研究生数学基础课程系列教材

何灿芝 罗 汉 主编

应用统计学

Yingyong Tongjixue

Yingyong Tongjixue

湖南大学出版社

研究生数学基础课程系列教材

应 用 统 计 学

主 编 何灿芝 罗 汉
参 编 喻胜华 邓爱珍

湖南大学出版社
2004年·长沙

内 容 简 介

本书系统地阐述了应用统计的基本概念，基本理论和基本方法．内容包括抽样分布、参数估计、假设检验、回归分析、方差分析、正交试验和多元统计分析等．各章节配有典型例题，各章之后配有适量习题，且在书后附有答案．

读者只要具备大学微积分、初等概率论和线性代数知识就可由浅入深地学习本书．

本书是理工科各专业硕士研究生《应用统计学》课程的教材，也可作为理工科本科各专业相应课程的教材，同时还可以作为广大科学工作者和工程技术人员学习应用统计知识的参考书．

图书在版编目(CIP)数据

应用统计学/何灿芝，罗汉主编．—长沙：湖南大学出版社，2004.3

ISBN 7-81053-731-8

Ⅰ.应… Ⅱ.①何…②罗… Ⅲ.应用统计学—研究生—教材 Ⅳ.C8

中国版本图书馆 CIP 数据核字(2004)第 013467 号

应用统计学

Yingyong Tongjixue

何灿芝 罗汉 主编

□ 责任编辑 厉 王
□ 特约编辑 刘灿文
□ 封面设计 张 毅
□ 出版发行 湖南大学出版社
社址 长沙市岳麓山 邮码 410082
电话 0731-8821691 0731-8821315
□ 经 销 湖南省新华书店
□ 印 装 长沙银都印务有限公司

□ 开本 720×960 16 开 □ 印张 17.25 □ 字数 300 千
□ 版次 2004 年 5 月第 1 版 □ 2004 年 5 月第 1 次印刷
□ 印数 1～5 000 册
□ 书号 ISBN 7-81053-731-8/C·60
□ 定价 30.00 元

前　言

概率论和应用统计(数理统计)是数学的两个密切联系的分支学科.概率论是研究各种描述随机现象的数学模型的概率特征,着重于理论上的探讨;应用统计是研究如何有效地收集、整理和分析受到随机影响的数据,从而对所考察的实际应用问题作出统计推断,提供解决问题的各种方法.它的理论基础是概率论.

应用统计的内容十分广泛,大致可分为抽样理论和统计推断两大类.属于前者的内容有抽样技术,试验设计等;属于后者的内容有参数估计和假设检验,还有非参数估计,回归分析,方差分析,多元统计分析等.

应用统计已广泛地应用在工业、农业、管理、经济、军事、医学、工程技术以及自然科学和社会科学的各个领域,并发挥着越来越重要的作用.因此,应用统计是每一个科学工作者和工程技术人员的必修课.

考虑到目前理工科硕士研究生入学的数学基础,本书着重介绍基本概念,基本理论和基本方法,着重于应用,对于繁琐或涉及过深的数学的证明加上"*"号或者证明从略.内容叙述力求做到简明易懂,例题和习题的选择做到具有典型性,便于教学与自学.

本书是在作者的近些年来用于理工科硕士研究生教学的《应用统计》一书的基础上进一步修改、补充而成的.这次出版得到了湖南大学数学与计量经济学院、湖南大学研究生院以及湖南大学出版社的大力支持,得到了数学与计量经济学院概率统计系各位同行的关心和帮助.刘炳文先生和本书责任编辑厉亚女士提出了许多宝贵的修改意见.李笋老师绘制了本书的全部插图.在此一并表示衷心的感谢.

由于作者水平有限,不妥之处恳请读者批评指正.

作　者

2004 年 1 月于湖南大学

目　次

第一章　样本及其分布

第一节　总体和样本

一、总体和个体

数理统计中把研究对象的“全体”称为总体，又叫母体，把组成该总体的每一个“元素”称为个体.

例如要考察一批灯泡寿命，若这批灯泡有 1 万个，这就是总体，而每个灯泡都是个体. 不过我们现在关心的不是灯泡本身，而是其“寿命”这个量的大小的分布情况. 因此，在应用上总体是指研究对象的某个数量指标 X（如灯泡的寿命）的取值的全体. 显然，这里的 X 是个随机变量.

定义 1　一个随机变量 X 或其相应的分布叫做一个总体，X 的每一个可能取值叫做一个个体.

在实际问题中，X 的分布未知或部分未知，故它正是统计推断的对象.

二、抽样和样本

1. 简单抽样

当我们研究某个总体的某个特性（如灯泡的寿命）时，若将总体中每一个个体都进行试验，这在实际中一般是不可能的. 这是因为：第一，试验往往带有破坏性，如寿命测试就是破坏性试验；第二，即使试验无破坏性，但由于总体数量大，限于人力和物力也不能进行全部试验. 因此，需要采用由局部推断总体的方法，即从总体 X 中抽取部分个体，如 n 个：$X_1, X_2, \cdots, X_n$，由这 n 个个体所获得的信息来推断总体的情况.

由于 $X_1, X_2, \cdots, X_n$ 是从 X 中抽取的，故每个 $X_i (i=1,2,\cdots,n)$ 都是一个随机变量，我们称 $X_1, X_2, \cdots, X_n$ 为来自总体的样本. 从总体中抽取样本的过程称为抽样，最实用的抽样是简单抽样，它必须满足以下两个要求：

(1) 样本具有代表性：要求 $X_1, X_2, \cdots, X_n$ 的分布与总体 X 的分布相同. 这就要求总体中每个个体被抽到的机会是均等的.

(2) 样本具有独立性：要求 $X_1, X_2, \cdots, X_n$ 相互独立. 这就要求每抽出一个个体之后，总体的元素不变(或近似不变).

实际上，简单随机抽样就是独立地、重复地从总体中抽取个体进行随机试验.

由简单抽样得到的样本，称为简单样本. 今后，如无特别说明，样本都是指简单样本，于是我们可以给样本一个严格的数学定义.

2. 样本

定义 2 若随机变量 $X_1, X_2, \cdots, X_n$ 相互独立，且每个 $X_i (i=1,2,\cdots,n)$ 与总体 X 有相同的分布，则称随机向量 $(X_1, X_2, \cdots, X_n)$ 为总体 X 的一个样本，n 称为样本容量，$(X_1, X_2, \cdots, X_n)$ 的所有可能取值的集合称为样本空间. 在一次具体的抽样中所得到的数值 $(x_1, x_2, \cdots, x_n)$ 称为样本 $(X_1, X_2, \cdots, X_n)$ 的一个观察值，简称样本观察值，它是样本空间中的一个点，故又称样本点. 一般说来，不同次数的抽样，所得到的样本点是不同的.

3. 样本的联合分布

设 $(X_1, X_2, \cdots, X_n)$ 是总体 X 的一个样本，若 X 的分布函数为 $F(x)$，概率密度函数为 $f(x)$，则根据样本的定义，个体 $X_i (i=1,2,\cdots,n)$ 与总体同分布，即 X_i 具有分布函数 $F(x_i)$ 和概率密度函数 $f(x_i)$，且 $X_1, X_2, \cdots, X_n$ 是相互独立的，故它们的联合分布函数和联合概率密度函数分别为：

$$F^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F(x_i); \tag{1.1}$$

$$f^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i). \tag{1.2}$$

注 (Ⅰ) 样本具有两重性：抽样试验之前，它是随机变量，抽样试验之后，得到的是确定的观察值.

(Ⅱ) 样本是相互独立的，是与总体同分布的一组随机变量或一个随机向量.

第二节 统计量

一、统计量

在应用统计中，样本是对总体进行估计和推断的依据，而经常用到的又是样本所构成的函数，假若这些函数中不含未知参数，则称为统计量.

定义 3　设 $X_1, X_2, \cdots, X_n$ 是总体 X 的样本，$g(X_1, X_2, \cdots, X_n)$ 是定义在样本空间上的不含未知参数的连续函数，则称 $g(X_1, X_2, \cdots, X_n)$ 是一个统计量. 若 $x_1, x_2, \cdots, x_n$ 是 $X_1, X_2, \cdots, X_n$ 的一个观察值，则 $g(x_1, x_2, \cdots, x_n)$ 是 $g(X_1, X_2, \cdots, X_n)$ 的一个观察值. 显然，统计量是随机变量.

例 1　判断下列随机变量哪些是统计量？

设总体 $X \sim N(\mu, \sigma^2)$，其中 μ 已知，σ^2 未知，又 $X_1, X_2, \cdots, X_n$ 是 X 的样本，随机变量为：

(1) $Y_1 = \sum_{i=1}^{n}(X_i - \mu)^2$；(2) $Y_2 = \left(\sum_{i=1}^{n} X_i\right)/\sigma^2$.

解　Y_1 是不含未知参数的样本的连续函数，故 Y_1 是一个统计量；Y_2 中含未知参数，故不是统计量.

二、样本矩——常用的统计量

设 $X_1, X_2, \cdots, X_n$ 是总体 X 的样本，$x_1, x_2, \cdots, x_n$ 为样本观察值，则有下列定义.

定义 4　(1) 样本均值：统计量 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 称为样本均值，$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ 为其观察值.

(2) 样本方差：统计量 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ 称为样本方差，它的观察值是 $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$；$S$ 称为样本均方差.

(3) 样本 k 阶原点矩：统计量 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$　$(k = 1, 2, \cdots, n)$ 称为样本 k 阶原点矩，其观察值为 $a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$　$(k = 1, 2, \cdots, n)$.

(4) 样本 k 阶中心矩：统计量 $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$　$(k = 1, 2, \cdots, n)$ 称为样本 k 阶中心矩，它的观察值为 $b_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k$　$(k = 1, 2, \cdots, n)$.

注　(Ⅰ) 样本均值是一阶原点矩.

(Ⅱ) 样本方差本应是二阶中心矩，但我们常把二阶中心矩的修正值 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ 定义为样本方差.

例 2　设总体 X 有有限的期望和方差：$EX=\mu, DX=\sigma^2$.

(1) 求 $E\overline{X}, D\overline{X}$；(2) 求 ES^2, EB_2.

解　(1) $E\overline{X}=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}EX_i=\frac{1}{n}\sum_{i=1}^{n}\mu=\mu$,

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

(2) $$ES^2=E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right]$$

$$=\frac{1}{n-1}E\left\{\sum_{i=1}^{n}[(X_i-\mu)-(\overline{X}-\mu)]^2\right\}$$

$$=\frac{1}{n-1}E\left\{\sum_{i=1}^{n}[(X_i-\mu)^2-2(X_i-\mu)(\overline{X}-\mu)+(\overline{X}-\mu)^2]\right\}$$

$$=\frac{1}{n-1}E\left\{\sum_{i=1}^{n}(X_i-\mu)^2-2n(\overline{X}-\mu)^2+n(\overline{X}-\mu)^2\right\}$$

$$=\frac{1}{n-1}E\left\{\sum_{i=1}^{n}(X_i-\mu)^2-n(\overline{X}-\mu)^2\right\}$$

$$=\frac{1}{n-1}\left\{\sum_{i=1}^{n}DX_i-nD\overline{X}\right\}$$

$$=\frac{1}{n-1}\left\{\sum_{i=1}^{n}\sigma^2-n\frac{\sigma^2}{n}\right\}=\sigma^2.$$

因为　$B_2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2=\frac{n-1}{n}S^2$,

所以　$EB_2=\frac{n-1}{n}ES^2=\frac{n-1}{n}\sigma^2$.

可见，σ^2 是 S^2 取值的集中位置，而不是 B_2 取值的集中位置.

第三节　经验分布函数

一、经验分布函数

定义 5　设总体 X 的样本 $(X_1, X_2, \cdots, X_n)$ 的取值 $(x_1, \cdots, x_n)$，将其按大小排列：

$$x_1^* \leqslant x_2^* \leqslant \cdots \leqslant x_n^*,$$

则称函数

$$F_n(x)=\begin{cases}0, & x<x_1^*;\\ \dfrac{k}{n}, & x_k^*\leqslant x<x_{k+1}^* \quad (k=1,2,\cdots,n);\\ 1, & x\geqslant x_n^*\end{cases}$$

为总体 X 的经验分布函数(或样本函数).

该函数是一个阶梯函数,跃度为 $\frac{1}{n}$(重合点的跃度合并),见图 1-1.

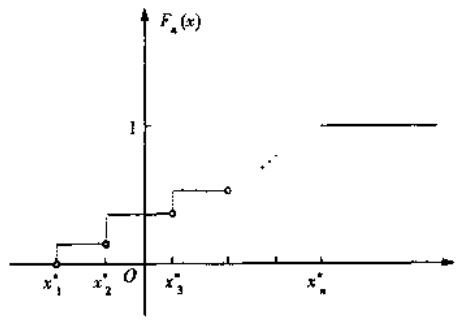

图 1-1

1. 经验分布函数的性质

(1) $0\leqslant F_n(x)\leqslant 1$;

(2) $F_n(-\infty)=0$,

$F_n(+\infty)=1$;

(3) $F_n(x)$单调不减;

(4) $F_n(x)$右连续.

2. $F_n(x)$的极限分布

$F_n(x)$是样本观察值 $x_i(i=1,2,\cdots,n)$的函数,观察值不同,$F_n(x)$也不同,故 $F_n(x)$是一个随机变量(x 为参数). 当 $x_k^*\leqslant x<x_{k+1}^*$时,则不大于 x 的观察值出现的频率是 $\frac{k}{n}$,因为 $F_n(x)$等于 n 重贝努利(Bernoulli)试验中事件 $A=\{X\leqslant x\}$所发生的频率,即

$$F_n(x)=\frac{k}{n}=\frac{x_1,\cdots,x_n\text{ 中不大于 }x\text{ 的个数}}{n},$$

因而 $nF_n(x)$表示事件 A 在 n 重贝努利试验中出现的次数,所以 $nF_n(x)$服从二项分布,即

$$nF_n(x)\sim B(n,p),$$

其中 $p=P(A)=P(X\leqslant x)=F(x)$,$F(x)$是总体 X 的分布函数.

根据贝努利大数定律:对 $\forall\varepsilon>0$,$|x|<+\infty$,有

$$\lim_{n\to\infty}\{|F_n(x)-F(x)|<\varepsilon\}=1, \tag{1.3}$$

即 $F_n(x)$依概率 1 收敛于 $F(x)$. 可见对每个固定的 x,当 n 充分大时,事件 $\{|F_n(x)-F(x)|<\varepsilon\}$是大概率事件,按照实际推断原理,在一次抽样中,此事件几乎必然会发生,故可利用一次抽样所得到的经验分布函数 $F_n(x)$值来近似 $F(x)$. 同时 $F_n(x)$是一个统计量,因此也常用它来估计 X 的理论分布函

数 $F(x)$.

由于 $F_n(x)$依赖于 x,故(1.3)式有局限性,格利文科(Гливенко)在 1933 年给出了一个更深入的结果.

定理 1 (格利文科定理) 设 $F(x)$是总体 X 的理论分布函数,$F_n(x)$是经验分布函数,则

$$P\{\lim_{n\to\infty}\operatorname{Sup}_{|x|<+\infty}|F_n(x)-F(x)|=0\}=1. \tag{1.4}$$

该定理说明,当 n 很大时,$F_n(x)$非常接近 $F(x)$,这就说明用样本可以推断总体.

二、频率直方图

前面讲了经验分布函数 $F_n(x)$可用来近似求理论分布函数 $F(x)$. 下面介绍一种关于连续型随机变量概率密度函数的近似求法——直方图法.

用频率直方图求概率密度函数的步骤:

1. 设 $f(x)$是总体 X 的概率密度函数,$X_1,\cdots,X_n$ 是样本,$x_1,\cdots,x_n$ 是样本值.

(1) 将 $x_1,\cdots,x_n$ 按大小排列:

$$x_1^*\leqslant x_2^*\leqslant\cdots\leqslant x_n^*,$$

找出 x_1^* 和 x_n^*,其中 $x_1^*=\min\{x_1,\cdots,x_n\}$,$x_n^*=\max\{x_1,\cdots,x_n\}$,得 $R=x_n^*-x_1^*$(称为样本极差).

选取 a(它略小于 x_1^*)和 b(它略大于 x_n^*),使样本值全部落入$[a,b]$区间内.

(2) 将$[a,b]$$m$ 等分,得分点

$$a=c_1\leqslant c_2\leqslant\cdots\leqslant c_{m+1}=b,$$

区间$[c_i,c_{i+1}]$的长度 $h=\dfrac{b-a}{m}$称为组距.

(3) 用唱票的办法,求出样本值落在每个小区间$[c_i,c_{i+1}]$中的个数,称为频数,记作 n_i,再求出频率:$f_i=\dfrac{n_i}{n}(i=1,2,\cdots,m)$,并作频数分布表 1-1.

(4) 列出频率、组距比例表 1-2.

(5) 作频率直方图.

在 xOy 平面上,对每个 $i(i=1,2,\cdots,m)$,以$\overline{c_ic_{i+1}}$为底,以 $y_i=\dfrac{f_i}{h}$为高,画一排竖的长方形,称为频率直方图,如图 1-2.

表 1－1　　频数分布表

分组 $c_i\sim c_{i+1}$	频数 n_i	频率 $f_i=\frac{n_i}{n}$
$c_1\sim c_2$	n_1	$\frac{n_1}{n}$
$c_2\sim c_3$	n_2	$\frac{n_2}{n}$
⋮	⋮	⋮
$c_m\sim c_{m+1}$	n_m	$\frac{n_m}{n}$
$\sum$	n	1

表 1－2　　频率与组距比例表

$c_i\sim c_{i+1}$	组中值 $\frac{c_i+c_{i+1}}{2}$	频数 n_i	频率 $f_i=\frac{n_i}{n}$	比例 $y_i=\frac{f_i}{h}$
$c_1\sim c_2$	$\frac{c_1+c_2}{2}$	n_1	$\frac{n_1}{n}$	$\frac{f_1}{h}$
$c_2\sim c_3$	$\frac{c_2+c_3}{2}$	n_2	$\frac{n_2}{n}$	$\frac{f_2}{h}$
⋮	⋮	⋮	⋮	⋮
$c_m\sim c_{m+1}$	$\frac{c_m+c_{m+1}}{2}$	n_m	$\frac{n_m}{n}$	$\frac{f_m}{h}$
$\sum$	—	n	1	—

2. 求 $f(x)$的近似 $f_n(x)$.

将每个矩形“顶边”的中点用折线联结起来，得频率多边形，再将它修饰成光滑曲线(见图 1－2)：

$$y=f_n(x),$$

则 $f_n(x)$就是总体 X 的概率密度函数 $f(x)$的一种近似，为什么呢?

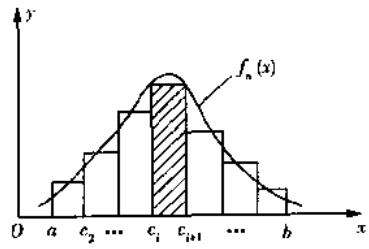

图 1－2

因为 n 个样本的抽取是独立的，由概率的统计定义知，n 个样本值中，落在$[c_i,c_{i+1}]$上的频率 f_i 近似地等于随机变量 X 落入$[c_i,c_{i+1}]$上的概率，即

$$f_i\approx P\{c_i\leqslant X\leqslant c_{i+1}\}=\int_{c_i}^{c_{i+1}}f(x)\mathrm{d}x.$$

又　　$$f_i = y_i h \approx \int_{c_i}^{c_{i+1}} f_n(x)\mathrm{d}x,$$

故　　$$\int_{c_i}^{c_{i+1}} f(x)\mathrm{d}x \approx \int_{c_i}^{c_{i+1}} f_n(x)\mathrm{d}x.$$

因为　　$f(x) \geqslant 0, f_n(x) \geqslant 0,$

于是　　$f(x) \approx f_n(x).$

注　（Ⅰ）作频率直方图时，将$[a,b]$分成$m+1$个小区间，m的大小随样本容量n而定，n大时，m也大，如$n=100$时，m可取12.

（Ⅱ）分点c_i的选取一般比样本值多一位小数.

（Ⅲ）组距不一定要相等.

例3　从某维尼纶厂生产的一批维尼纶中抽取100件进行纤维度检查，得到的数据如下所示.

1.36	1.49	1.43	1.41	1.37	1.40	1.32	1.42	1.47	1.39
1.41	1.36	1.40	1.34	1.42	1.42	1.45	1.35	1.42	1.39
1.44	1.42	1.39	1.42	1.42	1.30	1.34	1.42	1.37	1.36
1.37	1.34	1.37	1.37	1.44	1.45	1.32	1.48	1.40	1.45
1.39	1.46	1.39	1.53	1.36	1.48	1.40	1.39	1.38	1.40
1.36	1.45	1.50	1.43	1.38	1.43	1.41	1.48	1.39	1.45
1.37	1.37	1.39	1.45	1.31	1.41	1.44	1.44	1.42	1.47
1.35	1.36	1.39	1.40	1.38	1.35	1.42	1.43	1.42	1.42
1.42	1.40	1.41	1.37	1.46	1.36	1.37	1.27*	1.37	1.38
1.42	1.34	1.43	1.42	1.41	1.41	1.44	1.48	1.55*	1.37

试作出频率直方图，并求$f_n(x)$.

解　(1) 分组

$x_1^* = 1.27, x_n^* = 1.55$. 取$a=1.265, b=1.565, m=10$. 组距$h=\dfrac{b-a}{m}=0.03$.

(2) 列表(见表1-3).

(3) 作频率直方图如图1-3.

可见总体近似服从正态分布.

表1-3　　**频数、频率表**

分　组	频数	频率	$y_i=\dfrac{f_i}{h}$
1.265～1.295	1	0.01	0.33
1.295～1.325	4	0.04	1.33

续表

分　组	频数	频率	$y_i=\frac{f_i}{h}$
1.325～1.355	7	0.07	2.33
1.355～1.385	22	0.22	7.33
1.385～1.415	23	0.23	7.67
1.415～1.445	25	0.25	8.33
1.445～1.475	10	0.10	3.33
1.475～1.505	6	0.06	2.00
1.505～1.535	1	0.01	0.33
1.535～1.565	1	0.01	0.33
$\sum$	100	1.00	—

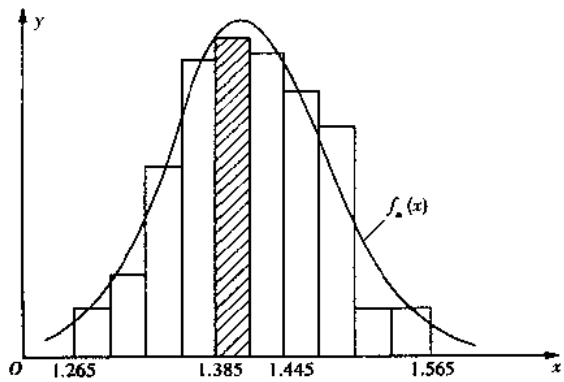

图 1－3

第四节　抽样分布

统计量的分布叫做抽样分布. 由于统计量是样本的函数，而样本与总体同分布，且相互独立，故统计量的分布由样本的联合分布惟一确定，即由总体分

布惟一确定.这一节我们将讨论正态总体下的一些重要统计量的精确分布.

一、正态分布

1. Z 统计量的分布.

定理 2　设总体 $X\sim N(\mu,\sigma^2)$,$X_1,X_2,\cdots,X_n$ 是 X 的样本,则

(1) $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\sim N\left(\mu,\frac{\sigma^2}{n}\right)$;

(2) $Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$.

事实上:

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

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

故(1)成立,将(1)标准化即得(2)成立.

2. 标准正态分布的分布函数和概率密度函数.

(1) 概率密度函数为

$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$,$|x|<+\infty$(见图 1-4).

(2) 分布函数为

$$\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt.$$

$\Phi(x)$的函数值有附表 2 可查.

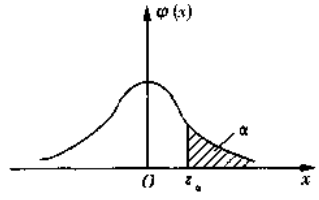

图 1-4

3. 标准正态分布的上 α 点

定义 6　若 $P\{Z>z_\alpha\}=\alpha$,则 z_α 称为 $N(0,1)$分布的上 100α 百分位点,简称上 α 点.

若　　$P\{|Z|>z_{\frac{\alpha}{2}}\}=\alpha$,

则称 $z_{\frac{\alpha}{2}}$ 为 $N(0,1)$的双侧 100α 百分位点,简称双侧 α 点(见图 1-5).

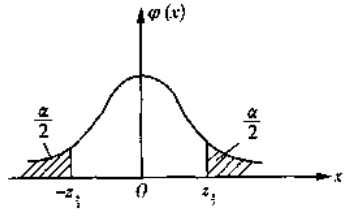

图 1-5

上 α 点 z_α 的求法:

因为　　$P\{Z>z_\alpha\}=\alpha$,

又　　$P\{Z>z_\alpha\}=1-P\{Z\leqslant z_\alpha\}=1-\Phi(z_\alpha)$,

所以　　$\Phi(z_\alpha)=1-\alpha$,反查附表 2 即得 z_α.

例 4　给定 $\alpha=0.05,0.025$，求 z_α.

解　(1) $\alpha=0.05$，

$$\Phi(z_\alpha)=1-\alpha=0.95,$$

所以

$$z_\alpha=\frac{1}{2}(1.64+1.65)=1.645.$$

(2) $\alpha=0.025,\Phi(z_\alpha)=1-\alpha=0.975$，

所以　　$z_\alpha=1.96$.

二、χ^2 分布

1. χ^2 统计量的分布.

定义 7　设总体 $X\sim N(0,1),X_1,X_2,\cdots,X_n$ 为样本.若

$$\chi^2=X_1^2+X_2^2+\cdots+X_n^2=\sum_{i=1}^{n}X_i^2,$$

则称统计量 χ^2 服从自由度为 n 的 χ^2 分布，记作 $\chi^2\sim\chi^2(n)$.

2. 自由度.

若对变量 $Y_1,Y_2,\cdots,Y_n$(随机的或非随机的)，存在一组不全为 0 的常数 $c_1,c_2,\cdots,c_n$，使得

$$c_1Y_1+c_2Y_2+\cdots+c_nY_n=0,$$

则称 $Y_1,Y_2,\cdots,Y_n$ 之间存在着一个线性的约束条件. 如果存在 k 个约束条件：

$$c_{i1}Y_1+\cdots+c_{in}Y_n=0,i=1,2,\cdots,k,$$

其系数矩阵$(c_{ij})_{kn}$的秩为 k，并且对于任何 $m(m\geqslant k)$个约束条件

$$d_{i1}Y_1+\cdots+d_{in}Y_n=0,i=1,2,\cdots,m,$$

系数矩阵$(d_{ij})_{mn}$的秩总不大于 k，则称 $Y_1,\cdots,Y_n$ 之间存在着 k 个独立的线性约束条件，由线性代数知，$Y_1,\cdots,Y_n$ 中有 $n-k$ 个独立变量(自由变量).

如果在 $\sum_{i=1}^{n}Y_i^2$ 中，$Y_1,\cdots,Y_n$ 之间存在着 k 个独立的线性约束条件，即有 $n-k$ 个自由变量，则称 $\sum_{i=1}^{n}Y_i^2$ 的自由度为 $n-k$. 由于 $X_1,X_2,\cdots,X_n$ 独立，故 $\chi^2=\sum_{i=1}^{n}X_i^2$ 的自由度为 n.

3. χ^2 的概率密度.

定理 3　设总体 $X\sim N(0,1),X_1,X_2,\cdots,X_n$ 为样本，则 $\chi^2=\sum_{i=1}^{n}X_i^2$ 的概