第六章　样本及抽样分布

前面五章我们讲述了概率论的基本内容，随后的四章将讲述数理统计. 数理统计是具有广泛应用的一个数学分支，它以概率论为理论基础，根据试验或观察得到的数据，来研究随机现象，对研究对象的客观规律性作出种种合理的估计和判断.

数理统计的内容包括：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、特点作出推断. 后者就是我们所说的统计推断问题. 本书只讲述统计推断的基本内容.

在概率论中，我们所研究的随机变量，它的分布都是假设已知的，在这一前提下去研究它的性质、特点和规律性，例如求出它的数字特征，讨论随机变量函数的分布，介绍常用的各种分布等. 在数理统计中，我们研究的随机变量，它的分布是未知的，或者是不完全知道的，人们是通过对所研究的随机变量进行重复独立的观察，得到许多观察值，对这些数据进行分析，从而对所研究的随机变量的分布作出种种推断的.

本章我们介绍总体、随机样本及统计量等基本概念，并着重介绍几个常用统计量及抽样分布.

§1　随机样本

我们知道，随机试验的结果很多是可以用数来表示的，另有一些试验的结果虽是定性的，但总可以将它数量化. 例如，检验某个学校学生的血型这一试验，其可能结果有 O 型、A 型、B 型、AB 型 4 种，是定性的. 如果分别以 1，2，3，4 依次记这 4 种血型，那么试验的结果就能用数来表示了.

在数理统计中，我们往往研究有关对象的某一项数量指标（例如研究某种型号灯泡的寿命这一数量指标）. 为此，考虑与这一数量指标相联系的随机试验，对这一数量指标进行试验或观察. 我们将试验的全部可能的观察值称为**总体**，这些值不一定都不相同，数目上也不一定是有限的，每一个可能观察值称为**个体**. 总体中所包含的个体的个数称为总体的**容量**. 容量为有限的称为**有限总体**，容量为无限的称为**无限总体**.

例如在考察某大学一年级男生的身高这一试验中，若一年级男生共 2 000 人，每个男生的身高是一个可能观察值，所形成的总体中共含 2 000 个可

能观察值，是一个有限总体. 又如考察某一湖泊中某种鱼的含汞量，所得总体也是有限总体. 观察并记录某一地点每天(包括以往、现在和将来)的最高气温，或者测量一湖泊任一地点的深度，所得总体是无限总体. 有些有限总体，它的容量很大，我们可以认为它是一个无限总体. 例如，考察全国正在使用的某种型号灯泡的寿命所形成的总体，由于可能观察值的个数很多，就可以认为是无限总体.

总体中的每一个个体是随机试验的一个观察值，因此它是某一随机变量 X 的值，这样，一个总体对应于一个随机变量 X. 我们对总体的研究就是对一个随机变量 X 的研究，X 的分布函数和数字特征就称为总体的分布函数和数字特征. 今后将不区分总体与相应的随机变量，笼统称为总体 X.

例如，我们检验自生产线出来的零件是次品还是正品，以 0 表示产品为正品，以 1 表示产品为次品. 设出现次品的概率为 p (常数)，那么总体是由一些“1”和一些“0”所组成，这一总体对应于一个具有参数为 p 的(0－1)分布：

$$P\{X=x\}=p^x(1-p)^{1-x},x=0,1$$

的随机变量. 我们就将它说成是(0－1)分布总体. 意指总体中的观察值是(0－1)分布随机变量的值. 又如上述灯泡寿命这一总体是指数分布总体，意指总体中的观察值是指数分布随机变量的值.

在实际中，总体的分布一般是未知的，或只知道它具有某种形式而其中包含着未知参数. 在数理统计中，人们都是通过从总体中抽取一部分个体，根据获得的数据来对总体分布作出推断的. 被抽出的部分个体叫做总体的一个样本.

所谓从总体抽取一个个体，就是对总体 X 进行一次观察并记录其结果. 我们在相同的条件下对总体 X 进行 n 次重复的、独立的观察. 将 n 次观察结果按试验的次序记为 $X_1,X_2,\cdots,X_n$. 由于 $X_1,X_2,\cdots,X_n$ 是对随机变量 X 观察的结果，且各次观察是在相同的条件下独立进行的，所以有理由认为 $X_1,X_2,\cdots,X_n$ 是相互独立的，且都是与 X 具有相同分布的随机变量. 这样得到的 $X_1,X_2,\cdots,X_n$ 称为来自总体 X 的一个简单随机样本，n 称为这个样本的容量. 以后如无特别说明，所提到的样本都是指简单随机样本.

当 n 次观察一经完成，我们就得到一组实数 $x_1,x_2,\cdots,x_n$，它们依次是随机变量 $X_1,X_2,\cdots,X_n$ 的观察值，称为样本值.

对于有限总体，采用放回抽样就能得到简单随机样本，但放回抽样使用起来不方便，当个体的总数 N 比要得到的样本的容量 n 大得多时，在实际中可将不放回抽样近似地当作放回抽样来处理.

至于无限总体，因抽取一个个体不影响它的分布，所以总是用不放回抽样. 例如，在生产过程中，每隔一定时间抽取一个个体，抽取 n 个就得到一个简单随机样本，实验室中的记录，水文、气象等观察资料都是样本. 试制新产品得到的样品的质量指标，也常被认为是样本.

综合上述，我们给出以下的定义.

定义 设 X 是具有分布函数 F 的随机变量，若 $X_1, X_2, \cdots, X_n$ 是具有同一分布函数 F 的、相互独立的随机变量，则称 $X_1, X_2, \cdots, X_n$ 为从分布函数 F（或总体 F、或总体 X）得到的**容量为 n 的简单随机样本**，简称**样本**，它们的观察值 $x_1, x_2, \cdots, x_n$ 称为**样本值**，又称为 X 的 n 个**独立的观察值**.

也可以将样本看成是一个随机向量，写成 $(X_1, X_2, \cdots, X_n)$，此时样本值相应地写成 $(x_1, x_2, \cdots, x_n)$. 若 $(x_1, x_2, \cdots, x_n)$ 与 $(y_1, y_2, \cdots, y_n)$ 都是相应于样本 $(X_1, X_2, \cdots, X_n)$ 的样本值，一般来说它们是不相同的.

由定义得：若 $X_1, X_2, \cdots, X_n$ 为 F 的一个样本，则 $X_1, X_2, \cdots, X_n$ 相互独立，且它们的分布函数都是 F，所以 $(X_1, X_2, \cdots, X_n)$ 的分布函数为

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

又若 X 具有概率密度 f，则 $(X_1, X_2, \cdots, X_n)$ 的概率密度为

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

§2 直方图和箱线图

为了研究总体分布的性质，人们通过试验得到许多观察值，一般来说这些数据是杂乱无章的. 为了利用它们进行统计分析，将这些数据加以整理，还常借助于表格或图形对它们加以描述. 本节将通过例子对连续型随机变量 X 引入“频率直方图.”接着介绍数据的“箱线图”. 它们使人们对总体 X 的分布有一个粗略的了解.

（一）直方图

例 1 下面列出了 84 个伊特拉斯坎（Etruscan）人男子的头颅的最大宽度（mm），现在来画这些数据的“频率直方图”.

141	148	132	138	154	142	150	146	155	158
150	140	147	148	144	150	149	145	149	158
143	141	144	144	126	140	144	142	141	140
145	135	147	146	141	136	140	146	142	137
148	154	137	139	143	140	131	143	141	149
148	135	148	152	143	144	141	143	147	146
150	132	142	142	143	153	149	146	149	138
142	149	142	137	134	144	146	147	140	142
140	137	152	145						

解 这些数据杂乱无章，先要将它们进行整理. 这些数据的最小值、最大值分别为 126、158，即所有数据落在区间[126，158]上，现取区间[124.5，159.5]，它能覆盖区间[126，158]. 将区间[124.5，159，5]等分为 7 个小区间①，小区间的长度记为 Δ，$\Delta=(159.5-124.5)/7=5$. Δ 称为组距. 小区间的端点称为组限. 数出落在每个小区间内的数据的频数 f_i，算出频率 f_i/n ($n=84, i=1,2,\cdots,7$) 如下表：

组　　限	频　数 f_i	频率 f_i/n	累积频率
124.5～129.5	1	0.011 9	0.011 9
129.5～134.5	4	0.047 6	0.059 5
134.5～139.5	10	0.119 1	0.178 6
139.5～144.5	33	0.392 9	0.571 5
144.5～149.5	24	0.285 7	0.857 2
149.5～154.5	9	0.107 1	0.952 4
154.5～159.5	3	0.035 7	1

现在自左至右依次在各个小区间上作以 $\frac{f_i}{n}\Big/\Delta$ 为高的小矩形. 如图 6－1 所示这样的图形叫**频率直方图**. 显然这种小矩形的面积就等于数据落在该小区间的频率 f_i/n. 由于当 n 很大时，频率接近于概率，因而一般来说，每个小区间上的小矩形面积接近于概率密度曲线之下该小区间之上的曲边梯形的面积. 于是，一般来说，直方图的外廓曲线接近于总体 X 的概率密度曲线. 从本例的直方图看(图 6－1)，它有一个峰，中间高，两头低，比较对称. 看起来样本很像来自某一正态总体 X (在第八章中将进一步讨论). 从直方图上还可以估计 X 落在某一区间的概率，例如从图上看到有 51.2%的人最大头颅宽度落在区间(134.5，144.5)之内，最大头颅宽度小于 129.5 的仅占 1.1%等等.

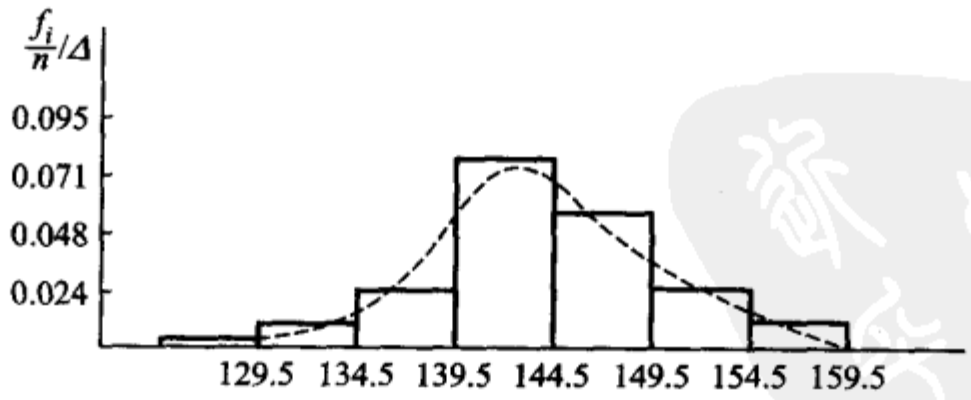

图 6－1

① 作直方图时，先取一个区间，其下限比最小的数据稍小，其上限比最大的数据稍大，然后将这一区间分为 k 个小区间，通常当 n 较大时 k 取 10～20，当 $n<50$ 时则 k 取 5～6. 若 k 取得过大，则会出现某些小区间内频数为零的情况(一般应设法避免). 分点通常取比数据精度高一位，以免数据落在分点上.

(二) 箱线图

先介绍样本分位数.

定义 设有容量为 n 的样本观察值 $x_1, x_2, \cdots, x_n$,样本 p 分位数($0<p<1$)记为 x_p,它具有以下的性质:(1)至少有 np 个观察值小于或等于 x_p;(2)至少有 $n(1-p)$ 个观察值大于或等于 x_p.

样本 p 分位数可按以下法则求得.将 $x_1, x_2, \cdots, x_n$ 按自小到大的次序排列成 $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$.

1° 若 np 不是整数,则只有一个数据满足定义中的两点要求,这一数据位于大于 np 的最小整数处,即为位于 $[np]+1$ 处的数.例如,$n=12, p=0.9, np=10.8, n(1-p)=1.2$,则 x_p 的位置应满足至少有 10.8 个数据$\leqslant x_p$ (x_p 应位于第 11 或大于第 11 处);且至少有 1.2 个数据$\geqslant x_p$ (x_p 应位于第 11 或小于第 11 处),故 x_p 应位于第 11 处.

2° 若 np 是整数.例如在 $n=20, p=0.95$ 时,x_p 的位置应满足至少有 19 个数据$\leqslant x_p$ (x_p 应位于第 19 或大于第 19 处)且至少有 1 个数据$\geqslant x_p$ (x_p 应位于第 20 或小于第 20 处),故第 19 或第 20 的数据均符合要求,就取这两个数的平均值作为 x_p.综上,

$$x_p=\begin{cases} x_{([np]+1)}, & \text{当 } np \text{ 不是整数,} \\ \dfrac{1}{2}\left[x_{(np)}+x_{(np+1)}\right], & \text{当 } np \text{ 是整数.} \end{cases}$$

特别,当 $p=0.5$ 时,0.5 分位数 $x_{0.5}$ 也记为 Q_2 或 M,称为**样本中位数**,即有

$$x_{0.5}=\begin{cases} x_{([\frac{n}{2}]+1)}, & \text{当 } n \text{ 是奇数,} \\ \dfrac{1}{2}\left[x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}\right], & \text{当 } n \text{ 是偶数.} \end{cases}$$

易知,当 n 是奇数时中位数 $x_{0.5}$ 就是 $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$ 这一数组最中间的一个数;而当 n 是偶数时中位数 $x_{0.5}$ 就是 $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$ 这一数组中最中间两个数的平均值.

0.25 分位数 $x_{0.25}$ 称为**第一四分位数**,又记为 Q_1;0.75 分位数 $x_{0.75}$ 称为**第三四分位数**,又记为 Q_3. $x_{0.25}, x_{0.5}, x_{0.75}$ 在统计中是很有用的.

例 2 设有一组容量为 18 的样本值如下(已经过排序)

122 126 133 140 145 145 149 150 157
162 166 175 177 177 183 188 199 212

求样本分位数:$x_{0.2}, x_{0.25}, x_{0.5}$.

解 (1) 因为 $np=18\times0.2=3.6$,$x_{0.2}$ 位于第 $[3.6]+1=4$ 处,即有 $x_{0.2}=x_{(4)}=140$.

(2) 因为 $np=18\times0.25=4.5$，$x_{0.25}$ 位于第 $[4.5]+1=5$ 处，即有 $x_{0.25}=145$.

(3) 因为 $np=18\times0.5=9$，$x_{0.5}$ 是这组数中间两个数的平均值，即有

$$x_{0.5}=\frac{1}{2}(157+162)=159.5.$$

下面介绍箱线图.

数据集的箱线图是由箱子和直线组成的图形，它是基于以下 5 个数的图形概括：最小值 Min，第一四分位数 Q_1，中位数 M，第三四分位数 Q_3 和最大值 Max. 它的作法如下：

(1) 画一水平数轴，在轴上标上 Min，Q_1，M，Q_3，Max. 在数轴上方画一个上、下侧平行于数轴的矩形箱子，箱子的左右两侧分别位于 Q_1，Q_3 的上方. 在 M 点的上方画一条垂直线段. 线段位于箱子内部.

(2) 自箱子左侧引一条水平线直至最小值 Min；在同一水平高度自箱子右侧引一条水平线直至最大值. 这样就将箱线图作好了，如图 6－2 所示. 箱线图也可以沿垂直数轴来作. 自箱线图可以形象地看出数据集的以下重要性质.

① 中心位置：中位数所在的位置就是数据集的中心.

② 散布程度：全部数据都落在[Min，Max]之内，在区间[Min，Q_1]，[Q_1，M]，[M，Q_3]，[Q_3，Max]的数据个数各占 1/4. 区间较短时，表示落在该区间的点较集中，反之较为分散.

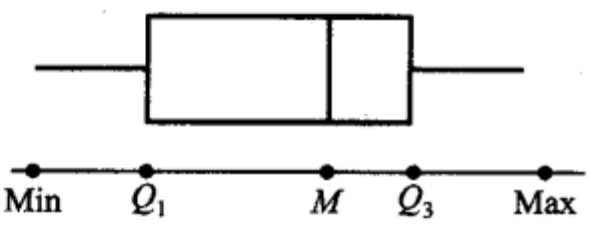

图 6－2

(3) 关于对称性：若中位数位于箱子的中间位置. 则数据分布较为对称. 又若 Min 离 M 的距离较 Max 离 M 的距离大，则表示数据分布向左倾斜，反之表示数据向右倾斜，且能看出分布尾部的长短.

例 3 以下是 8 个病人的血压(收缩压，mmHg)数据(已经过排序)，试作出箱线图.

102　110　117　118　122　123　132　150

解 因 $np=8\times0.25=2$，故 $Q_1=\frac{1}{2}(110+117)=113.5$.

因 $np=8\times0.5=4$，故 $x_{0.5}=Q_2=\frac{1}{2}(118+122)=120$.

因 $np=8\times0.75=6$，故 $x_{0.75}=Q_3=\frac{1}{2}(123+132)=127.5$.

Min＝102，Max＝150，作出箱线图如图 6－3 所示.

例 4 下面分别给出了 25 个男子和 25 个女子的肺活量(以升计. 数据已经

过排序）

女子组	2.7	2.8	2.9	3.1	3.1	3.1	3.2	3.4	3.4
	3.4	3.4	3.4	3.5	3.5	3.5	3.6	3.7	3.7
	3.7	3.8	3.8	4.0	4.1	4.2	4.2		
男子组	4.1	4.1	4.3	4.3	4.5	4.6	4.7	4.8	4.8
	5.1	5.3	5.3	5.3	5.4	5.4	5.5	5.6	5.7
	5.8	5.8	6.0	6.1	6.3	6.7	6.7		

试分别画出这两组数据的箱线图.

解　女子组 Min＝2.7，Max＝4.2，M＝3.5，

因 $np=25\times0.25=6.25$，$Q_1=3.2$.

因 $np=25\times0.75=18.75$，$Q_3=3.7$.

男子组 Min＝4.1，Max＝6.7，M＝5.3，

因 $np=25\times0.25=6.25$，$Q_1=4.7$.

因 $np=25\times0.75=18.75$，　$Q_3=5.8$.

作出箱线图如图 6－4 所示.　□

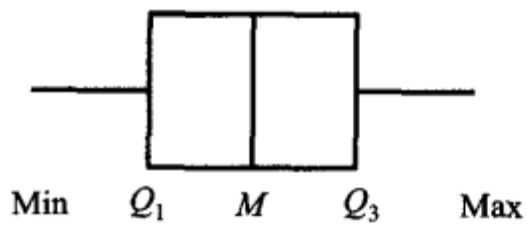

图 6－3

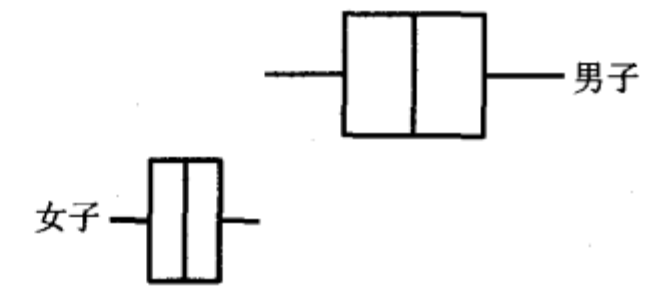

图 6－4

箱线图特别适用于比较两个或两个以上数据集的性质，为此，我们将几个数据集的箱线图画在同一个数轴上. 例如在例 3 中可以明显地看到男子的肺活量要比女子大，男子的肺活量较女子的肺活量为分散.

在数据集中某一个观察值不寻常地大于或小于该数集中的其他数据，称为**疑似异常值**. 疑似异常值的存在，会对随后的计算结果产生不适当的影响. 检查疑似异常值并加以适当的处理是十分重要的. 箱线图只要稍加修改，就能用来检测数据集是否存在疑似异常值.

第一四分位数 Q_1 与第三四分位数 Q_3 之间的距离：$Q_3-Q_1\xrightarrow{\text{记为}}IQR$，称为**四分位数间距**. 若数据小于 $Q_1-1.5IQR$ 或大于 $Q_3+1.5IQR$，就认为它是疑似异常值. 我们将上述箱线图的作法(1)、(2)、(3)作如下的改变：

(1′)同(1)

(2′)计算 $IQR=Q_3-Q_1$，若一个数据小于 $Q_1-1.5IQR$ 或大于 Q_3+

1.5IQR，则认为它是一个疑似异常值. 画出疑似异常值，并以 * 表示.

(3′)自箱子左侧引一水平线段直至数据集中除去疑似异常值后的最小值，又自箱子右侧引一水平线直至数据集中除去拟似异常值后的最大值. 按(1′)、(2′)、(3′)作出的图形称为**修正箱线图**.

例 5 下面给出了某医院 21 个病人的住院时间(以天计)，试画出修正箱线图(数据已经过排序).

1 2 3 3 4 4 5 6 6 7 7 9 9
10 12 12 13 15 18 23 55

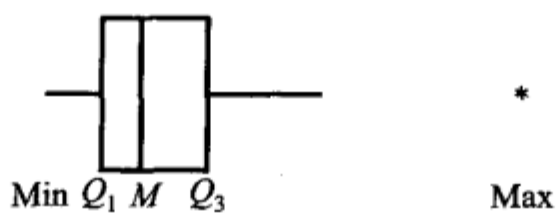

图 6−5

解 Min=1，Max=55，M=7，

因 21×0.25=5.25，得 Q_1=4，

又 21×0.75=15.75，得 Q_3=12，

故 $IQR=Q_3-Q_1=8$，

$Q_3+1.5IQR=12+1.5\times8=24$，$Q_1-1.5IQR=4-12=-8$.

观察值 55>24，故 55 是疑似异常值，且仅此一个疑似异常值. 作出修正箱线图如图 6−5 所示. 可见数据分布不对称，而向右倾斜，在中位数的右边较为分散. □

数据集中，疑似异常值的产生源于(1)数据的测量、记录或输入计算机时的错误；(2)数据来自不同的总体；(3)数据是正确的，但它只体现小概率事件. 当检测出疑似异常值时，人们需对疑似异常值出现的原因加以分析. 如果是由于测量或记录的错误，或某些其他明显的原因造成的，将这些疑似异常值从数据集中丢弃就可以了. 然而当出现的原因无法解释时要作出丢弃或保留这些值的决策无疑是困难的，此时我们在对数据集作分析时尽量选用稳健的方法，使得疑似异常值对我们的结论的影响较小. 例如我们采用中位数来描述数据集的中心趋势，而不使用数据集的平均值，因为后者受疑似异常值的影响较大.

§3 抽样分布

样本是进行统计推断的依据. 在应用时，往往不是直接使用样本本身，而是针对不同的问题构造样本的适当函数，利用这些样本的函数进行统计推断.

定义 设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本，$g(X_1, X_2, \cdots, X_n)$ 是 $X_1, X_2, \cdots, X_n$ 的函数，若 g 中不含未知参数，则称 $g(X_1, X_2, \cdots, X_n)$ 是**一统计量**.

因为 $X_1, X_2, \cdots, X_n$ 都是随机变量，而统计量 $g(X_1, X_2, \cdots, X_n)$ 是随机变量的函数，因此统计量是一个随机变量. 设 $x_1, x_2, \cdots, x_n$ 是相应于样本 $X_1, X_2, \cdots, X_n$ 的样本值，则称 $g(x_1, x_2, \cdots, x_n)$ 是 $g(X_1, X_2, \cdots, X_n)$ 的观察值.

下面列出几个常用的统计量. 设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本，$x_1, x_2, \cdots, x_n$ 是这一样本的观察值. 定义

样本平均值

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

样本方差

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

样本标准差

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

样本 k 阶(原点)矩

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

样本 k 阶中心矩

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k,\ k = 2, 3, \cdots.$$

它们的观察值分别为

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

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right);$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2};$$

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k,\quad k = 1, 2, \cdots;$$

$$b_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k,\quad k = 2, 3, \cdots.$$

这些观察值仍分别称为样本均值、样本方差、样本标准差、样本 k 阶(原点)矩以及样本 k 阶中心矩.

我们指出，若总体 X 的 k 阶矩 $E(X^k) \xrightarrow{\text{记成}} \mu_k$ 存在，则当 $n \to \infty$ 时，$A_k \xrightarrow{P}$

$\mu_k, k=1,2,\cdots$. 这是因为 $X_1, X_2, \cdots, X_n$ 独立且与 X 同分布，所以 $X_1^k, X_2^k, \cdots, X_n^k$ 独立且与 X^k 同分布. 故有

$$E(X_1^k)=E(X_2^k)=\cdots=E(X_n^k)=\mu_k.$$

从而由第五章的辛钦大数定理知

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k \xrightarrow{P} \mu_k, \quad k=1,2,\cdots.$$

进而由第五章中关于依概率收敛的序列的性质知道

$$g(A_1, A_2, \cdots, A_k) \xrightarrow{P} g(\mu_1, \mu_2, \cdots, \mu_k),$$

其中 g 为连续函数. 这就是下一章所要介绍的矩估计法的理论根据.

经验分布函数 我们还可以作出与总体分布函数 $F(x)$ 相应的统计量——经验分布函数. 它的作法如下：设 $X_1, X_2, \cdots, X_n$ 是总体 F 的一个样本，用 $S(x)$，$-\infty<x<\infty$ 表示 $X_1, X_2, \cdots, X_n$ 中不大于 x 的随机变量的个数. 定义经验分布函数 $F_n(x)$ 为

$$F_n(x)=\frac{1}{n}S(x), \quad -\infty<x<\infty.$$

对于一个样本值，那么经验分布函数 $F_n(x)$ 的观察值是很容易得到的（$F_n(x)$ 的观察值仍以 $F_n(x)$ 表示）. 例如

(1) 设总体 F 具有一个样本值 1,2,3，则经验分布函数 $F_3(x)$ 的观察值为

$$F_3(x)=\begin{cases}0, & \text{若 } x<1,\\ \frac{1}{3}, & \text{若 } 1\leqslant x<2,\\ \frac{2}{3}, & \text{若 } 2\leqslant x<3,\\ 1, & \text{若 } x\geqslant 3.\end{cases}$$

(2) 设总体 F 具有一个样本值 1,1,2，则经验分布函数 $F_3(x)$ 的观察值为

$$F_3(x)=\begin{cases}0, & \text{若 } x<1,\\ \frac{2}{3}, & \text{若 } 1\leqslant x<2,\\ 1, & \text{若 } x\geqslant 2.\end{cases}$$

一般，设 $x_1, x_2, \cdots, x_n$ 是总体 F 的一个容量为 n 的样本值. 先将 $x_1, x_2, \cdots, x_n$ 按自小到大的次序排列，并重新编号. 设为

$$x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}.$$

则经验分布函数 $F_n(x)$ 的观察值为

$$F_n(x)=\begin{cases}0, & \text{若 } x<x_{(1)},\\ \frac{k}{n}, & \text{若 } x_{(k)}\leqslant x<x_{(k+1)},\\ 1, & \text{若 } x\geqslant x_{(n)}.\end{cases}$$

对于经验分布函数 $F_n(x)$，格里汶科(Glivenko)在 1933 年证明了以下的结果：对于任一实数 x，当 $n\to\infty$ 时 $F_n(x)$ 以概率 1 一致收敛于分布函数 $F(x)$，即

$$P\{\lim_{n\to\infty}\sup_{-\infty<x<\infty}|F_n(x)-F(x)|=0\}=1.$$

因此，对于任一实数 x 当 n 充分大时，经验分布函数的任一个观察值 $F_n(x)$ 与总体分布函数 $F(x)$ 只有微小的差别，从而在实际上可当作 $F(x)$ 来使用①.

统计量的分布称为**抽样分布**. 在使用统计量进行统计推断时常需知道它的分布. 当总体的分布函数已知时，抽样分布是确定的，然而要求出统计量的精确分布，一般来说是困难的. 本节介绍来自正态总体的几个常用统计量的分布.

(一) χ^2 分布

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $N(0,1)$ 的样本，则称统计量

$$\chi^2=X_1^2+X_2^2+\cdots+X_n^2 \tag{3.1}$$

服从自由度为 n 的 χ^2 **分布**，记为 $\chi^2\sim\chi^2(n)$.

此处，自由度是指(3.1)式右端包含的独立变量的个数.

$\chi^2(n)$ 分布的概率密度为

$$f(y)=\begin{cases}\dfrac{1}{2^{n/2}\Gamma(n/2)}y^{n/2-1}e^{-y/2}, & y>0,\\ 0, & \text{其他.}\end{cases} \tag{3.2}$$

$f(y)$ 的图形如图 6－6 所示.

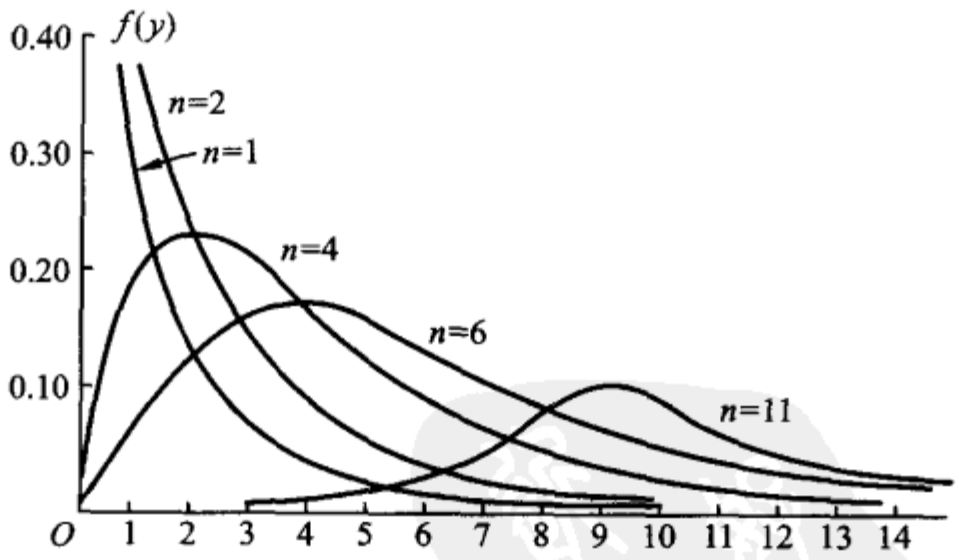

图 6－6

现在来推求(3.2)式.

---

① 对于任意固定的 x，$-\infty<x<\infty$，$S(x)\sim b(n,F(x))$，从而可知对于固定的 x，$E[F_n(x)]=E[\dfrac{S(x)}{n}]=\dfrac{1}{n}E[S(x)]=\dfrac{1}{n}[nF(x)]=F(x)$.

首先由第二章§5例3及第三章§5例3知 $\chi^2(1)$ 分布即为 $\Gamma\left(\frac{1}{2},2\right)$ 分布.

现 $X_i\sim N(0,1)$,由定义 $X_i^2\sim\chi^2(1)$,即 $X_i^2\sim\Gamma\left(\frac{1}{2},2\right)$,$i=1,2,\cdots,n$. 再由 X_1,$X_2,\cdots,X_n$ 的独立性知 $X_1^2,X_2^2,\cdots,X_n^2$ 相互独立,从而由 Γ 分布的可加性(见第三章§5例3)知

$$\chi^2=\sum_{i=1}^{n}X_i^2\sim\Gamma\left(\frac{n}{2},2\right),\tag{3.3}$$

即得 χ^2 的概率密度如(3.2)式所示. □

根据 Γ 分布的可加性易得 χ^2 分布的可加性如下:

χ^2 分布的可加性 设 $\chi_1^2\sim\chi^2(n_1)$,$\chi_2^2\sim\chi^2(n_2)$,并且 χ_1^2,χ_2^2 相互独立,则有

$$\chi_1^2+\chi_2^2\sim\chi^2(n_1+n_2).\tag{3.4}$$

χ^2 分布的数学期望和方差 若 $\chi^2\sim\chi^2(n)$,则有

$$E(\chi^2)=n,\quad D(\chi^2)=2n.\tag{3.5}$$

事实上,因 $X_i\sim N(0,1)$,故

$E(X_i^2)=D(X_i)=1$,

$D(X_i^2)=E(X_i^4)-[E(X_i^2)]^2=3-1=2,i=1,2,\cdots,n.$

于是

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

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

χ^2 分布的分位点 对于给定的正数 α,$0<\alpha<1$,称满足条件

$$P\{\chi^2>\chi_\alpha^2(n)\}=\int_{\chi_\alpha^2(n)}^{\infty}f(y)\mathrm{d}y=\alpha\tag{3.6}$$

的点 $\chi_\alpha^2(n)$ 为 $\chi^2(n)$ 分布的上 α 分位点,如图6—7所示. 对于不同的 α,n,上 α 分位点的值已制成表格,可以查用(参见附表5). 例如对于 $\alpha=0.1$,$n=25$,查得 $\chi_{0.1}^2(25)=34.382$. 但该表只详列到 $n=40$ 为止,费希尔(R. A. Fisher)曾证明,当 n 充分大时,近似地有

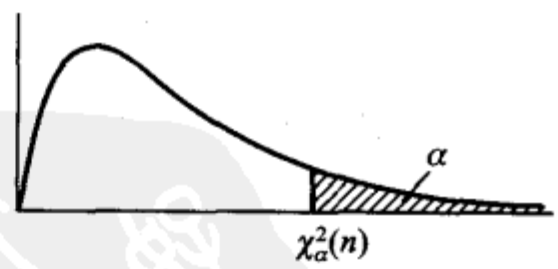

图 6—7

$$\chi_\alpha^2(n)\approx\frac{1}{2}(z_\alpha+\sqrt{2n-1})^2,\tag{3.7}$$

其中 z_α 是标准正态分布的上 α 分位点. 利用(3.7)式可以求得当 $n>40$ 时 $\chi^2(n)$ 分布的上 α 分位点的近似值.

例如,由(3.7)式可得 $\chi_{0.05}^2(50)\approx\frac{1}{2}(1.645+\sqrt{99})^2=67.221$(由更详细的表得 $\chi_{0.05}^2(50)=67.505$).

(二) t 分布

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

$$t=\frac{X}{\sqrt{Y/n}} \tag{3.8}$$

服从自由度为 n 的 t **分布**. 记为 $t\sim t(n)$.

t 分布又称**学生氏**(Student)**分布**. $t(n)$分布的概率密度函数为

$$h(t)=\frac{\Gamma[(n+1)/2]}{\sqrt{\pi n}\,\Gamma(n/2)}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2},\quad -\infty<t<\infty \tag{3.9}$$

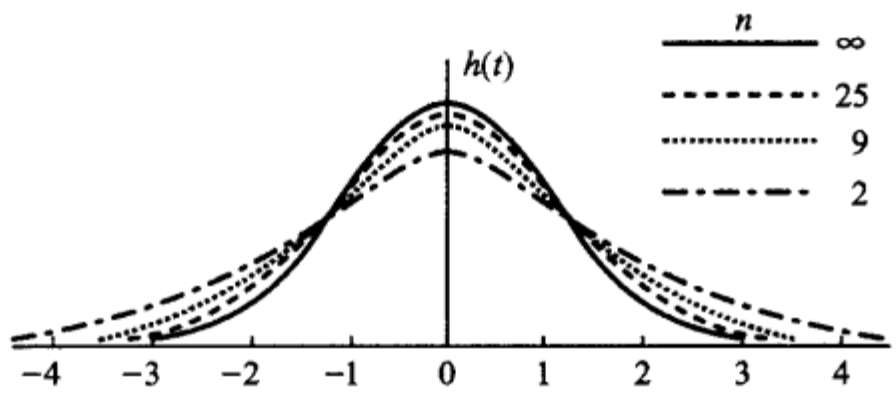

图 6—8

(证略). 图 6—8 中画出了 $h(t)$的图形. $h(t)$的图形关于 $t=0$ 对称，当 n 充分大时其图形类似于标准正态变量概率密度的图形. 事实上，利用 Γ 函数的性质可得

$$\lim_{n\to\infty}h(t)=\frac{1}{\sqrt{2\pi}}e^{-t^2/2}, \tag{3.10}$$

故当 n 足够大时 t 分布近似于 $N(0,1)$分布. 但对于较小的 n，t 分布与 $N(0,1)$分布相差较大(见附表 2 与附表 4).

t 分布的分位点　对于给定的 α，$0<\alpha<1$，称满足条件

$$P\{t>t_\alpha(n)\}=\int_{t_\alpha(n)}^{\infty}h(t)\mathrm{d}t=\alpha \tag{3.11}$$

的点 $t_\alpha(n)$为 $t(n)$分布的上 α 分位点(如图 6—9).

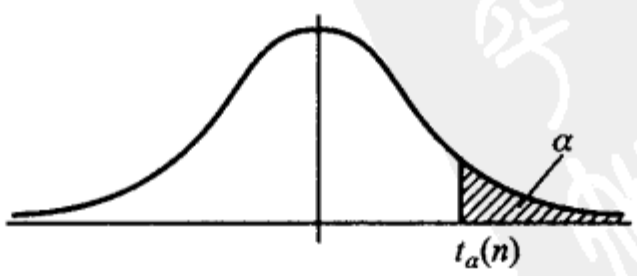

图 6—9

由 t 分布上 α 分位点的定义及 $h(t)$ 图形的对称性知

$$t_{1-\alpha}(n)=-t_{\alpha}(n). \tag{3.12}$$

t 分布的上 α 分位点可自附表 4 查得. 在 $n>45$ 时，对于常用的 α 的值，就用正态近似

$$t_{\alpha}(n)\approx z_{\alpha}. \tag{3.13}$$

(三) F 分布

设 $U\sim\chi^2(n_1)$，$V\sim\chi^2(n_2)$，且 U,V 相互独立，则称随机变量

$$F=\frac{U/n_1}{V/n_2} \tag{3.14}$$

服从自由度为 (n_1,n_2) 的 F **分布**，记为 $F\sim F(n_1,n_2)$.

$F(n_1,n_2)$ 分布的概率密度为

$$\psi(y)=\begin{cases}\dfrac{\Gamma[(n_1+n_2)/2](n_1/n_2)^{n_1/2}y^{(n_1/2)-1}}{\Gamma(n_1/2)\Gamma(n_2/2)[1+(n_1y/n_2)]^{(n_1+n_2)/2}}, & y>0,\\ 0, & \text{其他.}\end{cases} \tag{3.15}$$

(证略). 图 6—10 中画出了 $\psi(y)$ 的图形.

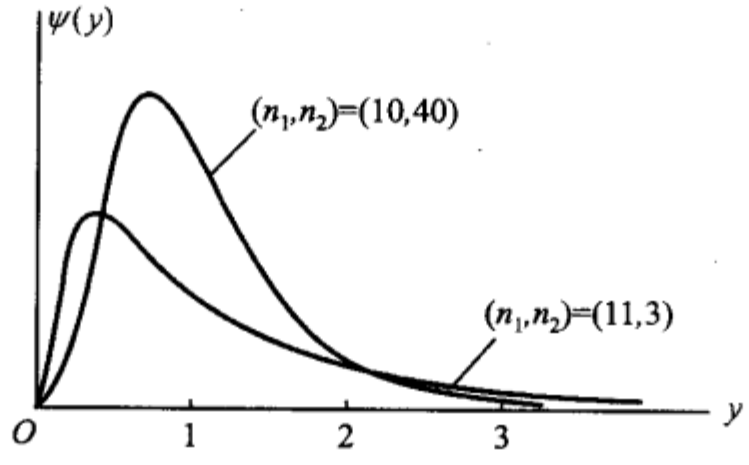

图 6—10

由定义可知，若 $F\sim F(n_1,n_2)$，则

$$\frac{1}{F}\sim F(n_2,n_1). \tag{3.16}$$

F 分布的分位点 对于给定的 α，$0<\alpha<1$，称满足条件

$$P\{F>F_{\alpha}(n_1,n_2)\}=\int_{F_{\alpha}(n_1,n_2)}^{\infty}\psi(y)\mathrm{d}y=\alpha \tag{3.17}$$

的点 $F_{\alpha}(n_1,n_2)$ 为 $F(n_1,n_2)$ 分布的上 α 分位点(图 6—11). F 分布的上 α 分位点有表格可查(见附表 6).

F 分布的上 α 分位点有如下的重要性质①：

$$F_{1-\alpha}(n_1,n_2)=\frac{1}{F_\alpha(n_2,n_1)}. \quad (3.18)$$

(3.18)式常用来求 F 分布表中未列出的常用的上 α 分位点. 例如，

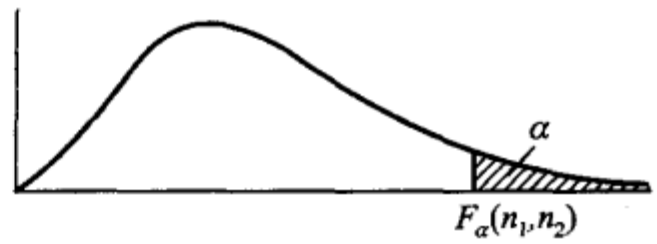

图 6－11

$$F_{0.95}(12,9)=\frac{1}{F_{0.05}(9,12)}=\frac{1}{2.80}=0.357.$$

(四) 正态总体的样本均值与样本方差的分布

设总体 X（不管服从什么分布，只要均值和方差存在）的均值为 μ，方差为 σ^2，$X_1,X_2,\cdots,X_n$ 是来自 X 的一个样本，$\overline{X}$，S^2 分别是样本均值和样本方差，则有

$$E(\overline{X})=\mu, \quad D(\overline{X})=\sigma^2/n. \quad (3.19)$$

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

$$=\frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2+\mu^2)-n(\sigma^2/n+\mu^2)\right]=\sigma^2,$$

即
$$E(S^2)=\sigma^2. \quad (3.20)$$

进而，设 $X\sim N(\mu,\sigma^2)$，由第四章 §2 的(2.8)式知 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 也服从正态分布，于是得到以下的定理：

定理一 设 $X_1,X_2,\cdots,X_n$ 是来自正态总体 $N(\mu,\sigma^2)$ 的样本，$\overline{X}$ 是样本均值，则有

$$\overline{X}\sim N(\mu,\sigma^2/n).$$

对于正态总体 $N(\mu,\sigma^2)$ 的样本均值 $\overline{X}$ 和样本方差 S^2，有以下两个重要定

---

① (3.18)式的证明如下：若 $F\sim F(n_1,n_2)$，按定义

$$1-\alpha=P\{F>F_{1-\alpha}(n_1,n_2)\}=P\left\{\frac{1}{F}<\frac{1}{F_{1-\alpha}(n_1,n_2)}\right\}$$

$$=1-P\left\{\frac{1}{F}\geqslant\frac{1}{F_{1-\alpha}(n_1,n_2)}\right\}=1-P\left\{\frac{1}{F}>\frac{1}{F_{1-\alpha}(n_1,n_2)}\right\},$$

于是
$$P\left\{\frac{1}{F}>\frac{1}{F_{1-\alpha}(n_1,n_2)}\right\}=\alpha. \quad (1)$$

再由 $\frac{1}{F}\sim F(n_2,n_1)$ 知
$$P\left\{\frac{1}{F}>F_\alpha(n_2,n_1)\right\}=\alpha. \quad (2)$$

比较(1)，(2)两式得

$$\frac{1}{F_{1-\alpha}(n_1,n_2)}=F_\alpha(n_2,n_1)，即\ F_{1-\alpha}(n_1,n_2)=\frac{1}{F_\alpha(n_2,n_1)}.$$

理.

定理二 设 $X_1, X_2, \cdots, X_n$ 是来自总体 $N(\mu, \sigma^2)$ 的样本，$\overline{X}, S^2$ 分别是样本均值和样本方差，则有

$$1^\circ \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1); \tag{3.21}$$

2° $\overline{X}$ 与 S^2 相互独立.

定理的证明见本章末附录.

定理三 设 $X_1, X_2, \cdots, X_n$ 是来自总体 $N(\mu, \sigma^2)$ 的样本，$\overline{X}, S^2$ 分别是样本均值和样本方差，则有

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1). \tag{3.22}$$

证 由定理一、定理二

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

且两者独立. 由 t 分布的定义知

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} \sim t(n-1).$$

化简上式左边，即得(3.22)式. □

对于两个正态总体的样本均值和样本方差有以下的定理.

定理四 设 $X_1, X_2, \cdots, X_{n_1}$ 与 $Y_1, Y_2, \cdots, Y_{n_2}$ 分别是来自正态总体 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 的样本，且这两个样本相互独立①. 设 $\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$，$\overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$ 分别是这两个样本的样本均值；$S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\overline{X})^2$，$S_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2$ 分别是这两个样本的样本方差，则有

$$1^\circ \quad \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1);$$

2° 当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 时，

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2),$$

其中 $$S_w^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}, \quad S_w = \sqrt{S_w^2}.$$

---

① 是指随机向量 $(X_1, X_2, \cdots, X_{n_1})$ 与 $(Y_1, Y_2, \cdots, Y_{n_2})$ 相互独立.

证 1° 由定理二

$$\frac{(n_1-1)S_1^2}{\sigma_1^2}\sim\chi^2(n_1-1),\frac{(n_2-1)S_2^2}{\sigma_2^2}\sim\chi^2(n_2-1).$$

由假设 S_1^2,S_2^2 相互独立,则由 F 分布的定义知

$$\frac{(n_1-1)S_1^2}{(n_1-1)\sigma_1^2}\Big/\frac{(n_2-1)S_2^2}{(n_2-1)\sigma_2^2}\sim F(n_1-1,n_2-1),$$

即

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}\sim F(n_1-1,n_2-1).$$

2° 易知 $\overline{X}-\overline{Y}\sim N\left(\mu_1-\mu_2,\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}\right)$,即有

$$U=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim N(0,1).$$

又由给定条件知

$$\frac{(n_1-1)S_1^2}{\sigma^2}\sim\chi^2(n_1-1),\quad\frac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_2-1),$$

且它们相互独立,故由 χ^2 分布的可加性知

$$V=\frac{(n_1-1)S_1^2}{\sigma^2}+\frac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_1+n_2-2).$$

由本章附录 2°知 U 与 V 相互独立.从而按 t 分布的定义知

$$\frac{U}{\sqrt{V/(n_1+n_2-2)}}=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2).$$ □

本节所介绍的几个分布以及四个定理,在下面各章中都起着重要的作用.应注意,它们都是在总体为正态这一基本假定下得到的.

小结

在数理统计中往往研究有关对象的某一项数量指标.对这一数量指标进行试验或观察,将试验的全部可能的观察值称为总体,每个观察值称为个体.总体中的每一个个体是某一随机变量 X 的值,因此一个总体对应一个随机变量 X.我们将不区分总体与相应的随机变量 X,笼统称为总体 X.随机变量 X 服从什么分布,就称总体服从什么分布.在实际中遇到的总体往往是有限总体,它对应一个离散型随机变量.当总体中包含的个体的个数很大时,在理论上可以认为它是一个无限总体.我们说某种型号的灯泡寿命总体服从指数分布,是指无限总体而言的.又如我们说某一年龄段的男性儿童的身高服从正态分布,也是指无限总体而言的.无限总体是人们对具体事物的抽象.无限总体的分布的形式较为简明,便于在数学上进行处理,使用方便.

在相同的条件下,对总体 X 进行 n 次重复的、独立的观察,得到 n 个结果 $X_1,X_2,\cdots,X_n$,

称随机变量 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的简单随机样本，它具有两条性质：

1° $X_1, X_2, \cdots, X_n$ 都与总体具有相同的分布；

2° $X_1, X_2, \cdots, X_n$ 相互独立．

我们就是利用来自样本的信息推断总体，得到有关总体分布的种种结论的．

样本 $X_1, X_2, \cdots, X_n$ 的函数 $g(X_1, X_2, \cdots, X_n)$，若不包含未知参数，则称为统计量．统计量是一个随机变量，它是完全由样本所确定的．统计量是进行统计推断的工具．样本均值

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

和样本方差

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2$$

是两个最重要的统计量．统计量的分布称为抽样分布．下面是三个来自正态分布的抽样分布：

$$\chi^2 \text{ 分布}，\quad t \text{ 分布}，\quad F \text{ 分布}．$$

这三个分布称为统计学的三大分布，它们在数理统计中有着广泛的应用．对于这三个分布，要求读者掌握它们的定义和密度函数图形的轮廓，还会使用分位点表写出分位点．

关于样本均值 $\overline{X}$、样本方差 S^2，有以下的结果．

1. 设 $X_1, X_2, \cdots, X_n$ 是来自总体 X（不管服从什么分布，只要它的均值和方差存在）的样本，且有 $E(X)=\mu, D(X)=\sigma^2$，则有

$$E(\overline{X})=\mu,\quad D(\overline{X})=\sigma^2/n.$$

2. 设总体 $X\sim N(\mu,\sigma^2)$，$X_1, X_2, \cdots, X_n$ 是来自 X 的样本，则有

1° $\overline{X}\sim N(\mu,\sigma^2/n)$；

2° $\dfrac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$；

3° $\overline{X}$ 与 S^2 相互独立；

4° $\dfrac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$.

3. 对于两个正态总体 $X\sim N(\mu_1,\sigma_1^2)$，$Y\sim N(\mu_2,\sigma_2^2)$，有定理四的重要结果．

■ 重要术语及主题

总体　简单随机样本　统计量

χ^2 分布、t 分布、F 分布的定义及它们的密度函数图形轮廓

上 α 分位点　$F_{1-\alpha}(n_1,n_2)=\dfrac{1}{F_\alpha(n_2,n_1)}$

小结中关于样本均值、样本方差的重要结果

附录

1° 定理二的证明

令 $Z_i=\dfrac{X_i-\mu}{\sigma}$，$i=1,2,\cdots,n$，则由定理二的假设知，$Z_1, Z_2, \cdots, Z_n$ 相互独立，且都服从 $N(0,1)$ 分布，而

$$\overline{Z}=\frac{1}{n}\sum_{i=1}^{n}Z_i=\frac{\overline{X}-\mu}{\sigma};$$

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

$$=\sum_{i=1}^{n}(Z_i-\overline{Z})^2=\sum_{i=1}^{n}Z_i^2-n\overline{Z}^2.$$

取一 n 阶正交矩阵 $\boldsymbol{A}=(a_{ij})$，其中第一行的元素均为 $1/\sqrt{n}$. 作正交变换

$$\boldsymbol{Y}=\boldsymbol{AZ},$$

其中

$$\boldsymbol{Y}=\begin{pmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{pmatrix},\quad \boldsymbol{Z}=\begin{pmatrix}Z_1\\Z_2\\\vdots\\Z_n\end{pmatrix}.$$

由于 $Y_i=\sum_{j=1}^{n}a_{ij}Z_j$，$i=1,2,\cdots,n$. 故 $Y_1,Y_2,\cdots,Y_n$ 仍为正态变量. 由 $Z_i\sim N(0,1)$，$i=1,2,\cdots,n$ 知

$$E(Y_i)=E\left(\sum_{j=1}^{n}a_{ij}Z_j\right)=\sum_{j=1}^{n}a_{ij}E(Z_j)=0.$$

又由 $\mathrm{Cov}(Z_i,Z_j)=\delta_{ij}$ ($\delta_{ij}=0$，当 $i\neq j$；$\delta_{ij}=1$，当 $i=j$)，$i,j=1,2,\cdots,n$，知

$$\mathrm{Cov}(Y_i,Y_k)=\mathrm{Cov}\left(\sum_{j=1}^{n}a_{ij}Z_j,\sum_{l=1}^{n}a_{kl}Z_l\right)$$

$$=\sum_{j=1}^{n}\sum_{l=1}^{n}a_{ij}a_{kl}\mathrm{Cov}(Z_j,Z_l)=\sum_{j=1}^{n}a_{ij}a_{kj}=\delta_{ik}$$

(由正交矩阵的性质)，故 $Y_1,Y_2,\cdots,Y_n$ 两两不相关. 又由于 n 维随机变量 $(Y_1,Y_2,\cdots,Y_n)$ 是由 n 维正态随机变量 $(X_1,X_2,\cdots,X_n)$ 经由线性变换而得到的，因此，$(Y_1,Y_2,\cdots,Y_n)$ 也是 n 维正态随机变量(参见第 4 章 § 4). 于是由 $Y_1,Y_2,\cdots,Y_n$ 两两不相关可推得 $Y_1,Y_2,\cdots,Y_n$ 相互独立(参见第 4 章 § 4)，且有 $Y_i\sim N(0,1)$，$i=1,2,\cdots,n$. 而

$$Y_1=\sum_{j=1}^{n}a_{1j}Z_j=\sum_{j=1}^{n}\frac{1}{\sqrt{n}}Z_j=\sqrt{n}\,\overline{Z};$$

$$\sum_{i=1}^{n}Y_i^2=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}=(\boldsymbol{AZ})^{\mathrm{T}}(\boldsymbol{AZ})=\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})\boldsymbol{Z}=\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{I}\boldsymbol{Z}=\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}=\sum_{i=1}^{n}Z_i^2,$$

于是

$$\frac{(n-1)S^2}{\sigma^2}=\sum_{i=1}^{n}Z_i^2-n\overline{Z}^2=\sum_{i=1}^{n}Y_i^2-Y_1^2=\sum_{i=2}^{n}Y_i^2.$$

由于 $Y_2,\cdots,Y_n$ 相互独立，且 $Y_i\sim N(0,1)$，$i=2,3,\cdots,n$，知 $\sum_{i=2}^{n}Y_i^2\sim\chi^2(n-1)$. 从而证得

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

再者，$\overline{X}=\sigma\overline{Z}+\mu=\frac{\sigma Y_1}{\sqrt{n}}+\mu$ 仅依赖于 Y_1，而 $S^2=\frac{\sigma^2}{n-1}\sum_{i=2}^{n}Y_i^2$ 仅依赖于 $Y_2,Y_3,\cdots,Y_n$. 再由 $Y_1,Y_2,\cdots,Y_n$ 的独立性，推知 $\overline{X}$ 与 S^2 相互独立.

2° 定理二的推广

定理二中 $\overline{X}$ 与 S^2 相互独立这一结论，还能推广到多个同方差正态总体的情形. 例如，对于两个同方差正态总体的情形. 设 $\overline{X},\overline{Y},S_1^2,S_2^2$ 是定理四 2°中所说的正态总体 $N(\mu_1,\sigma^2)$，$N(\mu_2,\sigma^2)$的样本均值和样本方差. 只要引入正交矩阵

$$\boldsymbol{T}=\begin{pmatrix}\boldsymbol{A}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{A}_2\end{pmatrix},$$

其中 $\boldsymbol{A}_i$ 为 n_i 阶正交矩阵，其第一行元素都是 $1/\sqrt{n_i}$ $(i=1,2)$，与上面同样的做法，考察向量

$$\boldsymbol{Z}=\boldsymbol{TV}$$

各分量的独立性，其中

$$\boldsymbol{V}^{\mathrm{T}}=(V_1,V_2,\cdots,V_n),$$

$$V_i=(X_i-\mu_1)/\sigma,\ i=1,2,\cdots,n_1,$$

$$V_{n_1+j}=(Y_j-\mu_2)/\sigma, j=1,2,\cdots,n_2, n_1+n_2=n.$$

就可证得 $\overline{X},\overline{Y},S_1^2,S_2^2$ 相互独立.

对于 m $(m\geqslant 2)$个同方差的正态总体的情形，设 $\overline{X}_i,S_i^2$ 分别是总体 $N(\mu_i,\sigma^2)$，$i=1,2,\cdots,m$ 的样本均值和样本方差，且设各样本相互独立，则 $\overline{X}_1,\overline{X}_2,\cdots,\overline{X}_m,S_1^2,S_2^2,\cdots,S_m^2$ 相互独立.

习题

1. 在总体 $N(52,6.3^2)$中随机抽取一容量为 36 的样本，求样本均值 $\overline{X}$ 落在 50.8 到53.8 之间的概率.

2. 在总体 $N(12,4)$中随机抽一容量为 5 的样本 X_1,X_2,X_3,X_4,X_5.

(1) 求样本均值与总体均值之差的绝对值大于 1 的概率.

(2) 求概率 $P\{\max\{X_1,X_2,X_3,X_4,X_5\}>15\}$；$P\{\min\{X_1,X_2,X_3,X_4,X_5\}<10\}$.

3. 求总体 $N(20,3)$的容量分别为 10，15 的两独立样本均值差的绝对值大于 0.3 的概率.

4. (1) 设样本 $X_1,X_2,\cdots,X_6$ 来自总体 $N(0,1)$，$Y=(X_1+X_2+X_3)^2+(X_4+X_5+X_6)^2$，试确定常数 C 使 CY 服从 χ^2 分布.

(2) 设样本 $X_1,X_2,\cdots,X_5$ 来自总体 $N(0,1)$，$Y=\dfrac{C(X_1+X_2)}{(X_3^2+X_4^2+X_5^2)^{1/2}}$，试确定常数 C 使 Y 服从 t 分布.

(3) 已知 $X\sim t(n)$，求证 $X^2\sim F(1,n)$.

5. (1) 已知某种能力测试的得分服从正态分布 $N(\mu,\sigma^2)$，随机取 10 个人参与这一测试. 求他们得分的联合概率密度，并求这 10 个人得分的平均值小于 μ 的概率.

(2) 在(1)中设 $\mu=62,\sigma^2=25$，若得分超过 70 就能得奖，求至少有一人得奖的概率.

6. 设总体 $X\sim b(1,p)$，$X_1,X_2,\cdots,X_n$ 是来自 X 的样本.

(1) 求$(X_1,X_2,\cdots,X_n)$的分布律.

(2) 求 $\sum\limits_{i=1}^{n}X_i$ 的分布律.

(3) 求 $E(\overline{X}),D(\overline{X}),E(S^2)$.

7. 设总体 $X\sim\chi^2(n)$，$X_1,X_2,\cdots,X_{10}$ 是来自 X 的样本，求 $E(\overline{X})$，$D(\overline{X})$，$E(S^2)$.

8. 设总体 $X\sim N(\mu,\sigma^2)$，$X_1,X_2,\cdots,X_{10}$ 是来自 X 的样本.

(1) 写出 $X_1,X_2,\cdots,X_{10}$ 的联合概率密度.

(2) 写出 $\overline{X}$ 的概率密度.

9. 设在总体 $N(\mu,\sigma^2)$ 中抽得一容量为 16 的样本，这里 μ,σ^2 均未知.

(1) 求 $P\{S^2/\sigma^2\leqslant 2.041\}$，其中 S^2 为样本方差.

(2) 求 $D(S^2)$.

10. 下面列出了 30 个美国 NBA 球员的体重(以磅计，1 磅 = 0.454kg)数据. 这些数据是从美国 NBA 球队 1990—1991 赛季的花名册中抽样得到的.

225	232	232	245	235	245	270	225	240	240
217	195	225	185	200	220	200	210	271	240
220	230	215	252	225	220	206	185	227	236

(1) 画出这些数据的频率直方图(提示：最大和最小观察值分别为 271 和 185，区间[184.5,271.5]包含所有数据，将整个区间分为 5 等份，为计算方便，将区间调整为(179.5,279.5).

(2) 作出这些数据的箱线图.

11. 截尾均值 设数据集包含 n 个数据，将这些数据自小到大排序为

$$x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)},$$

删去 $100\alpha\%$ 个数值小的数，同时删去 $100\alpha\%$ 个数值大的数，将留下的数据取算术平均，记为 $\overline{x}_\alpha$，即

$$\overline{x}_\alpha=\frac{x_{([n\alpha]+1)}+\cdots+x_{(n-[n\alpha])}}{n-2[n\alpha]}$$

其中 $[n\alpha]$ 是小于或等于 $n\alpha$ 的最大整数(一般取 α 为 0.1～0.2). $\overline{x}_\alpha$ 称为 $100\alpha\%$ 截尾均值. 例如对于第 10 题中的数据，取 $\alpha=0.1$，则有 $[n\alpha]=[30\times 0.1]=3$，得 $100\times 0.1\%$ 截尾均值

$$\overline{x}_\alpha=\frac{200+200+\cdots+245+245}{30-6}=225.416\ 7.$$

若数据来自某一总体的样本，则 $\overline{x}_\alpha$ 是一个统计量. $\overline{x}_\alpha$ 不受样本的极端值的影响. 截尾均值在实际应用问题中是常会用到的.

试求第 10 题的 30 个数据的 $\alpha=0.2$ 的截尾均值.