数据。例如：生产函数是某一生产过程中投入要素与产出量之间关系的定量描述，那么作为投入要素之一的资本的数据必须是真正反映投入到生产过程的资本的数量，如果用固定资产原值作为投入的固定资本要素数据，则不能算是准确的，因为原值中可能有一部分是闲置的，没有真正投入，尽管它本身的统计可能是准确的。

(3)可比性：往往人们可以得到的经济统计数据，却具有比较差的可比性，其原因在于统计范围口径的变化和价格口径的变化，必须进行处理后才能为研究所用。计量经济方法是从历史的数据或同一时间截面的不同点的数据中寻找其规律性，如果数据是不可比的，当然找出的规律并不反映经济活动本身固有的规律。比如在研究生产函数时，产值应采用统一的不变价，固定资产原值应经过价格调整等等。

(4)一致性：选择的样本数据必须与母体一致。比如：企业截面数据只能用于企业生产函数模型的估计，不能作为行业模型的样本数据。因为企业截面数据与行业截面数据不属于同一母体。

## 三、理论模型的设定

通过理论分析知道影响所研究问题的各因素，确定了解释变量、被解释变量。然后根据所收集的数据，利用样本数据作出解释变量与被解释变量之间关系的散布图，由散布图显示的变量之间的函数关系作为理论模型的数字形式。

一般情况下，很难事先确定模型的数学形式，那么就采用不同的形式进行试模拟，然后选择模拟结果较好的一种。

当理论模型的数学形式确定以后，再对模型中的参数用经济理论及根据对所研究经济活动的认识事先给出待估参数的符号及大小范围。并可用以检验模型的估计结果。

## 四、参数估计

当理论模型的数学形式确定以后，就要利用选择的数据对模型中的参数进行估计，参数是模型中表示变量之间关系的常系数。它将各种变量连接在模型中，具体说明解释变量对因变量的影响程度。模型参数的估计是一个纯技术过程，包括对模型识别问题的研究(对联立方程计量经济学模型)、解释变量之间相关程度的研究、适当的计量经济学估计方法的选择、应用软件的使用等方面。

## 五、模型检验

模型参数估计值得到后，一个计量经济学模型可以说初步建立了，但模型是否符合实际，能否解释实际经济过程，还需要进行检验。一般来讲，计量经济模型的检验分四个部分，即经济意义的检验、统计性检验、计量经济学检验和模型的预测检验。

1．经济意义的检验

主要是检验各个参数值是否与经济理论和实际经验相符。

2．统计性检验

是利用统计推断原理，检验参数估计值的可靠程度、观察数据的拟合程度。最常用的有拟合优度检验、变量和方程的显著性检验和估计值的标准差检验。

3．计量经济学检验

检验各种计量经济假设的合理性，主要包括随机扰动项的序列相关检验、异方差性检验、解释变量的多重共线性检验等等。

4．模型的预测检验

预测检验主要检验估计值的稳定性以及相对样本容量变化时的灵敏度，确定所建立的模型是否可以用于样本观测值以外的范围，即模型的所谓超样本特性。具体做法是：

(1)利用扩大了的样本重新估计模型的参数，将估计值与原来的估计值进行比较，并检验其间差异的显著性。

(2)将所建立的模型用于样本以外某一时期的实际预测，并将这个预测值和实际的观测值进行比较，然后检验其间差异的显著性。

## 六、模型应用

计量经济模型主要应用于分析经济结构、评价政策决策、仿真经济系统以及预测经济发展等等。模型的应用过程，也是检验模型和理论的过程。如果预测误差小，表明模型精度高，质量好，对现实解释能力强，理论符合实际；反之，就要对模型以及对建模所依据的经济理论进行修正。

计量经济工作的过程是一个不断修改、信息反馈的过程，其程序大致如图 1-2-1 所示：

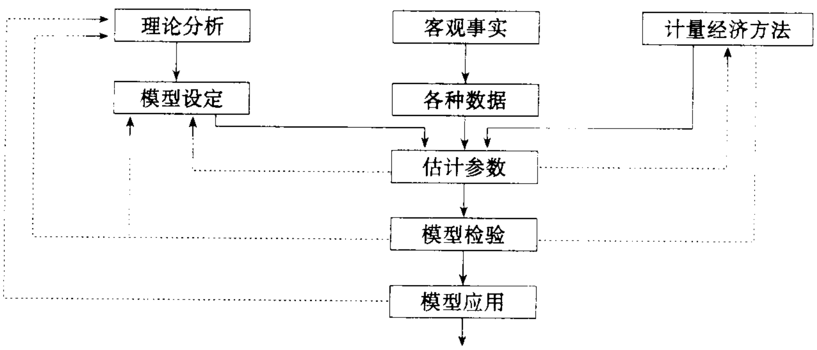


图 1-2-1

模型的建立和实际应用，离开了计算机几乎是不可能的，目前已有很多专门的计量经济学软件包可供使用。

# 第三节　数理统计基础

在计量经济学模型的研究中,研究的是具有随机性特征的经济变量关系,因此学习和运用计量经济学方法必须掌握一定的数理统计知识。本节将对数理统计中常用的一些基本概念、常用的随机变量分布及其分布参数作一简单回顾和介绍。

## 一、随机事件的概念

每次试验中,可能发生也可能不发生,而在大量试验中具有某种规律性的事件称为随机事件,简称事件。通常用大写拉丁字母 $A$、$B$、$C$ 等表示。

1. 基本事件

不能分解为其它事件组合的最简单的随机事件称为基本事件。例如:掷一颗骰子的试验中,其出现的点数,“1 点”、“2 点”、……、“6 点”,都是基本事件。

2. 复合事件

由几个基本事件组合的事件称为复合事件。例如,出现点数为“奇数点”,奇数点是由“1 点”、“3 点”、“5 点”这三个基本事件组成的,只要这三个基本事件中的一个发生,“奇数点”这个事件就发生,所以出现点数为“奇数点”这个事件就是复合事件。

3. 必然事件

每次试验中必然发生的事件称为必然事件,用符号 $\Omega$ 表示。例如,在上面的掷骰子试验中,“点数小于 7”是必然事件。

4. 不可能事件

每次试验中一定不发生的事件称为不可能事件,用符号 $\Phi$ 表示。例如,“点数大于 6”是不可能事件。

5. 样本空间

所有可能出现的基本事件的全体,称为样本空间。样本空间作为一个事件是必然事件,所以记为 $\Omega$。

## 二、事件间的关系

1. 事件的包含

如果事件 $A$ 发生必然导致事件 $B$ 发生,即属于 $A$ 的某一个基本事件也都属于 $B$,则称事件 $B$ 包含事件 $A$,或称事件 $A$ 含于事件 $B$。记作:$B \supset A$ 或 $A \subset B$。

2. 事件的相等

如果事件 $A$ 包含事件 $B$, 事件 $B$ 也包含事件 $A$,称事件 $A$ 与 $B$ 相等。记作:$A = B$。

3. 事件的并(和)

两个事件 $A$、$B$ 中至少有一个发生,即“$A$ 或 $B$”是一个事件,称为事件 $A$ 与 $B$ 的并(和)。记作:$A + B$ 或 $A \cup B$。

4. 事件的交(积)

两个事件 $A$ 与 $B$ 同时发生,即“$A$ 且 $B$”是一个事件,称为事件 $A$ 与 $B$ 之交。它是既属于 $A$ 又属于 $B$ 的所有公共基本事件构成的集合。记作:$AB$ 或 $A \cap B$

5. 事件的差

事件 $A$ 发生而事件 $B$ 不发生，是一个事件，称为事件 $A$ 与 $B$ 的差。它是由属于 $A$ 但不属于 $B$ 的那些基本事件组成的集合。记作：$A-B$。

6. 互不相容事件

如果事件 $A$ 与 $B$ 不能同时发生，即 $AB=\Phi$，称事件 $A$ 与 $B$ 互不相容（或称互斥）。互不相容的事件 $A$ 与 $B$ 没有公共的基本事件。显然基本事件间是互不相容的。

7. 对立事件

事件"非 $A$"称为事件 $A$ 的对立事件（或逆事件），它是由样本空间中不属于 $A$ 的基本事件组成的集合。记作：$\bar{A}$。显然

$$A\bar{A}=\Phi;A+\bar{A}=\Omega;\bar{\bar{A}}=A;\bar{A}=\Omega-A$$

8. 完备事件组

若事件 $A_1,A_2,\cdots,A_n$ 为两两互不相容的事件，且 $A_1+A_2+\cdots+A_n=\Omega$，称 $A_1,A_2,\cdots,A_n$ 构成一个完备事件组。

## 三、概率

除了必然事件和不可能事件这两种极端情况外，各种随机事件发生可能性的大小往往是不同的。概率，便是对随机事件发生可能性大小的测定。随机事件 $A$ 发生可能性的大小，用一个确定的数值表示，这个数值就称为事件 $A$ 发生的概率，记为 $P(A)$。正确的理解和计算随机事件的概率是进行统计推断和统计决策的基础。按不同的观点和不同的情况，概率可以分为古典概率、试验概率和主观概率。

1. 古典概率

样本空间 $\Omega$ 中的基本事件个数是有限个，每个基本事件出现的可能性相同，设事件 $A$ 是该样本空间 $\Omega$ 中的一个随机事件，事件 $A$ 的古典概率定义为：

$$P(A)=\frac{A\text{ 所含基本事件数}}{\Omega\text{ 所含基本事件数}}$$

**例 1-3-1**　一批产品共有 100 件，其中有 4 件不合格产品，试求随机抽取 1 件为不合格产品的概率。

**解**：不妨将 4 件不合格产品依次编号为 1，2，3，4，把其余 96 件产品依次编号为 5，6，…，100。取样本空间为 $\Omega=\{1,2,\cdots,100\}$。由于 $\Omega$ 中的基本事件个数是有限个，并且每一个产品抽到的可能性相同，因此是可以用古典概率公式计算的。

令 $A=\{$抽到一件不合格产品$\}$

于是

$$P(A)=\frac{A\text{ 所含基本事件数}}{\Omega\text{ 所含基本事件数}}=\frac{4}{100}=0.04$$

2. 试验概率

古典概率在应用上受到两个条件的限制：一是随机试验的结果只有有限个，二是这些结果出现的可能性相同。这些条件常常不能全部满足。例如，保险公司要了解人口死亡率，这就不能用古典概率来计算。如果用试验概率就可不受上述条件的限制。

设在相同条件下重复 $n$ 次试验，事件 $A$ 发生的次数为 $m$，则事件 $A$ 发生的频率为 $m/n$。虽然重复 $n$ 次试验得到的结果会有波动，但随着试验次数 $n$ 的无限增大，频率的波动会越来越小，并逐渐地稳定于某个常数 $P$。这个数 $P$ 就定义为事件 $A$ 发生的概率 $P(A)$。这样定义

的概率是依据统计试验的大量数据而得到的，因此称为试验概率，亦称统计概率。

试验概率虽然用频率来解释，但概率和频率毕竟是有区别的。频率是个试验值，它只能近似地反映事件发生可能性的大小；概率是个理论值，它精确地表示事件发生可能性的大小，它的值是唯一确定的。试验概率比较容易理解，但也存在一些问题。重复 $n+1$ 次试验的频率不一定比重复 $n$ 次试验的频率更接近真实的概率，而且试验也不可能无限制的进行下去。

3. 主观概率

在实际问题中，有些试验是无法在相同条件下重复进行的。例如，股价指数在未来一周内上升的可能性有多大等等。这样的事件无法进行统计试验，也就不可能计算事件发生的频率，只能凭经验进行主观的估计。这种主观估计的概率，称为主观概率。主观概率依人的看法和经验而定，对同一事件不同的人会给出不同的概率。

## 四、概率的基本性质及运算

1. 概率的基本性质

(1) $0 \leqslant P(A) \leqslant 1$

(2) $P(\Omega)=1, P(\Phi)=0$

(3) 设 $A$ 与 $B$ 是任意两个事件，则 $P(A \cup B)=P(A)+P(B)-P(AB)$

(4) 设 $A$ 与 $B$ 是两个互不相容事件，则 $P(A \cup B)=P(A)+P(B)$

可推广为：

设事件 $A_1, A_2, \cdots, A_n$ 两两互不相容，即 $A_iA_j=\Phi(1 \leqslant i \neq j \leqslant n)$，则

$$P(\sum_{i=1}^{n} A_i)=\sum_{i=1}^{n} P(A_i)$$

(5) $P(\bar{A})=1-P(A)$

(6) 设事件 $A$ 包含事件 $B$，即 $A \supset B$，则 $P(A-B)=P(A)-P(B)$；$P(A) \geqslant P(B)$

2. 条件概率

事件 $B$ 发生的条件下，事件 $A$ 发生的概率，称为条件概率，记为 $P(A/B)$，

$$P(A/B)=\frac{P(AB)}{P(B)} \text{ 或 } P(AB)=P(B)P(A/B)$$

其中：$P(B)>0$。

**例 1-3-2** 设一批产品共 $N$ 件，其中有 $M$ 件次品，不放回地抽取 2 件，问第 1 件抽到的是正品，第 2 件抽到的是次品的概率是多少？

**解**：令 $A=\{$第一件是正品$\}$

$B=\{$第二件是次品$\}$

$$P(AB)=P(A)P(B/A)=\frac{N-M}{N} \cdot \frac{M}{N-1}=\frac{(N-M)M}{N(N-1)}$$

3. 全概率公式

设 $B_1, B_2, \cdots, B_n$ 为 $n$ 个互不相容事件且 $\sum_{i=1}^{n} B_i=\Omega$，$P(B_i) \geqslant 0$，$i=1,2,\cdots,n$。则任一事件 $A$ 的概率为

$$P(A)=\sum_{i=1}^{n} P(B_i)P(A/B_i)$$

**例 1-3-3** 有 3 个工作人员被指定复制某种表格。第 1 个人复制了这种表格的 40%，第 2 个人复制了 35%，第 3 个人复制了 25%。第 1 个人、第 2 个人、第 3 个人的错误率分别为 0.04、

0.06、0.03。现随机抽取一份表格,问这份表格有错误的概率是多少?

**解**:令 $A$ = {抽取一份表格有错误}

$B_i$ = {抽取的表格是由第 $i$ 个工作人员复制的}, $i = 1,2,3$

显然 $B_1 \cup B_2 \cup B_3 = \Omega$ 且 $B_1, B_2, B_3$ 两两互不相容。所以由全概率公式得

$$\begin{aligned}P(A) &= P(B_1)P(A/B_1) + P(B_2)P(A/B_2) + P(B_3)P(A/B_3)\\ &= 0.4 \times 0.04 + 0.35 \times 0.06 + 0.25 \times 0.03\\ &= 0.0445\end{aligned}$$

故随机抽出的这份表格有错误的概率为0.0445。

4. 事件的独立性

对任意事件 $A$ 与 $B$,事件 $B$ 的发生往往对事件 $A$ 的概率产生影响,即 $P(A/B) \neq P(A)$。但如果事件 $B$ 的发生与否并不影响事件 $A$ 的概率,则有 $P(A/B) = P(A)$,这时

$$P(AB) = P(A)P(B)$$

可推广到一般。

设 $n$ 个事件 $A_1, A_2, \cdots, A_n$,如果其中任意一个事件的发生与否不受其余 $n-1$ 个事件发生与否的影响,那么就称 $n$ 个事件 $A_1, A_2, \cdots, A_n$ 相互独立。这时

$$P(A_1A_2\cdots A_n) = P(A_1)P(A_2)\cdots P(A_n)$$

应该指出,两个事件相互独立与互不相容是两个不同的概念。独立性是指两个事件的发生互不影响,互不相容是指两事件不能同时发生,它们之间并没有什么联系。

## 五、随机变量及其概率分布

1. 随机变量

随机变量是以一定的概率取不同的可能值的变量。

随机变量是随机事件的数量化。对于某一基本事件的集合中的每一个基本事件,指定某个实数与之对应,就可把随机事件转化为随机变量。

例如:掷骰子的例子中,指定数字1,2,…,6分别与出现点数"1点"、"2点"、…、"6点"这些基本事件对应,就得到可取6个可能值的随机变量。

随机变量取某值的概率,等于所对应的随机事件出现的概率。

随机变量分为离散型随机变量和连续型随机变量。

如果随机变量 $X$ 只取有限个或可列个可能值,而且以确定的概率取这些不同的值,则称 $X$ 为离散型随机变量。

如果随机变量 X 的取值范围是某个区间,则称为连续型随机变量。

2. 随机变量的概率分布

随机变量在各个可能值上出现概率的大小情况,称为随机变量的概率分布,简称分布。

(1) 离散型随机变量的概率分布——分布列

设离散型随机变量 $X$ 的可能取值为 $x_1, x_2, \cdots, x_n$,取这些值的概率分别为 $p_1, p_2, \cdots, p_n$,则

$$P(X = x_k) = p_k, k = 1,2,\cdots$$

称为离散型随机变量 $X$ 的概率分布或分布列。有时也用如表1-3-1所示的表格来表示 $X$ 的分布列:

表 1-3-1

| | | | | | |
|---|---|---|---|---|---|
| $x$ | $x_1$ | $x_2$ | … | $x_n$ | … |
| $p$ | $p_1$ | $p_2$ | … | $p_n$ | … |

由概率的性质知

$$p_k \geqslant 0, k = 1,2,\cdots$$

$$\sum_{k=1}^{\infty} p_k = 1$$

因此,分布列表明全部概率 1 在各个可能值之间分配的规律。

(2) 连续型随机变量的概率分布——概率密度函数

对于连续型随机变量的概率不能象离散型随机变量的概率那样表示为取各个可能值的概率。因为连续型随机变量的可能值有无限多个,每一个具体数值出现的概率为 0,所以连续型随机变量的概率分布用概率密度来表示,即反映在任意一个子区间上的密集程度。

连续型随机变量 $X$ 在 $x$ 处的概率密度为

$$f(x) = \lim_{\Delta x \to 0} \frac{P(x \leqslant X \leqslant x + \Delta x)}{\Delta x}$$

概率密度作为 $x$ 的函数,叫做密度函数。密度函数的图形是一条光滑的曲线,并满足

$$f(x) \geqslant 0$$

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1$$

即密度函数在横坐标的上方,它与横轴之间的面积等于 1,如图 1-3-1 所示。

有了密度函数就可以计算出连续型随机变量取任意区间$(a,b)$中的值的概率

$$P(a \leqslant X \leqslant b) = \int_a^b f(x)\,dx$$

相当于图中阴影部分的面积。

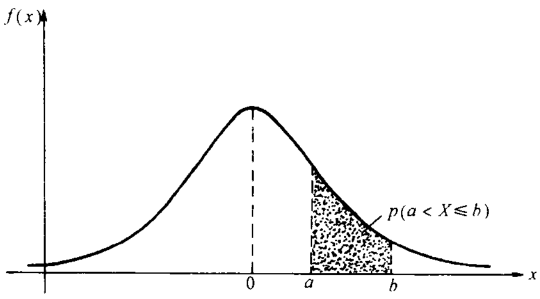

图 1-3-1 连续随机变量的密度函数

(3) 分布函数

随机变量的概率分部还可以用分布函数来表示。

随机变量 $X$ 取小于某值 $x$ 的概率,称为 $X = x$ 的分布函数,记为 $F(x)$,即

$$F(x) = P(X < x)$$

离散型随机变量 $X$ 的分布函数为

$$F(x) = \sum_{x_i < x} P(x_i)$$

连续型随机变量 $X$ 的分布函数为

$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

式中: $f(x)$——密度函数。

显然,分布函数是概率的累积。

3. 分布参数

反映随机变量概率分布特征的数字,称为分布参数。最重要的分布参数是数学期望、方差以及协方差。

(1) 数学期望(又称平均值或均值)

数学期望是随机变量各可能取值的概率为权数的加权算术平均数。通常用 $E$ 或 $M$ 表示。数学期望是反映随机变量 $X$ 的可能取值的平均水平或集中趋势的指标。是刻划随机变量性质的一个最重要的数字特征。

对于离散型随机变量 $X$ 的数学期望值为

$$E(X) = \sum_i x_i P(x_i)$$

**例 1-3-4** 某随机变量的可能取值为{1,2,3,4,5},取每个值的概率分别为:$P(x=1)=0.25$;$P(x=2)=0.375$;$P(x=3)=0.125$;$P(x=4)=0.125$;$P(x=5)=0.125$

该随机变量的数学期望值为

$$\begin{aligned} E(X) &= 1\times0.25+2\times0.375+3\times0.125+4\times0.125+5\times0.125 \\ &= 2.5 \end{aligned}$$

对于连续型随机变量 $X$ 的数学期望为

$$E(X) = \int_{-\infty}^{+\infty} xf(x)\,dx$$

式中:$f(x)$——$x$ 的密度函数。

数学期望具有下述性质:

①设 $c$ 是常数,则

$$E(c)=c$$

②设 $k$ 是常数,$X$ 是随机变量,则

$$E(kX) = kE(X)$$

③设 $X_1,X_2,\cdots,X_n$ 为 $n$ 个随机变量,则

$$E(X_1+X_2+\cdots+X_n) = E(X_1)+E(X_2)+\cdots+E(X_n)$$

(2) 方差

随机变量的方差是随机变量各个可能值对数学期望的离差平方以概率加权的算术平均数。记为 $D(X)$(或 $\mathrm{var}(X)$或 $\sigma^2$)。即

$$D(X) = \sigma^2 = E[X-E(X)]^2$$

方差是反映随机变量的分散程度或变异程度的数字,是反映随机变量与其期望值的离散程度的数字。

离散型随机变量的方差

$$D(X) = E[X-E(X)]^2 = \sum_i^n [x_i - E(X)]^2 P(x_i)$$

连续型随机变量的方差

$$D(X) = \int_{-\infty}^{+\infty} [X-E(X)]^2 f(x)\,dx$$

方差具有下述性质:

①设 $c$ 是常数,则

$$D(c)=0$$

②设 $k$ 是常数,$X$ 是随机变量,则

$$D(kX) = k^2 D(X)$$

③设 $X_1,X_2,\cdots,X_n$ 为 $n$ 个相互独立的随机变量(即这 $n$ 个随机变量中任一个随机变量的取值皆不受其余 $n-1$ 个随机变量取值的影响),则

$$D(X_1+X_2+\cdots+X_n)=D(X_1)+D(X_2)+\cdots+D(X_n)$$

(3) 标准差

方差的平方根称为标准差,记为 $\sigma$。

(4) 协方差

协方差是两个随机变量与各自数学期望离差之积的期望值,即 $E[X-E(X)][Y-E(Y)]$ 称为随机变量 $X$ 与 $Y$ 的协方差,记为 $\text{cov}(X,Y)$或 $\sigma_{xy}$。协方差可用于度量两个随机变量之间相关关系的密切程度。

离散型随机变量的协方差

$$E[X-E(X)][Y-E(Y)]=\sum_{i=1}^{n}[x_i-E(X)][y_i-E(Y)]P(x_iy_i)$$

连续型随机变量的协方差

$$E[X-E(X)][Y-E(Y)]=\iint_{-\infty}^{+\infty}[X-E(X)][Y-E(Y)]f(x,y)\,dxdy$$

(5) 协方差与相关系数的意义

协方差:$E[X-E(X)][Y-E(Y)]$,称为 $Y$ 与 $X$ 的协方差,记作 $\text{cov}(X,Y)$或 $\sigma_{xy}$。

相关系数:$\rho=\dfrac{E[X-E(X)][Y-E(Y)]}{\sqrt{E[X-E(X)]^2E[Y-E(Y)]^2}}$,是反映 $X$ 与 $Y$ 之间线性关系的密切程度的指标。

**例 1-3-5** $n$ 个家庭的收入和支出情况如图 1-3-2 所示:

横坐标 $x_i$ 代表收入,纵坐标 $y_i$ 代表支出,图中纵横交叉的两条直线分别代表这些家庭的平均收入 $E(x)=\bar{x}=\dfrac{1}{n}\sum_{i=1}^{n}x_i$ 和平均支出 $E(y)=\bar{y}=\dfrac{1}{n}\sum_{i=1}^{n}y_i$。这两条线将该图划分为 4 个区域。

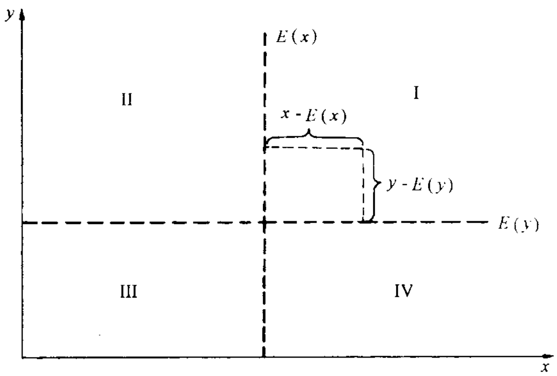


图 1-3-2

考察乘积$(x_i-\bar{x})(y_i-\bar{y})$在这四个区域的正负号,显然在第 I、III 区域为正值(即$(x_i-\bar{x})$与$(y_i-\bar{y})$为同号);在第 II、IV 区域为负值(即$(x_i-\bar{x})$与$(y_i-\bar{y})$为异号)。

当 $x$ 与 $y$ 具有正向关系时,大多数点必落在第 I、III 区域,从而 $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$为正。

当 $x$ 与 $y$ 具有反向关系时,大多数点必落在第 II、IV 区域,从而 $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$为负。

由此可见,$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$的值反映着 $x$ 的变动与 $y$ 的变动之间的关系。而 $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$的值的大小受观察点数的影响,和受 $x$ 与 $y$ 取什么测量单位的影响,为了克服这两个缺点,可对它进行改进。

①消除观察点数的影响

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\qquad(=E(x_i-\bar{x})(y_i-\bar{y}))$$

即是 $n$ 个家庭收入 $x$ 与支出 $y$ 的协方差。

②消除计量单位不同的影响

用 $x$ 和 $y$ 的标准差 $\sigma_x=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$；$\sigma_y=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}$ 作为计量单位，即

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\rho$$

即为这 $n$ 个家庭收入 $x$ 与支出 $y$ 的相关系数($-1\leqslant\rho\leqslant1$)。

当 $\rho>0$ 时，称 $x$ 与 $y$ 正相关；$\rho<0$ 时，称 $x$ 与 $y$ 负相关；$\rho=0$ 时，称 $x$ 与 $y$ 不相关。

注：$\rho$ 的绝对值越大，$x$ 与 $y$ 的关系越密切。

4．总体和样本

把所研究的对象的全部单位所组成的集合叫做总体，从总体中抽出部分单位所组成的集合，叫做样本。总体所含单位个数称为总体单位数。样本所含单位的个数称为样本容量。

5．样本统计量

如果所研究的总体充分大，总体分布参数 $E(x)$，$D(x)$ 以及 $\mathrm{cov}(x,y)$ 未知，抽样是获得总体信息的唯一有效方法。我们常常采用由样本数据推断估计总体参数，并且进行假设检验，对总体参数真值作出评价。抽样推断的工具就是样本统计量。常用的样本统计量为：

(1) 样本均值

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

(2) 样本方差

$$S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

(3) 样本协方差

$$\mathrm{cov}(x,y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

式中：$n$——样本容量。

由于抽样的随机性，对于不同的样本，样本统计量的值不同，因此就所有可能的样本来说，样本统计量也是一个随机变量。

## 六、常用的随机变量的概率分布

### (一) 正态分布

在客观世界中，有许多随机变量(现象)的分布表现为中间大两头小的形状，例如人们的身高、体重、同一收入水平下的消费支出、观测误差等等，这些随机变量的分布形态都可以用正态分布来描述。

1．定义

当随机变量 $X$ 的密度函数形式为

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}\qquad(\sigma>0)$$

时，称 $X$ 的分布为正态分布，记为 $X\sim N(m,\sigma^2)$

2．形状

如图 1-3-3 所示。

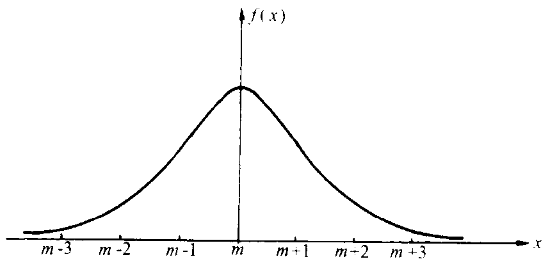


图 1-3-3

3. 特点

(1) 图形是左右对称的,对称轴为 $x = m$;

(2) 曲线有两个拐点,其位置在 $m \pm \sigma$ 处;

(3) $\sigma$ 的大小决定曲线低矮还是瘦高;

(4) 参数 $m$ 是随机变量 $X$ 的期望值,因为

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_{-\infty}^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} dx = m$$

(5) 参数 $\sigma^2$ 是随机变量 $X$ 的方差,因为

$$D(X) = \int_{-\infty}^{+\infty} (x - m)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \sigma^2$$

4. 标准正态分布

当正态分布的随机变量 $X$ 的数学期望为 0,方差为 1 时,称 $X$ 服从标准正态分布,记为 $X \sim N(0,1)$。

可将任何一个非标准的正态分布 $X \sim N(m,\sigma^2)$ 变换成标准的正态分布,这时只需做变换 $u = \frac{x-m}{\sigma}$,此时随机变量 $u$ 服从标准正态分布,即 $u \sim N(0,1)$,这时 $u$ 分布的密度函数为

$$\psi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \qquad (\mu \in R)$$

这时的对称轴为 $x = 0$。

标准正态分布有现成的表可查。

正态分布在数理统计中占有非常重要的地位,很多其它类型的概率分布的极限是正态分布。

(二) 抽样分布

在统计工作中,常常要根据研究对象的部分单位来判断总体情况。

例如:从一大批产品中抽取部分产品检查其质量,用这部分产品的质量来推断总体产品的质量;又如从城乡居民家庭中抽取若干家庭进行统计调查,用所得到的这部分家庭的生活情况来推断全体居民家庭的生活情况等等。这种推论是以抽样分布为依据的。

1. 随机抽样

从总体中抽取样本,如果总体的各个单位被抽中的机会是相等的,称这样的抽样为随机抽样。

在随机抽样过程中总体分布总是保持不变,叫做简单随机抽样。所以当总体单位数有限时,为了取得严格的简单随机抽样,必须是放回抽样(除非总体单位数很大,而样本容量(样本所含的单位数)又很小时)。

2. 统计量

根据样本资料计算的一些指标。如:

平均数:$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

样本方差:$S^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$

样本极差:$R = \max(x_i) - \min(x_i)$

等等。

由于抽样分布的随机性,只要样本容量小于总体单位数,可能的样本数有多个,这样的指标的数值也不同,所以样本指标也是一个随机变量,我们将这样的指标称为统计量。