第二章 统计调查

本章讲解统计调查的各种方法，重点理解统计调查的目的就是以样本提供的信息推断总体；理解随机误差与系统误差的概念，掌握两者的差异；了解各阶样本矩的计算；会查各种抽样分布的分位数。

2.1统计调查的意义和种类

一、统计调查的意义,了解

二、统计调查的要求：及时性、准确性、完整性

三、统计调查的种类

- 统计调查按调查范围不同，分为全面调查和非全面调查。
- 统计调查按调查的时间间隔长短，分为经常性调查和—次性调查。
- 统计调查按组织形式，分为统计报表和专门调查。
- 统计调查按搜集资料的方法不同，有：1、直接观察法 2、报告法 3、通讯法4、采访法5、问卷调查法

2.2非随机调查

一、统计报表

二、普查

三、重点调查

四、典型调查

一、统计报表

统计报表是按照统一规定的表式、指标解释、报送时间和报送程序，以企事业单位原始记录为基础、自下而上地搜集统计资料的一种调查方式。

统计报表根据不同的依据可以分为不同的类型。

虽然统计报表具有其他调查方法不可替代的好处，然而，在社会主义市场经济条件下存在一些不可克服的弊端。

许多统计报表已逐渐被抽样调查所替代。

二、普查

普查即普遍调查，是一种对调查对象全部个体逐一地、全面地进行调查，以达到了解调查对象总体情况的调查方式。

普查的规则：

1. 普查时间必须统一
2. 正确选择调查的具体时间
3. 调查项目应少而精
4. 普查应按一定周期进行

三、重点调查

重点调查是为了解社会经济现象的基本情况而专门组织的一种非全面调查。它是从调查对象的总体中只选择一部分重点单位（个体）的调查。

重点单位：是指这些单位在全部总体中虽然数目不多，但就调查的指标值来说却在总体中占很大比重。

四、典型调查

典型调查是根据调查目的和任务，在对调查对象初步分析的基础上，有意识地从中选取少数几个具有代表性的典型单位进行深入细致的调查研究，借以认识事物的本质及其规律性的一种专门组织的非全面调查。

注：按本教程关于统计学的定义，重点调查与典型调查不属于统计调查。

2.3随机抽样调查

一、抽样调查概念

它是按照随机原则从调查对象(即总体)中抽取部分个体进行调查，用调查所得指标数值对调查对象相应指标数值做出具有一定可靠性的估计和判断的一种统计调查方法。

注：调查（试验）有时是破坏性的、高成本的，有时也无必要逐一调查。

二、抽样调查的特点

抽样调查遵循随机原则。抽样调查是用部分个体的指标数值去推断和估计总体的指标数值，因此，抽样调查也叫抽样推断。

抽样调查必然会产生抽样误差。

注：抽样误差（随机误差）与系统误差的区别。

三、抽样调查类型及理论依据

由样品是否考虑顺序，抽中后是否放回，相互搭配随机抽样就有四种。

表2.1　不同抽样方法的可能样本个数

抽 样 方 式	可 能 样 本 个 数
考虑顺序的不重复抽样	$A_N^n = N!/(N-n)!$
考虑顺序的重复抽样	$B_N^n = N^n$
不考虑顺序的不重复抽样	$C_N^n = N!/[n!\times(N-n)!]$
不考虑顺序的重复抽样	$D_N^n = C_{N+n-1}^n$

抽样调查的理论依据是大数定律和中心极限定理。

例2.1 设总体$X \sim \begin{pmatrix} 3 & 4 & 5 & 6 & 7 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{pmatrix}$，$X_1$，$X_2$ 是有放回取自 X 的有序样本，求 $\bar{X}$ 的分布，并与总体 X 相比较。

解： 本例总体容量N=5，样本容量$n=2$，X_1与X_2相互独立与 X 有相同的分布，$\bar{X}=\frac{1}{2}(X_1+X_2)$ 也是随机变量。

总体均值 $EX=5$，方差$DX=2_i$ X_1，X_2的所有可能取值的组合（25种，即样本空间有25个元素，见表2.2）。

$$\bar{X} \sim \begin{pmatrix} 3 & 3.5 & 4 & 4.5 & 5 & 5.5 & 6 & 6.5 & 7 \\ 0.04 & 0.08 & 0.12 & 0.16 & 0.2 & 0.16 & 0.12 & 0.08 & 0.04 \end{pmatrix}$$

通过这个例子的学习，读者应该理解。

样本均值$\bar{x}$ 是随机变量，相对于样本观测值$(3,4)$，样本均值的观测值 $\bar{x}=\frac{1}{2}(3+4)=3.5$ 。

样本均值 $\bar{x}$ 的数学期望 $E\bar{X}=5$ 与总体均值（期望）相等，而 $\bar{x}$ 的方差 $D\bar{X}=\frac{DX}{n}=\frac{2}{2}=1$ ，即与总体方差相比，缩小了n 倍。

2.4抽样分布

一、经验分布函数

对于样本$X_1,\cdots,X_n$ 的观测值 $x_1,\cdots,x_n$ 定义新的随机变量X_*且其分布为 $P(X_*=x_i)=\frac{1}{n}, i=1,\cdots,n$

X_* 的分布函数$F_*(x)=P(X_*\le x)$ 就是

$$F_*(x)=\begin{cases}0, & x<x_{(1)}\\ \frac{k}{n}, & x_{(k)}\le x<x_{(k+1)}\\ 1, & x\ge x<x_{(n)}\end{cases} \qquad (2.3)$$

二、样本（中心、绝对）矩

1. 原点矩

设X为随机变量，对任意正整数m，称$E(X^m)$为随机变量X的m阶原点矩。当m=1时，一阶原点矩为$E(X)=\mu$，即随机变量X的数学期望。相应的样本m阶原点矩为：

$$\frac{1}{n}\sum X_i^m \qquad (2.4)$$

（2.5）

2. 中心矩

设X为随机变量，称$E[X-E(X)]^m$为以$E(X)$为中心的m阶中心矩，当m=2时，二阶中心矩为 $E[X-E(X)]^2$即为随机变量X的方差σ^2。

相应的样本m阶中心矩为,

$$\frac{1}{n}\sum(x_i-\bar{x})^m$$

在计算偏度和峰度时，要用到3阶中心矩。

三、样本均值、样本方差的分布

定义2.1 $X_1,\cdots,X_n$ 为来自正态总体 $N(0,1)$ 的样本，则称随机变量 $\chi^2=\sum X_i^2$ 服从自由度为 n 的 χ^2 分布，记作 $\chi^2\sim\chi^2(n)$ ，其密度曲线见图2.2，密度函数为，

$$f(x)=\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-n/2} \qquad x\geq 0$$

χ^2 分布的均值与方差与自由度有关，即

$$E\chi^2(n)=n,\ D\chi^2(n)=2n$$

四、分布的特征

1．分布是一种连续型的对称分布。它是关于 $t=0$ 对称的，且形状同标准正态分布类似，随机变量 t 取值从 $-\infty$ 到 $+\infty$。

2．一般说来，在小样本时，t分布的方差大于1，而当n无限增大时，t分布的方差趋于1，且t分布逐渐逼近标准正态分布。

3．t分布决定于自由度，只有知道其自由度，才能在t分布表上查得相应的概率。

4．t分布的均值和方差分别为：

$$E\left[t\left(n\right)\right]=0 \qquad D\left[t\left(n\right)\right]=\frac{n}{n-2}$$

五、F分布

定义2.3若随机变量 $X \sim \chi^2(n)$ 与随机变量 $Y \sim \chi^2(m)$ 独立，则称 $\xi = \frac{X/n}{Y/m}$ 服从第一自由度为n，第二自由度为m的F分布，记作 $\xi \sim F(n,m)$ ，其密度曲线见图2.4，

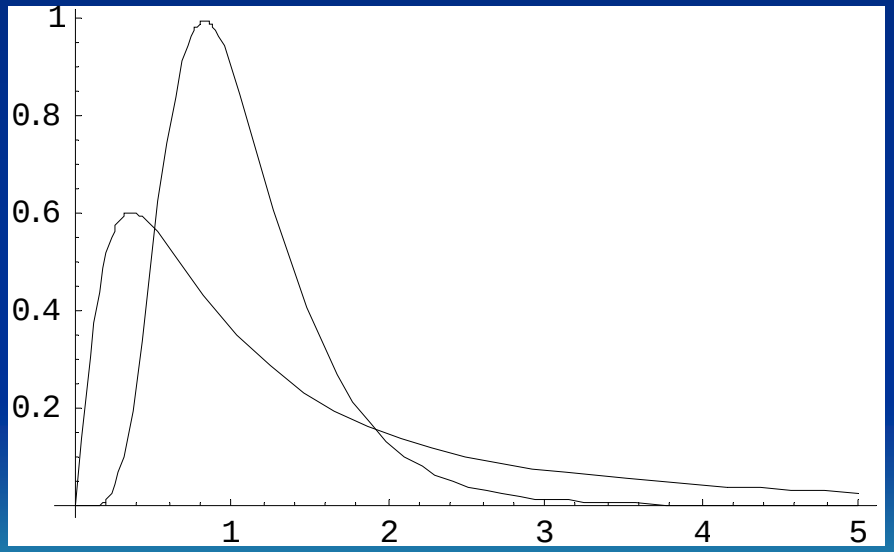


图**2.4** $F(5,3)$, $F(20,24)$ 的密度曲线

六、重要抽样分布定理

设$X_1,\cdots,X_n$为来自正态总体 $X \sim N(\mu,\sigma^2)$ 的样本，样本方差为 S^2，修正样本方差 $\tilde{S}^2$，则

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

(2) $\frac{nS^2}{\sigma^2}=\frac{\sum(X_i-\overline{X})^2}{\sigma^2} \sim \chi^2(n-1)$

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

若另有正态总体 $Y \sim N(\mu_Y,\sigma^2)$ (容量为m)的样本，则

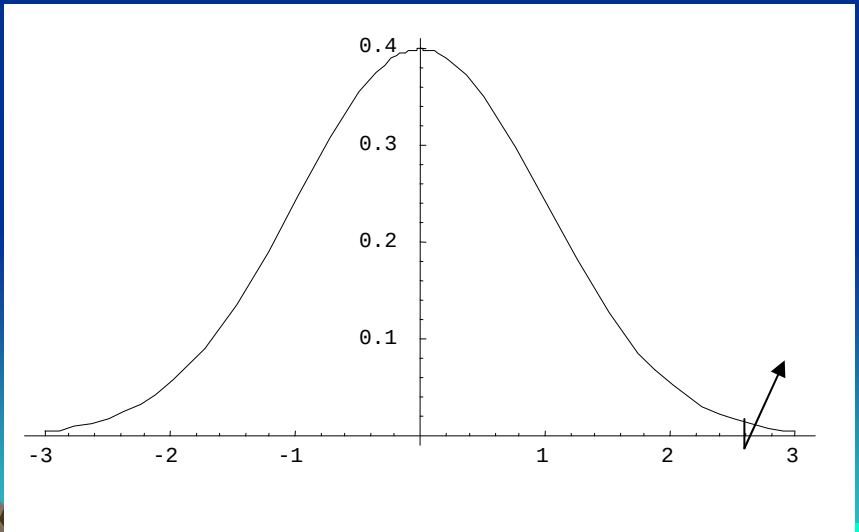
$$\frac{\tilde{S}_X^2}{\tilde{S}_Y^2} \sim F(n-1,m-1)$$

七、分位数

1. 正态分布分位数

设有随机变量ξ，对于满足$P\{\xi \geq v_a\} = a$，的v_a称为随机变量ξ的上侧a分位数，类似地可以定义的下侧a分位数μ_a，即若$P\{\xi \leq \mu_a\} = a$则μ_a称为随机变量ξ的下侧a分位数。显然v_a和μ_a之间满足$\mu_a = v_{1-a}$，以后一概用v_{1-a}

- 表示下侧a分位数。对于标准正态Z（$\sim N(0,1)$）的上侧a分位数用z_α表示，即$P\{Z \geq z_a\} = a$
- （如图2.5）。



2. χ^2- 分布分位数

χ^2-分布不是对称分布，使用的分布表必须既将上侧a分位数$x_\alpha^2(n)$，也将下侧分位数$x_{1-\alpha}^2(n)$同时列出,见附表3，其中 $P(x^2 \geq x_\alpha^2(n)) = \alpha$

并且分位数示意图见图2.6。

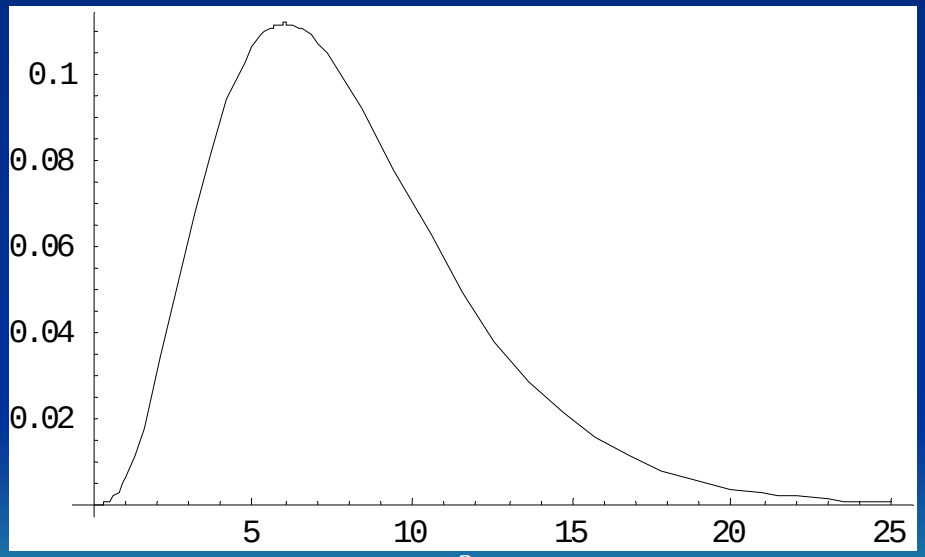


图2.6 χ^2 分布分位数

3.t－分布分位数

t 分布和正态分布相同（见图2.7），是对称的，因此

$$P(t > t_\alpha(n)) = P(t < -t_\alpha(n)) = \alpha$$

并且，列分位数表时只给出上侧a分位数表（见附表4），

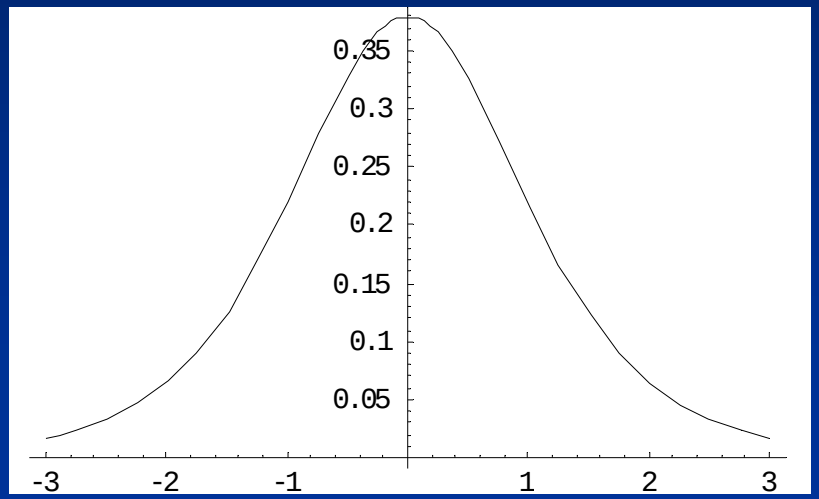


图2.7 t-分布分位数

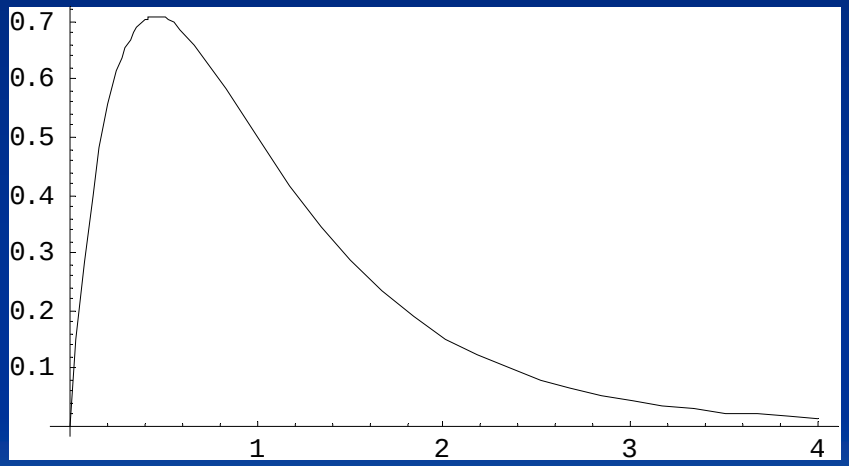
当n>45时用相应正态分布分位数代替，即

$$t_\alpha(n) = z_\alpha \ , n > 45$$

4．F分布分位数

对于*F*—分布，读者可以根据其定义验证关系式 $F_{1-a}(n,m)=[F_a(m,n)]^{-1}$

附表5只给出上侧*a*分位数，下侧*a*分位数利用上式换算得到，示意图见图2.8。



例如查附表5可得

$F_{0.05}(3,12)=3.49$ ，$F_{0.05}(10,10)=2.32$ ， $F_{0.05}(12,3)=8.74$

2.5抽抽样调查的组织方式

一、简单随机抽样

抽样框，随机数表

二、等距抽样

1. 随机起点等距抽祥
2. 半距起点等距抽样
3. 对称起点等距抽样

三、分层抽样

四、整群抽样

五、多阶段抽样