数 理 统 计

上海财经大学　统计与管理学院

Contents

第二章 抽样分布

第二章 抽样分布

- **例2.1** 某公司要采购一批产品，每件产品不是合格品就是不合格品，但该批产品总有一 个不合格品率p。由此，若从该批产品中随机抽取一件，用表示这一批产品的不合格数，不难看出服从一个二点分布$b(1,p)$，但分布中的参数 p 是不知道的。一些问题：
  - p 的大小如何；
  - p 大概落在什么范围内；
  - 能否认为 p 满足设定要求（如 $p \leq 0.05$）。

§2.1 总体与样本

2.1.1 总体与个体

- 总体(population)的三层含义：

- 研究对象的全体；
- 数据；
- 分布。

2.1.1 总体与个体

- 例2.2 磁带的一个质量指标是一卷磁带（20m）上的伤痕数。每卷磁带都有一个伤痕数，全部磁带的伤痕数构成一个总体。这个总体中相当一部分是0（无伤痕，合格品），但也有1,2,3等，但多于8个的伤痕数非常少见。研究表明，一卷磁带上的伤痕数X服从泊松分布$P(\lambda)$,但分布中的参数λ却是不知道的。显然, λ的大小决定了一批产品的质量，它直接影响生产方的经济效益。

- 本例中总体分布的类型是明确的，是泊松分布，但总体还有未知参数λ，故总体还不是一个特定的泊松分布。要最终确定总体分布，就要确定λ。

2.1.1 总体与个体

- 例2.3 对网络购物时支付方式使用情况进行分析，支付方式共有以下八种方式: 第三方支付平台、网上银行、货到付款、银行柜台转账付款、邮寄汇款、手机银行、使用个人手机卡（账号）支付、通过网点用户终端支付，设这八种支付方式在网络购物中所占的比例分别为$p_1, \cdots\cdots, p_8$,
- $p_1, \cdots\cdots, p_8$的大小如何;
- $p_1, \cdots\cdots, p_8$的范围;
- $p_1, \cdots\cdots, p_8$能否满足一定的要求(如$p_1 < 0.4$).

2.1.1 总体与个体

- 例2.4 考察常见的测量问题。一个测量者对一个物理量 μ 进行重复测量，此时一切可能的测量结果是$(-\infty,\infty)$，因此总体是一个取值于$(-\infty,\infty)$的随机变量X，关于该总体的分布我们可以知道些什么？
- 测量结果 X 可以看作物理量 μ 与测量误差 ε 的叠加，即

$$X=\mu+\varepsilon,$$

这里μ是一个确定的但未知的量，我们称之为参数，于是关于总体分布的假定主要是关于ε的分布的假定。如下几种假定各在一些场合是合理的。

2.1.1 总体与个体

（1）由中心极限定理，最常见的是假定随机误差$\varepsilon \sim N(0,\sigma^2)$，于是测量值的总体就是一个正态分布，即$X \sim N(\mu,\sigma^2)$，这里总体中有两个未知参数$\mu,\sigma^2$.

（2）假如不仅知道误差服从正态分布，还知道分布的方差，于是，就可假定$\varepsilon \sim N(0,\sigma_0^2)$,其中$\sigma_0$是一个已知的常数，如此，总体仍是一个正态分布族$N(\mu,\sigma_0^2)$,但总体只有一个未知参数$\mu$。

（3）假如并没有理由认定误差服从正态分布，但可以认为误差的分布是关于0对称的，则总体分布就变成一个分布类型未知但带有某种限制的分布，通常它不能被有限个参数所描述，常称为非参数分布。

2.1.2 样本

样本(sample)、样本量(sample size)

样本具有两重性:

- 一方面，由于样本是从总体中随机抽取的，抽取前无法预知它们的数值，因此，样本是随机变量，用大写字母 $X_1, X_2, \ldots, X_n$ 表示；
- 另一方面，样本在抽取以后经观测就有确定的观测值，因此，样本又是一组数值。此时用小写字母 $x_1, x_2, \ldots, x_n$表示是恰当的。

简单起见，无论是样本还是其观测值，样本一般均用 $x_1, x_2, \ldots, x_n$ 表示，应能从上下文中加以区别。

2.1.2 样本

- **例2.5** 啤酒厂生产的瓶装啤酒规定净含量为640 克。由于随机性，事实上不可能使得所有的啤酒 净含量均为640克。现从某厂生产的啤酒中随机抽取10瓶测定其净含量，得到如下结果：

  641, 635, 640, 637, 642, 638, 645, 643, 639, 640

这是一个容量为10的样本的观测值，

对应的总体为该厂生产的瓶装啤酒的净含量。

这样的样本称为**完全样本。**

2.1.2 样本

- **例2.6** 考察某厂生产的某种电子元件的寿命，选了100只进行寿命试验，得到如下数据：

寿命范围	元件数	寿命范围	元件数	寿命范围	元件数
(0 24]	4	(192 216]	6	(384 408]	4
(24 48]	8	(216 240]	3	(408 432]	4
(48 72]	6	(240 264]	3	(432 456]	1
(72 96]	5	(264 288]	5	(456 480]	2
(96 120]	3	(288 312]	5	(480 504]	2
(120 144]	4	(312 336]	3	(504 528]	3
(144 168]	5	(336 360]	5	(528 552]	1
(168 192]	4	(360 184]	1	>552	13

- 上表中的样本观测值没有具体的数值，只有一个范围，这样的样本称为**分组样本。**

2.1.2 样本

- 例2.7 对某大学新生进行身高统计，随机抽取40名学生的身高(厘米)，其结果如下

身高范围	人数	身高范围	人数
(140,145]	1	(160,165]	13
(145,150]	2	(165,170]	6
(150,155]	5	(170,175]	3
(155,160]	9	(175,180]	1

这也是一个分组样本的例子。

2.1.2 样本

- **样本的要求：简单随机样本**

- 要使得推断可靠，对样本就有要求，使样本能很好地代表总体。通常有如下两个要求：

  - 随机性：总体中每一个个体都有同等机会被选入样本 -- X_i 与总体X有相同的分布。
  - 独立性：样本中每一样品的取值不影响其它样品的取值 -- $X_1, X_2, \ldots, X_n$相互独立。

2.1.2 样本

- 用简单随机抽样方法得到的样本称为**简单随机样本(random sample)**，也简称**样本**。
- 于是，样本 $X_1, X_2, \ldots, X_n$ 可以看成是独立同分布 (iid) 的随机变量，其共同分布即为**总体分布**。
- 设总体X具有分布函数$F(x)$, $X_1, X_2, \ldots, X_n$为取自该总体的容量为n的样本，则样本联合分布函数为

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

2.1.2 样本

- 总体分为**有限总体**与**无限总体**。
- 实际中总体中的个体数大多是有限的。当个体数充分大时，将有限总体看作无限总体是一种合理的抽象。
- 对无限总体，随机性与独立性容易实现，困难在于排除有意或无意的人为干扰。
- 对有限总体，只要总体所含个体数很大，特别是与样本量相比很大，则独立性也可基本得到满足。

2.1.2 样本

- **例2.8** 设有一批产品共N个，需要进行抽样检验以了解其不合格品率p。如果把合格品记为0 ，不合格品记为1，则总体为一个二点分布。现从中采取不放回抽样抽出2个产品，这时，第二次抽到不合格品的概率依赖于第一次抽到的是否是不合格品，如果第一次抽到不合格品，则

$$P(X_2 = 1|X_1 = 1) = (Np - 1)/(N - 1)$$

而若第一次抽到的是合格品，则第二次抽到不合格品的概率为

$$P(X_2 = 1|X_1 = 0) = (Np)/(N - 1)$$

2.1.2 样本

- 显然，如此得到的样本不是简单随机样本。但是，当N 很大时，我们可以看到上述两种情形的概率都近似等于p 。所以当N 很大，而n不大（一个经验法则是 $n / N \leq 0.1$）时可以把该样本近似地看成简单随机样本。

- **思考：**
  若总体的密度函数为$f(x)$，则其样本的（联合）密度函数是什么？

§2.2 样本数据的整理与显示

2.2.1 经验分布函数

- 设$x_1, x_2, \dots, x_n$是取自总体分布函数为$F(x)$的样本，若将样本观测值由小到大进行排列，为$x_{(1)}, x_{(2)}, \dots, x_{(n)}$，则$x_{(1)}, x_{(2)}, \dots, x_{(n)}$为**有序样本**，用有序样本定义如下函数：

$$F_n(x) = \begin{cases} 0 & x < x_{(1)} \\ k/n & x_{(k)} \le x < x_{(k+1)}, k = 1, \dots, n-1. \\ 1 & x_{(n)} \le x \end{cases}$$

则$F_n(x)$是一非减右连续函数，且满足

$$F_n(-\infty) = 0 \text{ 和 } F_n(+\infty) = 1$$

由此可见，$F_n(x)$是一个分布函数，并称$F_n(x)$为**经验分布函数(empirical distribution function)**。

2.2.1 经验分布函数

- **例2.9** 某食品厂生产听装饮料，现从生产线上随机抽取5听饮料，称得其净重（单位：克）

351　347　355　344　351

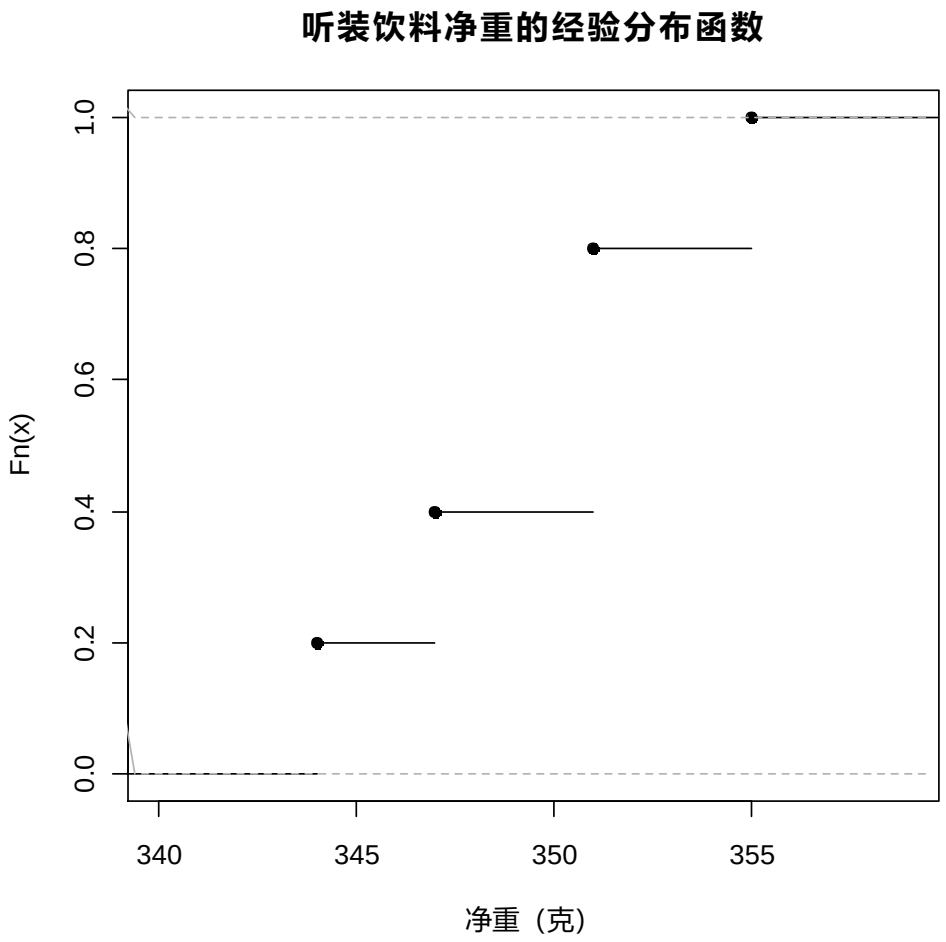
这是一个容量为5的样本，经排序可得有序样本：

$x_{(1)}= 344,\ x_{(2)}= 347,\ x_{(3)}= x_{(4)}= 351,\ x_{(5)}= 355$

2.2.1 经验分布函数

其经验分布函数为

$$F_n(x)=\begin{cases}0, & x<344\\ 0.2, & 344\le x<347\\ 0.4, & 347\le x<351\\ 0.8, & 351\le x<355\\ 1, & 355\le x\end{cases}$$



2.2.1 经验分布函数

- **2.10** 我国自1984年23届洛杉矶奥运会以来历届奥运会的金牌数如表所示：

奥运会	金牌数
第23届洛杉矶奥运会	15
第24届汉城奥运会	5
第25届巴塞罗那奥运会	16
第26届亚特兰大奥运会	16
第27届悉尼奥运会	28
第28届雅典奥运会	32
第29届北京奥运会	51

这是一个容量为7的样本，经排序可得有序样本：

2.2.1 经验分布函数

$x_{(1)}=5$，$x_{(2)}=15$，$x_{(3)}=x_{(4)}=16$，$x_{(5)}=28$，$x_{(6)}=32$，$x_{(7)}=51$

则金牌数的经验分布函数为：

$$F_n(x)=\begin{cases}0 & x<5\\ \frac{1}{7} & 5\le x<15\\ \frac{2}{7} & 15\le x<16\\ \frac{4}{7} & 16\le x<28\\ \frac{5}{7} & 28\le x<32\\ \frac{6}{7} & 32\le x<51\\ 1 & x\ge 51\end{cases}$$

2.2.1 经验分布函数

由伯努里大数定律：只要n相当大，$F_n(x)$依概率收敛于$F(x)$ 。更深刻的结果也是存在的，这就是格里纹科定理。

- **定理2.1（格里纹科定理）** 设$X_1, X_2, \ldots, X_n$是取自总体分布函数为$F(x)$的样本，$F_n(x)$ 是其经验分布函数，当$n \to \infty$时，有

$$P\{sup|F_n(x) - F(x)| \to 0\} = 1$$

- **格里纹科定理表明：** 当n 相当大时，经验分布函数是总体分布函数$F(x)$的一个良好的近似。经典的统计学中一切统计推断都以样本为依据，其理由就在于此。

2.2.2 频数频率表

- 样本数据的整理是统计研究的基础，整理数据的最常用方法之一是给出其频数分布表或频率分布表。
- **例2.11** 为研究某厂工人生产某种产品的能力，我们随机调查了20位工人某天生产的该种产品的数量，数据如下

160	196	164	148	170
175	178	166	181	162
161	168	166	162	172
156	170	157	162	154

2.2.2 频数频率表

- 对这20个数据(样本)进行整理,具体步骤如下:

(1) 对样本进行分组：作为一般性的原则，组数通常在5~20个，对容量较小的样本，通常分5组或6组；

(2) 确定每组组距：近似公式为

组距d = (最大观测值 – 最小观测值)/组数；

(3) 确定每组组限： 各组区间端点为

$$a_0, a_1 = a_0 + d, a_2 = a_0 + 2d, \ldots, a_k = a_0 + kd,$$

形成如下的分组区间

$$(a_0, a_1], (a_1, a_2], \ldots, (a_{k-1}, a_k]$$

其中a_0 略小于最小观测值， a_k 略大于最大观测值.

(4) 统计样本数据落入每个区间的个数——频数，并列出其频数频率分布表。

2.2.2 频数频率表

例2.11 的频数频率表

组序	分组区间	组中值	频数	频率	累计频率(%)
1	(147，157]	152	4	0.2	20
2	(157，167]	162	8	0.4	60
3	(167，177]	172	5	0.25	85
4	(177，187]	182	2	0.1	95
5	(187，197]	192	1	0.05	100
合计			20	1	

2.2.3 样本数据的图形显示

一、直方图(histogram)

直方图是频数分布的图形表示，它的横坐标表示所关心变量的取值区间，纵坐标有三种表示方法：**频数，频率，**最准确的是**频率/组距，**它可使得诸长条矩形面积和为1。凡此三种直方图的差别仅在于纵轴刻度的选择，直方图本身并无变化。

2.2.3 样本数据的图形显示

❖例2.12 1978-2012年我国的人口出生率如下表所示，

年份	出生率(‰)	年份	出生率(‰)	年份	出生率(‰)	年份	出生率(‰)
2012	12.1	2003	12.41	1994	17.7	1985	21.04
2011	11.93	2002	12.86	1993	18.09	1984	19.9
2010	11.9	2001	13.38	1992	18.24	1983	20.19
2009	11.95	2000	14.03	1991	19.68	1982	22.28
2008	12.14	1999	14.64	1990	21.06	1981	20.91
2007	12.1	1998	15.64	1989	21.58	1980	18.21
2006	12.09	1997	16.57	1988	22.37	1979	17.82
2005	12.4	1996	16.98	1987	23.33	1978	18.25
2004	12.29	1995	17.12	1986	22.43		

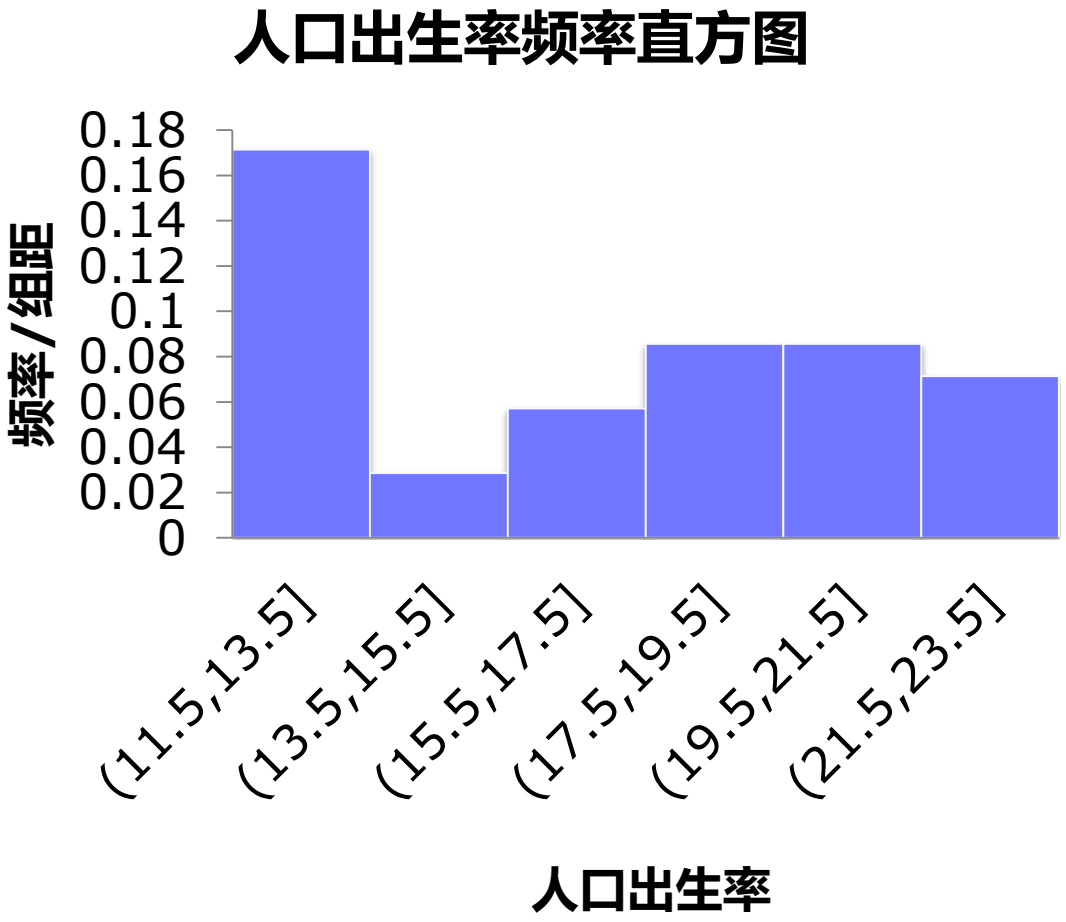
2.2.3 样本数据的图形显示

则1978-2012年我国的人口出生率的频数频率表

序数	分组区间	组中值	频数	频率	累计频率	频率/组距
1	(11.5,13.5]	12.5	12	0.342857	0.342857	0.171429
2	(13.5,15.5]	14.5	2	0.057143	0.4	0.028571
3	(15.5,17.5]	16.5	4	0.114286	0.514286	0.057143
4	(17.5,19.5]	18.5	6	0.171429	0.685714	0.085714
5	(19.5,21.5]	20.5	6	0.171429	0.857143	0.085714
6	(21.5,23.5]	22.5	5	0.142857	1	0.071429
合计			35	1		

2.2.3 样本数据的图形显示

由上表可得到我国1978-2012年人口出生率的频率直方图为：



2.2.3 样本数据的图形显示

二、茎叶图(stem-and-leaf plot)

把每一个数值分为两部分，前面一部分（百位和十位）称为茎，后面部分（个位）称为叶，然后画一条竖线，在竖线的左侧写上茎，右侧写上叶，就形成了茎叶图。如：

数值	分开	茎	和	叶
112 →	11 | 2 →	11	和	2

2.2.3 样本数据的图形显示

- 例2.13 某公司对应聘人员进行能力测试，测试成绩总分为150分。下面是50位应聘人员的测试成绩（已经过排序）：

64	67	70	72	74	76	76	79	80	81
82	82	83	85	86	88	91	91	92	93
93	93	95	95	95	97	97	99	100	100
102	104	106	106	107	108	108	112	112	114
116	118	119	119	122	123	125	126	128	133

- 我们用这批数据给出一个茎叶图，见下页。

2.2.3 样本数据的图形显示

6	4 7
7	0 2 4 6 6 9
8	0 1 2 2 3 5 6 8
9	1 1 2 3 3 3 5 6 6 7 7 9
10	0 0 2 4 6 6 7 8 8
11	2 2 4 6 8 9 9
12	2 3 5 6 8
13	3

测试成绩的茎叶图

2.2.3 样本数据的图形显示

- 在要比较两组样本时，可画出它们的背靠背的茎叶图。例2.11中需要对两个车间某天各40名员工生产的产品数量进行比较：

甲车间		乙车间
	5	6
8 7 7 7 5 5 5 4 2 1 1	6	6 7 7 8 8
8 7 7 6 6 4 4 2 1	7	2 2 4 5 5 5 5 6 6 6 8 8 9
8 7 6 6 5 3 2	8	0 1 1 3 3 3 4 4 4 6 6 7 7 8
7 3 2 1 0	9	0 2 3 5 8
5 3 0 0	10	7

- **注意：** 茎叶图保留数据中全部信息。当样本量较大，数据很分散，横跨二、三个数量级时， 茎叶图并不适用。

2.2.3 样本数据的图形显示

- 例2.14 1959-1992历届奥斯卡奖最佳女 / 男演员奖得主的年龄如下表所示：

女演员年龄 （1959-1992）									
38	28	27	31	37	30	24	34	60	61
26	35	34	34	26	37	42	41	35	31
41	33	30	74	33	49	38	61	21	41
26	80	42	29						

男演员年龄（1959-1992）									
35	47	31	46	39	56	41	44	42	43
62	43	40	48	48	56	38	60	32	41
42	37	76	39	55	45	35	61	33	51
31	42	62	62						

2.2.3 样本数据的图形显示

则历届奥斯卡奖女 / 男最佳演员奖得主的年龄的茎叶图如图所示：

女演员		男演员
9 8 7 6 6 6 4 1	2	
8 8 7 7 5 5 4 4 4 3 3 1 1 0 0	3	1 1 2 3 5 5 7 8 9 9
9 2 2 1 1 1	4	0 1 1 2 2 2 3 3 4 5 6 7 8 8
	5	1 5 6 6
1 1 0	6	0 1 2 2
4	7	6
0	8	

最佳演员年龄茎叶图

§2.3 统计量及其分布

2.3.1 统计量与抽样分布

- 当人们需要从样本获得对总体各种参数的认识时，最好的方法是构造样本的函数，不同的函数反映总体的不同特征。

- **定义2.1** 设$X_1, X_2, \ldots, X_n$为取自某总体的样本，若样本函数$T = T(X_1, X_2, \ldots, X_n)$中不含有任何未知参数。则称$T$为统计量(statistic)。统计量的分布称为抽样分布(sampling distribution)。

2.3.1 统计量与抽样分布

- 按照这一定义：若$X_1, X_2, \ldots, X_n$为样本，则$\sum_{i=1}^{n} X_i$，$\sum_{i=1}^{n} X_i^2$以及经验分布函数$F_n(x)$都是统计量。而当μ, σ^2未知时，$X_1 - \mu, X_1/\sigma$等均不是统计量。
- 尽管统计量不依赖于未知参数，但是它的分布一般是依赖于未知参数的。
- 下面介绍一些常见的统计量及其抽样分布。

2.3.2 样本均值及其抽样分布

- **定义2.2** 设$X_1, X_2, \ldots, X_n$为取自某总体的样本，其算术平均值称为样本均值(sample mean)，一般用$\bar{X}$表示，即

$$\bar{X} = (X_1 + \cdots + X_n)/n$$

- **思考：**在分组样本场合，样本均值如何计算？二者结果相同吗？

2.3.2 样本均值及其抽样分布

样本均值的基本性质：

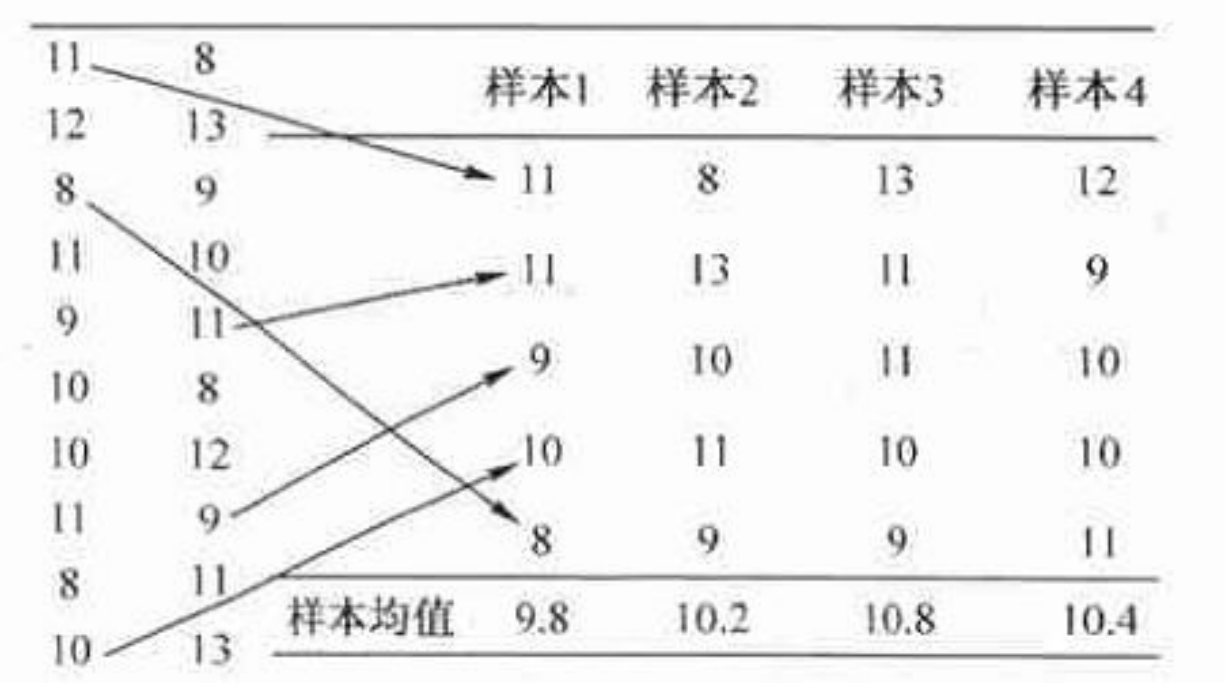
- **定理2.2** 若把样本中的数据与样本均值之差称为偏差，则样本所有偏差之和为0，即

$$\sum_{i=1}^{n}(X_i-\bar{X})=0$$

- **定理2.3** 数据观测值与均值的偏差平方和最小，即在形如 $\sum(X_i-c)^2$ 的函数中， $\sum(X_i-\bar{X})^2$ 最小，其中c为任意给定常数。

2.3.2 样本均值及其抽样分布

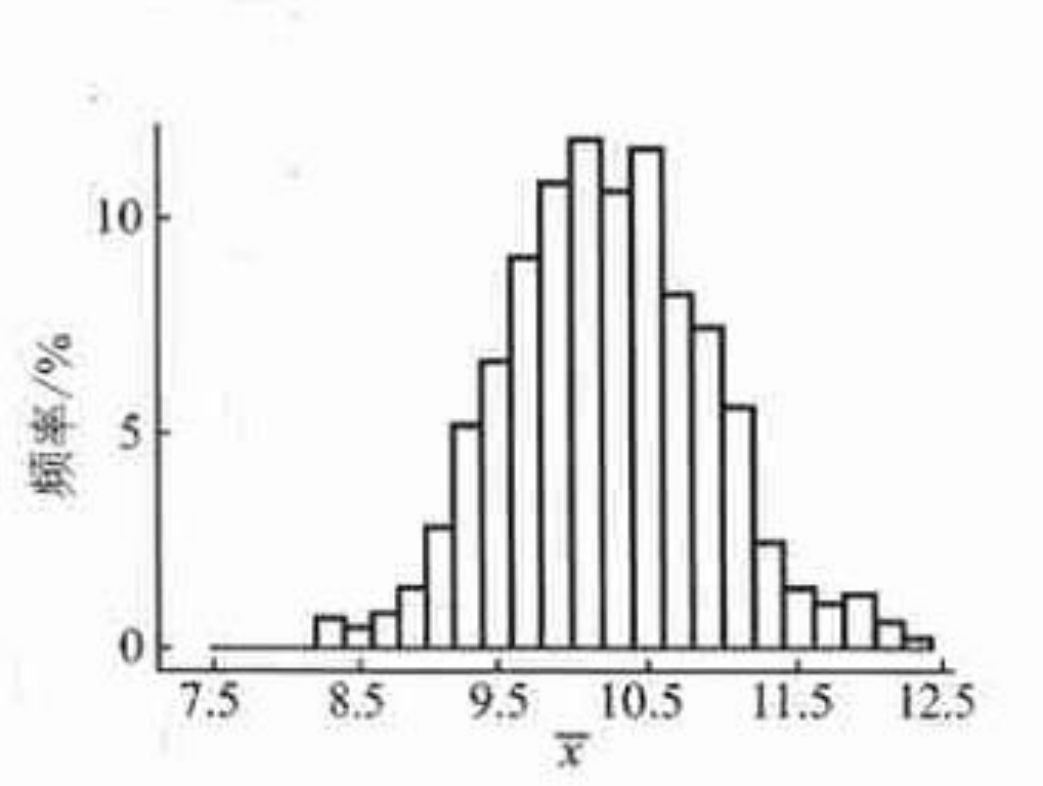
- **例2.15** 设有一个由20个数组成的总体，现从该总体同时取出容量为5的样本，记录后放回，再抽第二个样本 。下图记录了前四个样本及其样本均值。



		样本1	样本2	样本3	样本4
11	8	11	8	13	12
12	13				
8	9				
11	10	11	13	11	9
9	11				
10	8	9	10	11	10
10	12	10	11	10	10
11	9				
8	11	8	9	9	11
10	13				
样本均值		9.8	10.2	10.8	10.4

2.3.2 样本均值及其抽样分布

上述抽样过程可以无限重复，从而得到大量的 $\bar{x}$ 的值。下图给出前500个$\bar{x}$的值所形成的直方图，它反映了$\bar{x}$的抽样分布。



2.3.2 样本均值及其抽样分布

样本均值的抽样分布：

❖ **定理2.4** 设$X_1, X_2, \ldots, X_n$是来自某个总体的样本，$\bar{X}$为样本均值。

（1）若总体分布为$N(\mu, \sigma^2)$，则 $\bar{X}$ 的精确分布为$N(\mu, \sigma^2/n)$;

（2）若总体分布未知或不是正态分布，但$E(X) = \mu, Var(X) = \sigma^2$, 则$n$ 较大时$\bar{X}$的渐近分布为$N(\mu, \sigma^2/n)$, 常记为$\bar{X} \sim AN\left(\mu, \frac{\sigma^2}{n}\right)$，这里渐近分布是指$n$ 较大时的近似分布.

2.3.2 样本均值及其抽样分布

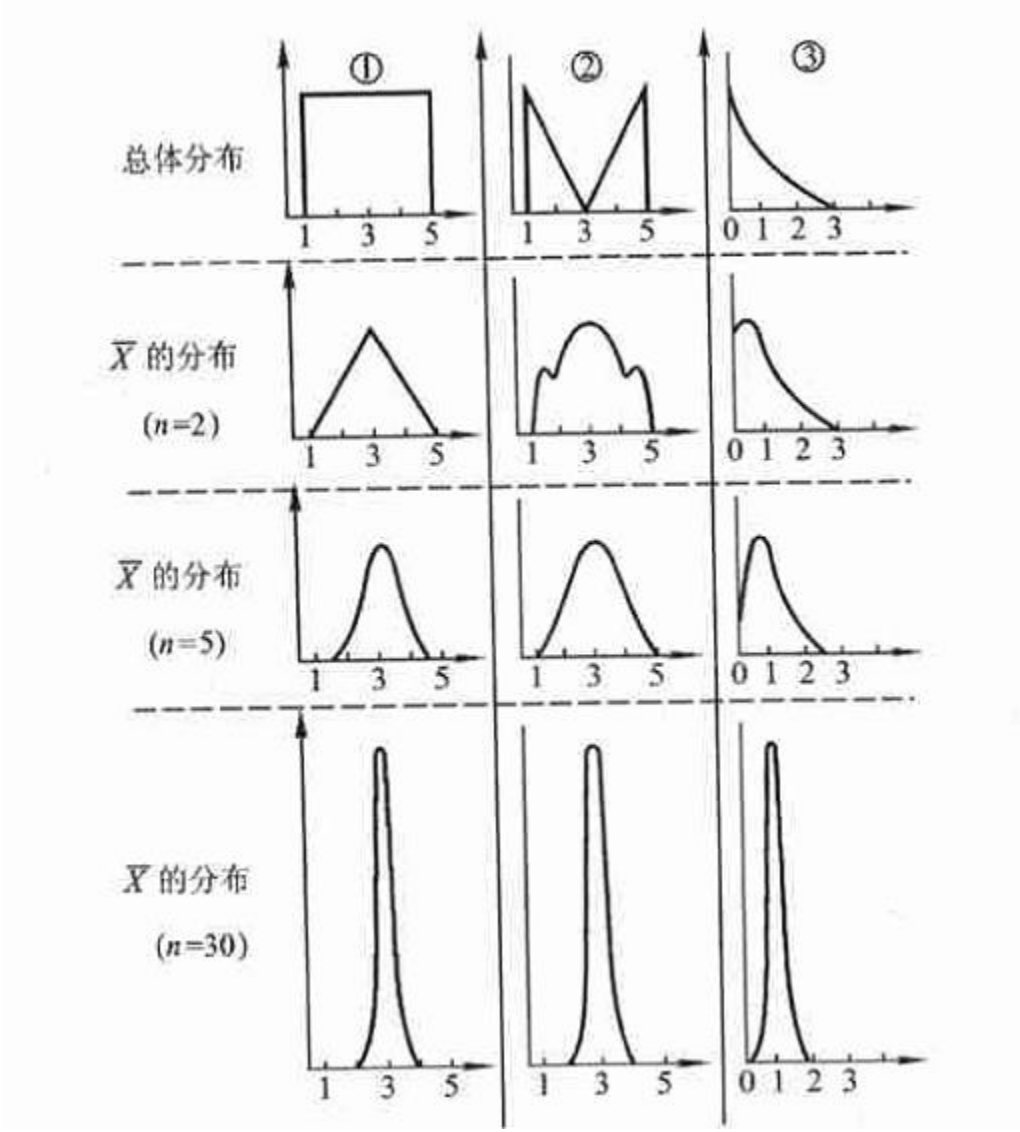
❖ **例2.16**

右图给出三个不同的总体样本均值的分布：

① 均匀分布

② 倒三角分布

③ 指数分布



2.3.3 样本方差与样本标准差

- 定义2.3 $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$称为样本方差(sample variance)，其算术平方根 $S_n = \sqrt{S_n^2}$称为样本标准差(sample standard deviation)。

- 在n 不大时，常用 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 作为样本方差（也称无偏方差，其含义在第六章讲述），其算术平方根也称为样本标准差。

2.3.3 样本方差与样本标准差

- 在这个定义中，$\sum_{i=1}^{n}(X_i-\bar{X})^2$称为偏差平方和(sum of square of deviations)，n–1称为偏差平方和的自由度(degree of freedom)。其含义是：在$\bar{X}$确定后，n 个偏差$X_1-\bar{X}, X_2-\bar{X}, \ldots, X_n-\bar{X}$中只有$n$–1个数据可以自由变动，而第$n$个则不能自由取值，因为$\sum(X_i-\bar{X})=0$。
- 样本偏差平方和有三个不同的表达式：

$$\sum(X_i-\bar{X})^2=\sum X_i^2-\frac{(\sum X_i)^2}{n}=\sum X_i^2-n\bar{X}^2$$

它们都可用来计算样本方差。

- **思考：**分组样本如何计算样本方差？

2.3.3 样本方差与样本标准差

- 样本均值的数学期望和方差，以及样本方差的数学期望都不依赖于总体的分布形式。

- **定理2.5** 设总体 X 具有二阶矩，即

$$E(X)=\mu<\infty, Var(X)=\sigma^2<\infty$$

$X_1, X_2, \ldots, X_n$ 为从该总体得到的样本，$\bar{X}$和S^2分别是样本均值和样本方差，则

$$E(\bar{X})=\mu \ ， \ Var(\bar{X})=\sigma^2/n \ ， \quad E(S^2)=\sigma^2$$

2.3.4 样本矩及其函数

- 样本均值和样本方差的更一般的推广是样本矩，这是一类常见的统计量。

- **定义2.4** $A_k = (\sum X_i^k)/n$称为样本 k 阶原点矩（origin moment）。样本一阶原点矩就是样本均值。

$$B_k = \sum (X_i - \bar{X})^k / n$$称为样本 k 阶中心矩（central moment）。样本二阶中心矩就是样本方差。

2.3.4 样本矩及其函数

- 当总体关于分布中心对称时，我们用$\bar{X}$和S刻画样本特征很有代表性，而当其不对称时，只用$\bar{X}$和S就显得很不够。为此，需要一些刻画分布形状的统计量，如样本偏度和样本峰度，它们都是样本中心矩的函数。

- **定义2.5** $\hat{\beta}_s = B_3/B_2^{3/2}$称为样本偏度(sample skewness)，

$$\hat{\beta}_k = \frac{B_4}{B_2^2} - 3$$称为样本峰度(sample kurtosis)。

- 样本偏度$\hat{\beta}_s$反映了样本数据与对称性偏离程度和偏离方向。样本峰度$\hat{\beta}_k$是反映总体分布密度曲线在其峰值附近的陡峭程度和尾部粗细的统计量。

2.3.5 次序统计量及其分布

一、定义

另一类常见的统计量是次序统计量。

❖ 定义2.6

设$X_1, X_2, \ldots, X_n$ 是取自总体X的样本, $X_{(i)}$称为该样本的第i 个次序统计量(order statistic)，它的取值是将样本观测值由小到大排列后得到的第 i 个观测值。其中$X_{(1)} = min\{X_1, X_2, \ldots, X_n\}$ 称为该样本的最小次序统计量，称$X_{(n)} = max\{X_1, X_2, \ldots, X_n\}$为该样本的最大次序统计量。

2.3.5 次序统计量及其分布

- 我们知道，在一个样本中，$X_1, X_2, \ldots, X_n$是独立同分布的，而次序统计量$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$则既不独立，分布也不相同，看下例。

- **例2.17** 设总体X 的分布为仅取0，1，2的离散均匀分布，分布列为

X	0	1	2
p	1/3	1/3	1/3

现从中抽取容量为3的样本，其一切可能取值有3^3=27种，表2.3.6（见教材272页）列出了这些值。

2.3.5 次序统计量及其分布

由此可给出 $X_{(1)}, X_{(2)}, x_{(3)}$ 分布列如下：

$X_{(1)}$	0	1	2
p	19/27	7/27	1/27

$X_{(2)}$	0	1	2
p	7/27	13/27	7/27

$X_{(3)}$	0	1	2
p	1/27	7/27	19/27

我们可以清楚地看到这三个次序统计量的分布是不相同的。

2.3.5 次序统计量及其分布

进一步，我们可以给出两个次序统计量的联合分布，如，$X_{(1)}$和 $X_{(2)}$的联合分布列为

$X_{(1)}$ \ $X_{(2)}$	0	1	2
0	7/27	9/27	3/27
1	0	4/27	3/27
2	0	0	1/27

因为 $P(X_{(1)} = 0, X_{(2)} = 0) = 7/27$ ，

而 $P(X_{(1)} = 0) \times P(X_{(2)} = 0) = (19/27) \times (7/27)$，

二者不等，由此可看出$X_{(1)}$和 $X_{(2)}$是不独立的。

2.3.5 次序统计量及其分布

二、单个次序统计量的分布

❖ **定理2.6** 设总体X的密度函数为$f(x)$，分布 函数为$F(x)$，$X_1, X_2, \ldots, X_n$为样本，则第k个次序统计量$X_{(k)}$的密度函数为

$$f_k(x) = \frac{n!}{(k-1)!\,(n-k)!}\big(F(x)\big)^{k-1}\big(1-F(x)\big)^{n-k}f(x)$$

2.3.5 次序统计量及其分布

- **例2.18** 设总体密度函数为 $f(x)=3x^2,\ 0<x<1$.从该总体抽得一个容量为5的样本，试计算 $P(X_{(2)}<1/2)$。

- **解：** 从次序统计量密度函数出发。

当 $0<x<1, F(x)=x^3$

当 $0<x<1$, $$f_{x_{(2)}}(x)=\frac{5!}{(5-2)!}[F(x)]^{2-1}\left(1-F(x)\right)^{5-2}f(x)$$

$$=20x^3\cdot 3x^2\cdot(1-x^3)^3$$

$$=60x^5(1-x^3)^3$$

$$P(X_{(2)}<1/2)=\int_0^{1/2}60x^5(1-x^3)^3dx=0.1207$$

2.3.5 次序统计量及其分布

❖**例2.19** 设总体密度函数为$f(x)=e^{-x},0<x<\infty,$从该总体中抽得一个容量为4的样本，求$P(X_{(4)}\geq 3)$。

❖**解：**有两种求法：

法一、从古典概型出发；

$$P\left(X_{(4)}\geq 3\right)=1-P\left(X_{(4)}<3\right)$$

$$=1-P\left(X_1<3,X_2<3,X_3<3,X_4<3\right)$$

$$=1-\{P(X_1<3)\}^4$$

又$P(X_1<3)=\int_0^3 e^{-x}dx=1-e^{-3}$

故$P\left(X_{(4)}\geq 3\right)=1-(1-e^{-3})^4$

2.3.5 次序统计量及其分布

❖**解（续）：**

法二、从次序统计量密度函数出发。

当 $x>0,\ F(x)=\int_0^x e^{-u}du=1-e^{-x}$

当 $x>0, f_{x_{(4)}}(x)=\frac{4!}{(4-1)!}[F(x)]^{4-1}f(x)$

$$=4(1-e^{-x})^3e^{-x}$$

$$P(X_{(4)}\geq 3)=\int_3^{+\infty}4(1-e^{-x})^3e^{-x}dx$$

$$=1-(1-e^{-3})^4$$

2.3.5 次序统计量及其分布

- **例2.20** 设总体分布为 $f(x)=\frac{1}{6},\ \ x=1,2,3,4,5,6$ (i.e. 离散的均匀分布)。从该总体中抽得一个容量为5的样本，求证$X_{(1)}$的分布为

$$f_{(1)}(x)=\left(\frac{7-x}{6}\right)^5-\left(\frac{6-x}{6}\right)^5,x=1,2,3,4,5,6$$

2.3.5 次序统计量及其分布

❖ 解：

$$
\begin{aligned}
f_{(1)}(x) &= P\left(X_{(1)} \le x\right) - P\left(X_{(1)} \le x-1\right) \\
&= \left(1 - P\left(X_{(1)} > x\right)\right) - \left(1 - P\left(X_{(1)} > x-1\right)\right) \\
&= P\left(X_{(1)} > x-1\right) - P\left(X_{(1)} > x\right) \\
&= \left(\frac{7-x}{6}\right)^5 - \left(\frac{6-x}{6}\right)^5
\end{aligned}
$$

❖ 注：次序统计量的密度函数公式不适用。Why?

2.3.5 次序统计量及其分布*

三、多个次序统计量的联合分布

对任意多个次序统计量可给出其联合分布，以两个为例说明：

❖ **定理2.7** 在定理2.6的记号下，次序统计量 $(X_{(i)}, X_{(j)}), (i<j)$ 的联合分布密度函数为

$$f_{i,j}(y,z)=\frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F(y)]^{i-1}$$
$$[F(z)-F(y)]^{j-i-1}[1-F(z)]^{n-j}f(y)f(z) \quad y\le z$$

2.3.5 次序统计量及其分布*

- 次序统计量的函数在实际中经常用到。

  如 样本极差（sample range） $R_n = X_{(n)} - X_{(1)}$，

  样本中程（sample mid-range） $\left[X_{(n)} + X_{(1)}\right]/2$。

- 样本极差是一个很常用的统计量，其分布只在很少几种场合可用初等函数表示。

2.3.5 次序统计量及其分布*

- **例2.3.9** 设总体分布为$U(0,1)$, $X_1, X_2, \ldots, X_n$为样本，则$(X_{(1)}, X_{(n)})$的联合密度函数为

$$f_{1,n}(y,z) = n(n-1)(z-y)^{n-2}, \qquad 0 < y < z < 1$$

- **解：** 令$R_n = X_{(n)} - X_{(1)}$，由$R > 0$，可以推出

$$0 < X_{(1)} = X_{(n)} - R \le 1 - R,$$

则

$$f_R(r) = \int_0^{1-r} n(n-1)[(y+r)-y]^{n-2}dy = n(n-1)r^{n-2}(1-r)$$

这正是参数为$(n-1, 2)$的贝塔分布。

2.3.5 次序统计量及其分布*

❖ **例2.21** 设总体密度函数为$f(x)=e^{-x},0<x<\infty$,从该总体中抽得一个容量为4的样本，求样本极差$R_4\geq 2$的概率。

❖ **解：** 设$X_{(1)}=y$和$X_{(n)}=z$的联合密度函数为：

$$f(y,z)=\frac{n!}{(n-2)!}e^{-y}e^{-z}(e^{-y}-e^{-z})^{n-2}$$

做变换$\begin{cases}R_n=z-y\\ Y=y\end{cases}\Rightarrow\begin{cases}y=Y\\ z=R_n+Y\end{cases}$

雅克比行列式绝对值$|J|=1$.

2.3.5 次序统计量及其分布*

则R_n和Y的联合密度

$$f_{R_n,Y}(r,y)=\frac{n!}{(n-2)!}e^{-y}e^{-(r+y)}\left(e^{-y}-e^{-(r+y)}\right)^2$$

$$=n(n-1)\mathrm{e}^{-4y}e^{-r}(1-e^{-r})^2$$

$$f_{R_n}(r)=\int_0^{+\infty} n(n-1)\,\mathrm{e}^{-4y}e^{-r}(1-e^{-r})^2dy$$

$$=\int_0^{+\infty}4\times 3\,\mathrm{e}^{-4y}e^{-r}(1-e^{-r})^2dy$$

$$=3e^{-r}(1-e^{-r})^2$$

$$\therefore P(r\geq 2)=\int_2^{+\infty}3e^{-r}(1-e^{-r})^2dx=1-(1-e^{-2})^3$$

2.3.6 样本分位数与样本中位数

- 样本中位数(sample median)也是一个很常见的统计量，它也是次序统计量的函数，通常如下定义：

$$m_{0.5}=\begin{cases} X_{\left(\frac{n+1}{2}\right)}, & n\text{为奇数} \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}\right), & n\text{为偶数} \end{cases}$$

- 更一般地，样本p分位数m_p可如下定义：

$$m_p=\begin{cases} X_{([np+1])}, & np\text{不是整数} \\ \frac{1}{2}\left(X_{(np)}+X_{(np+1)}\right), & np\text{是整数} \end{cases}$$

2.3.6 样本分位数与样本中位数*

❖ **定理2.8** 设总体密度函数为$f(x)$，X_p为其p分位数，$f(x)$在X_p处连续且$f(X_p)>0$，则当$n\to\infty$时样本p分位数m_p的渐近分布为

$$m_p \dot{\sim} N\left(x_p, \frac{p(1-p)}{n\cdot f^2(x_p)}\right)$$

特别，对样本中位数，当$n\to\infty$时近似地有

$$m_{0.5} \dot{\sim} N\left(x_{0.5}, \frac{1}{4n\cdot f^2(x_{0.5})}\right)$$

2.3.6 样本分位数与样本中位数*

- 例2.22 设总体为柯西分布，密度函数为

$$f(x,\theta) = 1/[\pi(1+(x-\theta)^2)] \quad ,-\infty < x < +\infty$$

不难看出θ是该总体的中位数，即$X_{0.5} = \theta$ 。

设 $X_1, X_2, \ldots, X_n$是来自该总体的样本，当样本量n较大时，样本中位数$m_{0.5}$ 的渐近分布为

$$m_{0.5} \sim AN(\theta, \pi^2/4n) .$$

2.3.6 样本分位数与样本中位数*

- **例2.23** 设$X_1, X_2, \cdots, X_n$是来自正态总体$N(\mu, \sigma^2)$的样本,当n足够大时，求样本中位数$m_{0.5}$的分布。
- **解：** 对样本中位数，当$n \to \infty$时近似地有

$$m_{0.5} \dot{\sim} N\left(x_{0.5}, \frac{1}{4n \cdot f^2(x_{0.5})}\right)$$

$$\text{其中} x_{0.5} = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & n\text{为奇数} \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right), & n\text{为偶数} \end{cases}$$

2.3.6 样本分位数与样本中位数

- 通常，样本均值在概括数据方面具有一定的优势。但当数据中含有极端值时，使用中位数比使用均值更好，中位数的这种抗干扰性在统计中称为具有稳健性。
- 例如样本值为5个数，

  3， 5， 9， 10， 13，

  在计算机输入时不小心输为

  3， 5， 9， 10， 133.

  则计算得出样本均值由8变为32，而中位数为9保持不变。

2.3.7 五数概括与箱线图

- 次序统计量的应用之一是五数概括(five-number summary)与箱线图(boxplot)。在得到有序样本后，容易计算如下五个值：

  最小观测值 $X_{min} = X_{(1)}$，最大观测值 $X_{max} = X_{(n)}$，

  中位数 $m_{0.5}$，第一4分位数(first quartile) $Q_1 = m_{0.25}$，

  第三4分位数(third quartile) $Q_3 = m_{0.75}$.

- 所谓五数概括就是指用这五个数：

  $$X_{min}, Q_1, m_{0.5}, Q_3, X_{max}$$

  来大致描述一批数据的轮廓。

2.3.7 五数概括与箱线图

- 箱线图——五数概括的图形表示，其主要包括：
  - 一个箱体：

    箱体内部的竖线位置对应样本中位数 $m_{0.5}$

    箱体左、右边线的位置分别对应 Q_1 和 Q_3
  - 两根线：从箱体两侧分别延伸出一条线

    左边延伸至 max{$X_{(1)}$, Q_1-1.5IQR}

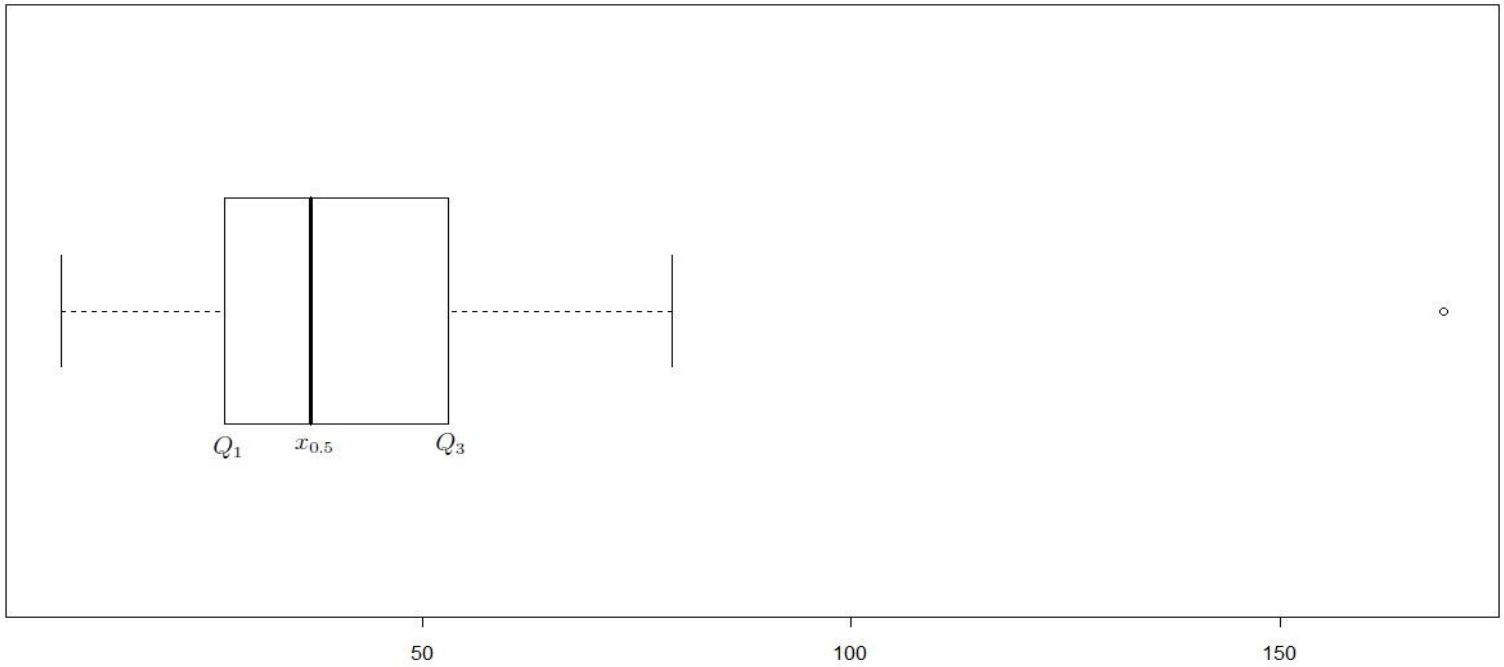
    右边延伸至 min{$X_{(n)}$, Q_3 +1.5IQR}

    其中 IQR=Q_3-Q_1 称为四分位数间距(interquartile range)，作为描述数据离散度统计量，IQR显然比极差更稳健。
  - 异常值点

    落在[Q_1-1.5IQR, Q_3 +1.5IQR]之外的数据称为异常值(outlier)，在图中单独标出。

2.3.7 五数概括与箱线图

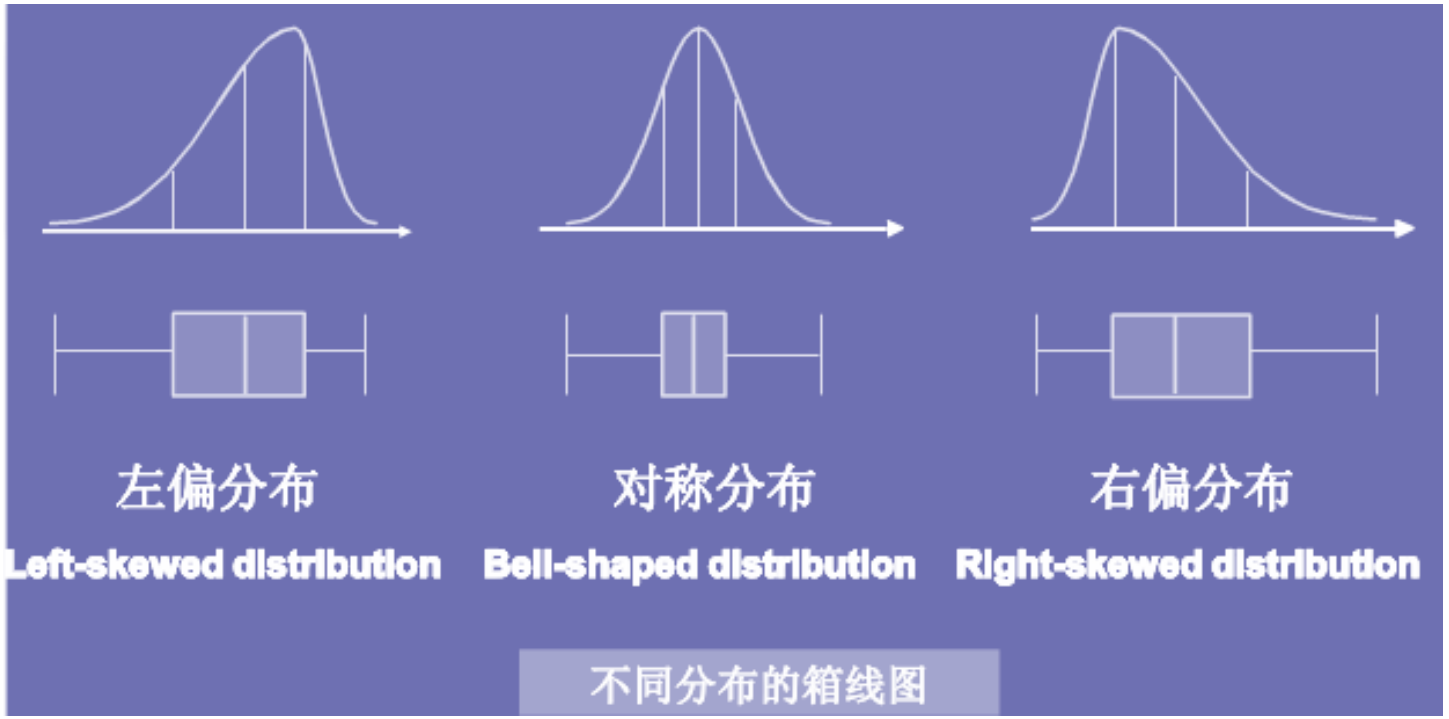
箱线图



箱线图可以看出数据分布的中心位置(location)、散布程度(dispersion)、对称性(symmetry)。

2.3.7 五数概括与箱线图

- 分布的形状与箱线图



不同分布的箱线图

§2.4 三大抽样分布

- 大家很快会看到，有很多统计推断是基于正态分布的假设的，以标准正态变量为基石而构造的三个著名统计量在实际中有广泛的应用，这是因为这三个统计量不仅有明确背景，而且其抽样分布的密度函数有明显表达式，它们被称为统计中的“**三大抽样分布**”。

2.4.1 χ^2 分布(卡方分布)

- **定义2.7** 设 $X_1, X_2, \ldots, X_n$, 独立同分布于标准正态分布$N(0,1)$，则

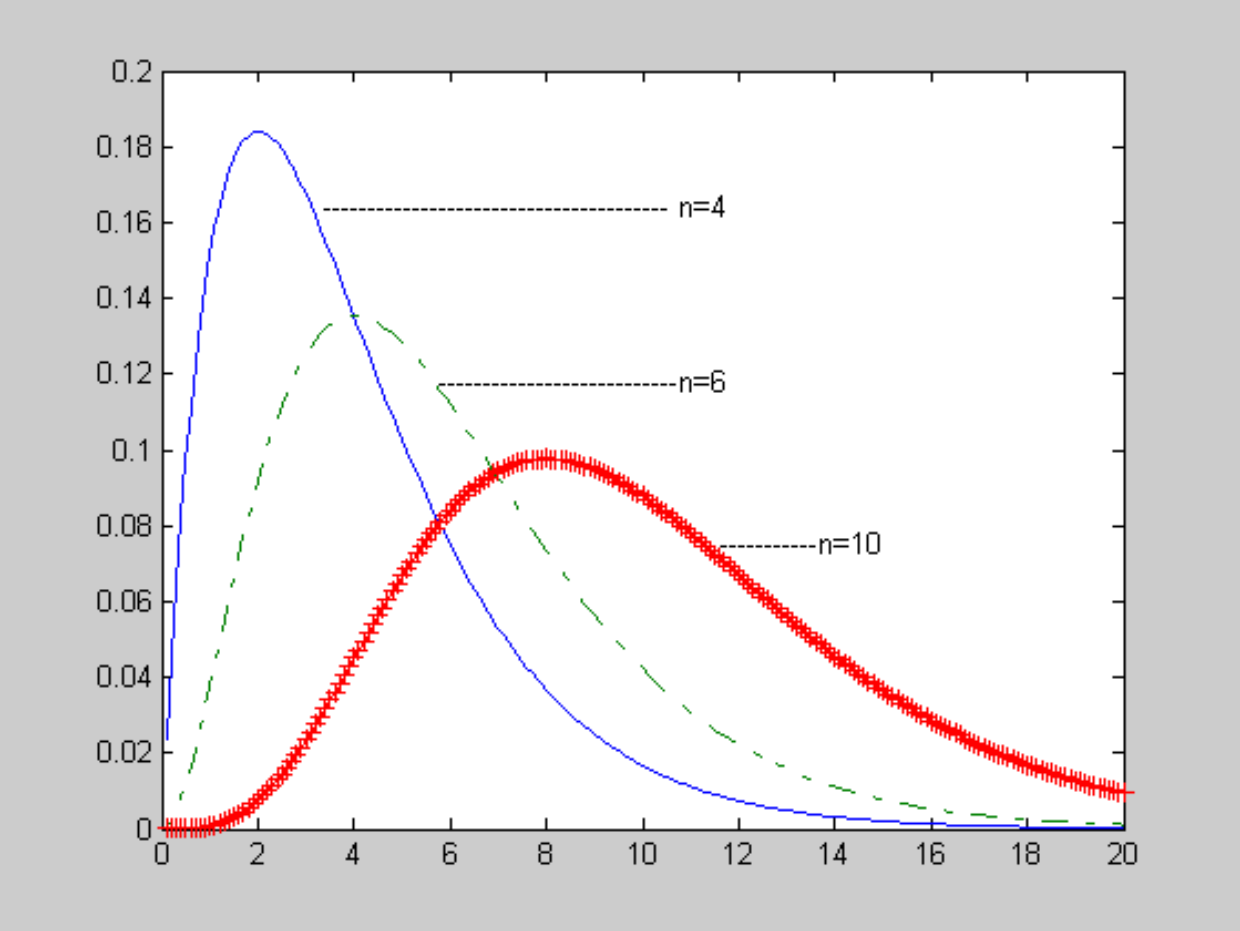
$$\chi^2 = X_1^2 + \cdots + X_n^2$$

的分布称为自由度为n 的χ^2分布，记为$\chi^2 \sim \chi^2(n)$.

- χ^2分布具有可加性。
- $E(\chi^2) = n$，$Var(\chi^2) = 2n$

2.4.1 χ^2 分布(卡方分布)

该密度函数的图像是一个只取非负值的偏态分布



2.4.1 χ^2 分布(卡方分布)

- 当随机变量 $\chi^2 \sim \chi^2(n)$时，对给定$\alpha(0 < \alpha < 1)$，称满足

$$P(\chi^2 \leq \chi^2_{1-\alpha}(n)) = 1-\alpha$$

的$\chi^2_{1-\alpha}(n)$是自由度为n的卡方分布的$1-\alpha$分位数。

分位数 $\chi^2_{1-\alpha}(n)$可以从附表3 中查到。

2.4.1 χ^2 分布(卡方分布)

- **定理2.9** 设$X_1, X_2, \ldots, X_n$ 是来自$N(\mu, \sigma^2)$ 的样本，其样本均值和样本方差分别为

$$\bar{X} = \sum X_i/n \text{ 和 } S^2 = \sum (X_i - \bar{X})^2/(n-1)$$

则有

(1) $\bar{X}$ 与 S^2 相互独立；

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

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

2.4.2 F 分布

❖ **定义2.8** 设$X_1 \sim \chi^2(m), X_2 \sim \chi^2(n), X_1$与$X_2$独立，则称

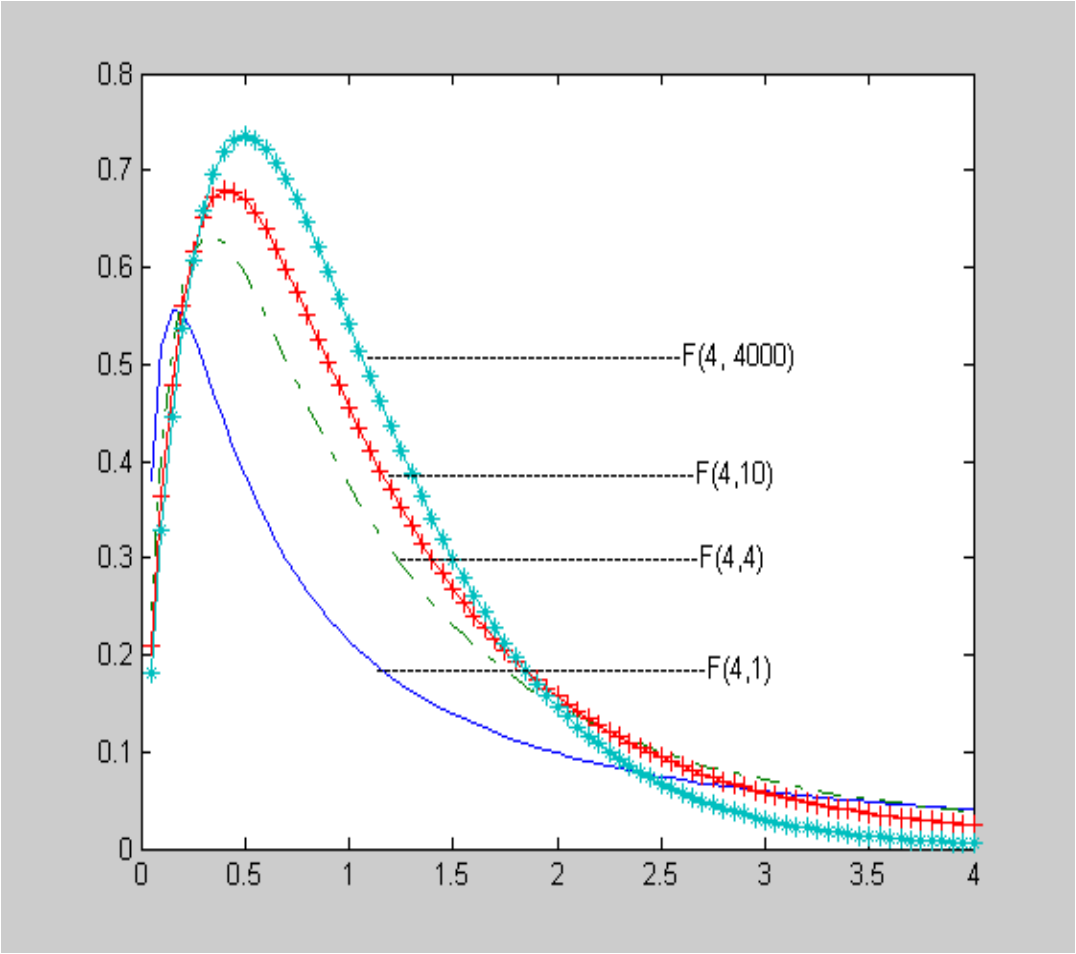
$$F = (X_1/m)/(X_2/n)$$

的分布是自由度为 m与 n的F分布，记为 $F \sim F(m,n)$，其中m 称为分子自由度，n 称为分母自由度。

❖ $1/F$也服从 F 分布，即 $\frac{1}{F} \sim F(n,m)$。

2.4.2 *F* 分布

该密度函数的图象也是一个只取非负值的偏态分布



2.4.2 F 分布

- 当随机变量 $F \sim F(m,n)$ 时，对给定 $\alpha(0<\alpha<1)$，称满足

$$P\left(F \le F_{1-\alpha}(m,n)\right) = 1-\alpha$$

的 $F_{1-\alpha}(m,n)$ 是自由度为 m 与 n 的 F 分布的 $1-\alpha$ 分位数。

- 由 F 分布的构造知 $F_{\alpha}(n,m) = 1/F_{1-\alpha}(m,n)$.

2.4.2 F 分布

- **推论2.1** 设 $X_1, X_2, \ldots, X_m$ 是来自 $N(\mu_1, \sigma_1^2)$ 的样本，$Y_1, Y_2, \ldots, Y_n$ 是来自 $N(\mu_2, \sigma_2^2)$ 的样本，且此两样本相互独立，则有

$$F = \frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} \sim F(m-1, n-1)$$

特别，若 $\sigma_1^2 = \sigma_2^2$，则

$$F = \frac{S_X^2}{S_Y^2} \sim F(m-1, n-1)$$

2.4.3 t分布

❖ **定义 2.9** 设随机变量X_1与X_2 独立，且$X_1 \sim N(0,1)$，$X_2 \sim \chi^2(n)$，则称

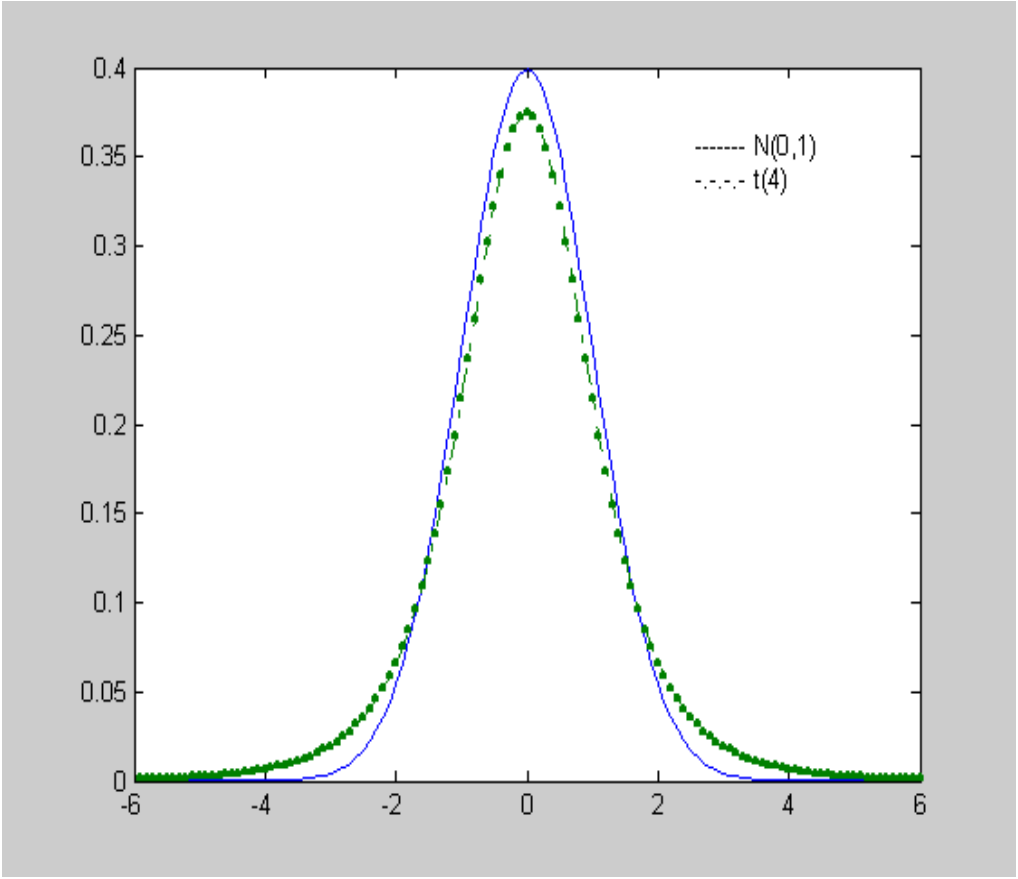
$$T = X_1 / \sqrt{X_2/n}$$

的分布为自由度为n的t分布，记为$T \sim t(n)$ 。

- 自由度为1的 t 分布就是标准柯西分布， 它的均值不存在；
- $n>1$时， t 分布的数学期望存在且为0；
- $n>2$时，t 分布的方差存在，且为$n/(n-2)$；
- 当自由度较大 (如$n \geq 30$) 时， t 分布可以用正态分布$N(0,1)$近似。

2.4.3 t分布

t 分布的密度函数的图象是一个关于纵轴对称的分布，与标准正态分布的密度函数形状类似，只是峰比标准正态分布低一些尾部的概率比标准正态分布的大一些。



2.4.3 t分布

- 当随机变量$T \sim t(n)$时，称满足

$$P(T \le t_{1-\alpha}(n)) = 1-\alpha$$

的$t_{1-\alpha}(n)$是自由度为n的t分布的$1-\alpha$分位数，分位数$t_{1-\alpha}(n)$可以从附表4中查到。

譬如 $n = 10, \alpha = 0.05$，那么从附表4上查得

$$t_{1-0.05}(10) = t_{0.95}(10) = 1.812\ .$$

由于t分布的密度函数关于0对称，故其分位数间有如下关系：

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

2.4.3 t分布

- **推论2.2** 设$X_1, X_2, \ldots, X_n$ 是来自正态分布$N(\mu, \sigma^2)$的一个样本，$\bar{X}$ 与 S^2 分别是该样本的样本均值与样本方差，则有

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t(n-1)$$

2.4.3 t分布

❖ **推论2.3** 在推论2.1的记号下，设

$$\sigma_1^2 = \sigma_2^2 = \sigma^2,$$

并记

$$S_w^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

$$= \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2}{m+n-2}$$

则

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2)$$

§2.4 三大抽样分布

❖ **例2.24** 设$X_1, X_2, \cdots, X_{2n}$取自正态分布总体$N(0,1)$的简单样本，试求：

$$Y_1 = \frac{(2n-3)\sum_{i=1}^{3} X_i^2}{3\sum_{i=4}^{2n} X_i^2},\ Y_2 = \frac{1}{2}\sum_{i=1}^{2n} X_i^2 + \sum_{i=1}^{n} X_{2i-1}X_{2i}$$的分布。

❖ **解：** (1)∵ $X_i \sim N(0,1)$, 且X_i独立同分布

$$\therefore \sum_{i=4}^{2n} X_i^2 \sim \chi^2(2n-3)$$

$$\sum_{i=1}^{3} X_i^2 \sim \chi^2(3)$$

且 $\sum_{i=4}^{2n} X_i^2$ 与 $\sum_{i=1}^{3} X_i^2$ 相互独立；

$$\therefore \quad Y_1 = \frac{(2n-3)\sum_{i=1}^{3} X_i^2}{3\sum_{i=4}^{2n} X_i^2} \sim F(3, 2n-3)$$

§2.4 三大抽样分布

（2）$\because Y_2 = \frac{1}{2}\sum_{i=1}^{2n} X_i^2 + \sum_{i=1}^{n} X_{2i-1}X_{2i}$

$$= \frac{1}{2}\sum_{i=1}^{n}(X_{2i-1} + X_{2i})^2$$

$$= \sum_{i=1}^{n}\left(\frac{X_{2i-1} + X_{2i}}{\sqrt{2}}\right)^2$$

$\because \frac{X_{2i-1} + X_{2i}}{\sqrt{2}} \sim N(0,1), \frac{X_{2i-1} + X_{2i}}{\sqrt{2}}$ 之间相互独立

$$\therefore Y_2 = \sum_{i=1}^{n}\left(\frac{X_{2i-1} + X_{2i}}{\sqrt{2}}\right)^2 \sim \chi^2(n)$$

§2.4 三大抽样分布

❖ **例2.25** 设随机变量$T\sim t(n)$,则$\frac{1}{T^2}\sim F(n,1)$.

❖ **证明：**

$\because T\sim t(n)$,则根据定义T可写$T=\frac{X}{\sqrt{\frac{Y}{n}}}$,

其中$X\sim N(0,1), Y\sim\chi^2(n)$,且$X$和$Y$相互独立,

又$X^2\sim\chi^2(1)$,

$\therefore \frac{1}{T^2}=\frac{Y/n}{X^2}\sim F(n,1)$.

§2.4 三大抽样分布

❖ **例2.26** 设样本$(X_1, \ldots, X_5)$来自于正态总体$N(0,1)$，

（1）试求常数a, 使得$a(X_1 + X_2)^2$服从χ^2分布，并指出自由度；

（2）试求常数b, 使得$\dfrac{b(X_1+X_2)}{\sqrt{X_3^2+X_4^2+X_5^2}}$服从$t$ 分布，并指出自由度；

§2.4 三大抽样分布

❖**解：** （1）

$$\because X_i \sim N(0,1)（i=1,2,3,4,5）且相互独立，$$

$$\therefore (X_1+X_2)\sim N(0,2) \Rightarrow \frac{(X_1+X_2)}{\sqrt{2}}\sim N(0,1)$$

$$\therefore \frac{(X_1+X_2)^2}{2}\sim \chi^2(1)$$

$$\therefore a=\frac{1}{2}, 自由度为1。$$

§2.4 三大抽样分布

（2）$\because (X_3^2+X_4^2+X_5^2)\sim\chi^2(3),\ (X_1+X_2)\sim N(0,2)$，

且$(X_3^2+X_4^2+X_5^2)$与(X_1+X_2)相互独立，

则由t分布的定义可得：$\dfrac{\frac{X_1+X_2}{\sqrt{2}}}{\sqrt{\frac{X_3^2+X_4^2+X_5^2}{3}}}\sim t(3)$

$\therefore b=\sqrt{\dfrac{3}{2}}$，自由度为3.

§2.4 三大抽样分布

- **例2.27** 设 $X_1, X_2 \cdots, X_n, X_{n+1}$ 是取自正态总 $N(\mu, \sigma^2)$ 的一个样本，记

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i\,,\quad S_i^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

求统计量 $\sqrt{\frac{n-1}{n+1}}\frac{X_{n+1}-\bar{X}_n}{S_n}$ 的分布。

§2.4 三大抽样分布

❖解：

$\because \bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$，且$X_{n+1}$与$\bar{X}_n$相互独立，

$$X_{n+1} \sim N(\mu, \sigma^2)$$

$$\therefore (X_{n+1} - \bar{X}_n) \sim N\left(0, \frac{n+1}{n}\sigma^2\right)$$

$$\therefore \frac{X_{n+1} - \bar{X}_n}{\sqrt{\frac{n+1}{n}\sigma^2}} \sim N(0,1)$$

§2.4 三大抽样分布

由$\frac{\sum(X_i-\bar{X})^2}{\sigma^2}\sim\chi^2(n-1)$得$\frac{nS_n^2}{\sigma^2}\sim\chi^2(n-1)$

$\because \bar{X}_n$与S_n^2相互独立，S_n^2与X_{n+1}相互独立$\therefore$ $\frac{X_{n+1}-\bar{X}_n}{\sqrt{\frac{n+1}{n}\sigma^2}}$

与$\frac{nS_n^2}{\sigma^2}$相互独立

$$\therefore \frac{\frac{X_{n+1}-\bar{X}_n}{\sqrt{\frac{n+1}{n}\sigma^2}}}{\sqrt{\frac{nS_n^2}{(n-1)\sigma^2}}}=\sqrt{\frac{n-1}{n+1}}\frac{X_{n+1}-\bar{X}_n}{S_n}\sim t(n-1)$$

§2.5 充分统计量*

2.5.1 充分性的概念

- 例2.28为研究某个运动员的打靶命中率，我们对该运动员进行测试，观测其10次，发现除第三、六次未命中外，其余8次都命中。这样的观测结果包含了两种信息：

(1) 打靶10次命中8次；

(2) 2次不命中分别出现在第3次和第6次打靶上。

2.5.1 充分性的概念*

- 第二种信息对了解该运动员的命中率是没有什么帮助的。一般地，设我们对该运动员进行n 次观测，得到 $X_1, X_2, \ldots, X_n$ ，每个 X_j 取值非0即1，命中为1，不命中为0。令 $T = X_1 + \cdots + X_n$ ，T为观测到的命中次数。在这种场合仅仅记录使用T 不会丢失任何与命中率 θ有关的信息，统计上将这种“样本加工不损失信息”称为“充分性”(sufficiency)。
- 样本$x = (X_1, X_2, \ldots, X_n)$ 有一个样本分布$F_\theta(x)$，这个分布包含了样本中一切有关θ的信息。

2.5.1 充分性的概念*

- 统计量$T = T(X_1, X_2, \ldots, X_n)$也有一个抽样分布$F_\theta^T(t)$，当我们期望用统计量$T$代替原始样本并且不损失任何有关$\theta$的信息时，也就是期望抽样分布$F_\theta^T(t)$像$F_\theta(x)$一样概括了有关$\theta$的一切信息，这即是说在统计量$T$的取值为$t$的情况下样本$X$的条件分布$F_\theta(X|T = t)$已不含$\theta$的信息，这正是统计量具有充分性的含义。

2.5.1 充分性的概念*

- **定义2.10** 设$X_1, X_2, \ldots, X_n$是来自某个总体的样本，总体分布函数为$F(x;\theta)$，统计量$T = T(X_1, X_2, \ldots, X_n)$称为$\theta$的充分统计量(sufficient statistic)，如果在给定T的取值后，$X_1, X_2, \ldots, X_n$的条件分布与θ无关.

2.5.1 充分性的概念*

- 例2.28（续） 设$X_1, X_2, \ldots, X_n$是来自总体$b(1, p)$的样本，证明命中次数，即统计量 $T = X_1 + X_2 + \cdots + X_n$ 为p的充分统计量.
- 解：$T \sim b(n, p)$

$$P(X_1 = x_1, \ldots, X_n = x_n | T = t)$$

$$= \frac{P(X_1 = x_1, \ldots, X_n = x_n, T = t)}{P(T = t)} = \frac{p^t(1-p)^{n-t}}{C_n^t p^t (1-p)^{n-t}}$$

$$= 1/C_n^t$$

$P(X_1 = x_1, \ldots, X_n = x_n | T = t)$与$p$无关。

2.5.2 因子分解定理*

- 充分性原则(sufficiency principle)：在统计学中有一个 基本原则--在充分统计量存在的场合，任何统计推断都可以基于充分统计量进行，这可以简化统计推断的程序。

- **定理2.10** 设总体概率函数为 $f(x;\theta)$，$X_1, X_2, \ldots, X_n$为样本，则 $T = T(X_1, X_2, \ldots, X_n)$为充分统计量的充分必要条件是：存在两个函数$g(t,\theta)$和$h(x_1, x_2, \ldots, x_n)$， 使得对任意的$\theta$和任一组观测值$x_1, x_2, \ldots, x_n$，有

$$f(x_1, x_2, \ldots, x_n; \theta) = g(T(x_1, x_2, \ldots, x_n), \theta)h(x_1, x_2, \ldots, x_n)$$

其中$g(t,\theta)$是通过统计量 T 的取值而依赖于样本的。

2.5.2 因子分解定理*

- 例2.29设$X_1, X_2, \ldots, X_n$是取自总体$U(0,\theta)$的样本，即总体的密度

函数为 $f(x;\theta)=\begin{cases}\frac{1}{\theta}, & 0<x<\theta\\ 0, & \text{其他}\end{cases}$

于是样本的联合密度函数为

$$f(x_1;\theta)\cdots f(x_n;\theta)=\begin{cases}\left(\frac{1}{\theta}\right)^n, & 0<\min\{x_i\}<\max\{x_i\}<\theta\\ 0, & \text{其他}\end{cases}$$

由于诸$x_i>0$，所以我们可将上式改写为

$$f(x_1;\theta)\cdots f(x_n;\theta)=(1/\theta)^n I_{\{x_{(n)}<\theta\}}$$

取$T=X_{(n)}$，并令 $g(t,\theta)=(1/\theta)^n I_{\{x_{(n)}<\theta\}}$，$h(X)=1$

由因子分解定理知$T=X_{(n)}$是θ的充分统计量。

2.5.2 因子分解定理*

- **例2.30** 设$X_1, X_2, \ldots, X_n$是取自总体$N(\mu, \sigma^2)$的样本，$\theta = (\mu, \sigma^2)$是未知的，则联合密度函数为

$$f(x_1, x_2, \ldots, x_n; \theta) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{n\mu^2}{2\sigma^2}\right\} exp\left\{-\frac{1}{2\sigma^2}(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=2}^{n} x_i)\right\}$$

取$t_1 = \sum x_i, t_2 = \sum x_i^2$,并令

$$g(t_1, t_2, \theta) = (2\pi\sigma^2)^{-n/2}\exp\{-n\mu^2/2\sigma^2\} \times \exp\{-(t_2 - 2\mu t_1)/2\sigma^2\},$$

其中$h(X) = 1$,

由因子分解定理，$T = (\sum X_i, \sum X_i^2)$是充分统计量。

2.5.2 因子分解定理*

- 进一步，我们指出这个统计量与$(\bar{X}, S^2)$是一一对应的，这说明在正态总体场合常用的$(\bar{X}, S^2)$是充分统计量。

- 一般地，有如下命题：
- **定理2.11** 若统计量T是充分统计量，统计量S与统计量T一一对应，则统计量S也是充分统计量。

Thank You !