

数理统计

一、什么是数理统计

二、小结

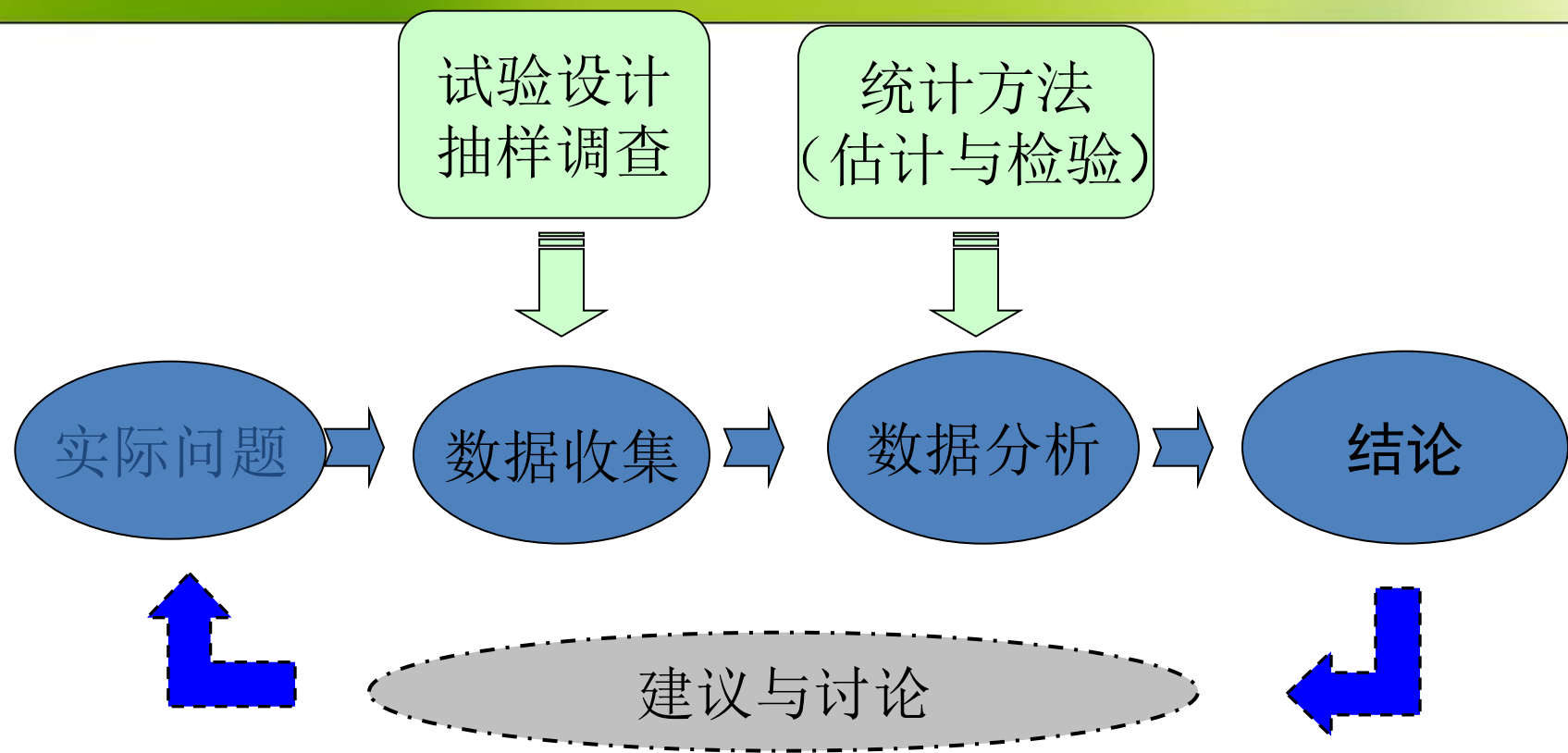


数理统计学

第二次世界大战军事上的需要以及大工业与管理的复杂化产生了运筹学、系统论、信息论、控制论与数理统计学等学科。数理统计学是一门研究怎样去有效地收集、整理和分析带有随机性的数据，以对所考察的问题做出推断或预测，直至为采取一定的决策和行动提供依据和建议的数学分支学科。统计方法的数学理论要用到很多近代数学知识，如函数论、拓扑学、矩阵代数、组合数学等等，但关系最密切的是概率论，故可以这样说：概率论是数理统计学的基础，数理统计学是概率论的一种应用。但是它们是两个并列的数学分支学科，并无从属关系。

数理统计的定义

- 统计学是收集和分析数据的科学与艺术
(不列颠百科全书)
- 统计学是数学的一个分支，它是一门用有效的方法收集和分析带有随机影响的数据的学科，且其目的是解决特定的问题（陈希孺院士）
- 统计学是一门应用性很强的学科，它是研究如何有效地收集、整理和分析受随机影响的数据，并对所考虑的问题作出推断或预测，直至为采取决策和行动提供依据和建议的一门学科。
(茆诗松)



参数统计 (parametric statistics)

已知总体分布类型，对未知参数 (μ 、 π) 进行统计推断

依赖于特定分布类型，比较的是参数

非参数统计 (nonparametric statistics)

对总体的分布类型不作任何要求

不受总体参数的影响，比较分布或分布位置

对于符合参数统计分析条件者，采用非参数统计分析，其检验效能较低



数理统计的内容

- 数理统计基本概念
- 统计量及其分布
- 参数估计
- 假设检验
- 方差分析与回归分析



二、小结

生活中最重要的问题,其中占大多数实际上只是概率问题。

-- 拉普拉斯

在终极的分析中,一切知识都是历史。
在抽象的意义下,一切科学都是数学。
在理性的世界里,所有的判断都是统计学。

-- C.R.劳



学习方法：

- 1、深刻理解，牢固掌握基本概念。
- 2、多做练习，狠抓解题基本功。
- 3、多实践，
学会使用统计软件处理统计数据。



第五章 统计量及其分布

- § 5.1 总体与样本
- § 5.2 样本数据的整理与显示
- § 5.3 统计量及其分布
- § 5.4 三大抽样分布
- § 5.5 充分统计量



数理统计的分类

描述统计学——

对随机现象进行观测、试验，以取得有代表性的观测值

推断统计学——

对已取得的观测值进行整理、分析，作出推断、决策，从而找出所研究的对象的规律性



一、总体与个体

1. **总体** 研究对象的全体称为总体.

2. **个体** 构成总体的每一个成员

实例

1、 在研究2000名学生的年龄时,这些学生的年龄的全体就构成一个总体,每个学生的年龄就是个体.

2、 考察某地区全体居民的身高情况,则该地区所有人的身高便构成一个总体,而每一个人的身高就是一个个体。




一般来说，我们只关心总体的某项数量指标，而这个数量指标的取值通常为一个随机变量，例如：

人的身高数据，其测量值在某个区间附近波动；

灯泡的寿命值总在0到无穷之间，在实验之后才能知道其确切取值；

所以可把这些总体的数量指标视为随机变量，称之为总体 X ，而相应的个体，即研究对象的每一个个体的数量指标也视为随机变量 X_1, X_2, \dots



注1：一个总体对应一个随机变量 X , 我们将不区分总体和相应的随机变量, 统称为总体 X .

X 的分布函数和数字特征称为**总体的分布函数和数字特征**。

注2：若研究的数量指标不止一个时, 则应分为几个总体来研究。



为研究总体的数量指标的统计特性，我们需要从总体中抽出若干个体来进行统计分析，这种方法叫做抽样，那么应当考虑下列两个问题：

1) 为什么要进行抽样，全部抽样行不行？

2) 抽样满足的基本要求是什么？



例5.1.3 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**。



例5.1.4 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：



表5.1.2 100只元件的寿命数据

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

表5.1.2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**。



二、简单随机样本(子样)

定义1.1

从总体中随机抽取的 n 个个体称为容量为 n 的样本,或子样,记为 X_1, X_2, \dots, X_n 。

记 x_i 为 X_i 的一次观察值,并称 (x_1, x_2, \dots, x_n) 为总体 X 的一个容量为 n 的样本观察值,或称样本的一个实现。



定义1.2 设 X_1, X_2, \dots, X_n 是来自总体 X 的容量为 n 的样本, 若

(1) **独立性**: X_1, X_2, \dots, X_n 相互独立;

(2) **同分布性**: X_1, X_2, \dots, X_n 与总体 X 有相同的分布。

则称 X_1, X_2, \dots, X_n 为总体 X 的**简单随机样本**, 或**简单随机子样**, 简称为**样本**或**子样**, 它们的观察值 x_1, x_2, \dots, x_n 称为**样本值**。



- 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, \dots, X_n 表示；

在简单随机抽样下，样本 X_1, X_2, \dots, 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**。



注1 样本容量 n 较大时,称样本为大样本, n 较小时,称为小样本。

注2 对样本 (X_1, X_2, \dots, X_n) 作一次观察所得实数值 (x_1, x_2, \dots, x_n) 称为样本值。



其他抽样方法

- 抽样
 - 从总体中抽取样本的过程
- 抽样方法
 - 概率抽样
 - 简单随机抽样、分层抽样、分群抽样
 - 非概率抽样
 - 便利抽样、判断抽样、配额抽样、滚雪球抽样



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。



例5.1.5 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 。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

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

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

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



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？



三、样本分布

若总体 X 的分布函数 $F(x)$, 则样本 (X_1, X_2, \dots, X_n) 的分布函数为

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



(1) 若总体 X 是离散型随机变量，其概率分布为
 $P\{X = x_i\} = p_i, i = 1, 2, \dots$ ，则样本 (X_1, X_2, \dots, X_n) 的
(联合)概率分布为

$$\begin{aligned} P\{X_1 = x_{k_1}, X_2 = x_{k_2}, \dots, X_n = x_{k_n}\} \\ = \prod_{i=1}^n P\{X = x_{k_i}\} = \prod_{i=1}^n p_{k_i} \end{aligned}$$

(2) 若总体 X 是连续型随机变量，其密度函数为
 $f(x)$ ，则样本 (X_1, X_2, \dots, X_n) 的(联合)密度函数为

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



例4.1 设总体 X 服从参数为 p 的0-1分布, 试求其样本 (X_1, X_2, \dots, X_n) 的概率分布。

解: 总体 X 服从0-1分布, 即

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

因为 X_1, X_2, \dots, X_n 相互独立, 且与 X 有相同的分布, 则样本 (X_1, X_2, \dots, X_n) 的概率分布为

$$\begin{aligned} & P\{(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)\} \\ &= \prod_{i=1}^n P\{X_i = x_i\} = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}, \quad x_i = 0, 1 \quad i = 1, 2, \dots, n \end{aligned}$$



例4.2 设总体 X 服从参数为 α 的指数分布, 试求其样本 (X_1, X_2, \dots, X_n) 的分布。

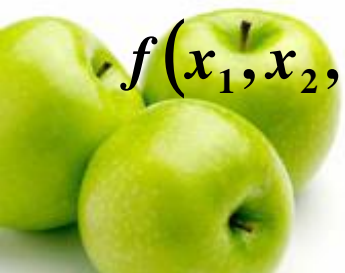
解: 因为 $X \sim E(\alpha)$

$$\text{故 } F(x) = \begin{cases} 1 - e^{-\alpha x} & x > 0 \\ 0 & x \leq 0 \end{cases}, \quad f(x) = \begin{cases} \alpha e^{-\alpha x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

因为 X_1, X_2, \dots, X_n 相互独立, 且与 X 有相同的分布, 故 (X_1, X_2, \dots, X_n) 的分布函数为:

$$F(x_1, x_2, \dots, x_n) = \prod_{i=1}^n F(x_i) = \begin{cases} \prod_{i=1}^n (1 - e^{-\alpha x_i}) & x_i > 0, i = 1, 2, \dots, n \\ 0 & \text{其它} \end{cases}$$

密度函数为:

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i) = \begin{cases} \prod_{i=1}^n \alpha e^{-\alpha x_i} = \alpha^n e^{-\alpha \sum_{i=1}^n x_i} & x_i > 0, i = 1, 2, \dots, n \\ 0 & \text{其它} \end{cases}$$


§ 5.2 样本数据的整理与显示

1. 经验分布函数
2. 频数—频率分布表



一、经验分布函数

定义2.1 设 X_1, X_2, \dots, X_n 为来自总体 X 的样本, $X_1^*, X_2^*, \dots, X_n^*$ 为其顺序统计量, $x_1^*, x_2^*, \dots, x_n^*$ 为顺序观察值, 对于任意实数, 称函数

$$F_n^*(x) = \begin{cases} 0 & x < x_1^* \\ k/n & x_k^* \leq x < x_{k+1}^* \quad k = 1, 2, \dots, n-1 \\ 1 & x \geq x_n^* \end{cases} \quad (2.1)$$

为总体 X 的经验分布函数,

单调不减, 右连续且 $F_n(-\infty) = 0$ 和 $F_n(+\infty) = 1$

2 求经验分布函数步骤

(1) 将所得数据 x_1, x_2, \dots, x_n 按从小到大排列为
顺序统计值 $x_1^* \leq x_2^* \leq \dots \leq x_n^*$

(2) 按 (2.1) 式写 $F_n(x)$

(3) 作出 $F_n(x)$ 的图形。



例

从一批标准重量为500克的罐头中，随机抽取8听，测得误差 X 如下(单位：克)

8、-4、6、-7、-2、1、0、1

试求经验分布函数；近似计算 $P\{X \geq 0\}$.



解：观测值排序 $-7 < -4 < -2 < 0 < 1 = 1 < 6 < 8$

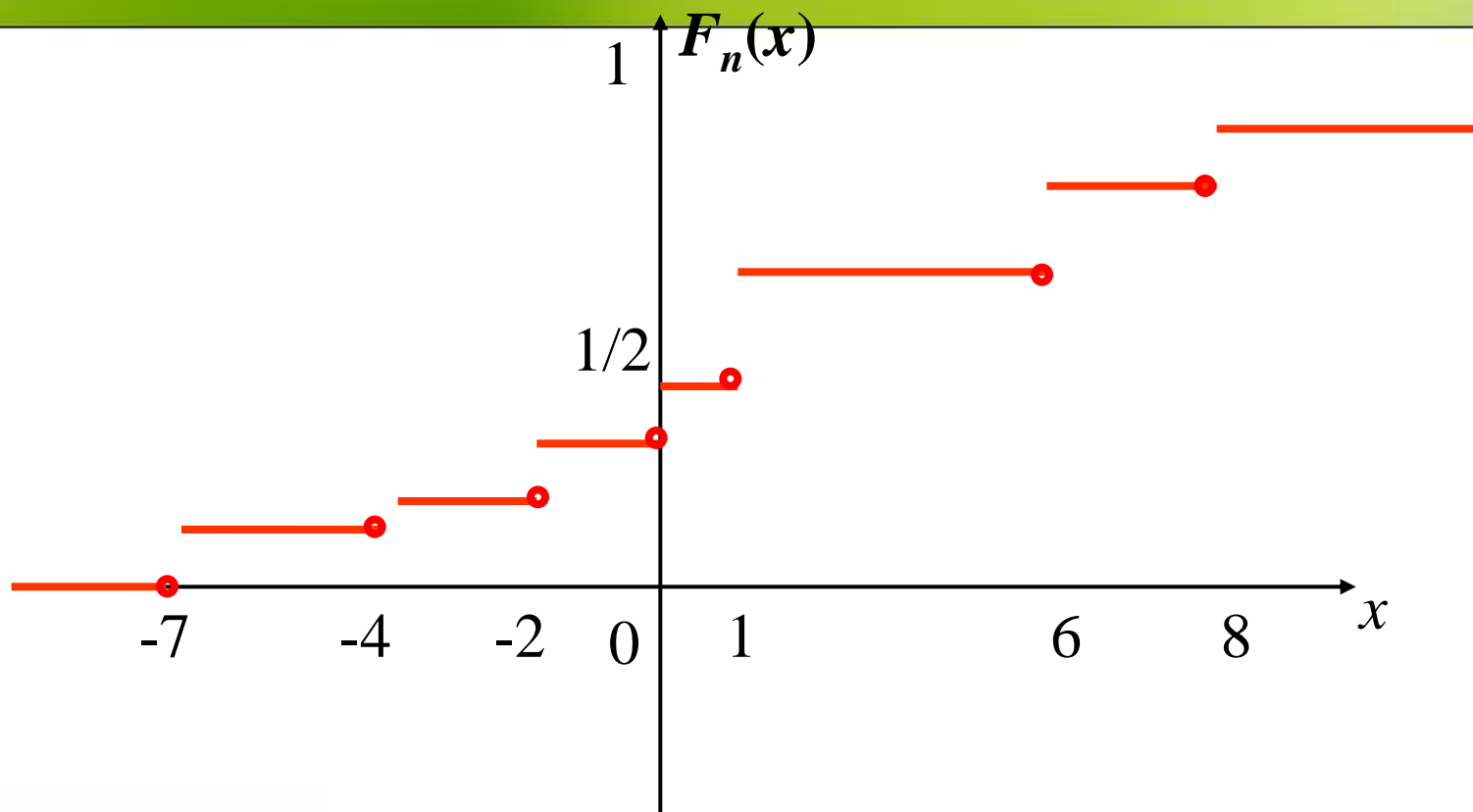
经验分布函数为

$$F_n(x) = \begin{cases} 0 & x < -7 \\ 1/8 & -7 \leq x < -4 \\ 2/8 & -4 \leq x < -2 \\ 3/8 & -2 \leq x < 0 \\ 4/8 & 0 \leq x < 1 \\ 6/8 & 1 \leq x < 6 \\ 7/8 & 6 \leq x < 8 \\ 1 & x \geq 8 \end{cases}$$

根据经验分布函数，有 $P\{X \geq 0\} \approx \frac{5}{8}$



(3) $F_n(x)$ 的图形如下



注1:由(2.1)式可见, $F_n(x)$ 单调,非降,右连续,
且在每个间断点 $x_{(k)}$ 上的跳跃量均为 $\frac{1}{n}$,且
 $0 \leq F_n(x) \leq 1$,具备分布函数的一切性质。



注3:由格列汶科定理知 对于任意实数 x ,有

$$P\{\lim_{n \rightarrow \infty} D_n = 0\} = 1 \quad (2.3)$$

其中 $D_n = \sup_{-\infty < x < \infty} |F_n(x) - F(x)|$

故可用经验分布函数 $F_n(x)$ 作为 $F(x)$ 的近似,
即作为总体 X 的近似分布函数即有

$$P\{a < X \leq b\} = F(b) - F(a) \approx F_n(b) - F_n(a)$$



5.2.2 频数—频率分布表

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5.2.2 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



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;

(2) 确定每组组距: 近似公式为
组距 $d = (\text{最大观测值} - \text{最小观测值}) / \text{组数}$;

(3) 确定每组组限: 各组区间端点为
 $a_0, a_1=a_0+d, a_2=a_0+2d, \dots, a_k=a_0+kd$,
形成如下的分组区间

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

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。

表5.2.1 例5.2.2 的频数频率分布表

组序	分组区间	组中值	频数	频率	累计频率(%)
1	(147, 157]	152	4	0.20	20
2	(157, 167]	162	8	0.40	60
3	(167, 177]	172	5	0.25	85
4	(177, 187]	182	2	0.10	95
5	(187, 197]	192	1	0.05	100
合计			20	1	

5.2.3 样本数据的图形显示

1. 直方图
2. 茎叶图
3. 折线图
4. 散点图



1.2.1 直方图

直方图分频数直方图和频率直方图两类。直方图用横轴表示观测值，并把横轴分成若干个区间（每个区间的宽度称作组距）；用纵轴表示落在相应区间内的观测值频数（个数）或频率，并用矩形（长条形）表示组频数或组频率的图形。

例 1-1： 20 个新生儿体重值（克）数据见表 1-1。画 20 个新生儿体重值的频数（频率）直方图。

表 1-1		新生儿体重值 x_i 数据		单位：克	
序号	体重值	序号	体重值		
1	2440	11	3180		
2	2620	12	3200		
3	2700	13	3200		
4	2880	14	3300		
5	2900	15	3420		
6	3000	16	3440		
7	3020	17	3500		
8	3040	18	3500		
9	3080	19	3600		
10	3100	20	3860		



1.2.1 直方图

例 1-1: 首先把这 20 个新生儿体重值按从小到大顺序排列如下:

2440, 2620, 2700, 2880, 2900, 3000, 3020, 3040, 3080, 3100, 3180, 3200, 3200, 3300, 3420, 3440, 3500, 3500, 3600, 3860。

知最小值是 2440 克, 最大值是 3860 克。把观测值的取值范围按 2400~2700, 2700~3000, 3000~3300, 3300~3600, 3600~3900 分成 5 组。记录这 20 个观测值分别落在这 5 个组内的频数 (个数)。结果分别是 2, 3, 8, 5, 2。用总观测值个数 20 除每个组频数, 得组频率值分别是 0.10, 0.15, 0.40, 0.25, 0.10。用上面的结果制成频数 (频率) 分布表 (见表 1-2)。

表 1-2 20 个新生儿体重值分组数据频数 (频率) 分布表

体重值 (克)	频数	频率	组中值 (克)
2400—2700 以下	2	0.10	2550
2700—3000 以下	3	0.15	2850
3000—3300 以下	8	0.40	3150
3300—3600 以下	5	0.25	3450
3600—3900 以下	2	0.10	3750
合计	20	1.00	——



例 1-1:

表 1-2 20 个新生儿体重值分组数据频数（频率）分布表

体重值（克）	频数	频率	组中值（克）
2400—2700 以下	2	0.10	2550
2700—3000 以下	3	0.15	2850
3000—3300 以下	8	0.40	3150
3300—3600 以下	5	0.25	3450
3600—3900 以下	2	0.10	3750
合计	20	1.00	——

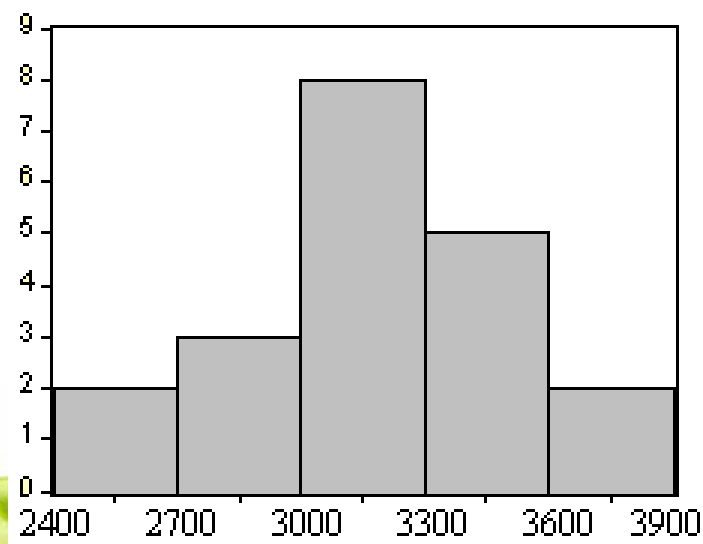


图 1-1 新生儿体重值的频数分布直方图

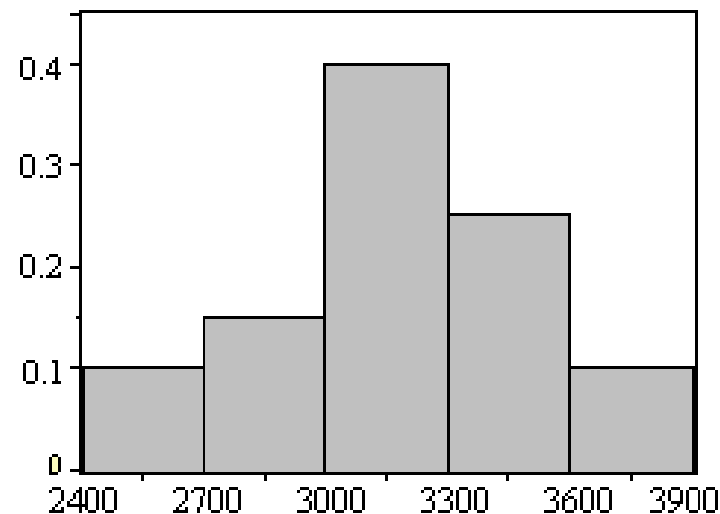


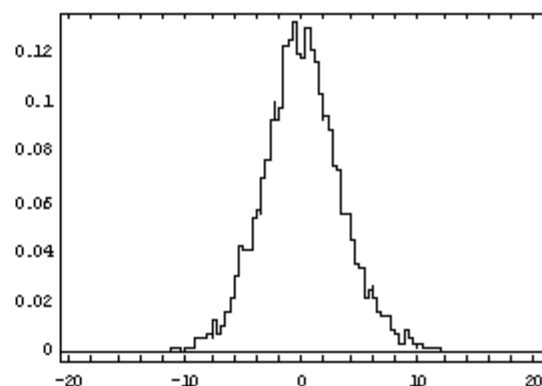
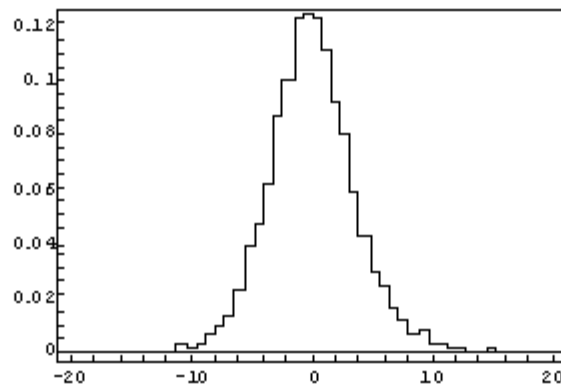
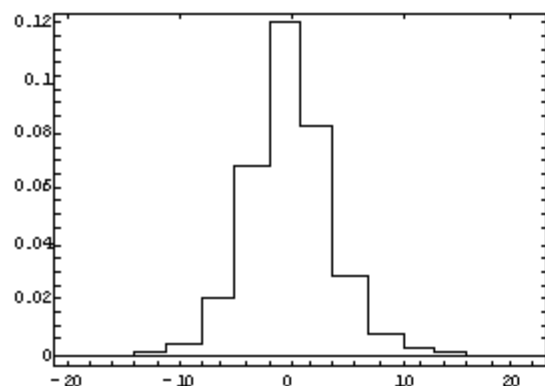
图 1-2 新生儿体重值的频率分布直方图

注意：

(1) 频数、频率直方图所展示的数据分布特征是一样的，只不过前者的纵轴表示的是频数，后者纵轴表示的是频率。

(2) 当观测值正巧等于组边界值时，注意不要在相邻两组中重复记录频数。以表 1-2 为例，记录组频数的规则是组下限值包括在本组内，组上限值不包括在本组内。比如观测值 2700 克正巧落在组边界值上。观测值 2700 克应该记录在第 2 组，而不是第 1 组中。观测值 3000 克也正巧落在组边界值上。观测值 3000 克应该记录在第 3 组，而不是第 2 组中。

(3) 同样一组数据由于分组数不同，所画频数（频率）直方图的特征会不一样。实际中应该选择一个最合适的分组数，以便充分展示数据的分布特征。一般分组数在 5~15 之间。



(4) 很多专用软件都有画直方图的功能，非常方便。

二、茎叶图

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



例5.2.3 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。



[illegible]

图5.2.3 测试成绩的茎叶图



在要比较两组样本时，
可画出它们的背靠背的茎叶图。

甲车间	6 2 0	5	6	乙车间
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。

1.2.2 折线图

折线图：把观测点按序号或时间顺序用直线连接起来的图形。

对于截面数据，横轴表示观测值的序号，纵轴表示观测值。对于时间序列数据，横轴表示时间，纵轴表示观测值。时间序列折线图也称时间序列图。

图 1-3 给出的是 2005 年 7 月 22 日至 2007 年 4 月 30 日 433 天的美元兑人民币元汇率值时间序列图。通过这张图可以清晰地看到在该期间人民币一直处于升值的大趋势中。

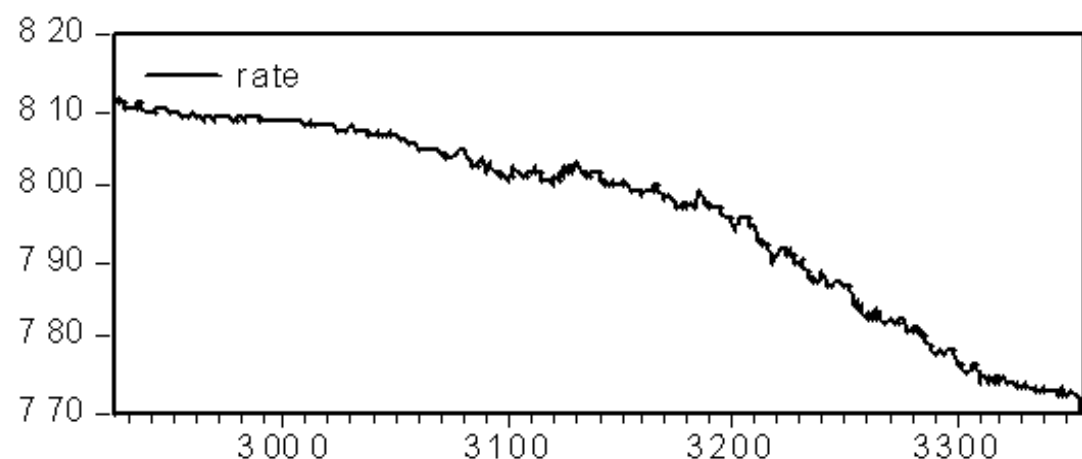


图 1-3 2005 年 7 月 22 日至 2007 年 4 月 30 日 433 天的人民币元兑美元汇率值时间序列图

1.2.3 散点图

散点图：用两个变量的成对观测值画出的观测点图。

通过散点图可以分析两个变量之间是否存在某种关系。如果存在关系，那么这种关系是线性的，还是非线性的。

图 1-4 给出的是 2002 年中国各地区城镇居民家庭人均消费性支出（Y2002，元）与可支配收入（X2002，元）数据散点图。右上方 4 个观测点分别代表北京、上海、浙江省和广东省。通过散点图可以清楚地看到经济相对发达地区的城镇居民家庭人均支出、可支配收入额都很高；经济相对欠发达地区，如甘肃、宁夏、青海、内蒙古等城镇居民家庭人均支出、可支配收入额都相对较低。

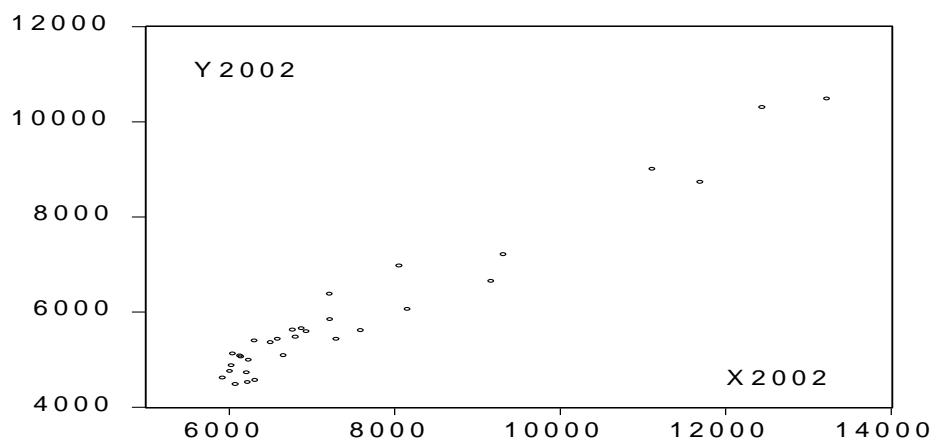


图 1-4 2002 年中国城镇居民家庭人均支出（ y_{2002} ）与可支配收入（ x_{2002} ）散点图

§ 5.3 统计量及其分布

定义5.3.1 设 (X_1, X_2, \dots, X_n) 为总体 X 的样本, 若样本的函数 $T=T(X_1, X_2, \dots, X_n)$ 中不含任何未知参数, 则 $T(X_1, X_2, \dots, X_n)$ 称为**统计量**, 其中 T 为连续函数。

若 (x_1, x_2, \dots, x_n) 为样本 (X_1, X_2, \dots, X_n) 的一次观察值, 则 T 的取值 $t=T(x_1, x_2, \dots, x_n)$ 称为**统计值**。



例1.4 设 (X_1, X_2, \dots, X_n) 是来自正态总体 $N(\mu, \sigma^2)$ 的容量为 n 的样本, 其中 μ 已知, σ 未知, 则

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

$$X_1 + X_2, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

均为 **统计量**。

$$(2) \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2, \quad \frac{X_1}{\sigma} + \frac{X_2}{\sigma},$$

均为样本 (X_1, X_2, \dots, X_n) 的函数,
当其中 σ 未知时, 它们都不是统计量。

2、描述样本的中心位置的统计量

(1) 样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ (1.6)

观察值为

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

样本均值处于样本的中间位置，
它可以反映总体分布的均值。

思考：在分组样本场合，样本均值如何计算？



例 下表是经过整理的分组数据表,给出了110个电子元件的失效时间

组中值 x_i	200	600	1000	1400
频数 f_i	6	28	37	23
组中值 x_i	1800	2200	2600	3000
频数 f_i	9	5	1	1

表中的组中值 x_i 为每一小组 $[t_{i-1}, t_i)$ 的中值,
即

$$x_i = \frac{t_{i-1} + t_i}{2}$$



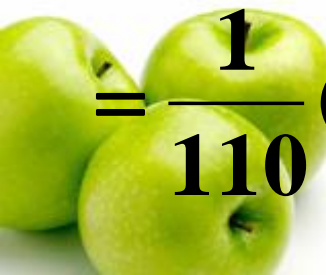
如此处第一个分组区间可视为 $[0,400)$,

故组中值为 $\frac{0+400}{2} = 200$

那么,平均失效时间近似为

$$\bar{X} = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i}$$

注意,此处为加权平均


$$= \frac{1}{110} (200 \times 6 + 600 \times 28 + \cdots + 3000 \times 1) = 1090.9$$

1、样本均值的基本性质：

定理5.3.1 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5.3.2 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) 样本中位数

$$M_n = \begin{cases} X_{k+1}^* & \text{当 } n = 2k + 1 \\ \frac{1}{2}(X_k^* + X_{k+1}^*) & \text{当 } n = 2k \end{cases}$$

例1.7 某工厂制作一种线圈, 为控制生产过程保持稳定, 从产品中任取10件, 测定其阻抗 X , 所得数据如下:

15.3, 13.0, 16.7, 14.2, 14.5, 14.5, 15.9, 15.0, 15.1, 16.4

试求: (1) 样本中位数 M_n 的值

(2) 若取出的第11件数据为15.2, 此时 M_n 又为何值?

解: 将所得数据按从小到大顺序排列为:

13.0, 14.2, 14.5, 14.5, 15.0, 15.1, 15.3, 15.9, 16.4, 16.7

(1) 可见 $n = 10 = 2k$ $k = 5$

此时, $M_{10} = (X_5^* + X_6^*) / 2 = (15.0 + 15.1) / 2 = 15.05$

(2) $X_{11} = 15.2$ 时, $n = 11 = 2k + 1$ $k = 5$

$M_{11} = X_6^* = 15.1$

算术平均数虽然对一组数据有代表性，但当数据分布不对称时，算术平均数的代表性很差。比如 1993 年 2 月至 1994 年 1 月人民币元对美元汇率数据 见图

12 个汇率值，有 11 个落在了 5.5~6.0 之间；1 个落在了 8.5~9.0 之间。

计算美元兑人民币元汇率数据的算术平均数

$$\bar{x} = \frac{5.7402 + 5.7190 + \dots + 8.7000}{12} = 6.00745$$

算术平均数 6.00745 显然没有代表性。因为这 12 个数据中，有 11 个值都小于 6.00745，此例说明算术平均数的计算受离群值（这里是 8.7000）影响很大。若没有 8.7000 这个极端值，平均数应该是 5.7627。为避免这种不合理，在数据分布不对称时，使用中位数评价数据更好些。

例 1-5：以表 1-4 中数据为例，从小到大整理后的数据如表 1-5。

表1-5 表1-4中12个汇率值的从小到大排列

序号	人民币元兑美元汇率值
1	5.7084
2	5.7190
3	5.7290
4	5.7402
5	5.7612
6	5.7612
7	5.7868
8	5.7868
9	5.7918
10	5.8000
11	5.8050
12	8.7000

$$Md = \frac{1}{2} (x_{12/2} + x_{(12/2)+1}) =$$

$$\frac{1}{2} (x_6 + x_7) = \frac{5.7612 + 5.7868}{2} = 5.774$$

中位数 5.774 比平均数 6.00745 的代表性要好。

(3)众数(mod): 数据中最常出现的值, 即样本中出现可能性最大的值, 不过, 众数可能不唯一。

例1.8 现有一数据集合: {2, 3, 3, 3, 3, 4, 4, 5, 6, 6, 6, 6, 6, 7, 7, 8}, 求其众数。

解: 集合中每一个值出现的次数如下表

数值	2	3	4	5	6	7	8
出现次数	1	4	2	1	5	2	1

故其众数为6。



3 描述样本数据分散程度的统计量

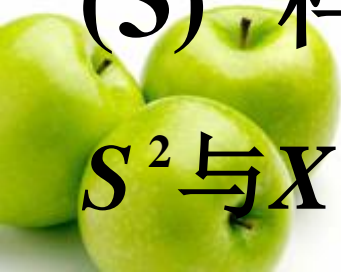
反映样本数据分散程度的统计量实际上反映了总体取值的分散程度常用统计量有以下几种

(1) 样本极差 $D_n^* = X_n^* - X_1^* \quad (1.8)$

(2) 样本方差 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

(3) 样本标准差 $S = \sqrt{S^2} \quad (1.9)$

S^2 与 X 的量纲不一致而 S 与 X 的量纲是一致的



(4) 变异系数

$$C_r = \frac{S}{\bar{X}} \quad (1.10)$$

变异系数用于不同数据集的分散程度的比较

例如测得北京到上海的平均距离1463公里, 测量误差的标准为公里, 而测得一张桌子的平均长度为1米, 测量误差的标准为0.01米, 两者的变异系数为

$$C_1 = \frac{1}{1463} = 0.000684 = 0.0684\%$$

$$C_2 = \frac{0.01}{1} = 0.01 = 1\%$$

故知前者测量的精度比后者高。



4 样本矩

对不同的总体矩有相应的样本矩

(1) $\overline{X^k} = \frac{1}{n} \sum_{i=1}^n X_i^k$ 为样本 k 阶原点矩

(2) $S_n^k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$ 为样本 k 阶中心矩

(3) $\overline{X^k Y^l} = \frac{1}{n} \sum_{i=1}^n X_i^k Y_i^l$ 为样本 $k + l$ 阶混合原点矩

(4) $S_{XY}^{k+l} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k (Y_i - \bar{Y})^l$

为样本 $k + l$ 阶混合中心矩



注1: 若用 $|X_i|$ 代替上式中 X_i , 所得式称为样本绝对矩

例如 $\frac{1}{n} \sum_{i=1}^n |X_i|^k$ 为样本 k 阶原点绝对矩



注2 样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

可作为总体均值的近似

样本方差 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2,$

可作为总体方差的近似

样本1+1阶混合中心矩 $S_{XY} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$

则可作为协方差 $Cov(X, Y)$ 的近似



样本偏度和样本峰度

当总体关于分布中心对称时，样本均值和样本方差刻画样本特征很有代表性，而当其不对称时，只用样本均值和样本方差就显得很不够。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： $\gamma_1 = b_3/b_2^{3/2}$ 称为样本偏度，
 $\gamma_2 = b_4/b_2^2$ 称为样本峰度。 $b_k = \sum (X_i - \bar{X})^k/n$

样本偏度 γ_1 反映了总体分布密度曲线的对称性信息。
样本峰度 γ_2 反映了总体分布密度曲线在其峰值附近的陡峭程度。



5 顺序统计量

设 (X_1, X_2, \dots, X_n) 为总体 X 的样本, 把

它们按从小到大的次序排列为

$$X_1^* \leq X_2^* \leq \dots \leq X_n^* \quad (1.4)$$

则称 $X_1^*, X_2^*, \dots, X_n^*$ 为原样本 (X_1, X_2, \dots, X_n) 的顺序统计量

称 X_k^* 为第 k 个顺序统计量 $(1 \leq k \leq n)$ 。

若样本值为 x_1, x_2, \dots, x_n , 则按从小到大顺序排列后得到顺序统计值

$$x_1^* \leq x_2^* \leq \dots \leq x_n^* \quad (1.5)$$



注1 顺序统计量保留了原样本的数据信息, 只去掉了不太重要的得到数据的顺序信息。

注2 X_k^* 意味着在 n 个数据中, 恰有 k 个数据不超过它, 即超过它的恰有 $n - k$ 个数据。

易见 $X_1^* = \min\{X_1, X_2, \dots, X_n\}$

$X_n^* = \max\{X_1, X_2, \dots, X_n\}$



注3 若已知总体 X 具有分布函数 $F(x)$,由概率论知识知

(1) X_1^* 的分布函数为

$$F_N(x) = 1 - (1 - F(x))^n$$

(2) X_n^* 的分布函数为

$$F_M(x) = [F(x)]^n$$



例4.6 设 X_1, X_2, \dots, X_5 是容量为5的样本, 今对样本作两次观察, 其值如下表

$x_i \backslash X_i$	X_1	X_2	X_3	X_4	X_5
1	3	1	10	5	6
2	2	6	7	2	8

试求两 观察的顺序统计值。

解：将上表中数据从小到大排列，即得顺序统计值。

$x_i \backslash X_{(i)}$	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$
1	1	3	5	6	10
2	2	2	6	7	8



注：样本方差的定义

样本方差定义中， $\sum (x_i - \bar{x})^2$ 称为**偏差平方和**， $n-1$ 称为偏差平方和的**自由度**。其含义是：在 \bar{x} 确定后， n 个偏差 $x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}$ 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：

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

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？



常见统计量的分布

统计量的分布称为抽样分布。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。



一、样本均值的分布

1、单个正态总体下的样本均值的分布

定理 设总体 X 服从正态总体 $N(\mu, \sigma^2)$, $X_1, X_2, \dots,$

X_n , 为来自 X 的一个样本, 则样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

服从均值为 μ , 方差为 $\frac{\sigma^2}{n}$ 的正态分布, 即

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad (3.1)$$



注1: 由于 X_1, X_2, \dots, X_n 与 X 相互独立同分布:

故 $E(X_i) = \mu, D(X_i) = \sigma^2, i = 1, 2, \dots, n,$

于是有

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \cdot n\mu = \mu \quad (3.2)$$

$$D(\bar{X}) = D\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n D(X_i) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n} \quad (3.3)$$



注2:由上述可知 \bar{X} 与 X 有相同的数学期望但 \bar{X} 的方差却只是 X 的方差的 $\frac{1}{n}$ 倍,即 \bar{X} 的取值于 μ 的集中程度远比总体要高且当 $n \rightarrow \infty$ 时, \bar{X} 取值几乎集中在数学期望 μ 这点处。



例 在总体 $N(52, 6.3^2)$ 中随机抽取一容量为36的样本，试求样本均值落在50.8到53.8之间的概率。

解：因 $X \sim N(52, 6.3^2)$, $n = 36$, 则

$$\bar{X} = \frac{1}{36} \sum_{i=1}^{36} X_i \sim N\left(52, \frac{6.3^2}{36}\right) = N(52, 1.05^2)$$

$$\begin{aligned} \text{故 } P\{50.8 < \bar{X} < 53.8\} &= \Phi\left(\frac{53.8 - 52}{1.05}\right) - \Phi\left(\frac{50.8 - 52}{1.05}\right) \\ &= \Phi(1.71) - \Phi(-1.14) = 0.8293 \end{aligned}$$



例 设 $X \sim N(72, 100)$, 为使样本均值大于70的概率不小于90%, 则样本容量至少取多少?

解 设样本容量为 n , 则 $\bar{X} \sim N(72, \frac{100}{n})$

$$\begin{aligned} \text{故 } P(\bar{X} > 70) &= 1 - P(\bar{X} \leq 70) \\ &= 1 - \Phi\left(\frac{70 - 72}{\frac{10}{\sqrt{n}}}\right) = \Phi(0.2\sqrt{n}) \end{aligned}$$

$$\text{令 } \Phi(0.2\sqrt{n}) \geq 0.9 \quad \text{得} \quad 0.2\sqrt{n} \geq 1.29$$

$$\text{即 } n \geq 41.6025 \quad \text{所以取 } n = 42$$



2、非正态总体下的样本均值的近似分布

定理3.3 设 X 为任意总体，其数学期望为 $E(X) = \mu$ ，方差为 $D(X) = \sigma^2$ ， X_1, X_2, \dots, X_n 为来自总体 X 的一个样本，当 n 较大时，近似的有

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad (3.6)$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad (3.7)$$

注：本定理利用中心极限定理可证。



例 若总体 X 的期望 $E(X) = \mu$, 方差 $D(X) = \sigma^2$, 任取 X 的容量为 n 的样本, 样本均值为 \bar{X} , 试问 n 多大时, 有

$$P\{|\bar{X} - \mu| < 0.1\sigma\} \geq 0.95$$

解: 由(3.5)式, \bar{X} 近似服从正态分布 $N(\mu, \frac{\sigma^2}{n})$, 故

$$P\{|\bar{X} - \mu| < 0.1\sigma\} = P\left\{\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} < 0.1\sqrt{n}\right\} \geq 0.95$$

得 $2\Phi(0.1\sqrt{n}) - 1 \geq 0.95$

$$\Phi(0.1\sqrt{n}) \geq 0.975$$

查表知 $0.1\sqrt{n} \geq 1.96$, 即 $n \geq 385$



3、两个正态总体下的样本均值的分布

定理3.2 设两个正态总体 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$, X 与 Y 相互独立 (X_1, X_2, \dots, X_n) 与 (Y_1, Y_2, \dots, Y_n) 为分别来自 X 与 Y 的简单随机样本 \bar{X}, \bar{Y} 分别为其样本均值, 则

$$\bar{X} \pm \bar{Y} \sim N(\mu_1 \pm \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}) \quad (3.4)$$

$$\frac{\bar{X} \pm \bar{Y} - (\mu_1 \pm \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0, 1) \quad (3.5)$$



例 试求总体 $N(20,3)$ 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。

解：设两独立样本均值分别为 \bar{X}, \bar{Y} , 则

$$\bar{X} \sim N(20, \frac{3}{10}) \quad \bar{Y} \sim N(20, \frac{3}{15})$$

由(3.3)式得

$$\bar{X} - \bar{Y} \sim N(0, \frac{3}{10} + \frac{3}{15}) = N(0, \frac{1}{2})$$

$$\text{所求概率 } P\{|\bar{X} - \bar{Y}| > 0.3\} = 1 - P\{|\bar{X} - \bar{Y}| \leq 0.3\}$$

$$= 1 - P\{-0.3 \leq \bar{X} - \bar{Y} \leq 0.3\}$$

$$= 1 - \left[\Phi\left(\frac{0.3 - 0}{\sqrt{1/2}}\right) - \Phi\left(\frac{-0.3 - 0}{\sqrt{1/2}}\right) \right]$$

$$= 1 - [2\Phi(0.3\sqrt{2}) - 1] = 2[1 - \Phi(0.3\sqrt{2})]$$

$$= 2[1 - \Phi(0.42)] = 2[1 - 0.6628] = 0.6744$$



二、样本方差的分布

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5.3.4 设总体 X 具有二阶矩，即

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

X_1, X_2, \dots, 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

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



三、次序统计量的分布

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

例5.3.6 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5.3.6列出了这些值，由此



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

$x_{(1)}$	0	1	2
p	$\frac{19}{27}$	$\frac{7}{27}$	$\frac{1}{27}$

$x_{(2)}$	0	1	2
p	$\frac{7}{27}$	$\frac{13}{27}$	$\frac{7}{27}$

$x_{(3)}$	0	1	2
p	$\frac{1}{27}$	$\frac{7}{27}$	$\frac{19}{27}$

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。



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

$X_{(1)} \backslash X_{(2)}$	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

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

二者不等,

由此可看出 $X_{(1)}$ 和 $X_{(2)}$ 是不独立的。



二、单个次序统计量的分布

定理5.3.5 设总体 X 的密度函数为 $f(x)$ ，分布函数为 $F(x)$ ， X_1, X_2, \dots, X_n 为样本，则第 k 个次序统计量 $X_{(k)}$ 的密度函数为

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



例5.3.7 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 $P(X_{(2)}<1/2)$ 。



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5.3.6 在定理5.3.5的记号下，次序统计量 $(X_{(i)}, X_{(j)}), (i < j)$ 的联合分布密度函数

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



次序统计量的函数在实际中经常用到。

如 样本极差 $R_n = X_{(n)} - X_{(1)}$,

样本中程 $[X_{(n)} - X_{(1)}]/2$ 。

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。



例5.3.9 设总体分布为 $U(0,1)$, X_1, X_2, \dots, X_n 为样本, 则 $(X_{(n)}, X_{(1)})$ 的联合密度函数为

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

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

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

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。



5.3.6 样本分位数与样本中位数

样本中位数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：

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



样本分位数与样本中位数的分布

定理5.3.7 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 $p(x_p) > 0$, 则当 $n \rightarrow \infty$ 时样本 p 分位数 m_p 的渐近分布为

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

特别, 对样本中位数, 当 $n \rightarrow \infty$ 时近似地有

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



例5.3.10 设总体为柯西分布，密度函数为

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

不难看出 θ 是该总体的中位数，即 $X_{0.5} = \theta$ 。

设 X_1, X_2, \dots, 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

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



5.3.7 五数概括与箱线图

次序统计量的应用之一是五数概括与箱线图。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 $m_{0.5}$ ，

第一四分位数 $Q_1 = m_{0.25}$ ，第三四分位数 $Q_3 =$

$m_{0.75}$ 。所谓五数概括就是指用这五个数：

$$x_{\min}, \quad Q_1, \quad m_{0.5}, \quad Q_3, \quad x_{\max}$$

来大致描述一批数据的轮廓。



§ 5.4 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”。



一、 χ^2 -分布

二、 t -分布

三、 F -分布



下分位点 设随机变量 X 的分布函数为

$$F(x) = P\{X \leq x\},$$

对于任一正数 α , ($0 < \alpha < 1$),

若 X 大于等于某实数 x_α 的概率为 α , 即

$$P\{X \leq x_\alpha\} = F(x_\alpha) = \alpha, \quad (0 < \alpha < 1)$$



一、 χ^2 -分布

1、 χ^2 -分布定义

定义3.1 若随机变量 X 的概率密度为

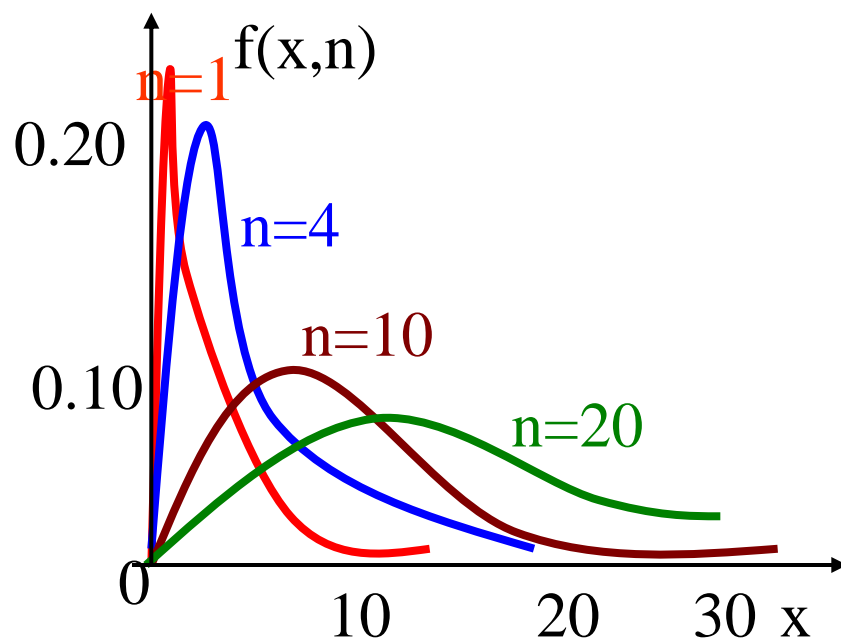
$$f(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & x > 0 \\ 0 & x \leq 0 \end{cases} \quad (3.8)$$

则称 X 服从自由度为 n 的 χ^2 分布,记为 $X \sim \chi^2(n)$ 。



注： $\chi^2(n)$ 分布实际上为参数是 $\frac{n}{2}, \frac{1}{2}$ 的 Γ 分布 $\Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$ 。

$\chi^2(n)$ 的密度图形如下，其形状与 n 有关。



2、 χ^2 - 分布的典型模式

定理 若随机变量 X_1, X_2, \dots, X_n 相互独立，且都服从标准正态分布 $N(0,1)$ ，则随机变量

$$Y = \sum_{i=1}^n X_i^2 \quad (3.9)$$

服从自由度为 n 的 χ^2 分布 $\chi^2(n)$ 。



3、 χ^2 -分布的性质

(1) $\chi^2(n)$ 的数字特征

$$E[\chi^2(n)] = n \quad D[\chi^2(n)] = 2n$$

(2) $\chi^2(n)$ 分布对参数 n 具有可加性

设 $\chi_1^2 \sim \chi_1^2(n_1), \chi_2^2 \sim \chi_2^2(n_2)$ 且相互独立

$$\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$$



例5.1 设总体 $X \sim N(0,1)$, 从总体中取一个容量为6的样本 X_1, X_2, \dots, X_6 , 设 $Y = (X_1 + X_2 + X_3)^2 + (X_4 + X_5 + X_6)^2$ 试确定常数 C , 使随机变量 CY 服从 χ^2 分布.

证明: 由题: $X_i \sim N(0,1)$ $X_1 + X_2 + X_3 \sim N(0,3)$

$$\frac{X_1 + X_2 + X_3}{\sqrt{3}} \sim N(0,1)$$

$$\frac{X_4 + X_5 + X_6}{\sqrt{3}} \sim N(0,1)$$



$$\frac{X_1 + X_2 + X_3}{\sqrt{3}} \sim N(0,1) \quad \frac{X_4 + X_5 + X_6}{\sqrt{3}} \sim N(0,1)$$

那么,

$$\begin{aligned} & \left(\frac{X_1 + X_2 + X_3}{\sqrt{3}} \right)^2 + \left(\frac{X_4 + X_5 + X_6}{\sqrt{3}} \right)^2 \\ &= \frac{1}{3} (X_1 + X_2 + X_3)^2 + \frac{1}{3} (X_4 + X_5 + X_6)^2 \\ &= \frac{1}{3} Y \sim \chi^2(2) \quad \text{故: } C=1/3 \end{aligned}$$



二、t-分布

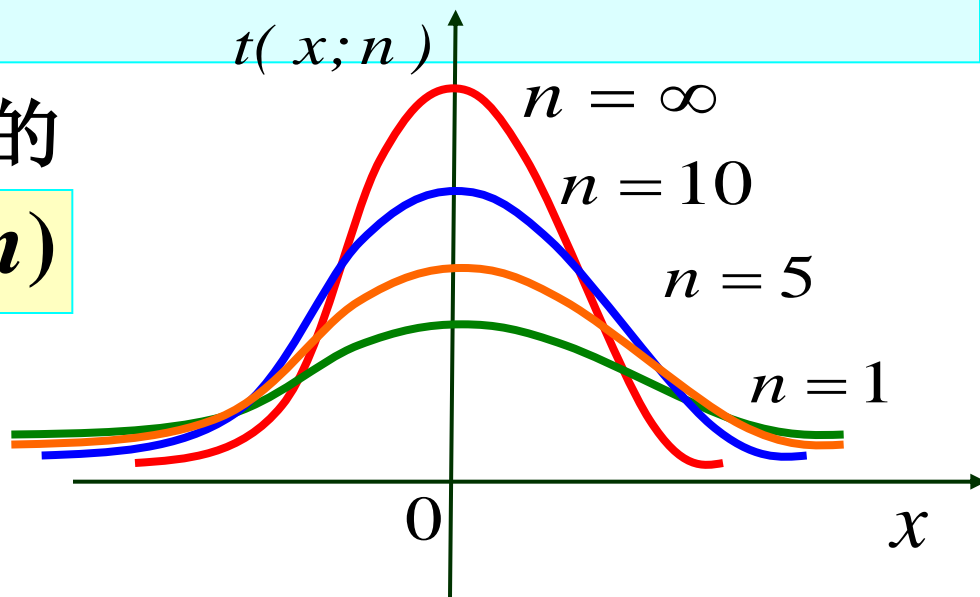
1、t 分布的定义

定义 若随机变量 X 的密度函数为

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad (-\infty < x < +\infty) \quad (3.12)$$

则称 X 服从自由度为 n 的
 t 分布，记作 $X \sim t(n)$

其密度函数图形如图



2、t(n)的典型模式

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

$$T = \frac{X}{\sqrt{Y/n}} \quad (3.15)$$

服从自由度为 n 的 t 分布, 记为 $T \sim t(n)$ 。



3、 $t(n)$ 的性质

(1) $t(n)$ 分布的密度曲线关于轴对称，即有

$$f(-x) = f(x)$$

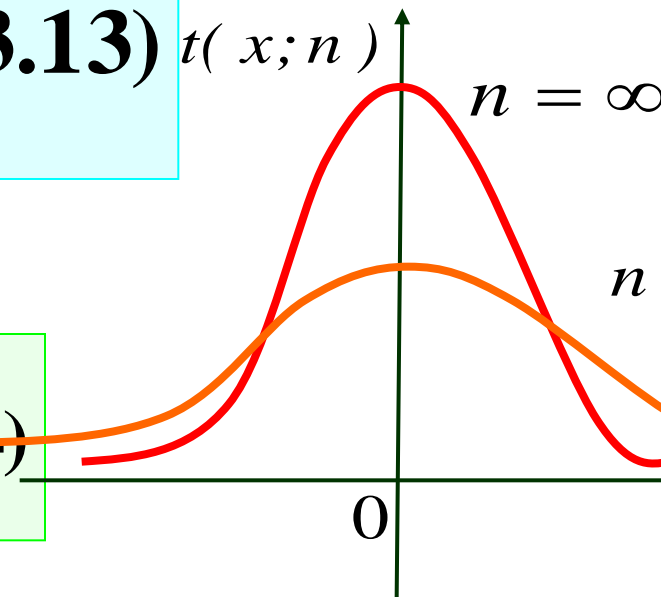
且与标准正态分布 $N(0,1)$ 密度曲线十分相近，但 $t(n)$ 分布密度曲线的峰顶要低，两端点较标准正态曲线较高。

(2) $t(n)$ 的极限分布为标准正态分布，即

$$\lim_{n \rightarrow \infty} f_{t(n)}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (3.13) \quad t(x; n)$$

(3) $t(n)$ 分布的数字特征

$$E[t(n)] = 0 \quad D[t(n)] = \frac{n}{n-2} \quad (3.14)$$



- 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。



例 设总体 X 服从标准正态分布 $N(0,1)$, 样本 (X_1, X_2, \dots, X_5) 来自总体 X , 试求常数 C , 使统计量

$$\frac{C(X_1 + X_2)}{\sqrt{X_3^2 + X_4^2 + X_5^2}} \quad \text{服从} t \text{分布。}$$

解: 因为 X_1, X_2, \dots, X_5 相互独立, 且同标准正态分布 $N(0,1)$

$$\text{故 } X_1 + X_2 \sim N(0, 2) \quad \frac{X_1 + X_2}{\sqrt{2}} \sim N(0, 1)$$

$$X_i^2 \sim \chi^2(1) (i = 1, 2, 3) \quad X_3^2 + X_4^2 + X_5^2 \sim \chi^2(3)$$

$$\text{故 } \frac{X_1 + X_2 / \sqrt{2}}{\sqrt{X_3^2 + X_4^2 + X_5^2} / \sqrt{3}} \sim t(3)$$

易见 $C = \sqrt{\frac{3}{2}}$ 。



三、F-分布

1、F分布的定义

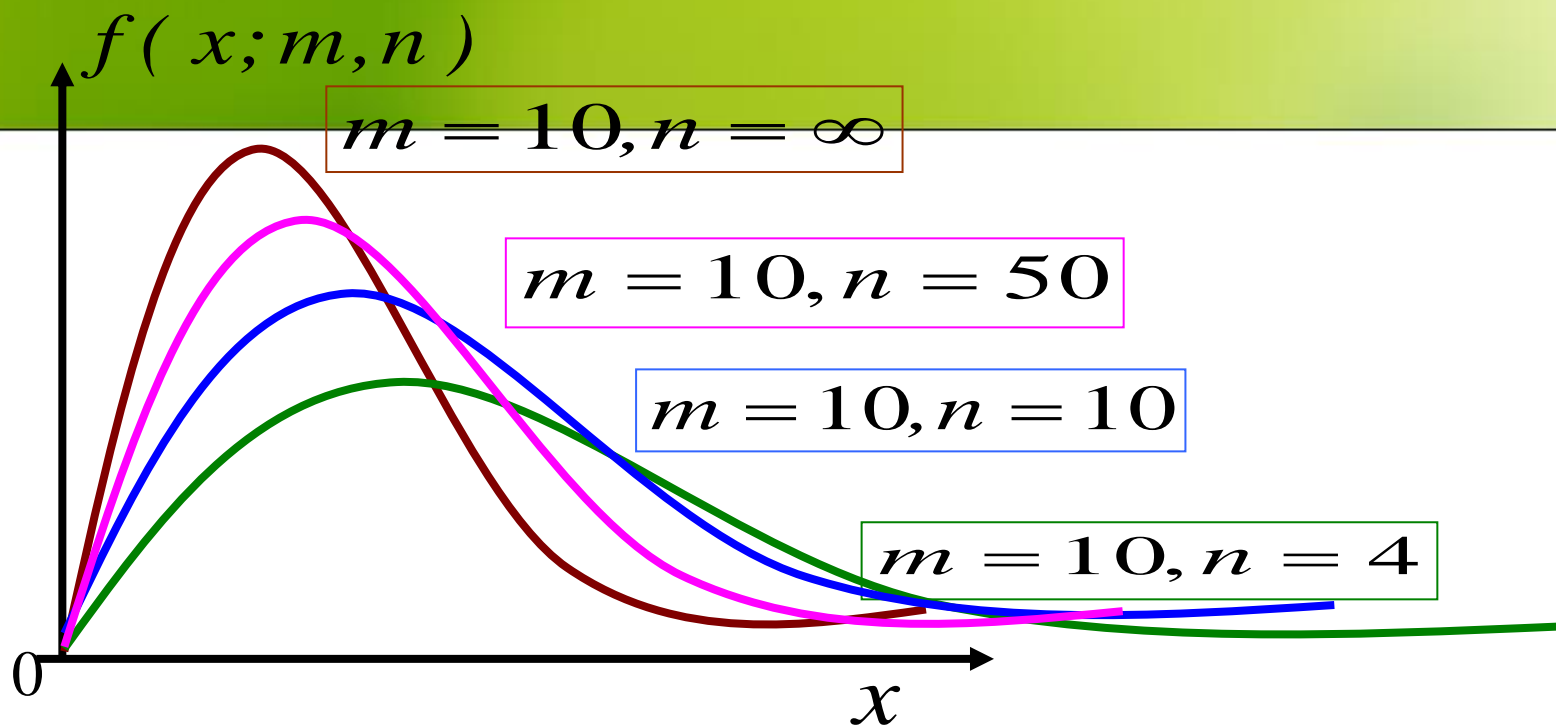
定义 若随机变量X的密度函数为

$$f(x) = \begin{cases} \frac{\Gamma\left(\frac{n_1 + n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1} \left(1 + \frac{n_1}{n_2} x\right)^{-\frac{n_1+n_2}{2}} & x > 0 \\ 0 & x \leq 0 \end{cases} \quad (3.20)$$

则称X服从自由度为 n_1 和 n_2 的F分布,其中 n_1 称为第一自由度, n_2 称为第二自由度记为 $X \sim F(n_1, n_2)$



其密度曲线见图。



注： F 分布是不对称分布其分布曲线向右偏斜， n_1, n_2 是它的两个参数，当 n_1, n_2 增大时， F 分布近似于对称。



2、F分布的典型模式

定理 设随机变量 X 和 Y 相互独立，分别服从自由度为 n_1 和 n_2 的 χ^2 分布，即 $X \sim \chi^2(n_1), Y \sim \chi^2(n_2)$ ，则

$$F = \frac{X/n_1}{Y/n_2} \sim F(n_1, n_2) \quad (3.21)$$

即 F 服从自由度为 n_1, n_2 的 F 分布 $F(n_1, n_2)$ 。



3、F分布的性质

(1) 服从 F 分布的随机变量的倒数也服从 F 分布：

若 F 服从 $F(n_1, n_2)$ 分布，则 $\frac{1}{F}$ 服从 $F(n_2, n_1)$ 分布。

(2) F 分布的数字特征

$$E[F(n_1, n_2)] = \frac{n_2}{n_2 - 2} \quad n_2 > 2$$

$$D[F(n_1, n_2)] = \frac{n_2^2(2n_1 + 2n_2 - 4)}{n_1(n_2 - 2)^2(n_2 - 4)} \quad n_2 > 4$$



5.4.4 一些重要结论



推论5.4.3 设 x_1, x_2, \dots, x_n 是来自 $N(\mu_1, \sigma_1^2)$ 的样本, y_1, y_2, \dots, 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

$$F = s_x^2 / s_y^2 \sim F(m-1, n-1)$$



推论5.4.4 在推论5.4.3的记号下, 设

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

并记

$$s_w^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2} = \frac{\sum_{i=1}^m (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2}{m+n-2}$$

则

$$\frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{s_w \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2)$$



三、单个正态总体下样本均值与样本方差的分布

定理3.7 设 X_1, X_2, \dots, X_n 是取自正态总体 $N(\mu, \sigma^2)$ 的样本，则有

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

$$2^0 \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

$3^0 \bar{X}$ 与 S^2 相互独立

$$4^0 \frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t(n-1)$$

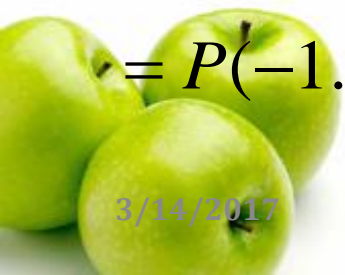
$$5^0 \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \sim \chi^2(n)$$

例， 设 X_1, X_2, \dots, X_{16} 为正态分布 $N(\mu, \sigma^2)$ 的一个样本，
试求样本均值 \bar{X} 与总体均值 μ 之差绝对值小于2的概率。

(1)已知 $\sigma^2 = 25$, (2) σ^2 未知, 但已知样本方差 $S_n^2 = 20.8$.

解: (1) 已知 $\sigma = 5$, 则 $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$

$$P(|\bar{X} - \mu| < 2) = P\left(\left|\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}\right| < \frac{2}{5 / \sqrt{16}}\right) \\ = P(-1.6 < \xi < 1.6) = 2\Phi(1.6) - 1 = 0.8904$$



解：(2) 未知 σ ，则 $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(N-1)$

$$P(|\bar{X} - \mu| < 2) = P\left(\left|\frac{\bar{X} - \mu}{S/\sqrt{n}}\right| < \frac{2}{\sqrt{20.8/16}}\right)$$

$$= P(|\eta| < 1.754) = 2P(\eta < 1.754) - 1 = 0.9$$



二、两个正态总体下样本均值差与样本方差比的分布

定理3.8 设 X_1, X_2, \dots, X_{n_1} 与 Y_1, Y_2, \dots, Y_{n_2} 分别为取自正态总体 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 的样本，且它们相互独立，则有

$$1^0 \quad \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$



证: 1⁰ 因为 $\bar{X} \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$, $\bar{Y} \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$

且 \bar{X} 与 \bar{Y} 相互独立, 故

$$\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

$$\text{故 } U = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$



$$2^0 \quad \text{当 } \sigma_1 = \sigma_2 \text{ 时, } T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

$$\text{其中 } S_w^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$\text{证明: 当 } \sigma_1 = \sigma_2 = \sigma \text{ 时 } U = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

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

且相互独立, 故由 χ^2 分布可加性知

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

又因 U 与 x^2 相互独立,故得

$$T = \frac{U}{\sqrt{x^2/n_1 + n_2 - 2}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$



3⁰ 当 σ_1, σ_2 未知, $\sigma_1 \neq \sigma_2$ 时,

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n-1} + \frac{S_2^2}{m-1}}} \sim N(0,1)$$



性质3⁰ $F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$

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



性质4⁰

$$\frac{n_2 \sigma_2^2 \sum_{i=1}^{n_1} (X_i - \mu_1)^2}{n_1 \sigma_1^2 \sum_{i=1}^{n_2} (X_i - \mu_2)^2} \sim F(n_1, n_2)$$

证明: $x_1^2 = \frac{\sum_{i=1}^{n_1} (X_i - \mu)^2}{\sigma_1^2} \sim x^2(n_1)$

$$x_2^2 = \frac{\sum_{i=1}^{n_2} (X_i - \mu)^2}{\sigma_2^2} \sim x^2(n_2)$$

且相互独立, 故由定理3.6知

$$F = \frac{x_1^2 / n_1}{x_2^2 / n_2} = \frac{n_2 \sigma_2^2 \sum_{i=1}^{n_1} (X_i - \mu)^2}{n_1 \sigma_1^2 \sum_{i=1}^{n_2} (X_i - \mu)^2} \sim F(n_1, n_2)$$



三、典型例题

例：设总体 X 服从 $N(u_1, \sigma_1^2)$, Y 服从 $N(u_2, \sigma_2^2)$,

从二总体中分别抽样，得到下列数据：

$$n_1 = 8, \bar{x} = 10.5, S_1^2 = 42.25; n_2 = 10, \bar{y} = 13.4, S_2^2 = 56.25;$$

求(1) $P(\frac{\sigma_2^2}{\sigma_1^2} < 4.40)$; (2) 假定 $\sigma_1^2 = \sigma_2^2$, 求 $P(u_1 < u_2)$ 。



解:(1)当 $\sigma_1^2 \neq \sigma_2^2$ 时, $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ 服从自由度为 (n_1-1, n_2-1) 的 F 分布

$$P\left(\frac{\sigma_2^2}{\sigma_1^2} < 4.40\right) = P\left(\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < \frac{S_1^2}{S_2^2} \times 4.40\right)$$

$$= P\left(\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} < \frac{42.25}{56.25} \times 4.40\right) = P\left(\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} < 3.3048\right) = 1 - P\left(\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \geq 3.3048\right)$$

$$\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \text{ 服从 } F(7,9) \text{ 分布。} \quad = 1 - 0.05 = 0.95$$



(2)当 $\sigma_1^2 = \sigma_2^2$ 时, $\frac{(\bar{X} - \bar{Y}) - (u_1 - u_2)}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2} S_w}}$ 服从自由度为

$(n_1 + n_2 - 2)$ 的 t 分布, 其中 $S_w^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}$

$$P(u_1 < u_2) = P(-(u_1 - u_2) > 0)$$

$$= P\left\{ \frac{(\bar{X} - \bar{Y}) - (u_1 - u_2)}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2} S_w}} > \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2} S_w}} \right\}$$



因为 $\frac{(\bar{X} - \bar{Y}) - (u_1 - u_2)}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2} S_w}}$ 服从 $t(16)$ 分布

$$\text{所以 } 1 - P\left\{\frac{(\bar{X} - \bar{Y}) - (u_1 - u_2)}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2} S_w}} < -0.8635\right\}$$

$$= 1 - 0.2 = 0.8$$



§ 5.5 充分统计量

5.5.1 充分性的概念

例5.5.1 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。



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, \dots, x_n ，每个 x_j 取值非0即1，命中为1，不命中为0。令 $T = x_1 + \dots + x_n$ ， 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”。

样本 $X=(X_1, X_2, \dots, X_n)$ 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。



统计量 $T = T(X_1, X_2, \dots, X_n)$ 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 θ 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 θ 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 θ 的信息，这正是统计量具有充分性的含义。



定义5.5.1 设 X_1, X_2, \dots, 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 $T = T(X_1, X_2, \dots, x_n)$ 称为 θ 的充分统计量，如果在给定 T 的取值后， X_1, X_2, \dots, X_n 的条件分布与 θ 无关.



5.5.2 因子分解定理

充分性原则： 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5.5.1 设总体概率函数为 $f(x; \theta)$ ， X_1, \dots, X_n 为样本，则 $T=T(X_1, \dots, X_n)$ 为充分统计量的充分必要条件是：存在两个函数 $g(t; \theta)$ 和 $h(x_1, \dots, x_n)$ ，使得对任意的 θ 和任一组观测值 x_1, x_2, \dots, x_n ，有

$$f(x_1, x_2, \dots, x_n; \theta) = g(T(x_1, x_2, \dots, x_n); \theta) h(x_1, x_2, \dots, x_n) \quad (5.5.1)$$


其中 $g(t, \theta)$ 是通过统计量 T 的取值而依赖于样本的。

例5.5.4 设 X_1, X_2, \dots, X_n 是取自总体 $U(0, \theta)$ 的样本，
即总体的密度函数为

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

于是样本的联合密度函数为



$$f(x_1; \theta) \dots f(x_n; \theta) = \begin{cases} (1/\theta)^n, & 0 < \min\{x_i\} < \max\{x_i\} < \theta \\ 0, & \text{其它} \end{cases}$$

由于诸 $x_i > 0$ ，所以我们将上式改写为

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

取 $T = x_{(n)}$ ，并令 $g(t; \theta) = (1/\theta)^n I_{\{t < \theta\}}$ ， $h(x) = 1$ ，
由因子分解定理知 $T = X_{(n)}$ 是 θ 的充分统计量。

例5.5.5 设 X_1, X_2, \dots, 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



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

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

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

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。



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。

