张生智 编

◎ 高职高专教育使用教材

数理统计基础讲义

SHULI TONGJI JICHU JIANGYI

甘肃民族出版社
GANSU NATIONALITIES PUBLISHING HOUSE

◎ 高职高专教育使用教材

数理统计基础讲义

SHULI TONGJI JICHU JIANGYI

张生智 编

甘肃民族出版社
GANSU NATIONALITIES PUBLISHING HOUSE

图书在版编目（CIP）数据

数理统计基础讲义/张生智编. —兰州：甘肃民族出版社，2008.11
ISBN 978-7-5421-1427-3

Ⅰ.数… Ⅱ.张… Ⅲ.数理统计—教学参考资料 Ⅳ.0212

中国版本图书馆CIP数据核字（2008）第170338号

书　　名:数理统计基础讲义
作　　者:张生智　编
责任编辑:张兰萍
封面设计:王林强
出　　版:甘肃民族出版社(730030　兰州市南滨河东路520号)
发　　行:甘肃民族出版社发行部(730030　兰州市南滨河东路520号)
印　　刷:甘肃天河印刷有限责任公司
开　　本:787毫米×1092毫米　1/16　**印张**:12.5　**插页**:2
字　　数:304千
版　　次:2008年12月第1版　　2008年12月第1次印刷
印　　数:1～1 000
书　　号:ISBN 978-7-5421-1427-3
定　　价:32.50元

甘肃民族出版社图书若有破损、缺页或无文字现象，可直接与本社联系调换。
邮编:730030　地址:兰州市南滨河东路520号　网址:http://www.gansumz.com
电话:0931－8773420(藏文编辑部　联系人:交巴李加　E－mail:melce@sina.com)
电话:0931－8773261(汉文编辑部　联系人:李青立　E－mail:Lili295@sohu.com)
电话:0931－8773219(策划部　联系人:桂谕　E－mail:lanzhougy@163.com)
电话:0931－8773271(经营部　联系人:葛慧　E－mail:gsmzgehui3271@tom.com)

前　言

《数理统计》是大学数学教育中最重要的基础课之一，是数学系各专业的一门重要课程。它不仅是一种工具，而且是一种思维模式；不仅是一种知识，而且是一种素养。在培养高素质科学技术人才中具有其独特的、不可替代的重要作用。1998 年教育部颁发的本科学科门类分类中，已将“统计学”与数学、物理、理论力学等学科并列为理科分类中的一级学科。它是以概率论为基础，根据试验或观察得到的数据，来研究随机现象统计规律性的学科。随着概率论的发展，应用概率论的结果更深入地分析研究统计资料，通过对某些现象的频率的观察来发现该现象的内在规律性，并做出一定精确程度的判断和预测，将这些研究的某些结果加以归纳整理，逐步形成一定的数学模型，这些组成了数理统计的内容。

《数理统计》在自然科学、工程技术、管理科学及人文社会科学中得到越来越广泛和深刻的应用，其研究的内容也随着科学技术和经济与社会的不断发展而逐步扩大，但概括地说可以分为两大类：①试验的设计和研究，即研究如何更合理更有效地获得观察资料的方法；②统计推断，即研究如何利用一定的资料对所关心的问题做出尽可能精确可靠的结论。当然这两部分内容有着密切的联系，在实际应用中更应前后兼顾。但按民族师专数学教育专业的总体设计，《数理统计》课程只讨论统计推断，统计推断主要讨论总体的参数估计和区间估计以及正态总体下参数的假设检验，考虑到民族师范院校学生毕业后绝大多数人所从事的教师教育工作，因此，在本书中还增加了描述性统计的内容。本课程的目的是让学生了解统计推断检验等方法并能够应用这些方法对研究对象的客观规律性作出种种合理的估计和判断。掌握数据处理的一般性方法；掌握总体参数的点估计和区间估计；掌握假设检验的基本方法与技巧，并能运用其方法和技巧进行统计推断。理解方差分析及一元线性回归分析的基本原理。

在本书的编写中还充分考虑到了目前国内中小学课程体系和课程内容改革对概率论与数理统计的要求，以及民族师范院校学生的实际情况，并结合了编者在民族院校多年的教学经验和研究成果。为了提高学生的对本课程的学习兴趣和针对性，在每一章开头有本章主要内容、本章重点与难点提示、教学要求等，内容编写中还穿插了部分统计学的发展小常识，在每一章中还配备了适量的典型例题，在重点部分还配置了课堂练习。每章最后附有本章重点与难点分析、本章小结、习题、自测题(除第五章外)。本书还取消了传统的参考答案，主要思路是借此培养学生自己动手的主动性和能力。

本书可供民族师专、普通师专、函授院校作为教材使用，也可供师范院校非数学专业作为教材使用，同时，也可以作为以上各类院校学生自学用书。本书同时也是数理统计精品课程建设的配套教材。

本书的编写得到了本系樊正恩、席进华、王大胄等老师的大力帮助，在此表示感谢。限于编者水平，本教材难免存在缺点和错误，敬请读者批评指正。

编者

2008 年 5 月于甘肃合作民族师专

目 录

第一章　数理统计的基本概念

在概率论的讨论中我们知道随机变量及其概率分布全面描述了随机现象的统计规律，在许多问题的讨论中通常总是已知随机变量所服从的概率分布，或者假设概率分布为已知的。概率论中一切的计算与推理就是在这个基础上进行。但在实际中，情况往往并非如此。一个随机现象所服从的分布是什么概型可能完全不知道，或者根据研究对象的特点知道其概型，但不知其分布函数中所含的参数。例如，一段时间内某地区气温服从什么分布是完全不知道的。再如某超市一段时间内售后商品的投诉次数服从什么分布是完全不知道的。又如，某件商品是合格品还是次品是服从两点分布的，但分布中的参数 p(合格率)是不知道的。因此，要对这些问题进行研究，就必须知道它们的概率分布或者概率分布中所含的参数。这就是数理统计所要解决的一个首要问题，而在解决这个问题时，就要在所研究的对象的全体中抽取一部分进行试验或观测以获得相关信息，从而对整体做出推断。而数理统计学从这个意义上讲，就是依据以试验或观测取得的有限的信息，以部分推断整体的科学。而正是这个特点决定了数理统计有时做出的结论不可能绝对正确，多少总有一定程度的错误发生，而这种错误发生的可能性在数理统计中就用概率来度量。这个概率小，推断就比较可靠，概率大，推断就比较不可靠。这种伴随有一定概率的推断称为统计推断。

本章主要内容有数理统计学的任务，数理统计学基本概念，总体、个体与样本，样本的产生法(随机数表法与抽签法)，χ^2－分布，F－分布，t－分布，分位点，统计量，抽样分布(费希尔定理及常见的几个抽样分布)。通过对本章内容的学习使读者理解数理统计的基本概念：总体、个体、样本、统计量；掌握样本均值、样本方差和样本矩的计算；掌握 χ^2－分布，F－分布，t－分布的定义及结构；掌握常用概率分布分位数的概念并会查分位数表；了解样本的产生法(随机数表法与抽签法)；熟悉统计量，抽样分布(费希尔定理及常见的几个抽样分布)的基本内容和要求. 本章重点是统计量、抽样分布。本章难点是抽样分布(常见统计量的分布)。

§1.1 数理统计学的任务及基本概念

(一)数理统计学的任务和基本内容

数理统计主要研究两类问题:①试验的设计和研究,即研究如何更合理更有效地获得观察资料的方法;②统计推断,即研究如何利用一定的资料对所关心的问题做出尽可能精确可靠的结论。本书只对统计推断进行讨论。其内容包括:如何收集、整理数据资料;如何对所得的数据资料进行分析、研究,从而对所研究的对象的性质、特点做出推断。数理统计的任务就是研究如何合理的用部分推断整体,这也就是我们所说的统计推断问题。统计推断问题就是数理统计的基本内容。

(二)总体、个体与样本

在数理统计学中,将我们研究的问题所涉及的对象的全体称为总体,而把总体中的每个成员称为个体。

例 1:我们研究一家工厂的某种产品的废品率,这种产品就是我们的总体,而每件产品则是个体。

例 2:研究某学校某年级学生学习"数学"的期末考试成绩情况,该学校该年级全体学生的期末数学考试成绩构成总体,而每个学生的期末数学考试成绩则为个体。

例 3:普查某地区某年龄段人群的身高,则该地区某年龄段的所有人构成总体,而该年龄段的每一个人则为个体。

这里所讲的产品的废品率、学生的期末数学考试成绩、人员的身高,它们的取值都是不同的,即每个个体所取的值是不同的。个体与总体就好像集合论中的元素与集合之间的关系。在试验中抽取某个个体所观察得到的数值 ξ 就是一个随机变量,因而我们用 ξ 的分布去描述总体分布情况。以后我们把总体与随机变量 ξ 可能取值的全体所组成的集合等同起来,并把随机变量 ξ 的分布称为总体的分布,即总体分布就是设定的表示总体的随机变量 ξ 的分布。总体的分布一般说来是未知的,有时虽已知总体分布的类型(如正态分布),但不知道分布中所含的参数,有时连分布所属的类型也不能肯定。统计学的任务就是依据概率论的有关理论利用通过各种方法获取的资料对总体的未知分布进行推断。

为了对总体 ξ 的未知分布进行推断,必须从总体中随机地抽取若干个个体来获取总体的部分信息。抽出的这部分个体称为样本。

假定从总体 ξ 中抽取了 n 个个体 $\xi_1,\xi_2,\cdots,\xi_n$,由于抽取是随机的,抽取之前并不知道这 n 个个体究竟是什么,因此,$\xi_1,\xi_2,\cdots,\xi_n$ 是随机变量。我们称$(\xi_1,\xi_2,\cdots,\xi_n)$为总体的一个样本,样本中个体的数目 n 称为样本容量,当一次抽样完成后,我们把得到的 n 个具体的数据$(x_1,x_2,\cdots,x_n)$称为样本$(\xi_1,\xi_2,\cdots,\xi_n)$的一个样本值(或试验值)。把样本$(\xi_1,\xi_2,\cdots,\xi_n)$的所有取值的全体称为样本空间。

样本的一个重要性质是它的二重性。假设$(\xi_1,\xi_2,\cdots,\xi_n)$是从总体$\xi$中抽取的样本，在一次具体的观测或试验中，它们是一批测量值，是一些已知的数。这就是说，样本具有数的属性。这一点比较容易理解。但是，另一方面，由于在具体的试验或观测中，受到各种随机因素的影响，在不同的观测中样本取值可能不同。因此，当脱离开特定的具体试验或观测时，我们并不知道样本$(\xi_1,\xi_2,\cdots,\xi_n)$的具体取值到底是多少，因此，可以把它们看成随机变量。这时，样本就具有随机变量的属性。样本$(\xi_1,\xi_2,\cdots,\xi_n)$既可被看成数又可被看成随机变量，这就是所谓的样本二重性。这里需要特别强调的是，以后凡是我们离开具体的一次观测或试验来谈及样本$(\xi_1,\xi_2,\cdots,\xi_n)$时，它们总是被看成随机变量，关于样本的这个基本的认识对理解后面的内容十分重要。

如果总体所包含的个体数量是有限的，则称该总体为有限总体，其分布是离散型的，如例1。如果总体所包含的个体数量是无限的，则称该总体为无限总体，其分布可以是连续型的，如例2、例3。在数理统计中，研究有限总体比较困难，因为它的分布是离散型的，且分布列与总体所含个体数量有关系。所以，通常在总体所含个体数量比较大时，我们就把它近似地视为无限总体，并且用连续型分布去逼近总体的分布，这样便于做进一步的统计分析。例如，我们研究某大城市年龄在1到10岁之间儿童的身高。显然，不管这个城市规模有多大，在这个年龄段的儿童数量总是有限的，因此，这个总体只能是有限总体，总体分布也只能是离散型分布。然而，为了便于处理问题，我们可以把它近似地看成一个无限总体，并且通常用正态分布来逼近这个总体的分布。当城市比较大，儿童数量比较多时，这种逼近所带来的误差，从应用观点来看，可以忽略不计。

样本$(\xi_1,\xi_2,\cdots,\xi_n)$的抽取是在在相同条件下对总体进行$n$次独立地重复观测，那么就可以认为所获得的样本$(\xi_1,\xi_2,\cdots,\xi_n)$是独立同分布的变量，这样的样本称为简单随机样本，简称为样本。它实际上满足以下两个条件：

(1)代表性：$\xi_1,\xi_2,\cdots,\xi_n$和总体ξ具有相同的分布；

(2)独立性：$\xi_1,\xi_2,\cdots,\xi_n$相互独立。

通常获取简单随机样本的方法有两种：

1. 抽签法

抽签法是利用抽签原理进行抽样的一种方法。具体作法是，先把总体中的每一个个体编上号并对应的写在签上，然后将签充分混合，从中随机抽取n个签，与被抽到的签号相应的个体作为样本的相应分量。

抽签法有有放回和无放回两种情形，当每次抽取一个个体进行考察后放回去，再抽取第二个，连续抽n次，这种抽样方法称为有放回抽样。有放回抽样所得到的样本是简单随机样本。当每次抽取一个个体不放回去，再抽取第二个，连续抽n次，这种抽样方法称为无放回抽样。当总体所含个体数目比较大，而样本容量比较小时，无放回抽样所得到的样本可近似看做是简单随机样本。

2. 随机数表法

随机数表法是借助于随机数表进行抽样的一种方法。随机数表是由0—9这十个数字随机排列而成的，第一张随机数表是由铁皮特(*Tippet*)在1927年给出。利用随机数表法所获取的样本可视为简单随机样本。

我们通过举例说明如何使用随机数表。

例 4 假定我们要从 50 名同学中抽 5 名同学考察其综合素质，可先将 50 名同学顺序编号：00，01，…，49，然后任意决定表中的一个数作为起始数，按向右顺序逐次取两位数，如果遇到超过 49 的或重复出现的数，就删去，当数取到表格边缘时，换下一行继续取，直到取到 00，01，…，49 中的 5 个数为止。

必须说明的是，本课程只对简单随机样本进行讨论，因此，若无特别说明，所说的样本都是简单随机样本。

(三)样本函数与统计量

1. 样本的联合分布函数

为了今后讨论方便，我们约定，以希腊字母 ξ 表示随机变量，而以小写英文字母 x_i 表示它的观察值，并称样本的一组具体的观察值 $(\xi_1,\xi_2,\cdots,\xi_n)$ 为样本值，全体样本值组成的集合称为样本空间 Ω。

设总体 ξ 具有分布函数 $F(x)$ 或密度函数 $f(x)$，则由概率论知识，我们知道样本 $(\xi_1,\xi_2,\cdots,\xi_n)$ 的分布函数为 $F(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}F(x_i)$，样本 $(\xi_1,\xi_2,\cdots,\xi_n)$ 的密度函数为 $f(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i)$，并称之为样本的联合分布函数或联合密度函数。

在数理统计中，总体或者说总体分布是我们研究的目标，而样本是从总体中随机抽取的一部分个体。通过对这些个体(即样本)进行具体的研究，我们所得到的统计结论以及对这些结论的统计解释，都反映或体现着总体的信息，也就是说，这些信息是对总体而言的。因此，我们总是着眼于总体，而着手于样本，用样本去推断总体。这种由已知推断未知，用具体推断抽象的思想，对我们后面的学习和研究是大有裨益的。

2. 统计量

样本 $(\xi_1,\xi_2,\cdots,\xi_n)$ 是总体 ξ 的一个反映，它体现出总体的许多性质，在获得了样本之后，下一步我们就要对样本进行统计分析，也就是对样本进行加工、整理，从中提取有用信息，并把这些有用信息集中起来。例如，当我们对总体 ξ 的均值感兴趣时，由样本 $(\xi_1,\xi_2,\cdots,\xi_n)$ 产生了一个函数 $\frac{1}{n}\sum_{i=1}^{n}\xi_i$，由概率论的相关理论知道，它较好的集中了样本中有关均值的信息，可以利用它对总体 ξ 的均值进行推断。把一个长度为 μ 的物体测量了 n 次，获得了一组样本 $(\xi_1,\xi_2,\cdots,\xi_n)$ 后，往往计算它们的算术平均值 $\bar{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i$，用来作为 μ 的估计，这 $\bar{\xi}$ 就是对样本 $(\xi_1,\xi_2,\cdots,\xi_n)$ 进行加工处理后得到的一个量，在统计学上称为统计量。

如此看来，统计推断往往要借助于样本的函数。因此，一般情况下，我们把样本 $(\xi_1,\xi_2,\cdots,\xi_n)$ 的不包含任何未知参数的单值函数称为统计量。这样，一旦有了样本，就可以算出统计量的试验值。例如，$\bar{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i$、$\sum_{i=1}^{n}(\frac{\xi_i}{2})^2$、$\frac{1}{n}\sum_{i=1}^{n}\xi_i^2$、$max\{\xi_1,\xi_2,\cdots,\xi_n\}$ 就是统计量，若 $\xi\sim N(\mu,\sigma^2)$，且 μ、σ^2 未知，则 $\sum_{i=1}^{n}(\xi_i-\mu)^2$、$\sum_{i=1}^{n}(\frac{\xi_i}{\sigma})^2$、$\sum_{i=1}^{n}(\frac{\xi_i-\mu}{\sigma})^2$ 就不是统计量，这是因为这三个函数中

包含了未知参数 μ 和 σ^2；若 μ 已知而 σ^2 未知，则 $\sum_{i=1}^{n}(\xi_i-\mu)^2$ 是统计量，$\sum_{i=1}^{n}(\frac{\xi_i}{\sigma})^2$、$\sum_{i=1}^{n}(\frac{\xi_i-\mu}{\sigma})^2$ 不是统计量，因为前者虽然包含了 μ，但 μ 已知，后两者包含了未知参数 σ^2。

统计量是用来对总体分布参数作估计或检验的，因此它应该包含了样本中有关参数的尽可能多的信息，在统计学中，根据不同的目的构造了许多不同的统计量。

下面介绍几个常用的统计量，设 $(\xi_1,\xi_2,\cdots,\xi_n)$ 是取自总体的一个样本：

(1) 样本均值

称 $\bar{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i$ 为样本均值。它的基本作用是估计总体分布的均值和对有关总体均值的假设做检验。

(2) 样本方差

称 $S^2=\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\bar{\xi})^2$ 为样本方差。它的基本作用是用来估计总体分布的方差 σ^2 和对有关总体分布的均值或方差的假设进行检验。需要特别说明的是，在一些统计著作中，有时把样本方差定义为 $S^{*2}=\frac{1}{n-1}\sum_{i=1}^{n}(\xi_i-\bar{\xi})^2$，这种定义的缺点是，它与概率论中方差的计算方式不一致，且在利用统计的思想进行构造反映数据集中与分散程度的统计量时，直接得到的是 S^2，而不是 S^{*2}。因此，本书中我们称 S^{*2} 为样本修正方差。称 S^2 的算术平方根 $S=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\bar{\xi})^2}$ 为样本标准差，它的基本作用是用来估计总体分布的标准差 σ. 必须注意，S 与样本具有相同的度量单位，而 S^2 则不然。通常我们称 S^{*2} 的算术平方根 $S^*=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\xi_i-\bar{\xi})^2}$ 为样本修正标准差，它的基本作用与 S 基本相同。

(3) 样本矩

一般包括样本 k 阶原点矩 $\hat{m}_k=\frac{1}{n}\sum_{i=1}^{n}\xi_i^k$ 和样本 k 阶中心矩 $\hat{c}_k=\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\bar{\xi})^k$，其中 $\hat{m}_1=\bar{\xi}$ 称为样本均值，$\hat{c}_2=s^2$ 称为样本方差，s 称为样本标准差。有时用到 $k+l$ 阶中心混合矩 $\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\bar{\xi})^k(\eta_i-\bar{\eta})^l$ 和 $k+l$ 阶原点混合矩 $\frac{1}{n}\sum_{i=1}^{n}\xi_i^k\eta_i^l$，其中 $(\xi_1,\xi_2,\cdots,\xi_n)$ 和 $(\eta_1,\eta_2,\cdots,\eta_n)$ 分别是总体 ξ 和 η 总体的样本。

(4)顺序统计量

设 $(\xi_1,\xi_2,\cdots,\xi_n)$ 是总体的一个样本，$(x_1,x_2,\cdots,x_n)$ 为样本值，将 $x_1,x_2,\cdots,x_n$ 按大小顺序排列成 $x_1^*\leqslant x_2^*\leqslant\cdots\leqslant x_{n-1}^*\leqslant x_n^*$，并定义统计量 ξ_k^* $(1\leqslant k\leqslant n)$ 如下：当 $(\xi_1,\xi_2,\cdots,\xi_n)$ 取 $(x_1,x_2,\cdots,x_n)$ 时 $\xi_k^*=x_k^*$，$(\xi_1^*,\xi_2^*,\cdots,\xi_n^*)$ 称为样本 $(\xi_1,\xi_2,\cdots,\xi_n)$ 的顺序统计量。

设 $\hat{x}_p=\begin{cases}\frac{\xi_{np}^*+\xi_{np+1}^*}{2}, & np \text{ 是整数}\\ \xi_{[np+1]}^*, & np \text{ 不是整数}\end{cases}$，则称 $\hat{x}_p$ 为样本 $(\xi_1,\xi_2,\cdots,\xi_n)$ 的 p 分位数。

特别 $p=\frac{1}{2}$ 当时，称 $\hat{x}_{\frac{1}{2}}$ 为样本中位数，且 $\hat{x}_{\frac{1}{2}}=\begin{cases}\frac{\xi_{\frac{n}{2}}^*+\xi_{\frac{n}{2}+1}^*}{2}, & n \text{ 为偶数}\\ \xi_{\frac{n+1}{2}}^*, & n \text{ 为奇数}\end{cases}$

$\hat{p}=\xi_n^*-\xi_1^*$ 称为样本极差。

为了使用方便，下面介绍几个常用的计算公式：

(1) $\sum_{i=1}^{n}(\xi_i-\bar{\xi})=0$；

(2) $\sum_{i=1}^{n}(\xi_i-c)^2=\sum_{i=1}^{n}(\xi_i-\bar{\xi})^2+n(\bar{\xi}-c)^2$；

$s^2=\frac{1}{n}\sum_{i=1}^{n}\xi_i^2-(\bar{\xi})^2$；

若 $\eta_i=\frac{\xi_i-a}{b}$，则 $\bar{\eta}=\frac{\bar{\xi}-a}{b}$，$s_\eta^2=\frac{1}{b^2}s_\xi^2$；特别地 $s_{\xi\pm a}^2=s_\xi^2$

(3) $s^{*2}=\frac{n}{n-1}S^2$；

(4)对样本$(\xi_1,\xi_2,\cdots,\xi_n)$做一个变换 $\eta_i=a\xi_i+b, i=1,2,\cdots,n$，这里 a、b 是已知常数，则样本$(\eta_1,\eta_2,\cdots,\eta_n)$的均值$\bar{\eta}=\frac{1}{n}\sum_{i=1}^{n}\eta_i$ 和$\bar{\xi}$有如下关系：

$\bar{\eta}=a\bar{\xi}+b$

这些公式的证明比较简单，留给读者自己练习。

$\bar{\xi}$与 s^2 的一些性质：

定理 1.1 设$(\xi_1,\xi_2,\cdots,\xi_n)$是总体 ξ 的一个样本，$E(\xi)=\mu$，$D(\xi)=\sigma^2$，则

(1) $E(\bar{\xi})=\mu$，$D(\bar{\xi})=\frac{\sigma^2}{n}$；

(2) $E(s^2)=\frac{n-1}{n}\sigma^2$，$E(s^{*2})=\sigma^2$。

证明：(1) $E(\bar{\xi})=E(\frac{1}{n}\sum_{i=1}^{n}\xi_i)=\frac{1}{n}\sum_{i=1}^{n}E(\xi_i)=\frac{1}{n}\sum_{i=1}^{n}\mu=\mu$

$$D(\bar{\xi})=D(\frac{1}{n}\sum_{i=1}^{n}\xi_i)$$

$$=\frac{1}{n^2}\sum_{i=1}^{n}D(\xi_i)$$

$$=\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2=\frac{\sigma^2}{n}$$

(2) $E(S^2)=E[\frac{1}{n}\sum_{i=1}^{n}\xi_i^2-(\bar{\xi})^2]=\frac{1}{n}\sum_{i=1}^{n}E(\xi_i^2)-E(\bar{\xi})^2$

$$=\frac{1}{n}\sum_{i=1}^{n}[D\xi_i+(E\xi_i)^2]-[D\bar{\xi}+(E\xi_i)^2]$$

$$=\frac{1}{n}(n\sigma^2+n\mu^2)-\frac{\sigma^2}{n}-\mu^2=\frac{n-1}{n}\sigma^2$$

$$E(S^{*2})=E(\frac{n}{n-1}S^2)=\frac{n}{n-1}E(S^2)=\frac{n}{n-1}\frac{n-1}{n}\sigma^2=\sigma^2$$

§1.2　抽样分布

(一)抽样分布的概念

总体的分布往往是未知的,或部分地未知。根据实际问题的需要,有时需对总体未知的重要数字特征或总体分布所含的未知参数进行推断。这类问题我们称作为参数的统计推断。在参数统计推断问题中,经常需要利用总体的样本构造出合适的统计量,并使其服从或渐进地服从已知的分布。讨论抽样分布的途径有两个:一是精确地求出抽样分布并称相应的统计推断为小样本统计推断;另一是让样本容量趋于无穷并求出抽样分布的极限分布,然后,在样本容量充分大时,再利用该极限分布作为抽样分布的近似分布,继而对未知参数进行统计推断,因此称相应的统计推断为大样本统计推断。本书只讨论小样本统计推断。

前面我们已经讲过,样本具有二重性。统计量作为样本的函数也具有二重性,即对一次具体的观测或试验,它们都是具体的数值。这时我们会说,样本均值$\bar{\xi}=20$,或样本方差$S^2=2.25$等等。但是离开具体的某次观测或试验,样本是随机变量。因此统计量也是随机变量,也有自己的概率分布,利用统计量对总体进行推断时,一般要借助于统计量的分布,统计量的分布即称为抽样分布。这个分布原则上可以从样本的概率分布计算出来。

一般说来,统计量的抽样分布的计算是很困难的。但是,如果总体服从正态分布,那么像样本均值和样本方差等常见的较简单的统计量的精确抽样分布是容易算出的。本书仅讨论总体服从正态分布情况下的抽样分布。

定理 1.2　(费希尔定理)设$(\xi_1,\xi_2,\cdots,\xi_n)$是来自正态总体$N(\mu,\sigma^2)$的样本,$\bar{\xi}$和$S^2$分别为此样本的样本均值和样本方差,则

(1) $\bar{\xi}\sim N(\mu,\frac{\sigma^2}{n})$;

(2) $\frac{nS^2}{\sigma^2}=\frac{(n-1)S^{*2}}{\sigma^2}\sim\chi^2(n-1)$;

(3) $\bar{\xi}$与S^2相互独立。

其中$\bar{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i$,$S^2=\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\bar{\xi})^2$。

证明:(1) 因$\bar{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i$为相互独立正态随机变量的线性和,所以$\bar{\xi}$仍是一个正态随机变量,又由于$E(\bar{\xi})=\mu$,$D(\bar{\xi})=\frac{\sigma^2}{n}$,所以$\bar{\xi}\sim N(\mu,\frac{\sigma^2}{n})$。

定理中结论(2)与(3)证明较复杂,略去。有了上述关于正态总体的样本均值和样本方差的抽样分布的基础性定理,再结合上一节中关于常用统计分布的有关论述,便可容易地构造出单个正态总体与两个正态总体中样本的一些统计量并使之服从确定的已知分布。

(二)三个常用分布

(1)χ^2—分布:设 $\xi_1,\xi_2,\cdots,\xi_n$ 是相互独立的随机变量,且 $\xi_i\sim N(0,1)(i=1,2,\cdots,n)$,则随机变量 $\chi^2=\xi_1^2+\xi_2^2+\cdots+\xi_n^2$ 所服从的分布称为自由度为 n 的 χ^2 分布,记作 $\chi^2\sim\chi^2(n)$,其中 n 称为自由度,其密度函数为:

$$f(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

图 1-1 描绘了 χ^2 分布密度函数的图形,它随着自由度的不同而有所改变。

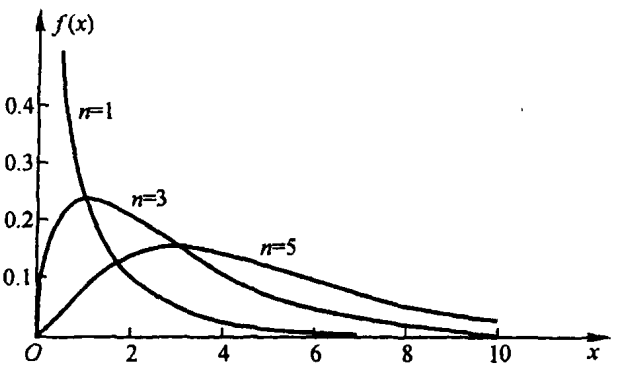

图 1-1

定理 1.3 χ^2-分布的具有性质:

(ⅰ)若 $\xi\sim N(0,1)$,则 $\xi^2\sim\chi^2(1)$;

(ⅱ)若 ξ 与 η 相互独立;且 $\xi\sim\chi^2(m)$,$\eta\sim\chi^2(n)$,则 $\xi+\eta\sim\chi^2(m+n)$;此性质称为 χ^2—分布的可加性。

(ⅲ)若 $\xi_1,\xi_2,\cdots,\xi_n$ 相互独立并且都服从 $\chi^2(1)$ 分布,则 $\sum\limits_{i=1}^{n}\xi_i\sim\chi^2(n)$;

(ⅳ)若 $\xi_1,\xi_2,\cdots,\xi_n$ 相互独立并且都服从 $N(a_i,\sigma_i^2)$,$i=1,2,\cdots,n$,则

$\chi^2=\sum\limits_{i=1}^{n}(\dfrac{\xi^2-a_i}{\sigma_i})^2\sim\chi^2(n)$;

(ⅴ)若 $\xi\sim\chi^2(n)$,则 $E(\xi)=n$,$D(\xi)=2n$

证明:(ⅰ)的证明见概率论。

(ⅱ)依 χ^2—分布的定义有

$$\xi=\sum_{i=1}^{n}\xi_i^2,\eta=\sum_{i=m+1}^{m+n}\xi_i^2$$

其中 $\xi_1,\xi_2,\cdots,\xi_m,\xi_{m+1},\cdots,\xi_{m+n}$ 是独立同分布的随机变量,且都服从 $N(0,1)$,于是

$$\xi+\eta=\sum_{i=1}^{m+n}\xi_i^2\sim x^2(m+n)$$

(ⅲ)由 χ^2—分布的定义及性质(ⅰ)立即得证。

(ⅳ)由正态分布的标准化知 $\dfrac{\xi_i-a_i}{\sigma_i}\sim N(0,1)$,于是根据性质(ⅰ)

$$(\frac{\xi_i-a_i}{\sigma_i})^2\sim x^2(1)$$

且根据 $\xi_1,\xi_2,\cdots,\xi_n$ 相互独立性，可知 $\frac{\xi_i-a_i}{\sigma_i},\frac{\xi_2-a_2}{\sigma_2},\cdots,\frac{\xi_n-a_n}{\sigma_n}$ 也相互独立，从而，由性质(ⅲ)即可得证。

(ⅴ)依 χ^2 一分布的定义，$\xi=\xi_1^2+\xi_2^2+\cdots+\xi_n^2$，其中，$\xi_i\sim N(0,1)(i=1,2,\cdots,n)$，$\xi_1,\xi_2,\cdots,\xi_n$ 相互独立，于是

$$E\xi_i^2=D\xi_i+(E\xi_i)^2=1+0^2=1$$

$$E\xi_i^4=\int_{-\infty}^{+\infty}x^4\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=3$$

$$D\xi_i^2=E\xi_i^4-(E\xi_i^2)^2=3-1^2=2$$

故

$$E\xi=E(\sum_{i=1}^{n}\xi_i^2)=\sum_{i=1}^{n}E\xi_i^2=n$$

$$D\xi=D(\sum_{i=1}^{n}\xi_i^2)=\sum_{i=1}^{n}D\xi_i^2=2n$$

用数学归纳法可把 χ^2 一分布的可加性推广到两个以上的相互独立的 χ^2 随机变量上去。

(2)t 一分布：设 $\xi\sim N(0,1)$，$\eta\sim\chi^2(n)$，且 ξ 与 η 相互独立，则称随机变量 $T=\frac{\xi}{\sqrt{\eta/n}}$ 所服从的分布为自由度为 n 的 t 分布，记作 $T\sim t(n)$，n 称为自由度. 其密度函数为

$$f(x)=\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1+\frac{x^2}{n})^{-\frac{n+1}{2}},-\infty<x<+\infty。$$

图 1-2 描绘了 t 一分布密度函数的图形，它随着自由度的不同而有所改变。

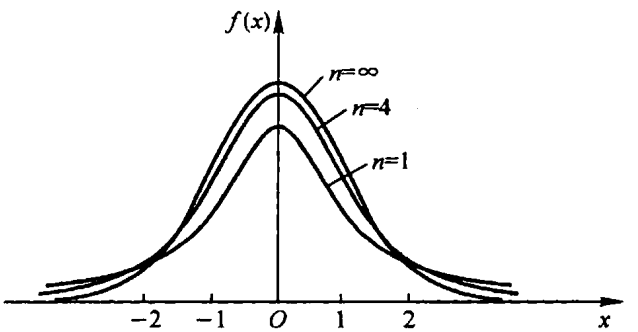

图 1-2

定理 1.4　t 一分布具有如下性质：

(ⅰ)t 一分布的密度函数为偶函数；

(ⅱ)若 $T\sim t(n)$，则 $E(T)=0$(当 $n\geqslant 2$ 时)；

证明：(ⅰ)由 t 一分布的密度函数表达式立即得证。

(ⅱ) 当 $n\geqslant 2$ 时，$E(T)=\int_{-\infty}^{+\infty}xf(x)dx=0$。

当 $n=1$ 时，$E(T)$不存在。

需要说明的是可以利用函数的性质证明，当 $n\to\infty$时，t－分布的极限分布为标准正态分布。即

$$\lim_{n\to\infty} f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \qquad (-\infty<x<+\infty)$$

t－分布是统计学中的一个重要分布，它与 $N(0,1)$的微小差别是戈塞特(*Gosset*, *W. S.* 1876－1937)提出的。他是英国一家酿酒厂的化学技师，在长期从事实验和数据分析工作中，发现了 t 分布，并在 1908 年以"*Student*"笔名发表此项结果，故后人又称它为"学生氏分布"。在当时正态分布一统天下的情况下，戈塞特的 t－分布没有被外界理解和接受，只能在他的酿酒厂中使用，直到 1923 年英国统计学家费西尔(*Fisher*, *R. A.* 1890－1962)给出分布的严格推导并于 1925 年编制了 t－分布表后，t 分布才得到学术界的承认，并获得迅速的传播、发展和应用。

(3)F－分布：设 $\xi\sim\chi^2(n_1)$，$\eta\sim\chi^2(n_2)$，且 ξ 与 η 相互独立，则称随机变量 $F=\dfrac{\xi/n_1}{\eta/n_2}$所服从的分布为第一自由度为 n_1，第二自由度为 n_2 的 F－分布，记作 $F\sim F(n_1,n_2)$. 其密度函数为

$$f(x)=\begin{cases}\dfrac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}n_1^{\frac{n_1}{2}}n_2^{\frac{n_2}{2}}x^{\frac{n_1}{2}-1}(n_1x+n_2)^{-\frac{n_1+n_2}{2}}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

图 1-3 描绘了 F－分布密度函数的图形，它随着自由度的不同而有所改变。

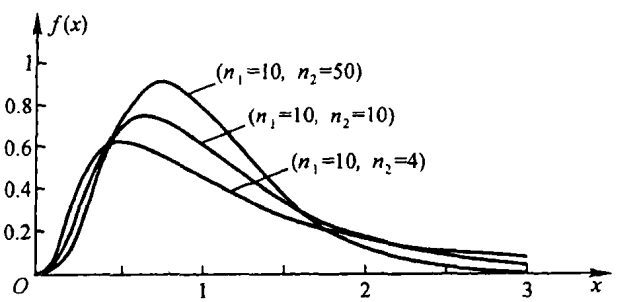

图 1-3

定理 1.5 F－分布具有性质：若 $F\sim F(n_1,n_2)$，则$\dfrac{1}{F}\sim F(n_2,n_1)$。

证明：因为 $F=\dfrac{\xi/n_1}{\eta/n_2}\sim F(n_1,n_2)$，所以

$$\frac{1}{F}=\frac{\eta/n_2}{\xi/n_1}\sim F(n_2,n_1)$$

定理 1.6 设$(\xi_1,\xi_2,\cdots,\xi_n)$是来自正态总体 $N(\mu,\sigma^2)$的样本，$\bar{\xi}$和 S^2 分别为此样本的样本均值和样本方差，则

(1) $U=\dfrac{\bar{\xi}-\mu}{\sigma}\sqrt{n}\sim N(0,1)$；

(2) $T=\frac{\bar{\xi}-\mu}{s}\sqrt{n-1}=\frac{\bar{\xi}-\mu}{s^*}\sqrt{n}\sim t(n-1)$;

其中$\bar{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i, S^2=\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\bar{\xi})^2, S^{*2}=\frac{1}{n-1}\sum_{i=1}^{n}(\xi_i-\bar{\xi})^2$.

证明:(1) 由定理 1.2,$\bar{\xi}\sim N(\mu,\frac{\sigma^2}{n})$,于是由正态分布的标准化

$$U=\frac{\bar{\xi}-\mu}{\sigma}\sqrt{n}\sim N(0,1)$$

(2) 由定理 1.2 知$\bar{\xi}$与 S^2 相互独立,且

$$\frac{nS^2}{\sigma^2}=\frac{(n-1)S^{*2}}{\sigma^2}\sim\chi^2(n-1)$$

又由本定理(1)$\frac{\bar{\xi}-\mu}{\sigma}\sqrt{n}\sim N(0,1)$,且$\frac{\bar{\xi}-\mu}{\sigma}\sqrt{n}$与$\frac{nS^2}{\sigma^2}$相互独立,所以由 t-分布的结构

$$\frac{\frac{\bar{\xi}-\mu}{\sigma}\sqrt{n}}{\sqrt{\frac{nS^2}{(n-1)\sigma^2}}}=\frac{\bar{\xi}-\mu}{S}\sqrt{n-1}=T\sim t(n-1)$$

定理 1.7　设$(\xi_1,\xi_2,\cdots,\xi_m)$是来自总体 $\xi\sim N(\mu_1,\sigma_1^2)$的样本,$(\eta_1,\eta_2,\cdots,\eta_n)$是来自总体 $\eta\sim N(\mu_2,\sigma_2^2)$的样本,且总体 ξ 与 η 相互独立,则

(1)$U=\frac{(\bar{\xi}-\bar{\eta})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}\sim N(0,1)$;

(2)$F=\frac{\frac{1}{m\sigma_1^2}\sum_{i=1}^{m}(\xi_i-\mu_1)^2}{\frac{1}{n\sigma_2^2}\sum_{i=1}^{n}(\eta_i-\mu_2)^2}\sim F(m,n)$;

(3)$F=\frac{\frac{mS_1^2}{(m-1)\sigma_1^2}}{\frac{nS_2^2}{(n-1)\sigma_2^2}}=\frac{\frac{S_1^{*2}}{\sigma_1^2}}{\frac{S_2^{*2}}{\sigma_2^2}}=\frac{S_1^{*2}}{S_2^{*2}}\frac{\sigma_2^2}{\sigma_1^2}\sim F(m-1,n-1)$.

特别当 $\sigma_1^2=\sigma_2^2$ 时, $F=S_1^{*2}/S_2^{*2}\sim F(m-1,n-1)$.

其中

$$\bar{\xi}=\frac{1}{m}\sum_{i=1}^{m}\xi_i,\ \bar{\eta}=\frac{1}{n}\sum_{i=1}^{n}\eta_i$$

$$S_1^2=\frac{1}{m}\sum_{i=1}^{m}(\xi_i-\bar{\xi})^2, S_2^2=\frac{1}{n}\sum_{i=1}^{n}(\eta_i-\bar{\eta})^2$$

$$S_1^{*2}=\frac{1}{m-1}\sum_{i=1}^{m}(\xi_i-\bar{\xi})^2, S_2^{*2}=\frac{1}{n-1}\sum_{i=1}^{n}(\eta_i-\bar{\eta})^2$$

证明:(1)由定理 1.2 得,$\bar{\xi}\sim N(\mu_1,\frac{\sigma_1^2}{m})$,$\bar{\eta}\sim N(\mu_2,\frac{\sigma_2^2}{n})$两者又相互独立,于是

$$\bar{\xi}-\bar{\eta}\sim N(\mu_1-\mu_2,\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n})$$