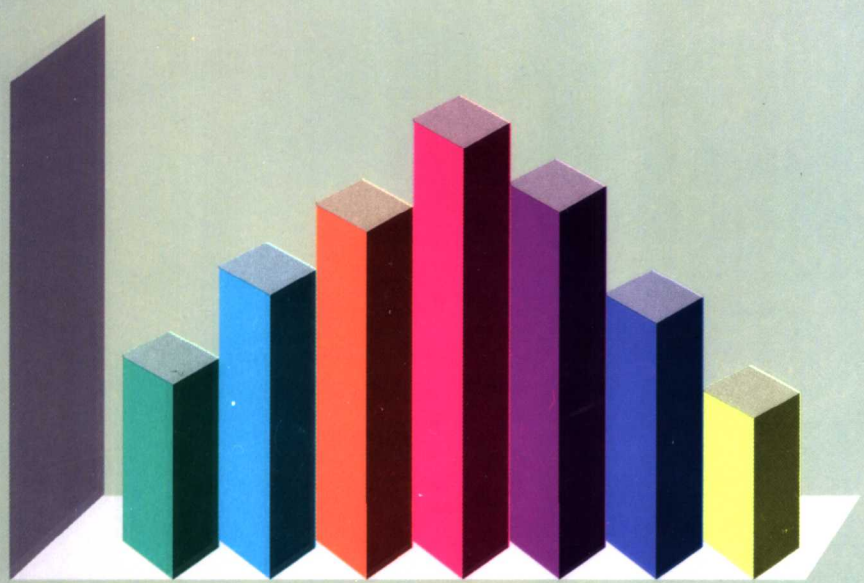
胡发胜 宿洁 编著

Mathematical Statistics

数理统计



山东大学出版社
Shandong University Press

数理统计

胡发胜　宿洁　编著

山东大学出版社

图书在版编目(CIP)数据

数理统计/胡发胜　宿洁编著.—济南:山东大学出版社,2004.9
ISBN 7-5607-2861-8

Ⅰ.数...
Ⅱ.①胡...②宿...
Ⅲ.数理统计
Ⅳ.0212

中国版本图书馆 CIP 数据核字(2004)第 091427 号

内容提要

本书的主要内容有:抽样分布、参数估计、假设检验、方差分析和回归分析.本书每章末附有习题,书后附有答案.

本书可作为数学类各专业的本科教材,也可以作为科研人员以及从事实际应用的工程技术人员的参考书.

山东大学出版社出版发行
(山东省济南市山大南路 27 号　邮政编码:250100)
山东省新华书店经销
日照报业印刷有限公司印刷
787×1092 毫米　1/16　9.5 印张　215 千字
2004 年 9 月第 1 版　　2004 年 9 月第 1 次印刷
印数:1-2000 册
定价:18.00 元

前　言

数理统计是一门研究随机现象数量规律的一门学科，主要是研究如何以有效的方式收集、整理和分析随机数据，并在此基础上，对随机性问题作出系统性的推断，从而为决策分析服务．该学科在工农业生产、经济管理、生命科学等诸多领域都有广泛的应用．

本书全面系统地介绍了数理统计的概念、理论和方法，详细论述了抽样分布、参数估计、假设检验、方差分析和回归分析等的基本概念、主要结论和具体操作方法等内容．

本书具有以下特点：一是内容全面、系统，突出数学思想，同时密切联系实际问题，适当反映统计方法在实际中的新进展；二是语言文字表达清晰、平实，便于读者接受和理解．

本书可作为数学类各专业的本科教材，也可以作为科研人员以及从事实际应用的工程技术人员的参考书．

本书共分五章，第一至四章由胡发胜编写，第五章及每章的习题由宿洁编写．

本书获得山东大学出版基金委员会资助．

在本书的编写过程中得到了山东大学教务处、山东大学出版基金委员会、山东大学数学与系统科学学院领导的鼎力支持．吴臻教授和林路教授对本书内容提出了许多宝贵修改意见和建议．谨此一并表示诚挚的感谢．

限于编者的水平，书中难免有不妥和错误之处，欢迎读者批评指正．

编　者

2004 年 8 月

目　录

第一章　数理统计的基本知识

§1.1　数理统计学

数理统计和概率论一样，也是研究随机现象数量规律性的一门科学. 但二者研究的内容和出发点有所不同.

在概率论中，为研究随机现象，引入了随机试验、随机事件、随机变量等概念，并且我们知道：只要对随机现象进行多次的观察，就可以得到随机变量的分布. 而有了随机变量的分布，就可以很容易地研究事件发生的概率和随机变量的数字特征等问题. 至于如何设计随机试验，以及对随机现象只能进行有限次甚至是少量次的观测时，如何去推断随机变量的分布等问题，概率论并未涉及.

例 1.1　一个灯泡厂希望了解本厂生产的一批灯泡的寿命，假定产量是 10 万只. 由于测定灯泡寿命是破坏性试验，因而不可能对每一灯泡进行测定，而只能抽取若干个(比如 100 只) 灯泡做试验，从这 100 个灯泡的寿命数据去推断该批灯泡的寿命特征，如平均寿命等.

例 1.2　根据去年的调查，某城市一个家庭每月的耗电量服从正态分布 $N(45,9^2)$，为了确定今年家庭平均每月的耗电量是否提高，随机抽查 100 个家庭，统计得他们每月的耗电量的平均值为 47.5，据此数据给出你的结论.

类似的问题在实际中经常会遇见，这些问题的共同点是：测量数据带有随机性，回答这些问题一般要涉及两个方面：① **试验的设计和研究**，即研究如何更合理、更有效地获取观测数据的方法. 在例 1.1 中，就是这 100 个测量灯泡如何去选取. ② **统计推断**，即研究如何利用测定的数据和其他知识对所关心的问题作出尽可能精确、可信的结论. 当然这两方面有密切联系，在实际问题中要二者兼顾. 它们都是数理统计所要研究的内容，本书仅讨论统计推断.

数理统计就是研究怎样以有效的方式**收集**、**整理**、**分析**带有随机性的数据，并在此基础上，对所研究的问题作出系统性的推断，从而为某种决策分析服务. 数理统计的理论基础是概率论.

本章主要介绍数理统计的一些基本概念和数据处理的一些初等方法.

§1.2 总体和样本

一、总体和个体

研究实际问题时，首先要明确所有的研究对象是什么. 由于我们关心的往往是研究对象的某种数量指标，因而还需明确研究的数量指标是什么. 例 1.1 中，该厂生产的 10 万只灯泡就是所有的研究对象，而灯泡寿命就是要研究的数量指标.

数理统计中将研究对象所有成员的某一数量指标取值的全体称为**总体**，而构成总体的每一个元素称为**个体**. 例 1.1 中，总体就是这 10 万只灯泡的寿命数据(尚未知道)组成的全体. 个体就是每只灯泡的寿命数据. 在例 1.2 中，总体是该城市每户家庭月耗电量数的全体，个体是每户家庭月耗电量数.

数量指标的取值随个体不同而不同，尽管在观测之前可以知道其一切可能取值，但事先无法准确预测每一个体的具体取值. 因而可以用**随机变量** X **去描述总体**，简称**总体** X，X 的分布函数 $F(x)$ 便是总体的分布函数，有时也用 $F(x)$ 去表示一个总体. 例如，描述总体的随机变量 X 服从正态分布时，可简称总体 X 为正态总体.

在有些问题中，人们关心的数量指标不止一个，例如：观察某市中学生的身高和体重状况，可用随机向量 (X_1, X_2) 去描述这两个数量指标. 再如，观察某省大型企业的经济效益状况时，有劳动生产率、资金利润率等多个数量指标. 一般地，可用多维随机向量 $(X_1, X_2, \cdots, X_p)'$ 去描述 p 个数量指标，也可用其联合分布函数 $F(x_1, x_2, \cdots, x_p)$ 去描述它们，这种总体称为 p **维总体**.

总体中的个体数量可能是有限的或无限的. 若总体含有有限个个体，称为**有限总体**，若总体含有无限个个体，称为**无限总体**. 有时，一个有限总体中的个体数目非常大，可近似看成是无限总体.

二、样　本

为了研究总体，就必须对个体进行试验与观测. 在多数情况下，对个体的观测往往要付出一定的人力、物力和财力，有些试验可能周期长或具有破坏性(例如：观测灯泡的使用寿命). 因此，通常人们只是从总体中抽取若干个体，通过测定这些个体的值对总体进行推断.

从总体中抽取的待测个体组成的集合称为**样本**，样本中的个体数目称为**样本容量**. 从总体 X 中抽出的容量为 n 的样本常记为 $X_1, X_2, \cdots, X_n$. 由于每个个体是否被抽到具有随机性，第 i 个样本 X_i 的值在观测前是无法知道的，因而每个 X_i 都被看成是随机变量.

样本$(X_1,X_2,\cdots,X_n)'$所有可能取值的全体称为**样本空间**.样本空间一般是R^n或R^n的一个子集,一个样本观测值$(x_1,x_2,\cdots,x_n)'$就是样本空间的一个点(元素).

抽样的目的是为了对总体进行推断,这就要求抽取的样本能很好地反映总体的信息,所以要有一个好的抽样方法,通常要求抽取出的样本满足以下两点:

(1)**代表性** 只要每一个体都有同等机会被选入样本,样本X_i就与总体X具有相同的分布函数,这样的样本就具有代表性.

(2)**独立性** 要求样本$X_1,X_2,\cdots,X_n$的取值彼此互不影响,也就是要求$X_1,X_2,\cdots,X_n$相互独立.

满足以上两条性质的样本称为**简单随机样本**,即简单随机样本$X_1,X_2,\cdots,X_n$是相互独立的具有相同分布的n个随机变量,简记为iid.本书涉及的样本都是指简单随机样本,样本的观测值通常记为$x_1,x_2,\cdots,x_n$.显然,简单随机样本有下列性质:

定理 1.2.1 设$X_1,X_2,\cdots,X_n$是来自总体X的简单随机样本,

(1)若总体X的分布函数为$F(x)$,则样本$X_1,X_2,\cdots,X_n$的联合分布函数为

$$F(x_1,\cdots,x_n)=\prod_{i=1}^{n}F(x_i) \tag{1.2.1}$$

(2)若总体X是连续型随机变量,概率密度函数为$P(x)$,则$X_1,X_2,\cdots,X_n$的联合密度函数为

$$P(x_1,\cdots,x_n)=\prod_{i=1}^{n}P(x_i) \tag{1.2.2}$$

(3)样本$X_1,\cdots,X_n$与总体X具有相同的各阶矩,特别有

$$E(X_i)=E(X),D(X_i)=D(X),\quad i=1,2,\cdots,n \tag{1.2.3}$$

§1.3 统计量与几种概率分布

样本来自于总体,是总体的代表和反映.对抽取的样本进行观测后,得到的是一些杂乱无章的数据,通常不能直接利用它们进行推断,而需要对它们进行加工和整理,把样本中所包含的我们所关心问题有关**信息集中**起来,也就是针对不同的问题构造出样本的某种函数,这种函数在数理统计中称为**统计量**.严格地说,一个统计量就是n元随机变量$(X_1,X_2,\cdots,X_n)$一个**波雷尔**(Borel)**可测函数**,且要求它不含有任何未知参数,记为$T(X_1,X_2,\cdots,X_n)$.因此**统计量也是一个随机变量**(或**向量**).

例如:设X_1,X_2,X_3是来自正态总体$N(\mu,\sigma^2)$的一组样本,其中μ已知,而σ^2未知,则$\frac{1}{3}(X_1+X_2+X_3)-\mu$,$\frac{1}{2}(X_1+X_2)$,$\frac{1}{3}(X_1^2+X_2^2+X_3^2)$都是统计量.而$X_1^2/\sigma^2$,$(X_2-\mu)/\sigma$都不是统计量,因为它们含有未知参数.当我们获得了样本的观测值,就可以得

到统计量的观测值，下面介绍常用的几个统计量.

二、常用统计量

设 $X_1,\cdots,X_n$ 是总体 X 的一组样本，则常用的统计量有

1. **样本均值**

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{1.3.1}$$

它反映了总体均值的信息.

2. **样本方差**

$$S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2 \tag{1.3.2}$$

它反映了总体方差的信息，另一个常用的是**样本无偏方差**

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \tag{1.3.3}$$

不难得出

$$S_n^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2,S^2=\frac{1}{n-1}\Big[\sum_{i=1}^{n}X_i^2-n\overline{X}^2\Big]$$

3. **样本标准差**

$$S_n=\sqrt{S_n^2}\text{ 或 }S=\sqrt{S^2} \tag{1.3.4}$$

它们反映了总体标准差的信息.

4. **样本的 k 阶原点矩**

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k\quad k=1,2,\cdots \tag{1.3.5}$$

它反映了总体 k 阶原点矩 μ_k 的信息. 显然 $A_1=\overline{X}$.

5. **样本的 k 阶中心矩**

$$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k\quad k=1,2,\cdots \tag{1.3.6}$$

它反映了总体 k 阶中心矩的信息. 显然 $B_1=0,B_2=S_n^2$.

另外，还有样本偏度及样本峰度等统计量.

设 $(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)$ 是来自二维总体 (X,Y) 的一组样本，则常用统计量有

6. **样本协方差**

$$S_{XY}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-Y\overline{X}) \tag{1.3.7}$$

它反映了总体的两个分量 X 和 Y 的协方差的信息.

7. **样本相关系数**

$$\rho_{XY}=\frac{S_{XY}^2}{S_X S_Y} \tag{1.3.8}$$

其中

$$S_X^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2, S_Y^2=\frac{1}{n}\sum_{i=1}^{n}(Y_i-Y\bar{X})^2.$$

它反映了总体的二个分量 X 和 Y 的相关系数的信息.

三、几种概率分布

下面介绍统计学中常用的三种概率分布,它们是 χ^2 分布、t 分布和 F 分布.

1. χ^2 **分布**

定义 1.3.1 设随机变量 $X_1,X_2,\cdots,X_n$ 相互独立且均服从正态分布 $N(0,1)$,则称随机变量

$$\chi^2=\sum_{i=1}^{n}X_i^2 \tag{1.3.9}$$

所服从的分布为**自由度是 n 的 χ^2 分布**,记作 $\chi^2\sim\chi^2(n)$.

定理 1.3.1 (χ^2 分布的可加性)

设 $Y\sim\chi^2(n_1)$, $Z\sim\chi^2(n_2)$,并且 Y,Z 相互独立,则

$$X=Y+Z\sim\chi^2(n_1+n_2)$$

证明 设 $X_1,X_2,\cdots,X_{n_1},X_{n_1+1},\cdots,X_{n_1+n_2}$ 是 n_1+n_2 个相互独立且均服从 $N(0,1)$ 的随机变量,记 $Y_1=\sum_{j=1}^{n_1}X_j^2, Z_1=\sum_{j=1}^{n_2}X_{n_1+j}^2$,则 $Y_1\sim\chi^2(n_1), Z_1\sim\chi^2(n_2)$. 并且 Y_1 与 Z_1 相互独立,因而随机变量 Y 与 Y_1 同分布,随机变量 Z 与 Z_1 同分布. 又 Y 与 Z 也相互独立,根据相互独立的随机变量和的密度函数公式知:随机变量 $X=Y+Z$ 与随机变量 Y_1+Z_1 同分布,而后者等于 $\sum_{j=1}^{n_1+n_2}X_j^2$ 服从 $\chi^2(n_1+n_2)$,故

$$X=Y+Z\sim\chi^2(n_1+n_2)$$

证毕.

定理 1.3.2 设 $X\sim\chi^2(n)$,则

(1)X 的概率密度函数为

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

(2)$E(X)=n, D(X)=2n$.

证明 (1) 用归纳法证明. 当 $n=1$ 时,X 的分布函数

$$F(x)=P(X\leqslant x)=\int_{y^2\leqslant x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}y\overset{y^2=t}{=\!=}\int_0^x\frac{1}{\sqrt{2\pi}}t^{-\frac{1}{2}}\mathrm{e}^{-\frac{t}{2}}\mathrm{d}t$$

两边对 x 求导,并利用 $\Gamma(\frac{1}{2})=\sqrt{\pi}$ 得

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

故结论成立,设 $n-1$ 时,结论也成立.

对于 n,不妨设 $Y\sim\chi^2(n-1)$,$Z\sim\chi^2(1)$,并且二者相互独立,由 χ^2 分布的可加性知 X 与 $Y+Z$ 同分布,利用归纳假设及独立和的密度函数公式可知

对于 $x>0$,有

$$\begin{aligned}\chi^2(x,n)&=\int_{-\infty}^{+\infty}P_Y(y)P_Z(x-y)\mathrm{d}y\\&=\int_0^x\frac{1}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})\Gamma(\frac{1}{2})}y^{\frac{n-1}{2}-1}\mathrm{e}^{-\frac{y}{2}}\frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})}(x-y)^{-\frac{1}{2}}\mathrm{e}^{-\frac{x-y}{2}}\mathrm{d}y\\&=\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n-1}{2})\Gamma(\frac{1}{2})}\mathrm{e}^{-\frac{x}{2}}\int_0^x y^{\frac{n-1}{2}-1}(x-y)^{-\frac{1}{2}}\mathrm{d}y\end{aligned}$$

令 $z=\dfrac{y}{x}$,则

$$\begin{aligned}\int_0^x y^{\frac{n-1}{2}-1}(x-y)^{-\frac{1}{2}}\mathrm{d}y&=\int_0^1 x^{\frac{n}{2}-1}z^{\frac{n-1}{2}-1}(1-z)^{-\frac{1}{2}}\mathrm{d}z\\&=x^{\frac{n}{2}-1}B\left(\frac{n-1}{2},\frac{1}{2}\right)\\&=\frac{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}\end{aligned}$$

代入上式得

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

(2) $X=\sum\limits_{i=1}^{n}X_i^2$,$X_i\sim N(0,1)$ 且 iid,则

$$E(X)=E\left(\sum_{i=1}^{n}X_i^2\right)=\sum_{i=1}^{n}E(X_i^2)=\sum_{i=1}^{n}D(X_i)=n$$

利用 Γ 函数的性质 $\Gamma(p)=(p-1)\Gamma(p-1)$，$\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$ 不难求得 $E(X_i^4)=3$，故

$$\begin{aligned}D(X)&=D\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}D(X_i^2)\\&=\sum_{i=1}^{n}[E(X_i^4)-(E(X_i^2))^2]\\&=\sum_{i=1}^{n}(3-1^2)=2n\end{aligned}$$

证毕.

下面给出一个比定理 1.3.1 更深刻的结论

定理 1.3.3（**柯赫伦定理**）设 $X_1,X_2,\cdots,X_n$ 是相互独立且服从正态分布 $N(0,1)$ 的随机变量，若

$$Q=Q_1+\cdots+Q_k=\sum_{i=1}^{n}X_i^2$$

其中 Q_i 是秩为 n_i 的 $X_1,X_2,\cdots,X_n$ 的二次型. 则 $Q_1,\cdots,Q_k$ 相互独立，且 $Q_i\sim\chi^2(n_i)$ 的充分必要条件是

$$\sum_{i=1}^{k}n_i=n$$

证明　必要性即为定理 1.3.1.

充分性. 由于 Q_i 是秩为 n_i 的 $X_1,X_2,\cdots,X_n$ 的二次型，根据线性代数的知识，Q_i 有下列形式的标准形

$$\begin{aligned}Q_1&=b_{11}Y_{11}^2+b_{12}Y_{12}^2+\cdots+b_{1n_1}Y_{1n_1}^2\\Q_2&=b_{21}Y_{21}^2+b_{22}Y_{22}^2+\cdots+b_{2n_2}Y_{2n_2}^2\\&\cdots\cdots\\Q_k&=b_{k1}Y_{k1}^2+b_{k2}Y_{k2}^2+\cdots+b_{kn_k}Y_{kn_k}^2\end{aligned}$$

其中 Y_{ij} 都是 $X_1,X_2,\cdots,X_n$ 的线性组合，系数 $b_{ij}=1$ 或 -1. 记

$$\begin{aligned}Y&=(Y_{11},\cdots,Y_{1n_1},\cdots,Y_{k1},\cdots,Y_{kn_k})'\\B&=\operatorname{diag}(b_{11},\cdots,b_{1n_1},\cdots,b_{k1},\cdots,b_{kn_k})\\X&=(X_1,X_2,\cdots,X_n)'\end{aligned}$$

并设 $Y=AX$，则

$$X'X=\sum_{i=1}^{n}X_i^2=Q=Q_1+\cdots+Q_k=Y'BY=X'A'BAX$$

得到 $A'BA=I$，所以 $B=(AA')^{-1}$ 是正定矩阵，从而所有的 b_{ij} 等于 1，于是 $B=I$ 及 $AA'=I$. 因而 A 是正交矩阵.

由于 $X \sim N_n(0, I_n)$，由多元正态分布的性质，$Y = AX \sim N_n(0, AA') = N_n(0, I_n)$. 故 Y 的 n 个分量相互独立且均服从 $N(0,1)$. 由 Q_i 的表达式知：Q_1、Q_2、…、Q_k 相互独立，且 $Q_i \sim \chi^2(n_i)$　$i = 1, 2, \cdots, k$. 证毕.

柯赫伦定理在方差分析的研究中起重要作用.

2. t **分布**

定义 1.3.2　设 $X \sim N(0,1)$，$Y \sim \chi^2(n)$，并且 X 与 Y 相互独立，则称随机变量

$$T = \frac{X}{\sqrt{Y/n}} \tag{1.3.10}$$

所服从的分布为**自由度是 n 的 t 分布**，记作 $T \sim t(n)$.

后面将看到，它在正态总体的抽样中是很自然地出现的. 利用独立随机变量商的密度函数公式，不难得到 $t(n)$ 分布的密度函数为

$$t(x,n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

可见，t 分布的密度函数关于 Y 轴对称，且

$$\lim_{x \to \infty} t(x,n) = 0, \quad \lim_{n \to \infty} t(x,n) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

故当 $n \to \infty$ 时，t 分布趋于正态分布. 一般说来，当 $n > 30$ 时，t 分布与正态分布 $N(0,1)$ 就非常接近了. 但对较小的 n 值，t 分布与正态分布之间有较大的差异，且

$$P(|T| \geqslant t_0) \geqslant P(|X| \geqslant t_0)$$

其中 $X \sim N(0,1)$，即 t 分布的尾部比在标准正态分布的尾部有更大的概率.

$t(n)$ 分布只存在阶数 $k < n$ 的矩，特别 $n = 1$ 时，t 分布即是标准化柯西分布，不存在任何阶矩. 当 $n > 2$ 时有

$$E(X) = 0, \quad D(X) = E(X^2) = \frac{n}{n-2}$$

3. F **分布**

定义 1.3.3　设 $X \sim \chi^2(m)$，$Y \sim \chi^2(n)$，且 X 与 Y 相互独立，则称随机变量

$$F = \frac{X/m}{Y/n} \tag{1.3.11}$$

所服从的分布为自由度是 (m,n) 的 F **分布**，记作 $F \sim F(m,n)$，其中 m 称为**第一自由度**，n **称为第二自由度**.

利用独立随机变量商的密度函数公式不难得到 $F(m,n)$ 分布的密度函数是

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

由 F 分布的定义，容易看出：

(1) 如果 $X\sim F(m,n)$，则 $\dfrac{1}{X}\sim F(n,m)$；

(2) 如果 $X\sim t(n)$，则 $X^2\sim F(1,n)$，即 $F(1,n)$ 与 $t^2(n)$ 同分布.

4. 概率分布的分位数

定义 1.4.4 设随机变量 X 的分布函数为 $F(x)$，如果实数 y_α 使得

$$P(X\leqslant y_\alpha)=F(y_\alpha)=\alpha \tag{1.3.12}$$

则称 y_α 为 X 的**下侧 α 分位数**. 如果实数 x_α 使得

$$P(X>x_\alpha)=1-F(x_\alpha)=\alpha \tag{1.3.13}$$

则称 x_α 为 X 的**上侧 α 分位数**.

分位数也叫**分位点**或临界值. 如果 X 是连续型随机变量，则 X 的**下侧 α 分位数**表明：**密度函数位于 y_α 左边图形的面积等于 α**. X 的**上侧 α 分位数**表明：**密度函数位于 x_α 右边图形的面积等于 α**. 不难证明分位数有下列性质：

(1) 对任何随机变量

$$x_\alpha=y_{1-\alpha},\quad y_\alpha=x_{1-\alpha}$$

(2) 对密度函数是偶函数的随机变量

$$x_{1-\alpha}=-x_\alpha,\quad y_{1-\alpha}=-y_\alpha$$

特别标准正态分布 $N(0,1)$ 和 $t(n)$ 分布有此性质.

(3) 设 $F_\alpha(m,n)$ 为 $F(m,n)$ 的上(或下)侧 α 分位数，则

$$F_\alpha(m,n)=\frac{1}{F_{1-\alpha}(n,m)}$$

读者可自己证明上述性质.

常用的正态分布 $N(0,1)$，$t(n)$，$\chi^2(n)$ 分布及 $F(m,n)$ 分布的上侧 α 分位数分别用 μ_α，t_α，χ^2_α 和 F_α 表示. 本书附表给出了上述四种分布的上侧 α 分位数，例如：$\mu_{0.05}=1.625$，$t_{0.05}(10)=3.1693$，$\chi^2_{0.01}(9)=21.666$，$F_{0.05}(3,4)=6.59$.

5. 几点说明

(1) 本书后面所提到的 α 分位数均为上侧 α 分位数.

(2) 如果 $\chi^2\sim\chi^2(n)$，且 $n>45$，一般没有表查 χ^2 分布的上侧 α 分位数 $\chi^2_\alpha(n)$. 可以证明：$n\to\infty$ 时，有 $\sqrt{2\chi^2(n)}-\sqrt{2n-1}\xrightarrow{L}N(0,1)$(读者自证)，故 n 较大时，有近似公式

$$\chi_\alpha^2(n) \approx \frac{1}{2}(\mu_\alpha + \sqrt{2n-1})^2$$

(3) 如果 $t \sim t(n)$，且 $n > 30$，一般没有表查 t 分布的上侧 α 分位数 $t_\alpha(n)$. 由于 n 充分大时，$t(n)$ 分布收敛于 $N(0,1)$，故 n 较大时，有近似公式

$$t_\alpha(n) \approx \mu_\alpha$$

(4) 如果 X 是离散型随机变量(例如：二项分布、泊松分布等)，则其分布函数 $F(x)$ 在 $(-\infty,\infty)$ 上不连续. 因而 X 的某些上(或下)侧 α 分位数将不存在. 这时上侧 α 分位数 x_α 和下侧 α 分位数 y_α 分别由下式定义

$$x_\alpha = \inf\{x: P(X > x) = 1 - F(x) \leqslant \alpha\}$$

$$y_\alpha = \inf\{y: P(Y \leqslant y) = F(y) \geqslant \alpha\}$$

此定义的上、下侧 α 分位数仍满足分位数的性质(1).

§1.4 抽样分布

一、抽样分布

我们知道，统计量是样本的函数，也是一个随机变量(向量). **统计量的分布称为抽样分布**，统计推断的好坏与所选择的统计量的分布有着密切的关系，因此寻求抽样分布是统计学的一项重要内容.

寻求抽样分布的主要有以下两种：

1. 当总体 X 的分布函数已知(可以含有未知参数)，如果对任一容量为 n 的样本 $X_1, X_2, \cdots, X_n$ 能求出给定统计量 $T = T(X_1, X_2, \cdots, X_n)$ 的分布函数的明显表达式，这种方法称为**精确方法**，所得分布称为 $T = T(X_1, X^2, \cdots, X_n)$ 的精确抽样分布. 求出统计量的精确分布，对于样本容量 n 较小的统计推断问题(小样本问题)特别有用. 目前精确抽样分布大多是在正态总体条件下得到的，本节将给予讨论.

2. 一般情况下，精确抽样分布不易求出，或者求出来过于复杂而不便于应用，这时人们可寻找 $n \to \infty$ 时统计量 T 的极限分布，如果该极限存在并能求出，则当 n 较大时，可用此极限当作 T 的近似分布，这种极限分布称为**渐近分布**. 渐近分布对于大样本问题是非常有用的.

二、正态总体的抽样分布

下面介绍以后各章中要用到的几个抽样分布定理.

定理 1.4.1 设 $(X_1, X_2, \cdots, X_n)$ 是来自正态总体 $X \sim N(\mu, \sigma^2)$ 的一组样本，则有下列结论：

(1) $$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\sim N(\mu,\frac{\sigma^2}{n})\text{ 或}\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}\sim N(0,1)\tag{1.4.1}$$

(2) $$\frac{(n-1)S^2}{\sigma^2}=\frac{nS_n^2}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2\sim\chi^2(n-1)\tag{1.4.2}$$

(3) $X\overline{X}$ 与 S^2 独立，且 $$\frac{\sqrt{n}(\overline{X}-\mu)}{S}\sim t(n-1)\tag{1.4.3}$$

证明 (1) 根据概率论的知识，独立正态随机变量的线性组合仍然服从正态分布，而 $\overline{X}$ 是 n 个独立正态随机变量 $X_1,X_2,\cdots,X_n$ 的线性组合，故 $\overline{X}$ 服从正态分布，其均值和方差分别为

$$E(\overline{X})=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}\sum_{i=1}^{n}\mu=\mu$$

$$D(\overline{X})=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i)=\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2=\frac{\sigma^2}{n}$$

故有 $\overline{X}\sim N(\mu,\frac{\sigma^2}{n})$. 将 $\overline{X}$ 标准化得

$$\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}\sim N(0,1)$$

(2) 和(3) 下面利用柯赫伦定理 1.3.3 进行证明，令

$$Y_i=\frac{X_i-\mu}{\sigma}\tag{1.4.4}$$

则 $Y_1,Y_2,\cdots,Y_n$ 相互独立且均服从 $N(0,1)$

$$\overline{Y}=\frac{\overline{X}-\mu}{\sigma},\quad \sqrt{n}\,\overline{Y}\sim N(0,1)$$

故有

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2=\sum_{i=1}^{n}(Y_i-\overline{Y})^2\triangleq Q_1$$

$$(\sqrt{n}\,\overline{Y})^2\triangleq Q_2$$

$$Q_1+Q_2=\sum_{i=1}^{n}(Y_i-\overline{Y})^2+(\sqrt{n}\,\overline{Y})^2=\sum_{i=1}^{n}Y_i^2$$

根据线性代数的知识，Q_1,Q_2 都是 $Y_1,Y_2,\cdots,Y_n$ 的非负定二次型，且 $Q_1=0$ 的充分必要条件是 $Y_1=Y_2=\cdots=Y_n$，而 $Q_2=0$ 的充分必要条件是 $Y_1+Y_2+\cdots+Y_n=0$，因而二次型 Q_1 和 Q_2 的秩分别为 $n-1$ 和 1. 由于二个秩之和等于 n，根据柯赫伦定理，$Q_1\sim\chi^2(n-1)$，$Q_2\sim\chi^2(1)$，且 Q_1 与 Q_2 相互独立. 即

$$\frac{1}{\sigma_2}\sum_{i=1}^{n}(X_i-\overline{X})^2\sim\chi^2(n-1)\tag{1.4.5}$$