

第六章 样本及抽样分布



随 机 样 本



直 方 图 和 箱 线 图



抽 样 分 布

第六章 样本及抽样分布

数理统计

——● 如何收集、整理数据资料

——● 如何对所得数据进行分析、研究，从而对研究对象的性质、特点作出判断






统计推断问题 ----- 数理统计的内容

以概率论为理论基础，根据试验观察得到的数据，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

§ 1 随机样本



一、总体与个体相关概念

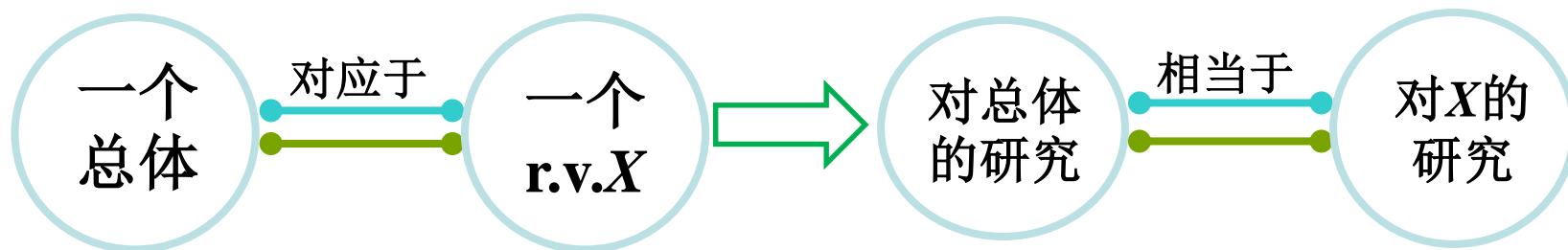
-  **总体** 试验的全部可能的观察值
-  **个体** 总体的每一个可能观察值
-  **容量** 总体中所包含个体的个数
-  **有限总体** 容量有限的总体
-  **无限总体** 容量无限的总体

§ 1 随机样本

一、总体与个体相关概念

例 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2000人，每个男生的身高是一个可能观察值，所形成的有限总体中共含2000个可能观察值。

例 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，所得总体为无限总体



可能观察值个数很多时，有限总体可以近似认为是无限总体

§ 1 随机样本



二、随机样本相关概念

 **样 本** 从总体中抽出的部分个体

 **总体 X 的一个简单随机样本**

相同条件下对总体 X 进行 n 次观察结果记为 X_1, X_2, \dots, X_n , 其中 X_1, X_2, \dots, X_n 相互独立, 且与 X 有相同分布, 则称这个这些结果为总体 X 的一个**简单随机样本**, n 称为这个**样本的容量**

 **样本值**

n 次观察一经完成, 得到一组实数 x_1, x_2, \dots, x_n , 它们依次是 X_1, X_2, \dots, X_n 的观察值, 称为**样本值**

§ 1 随机样本



三、随机样本定义



综上所述：

总体就是一个r.v. X

定 义

设 X 是具有分布函数 F 的随机变量，若 X_1, X_2, \dots, 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, \dots, 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, \dots, x_n 称为**样本值**，又称为 X 的 n 个**独立的观察值**

§ 1 随机样本



三、随机样本定义

若 X_1, X_2, \dots, X_n 为 F 的一个样本, 则 X_1, X_2, \dots, X_n 相互独立且它们有相同分布函数 $F(X)$, 所以 (X_1, X_2, \dots, X_n) 的分布函数为
$$F^*(x_1, x_2, \dots, x_n) = \prod_{i=1}^n F(x_i)$$

又若 X 具有概率密度 f , 则 (X_1, X_2, \dots, X_n) 的概率密度为
$$f^*(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i)$$

§ 2 直方图和箱线图



为了研究总体分布的性质，人们通过实验得到许多观察值，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。

通过引入**频率直方图**和**箱线图**可对总体的分布有一个粗略的了解

§ 2 直方图和箱线图

一、直方图



引例

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						

§ 2 直方图和箱线图



一、直方图



引例

解 ● 数据整理

最小值：126，最大值：158 \Rightarrow 数据落在区间[126,158]

- 现在取区间[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，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**。

§ 2 直方图和箱线图

一、直方图



引 例

- 数出落在每个小区间内的数据的频数 f ，算出频率 f_i/n ($n=84, i=1, 2, \dots, 7$)

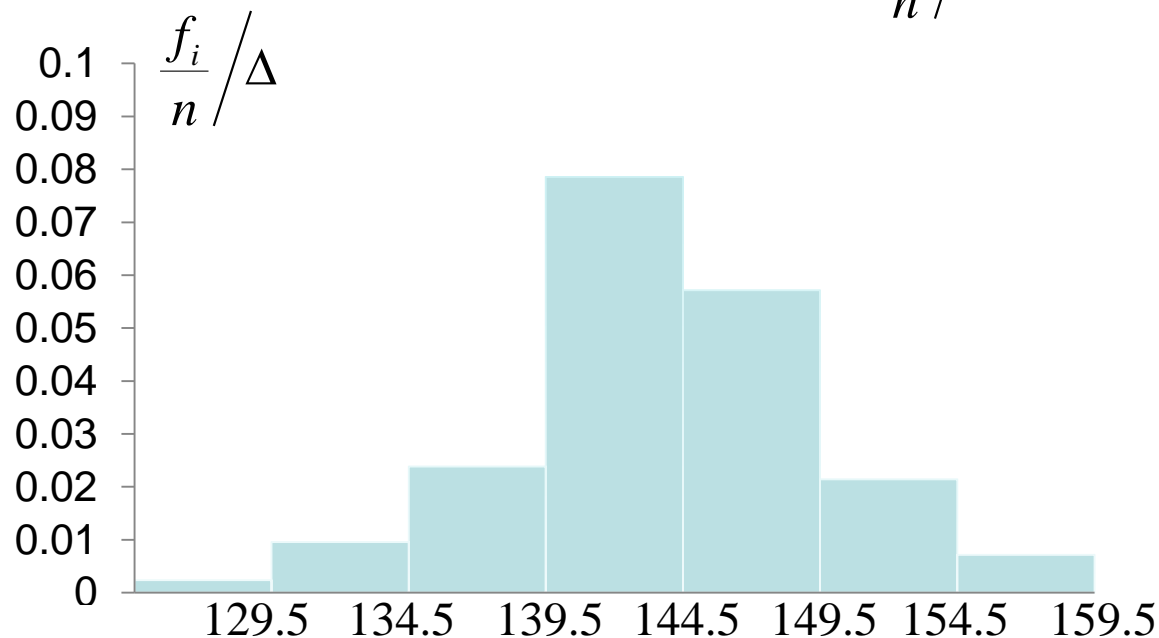
组 限	频数 f_i	频率 f_i/n	累计频率
124.5~129.5	1	0.0119	0.0119
129.5~134.5	4	0.0476	0.0595
134.5~139.5	10	0.1191	0.1786
139.5~144.5	33	0.3929	0.5715
144.5~149.5	24	0.2857	0.8572
149.5~154.5	9	0.1071	0.9643
154.5~159.5	3	0.0357	1

§ 2 直方图和箱线图



一、直方图

- 自左至右一次在各个小区间上作以 $\frac{f_i}{n}/\Delta$ 为高的小矩形。



如图所示的图形叫**频率直方图**

§ 2 直方图和箱线图



一、直方图

直方图的外廓曲线接近于总体 X 的概率密度曲线

本例的直方图有一个峰，中间高，两头低，比较对称  近似 某一正态总体 X

直方图上还可以估计 X 落在某一区间的概率

从本例图上看到有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

§ 2 直方图和箱线图

一、直方图



matlab 命令

命令	名称	输入	输出	注意事项
<code>[n,y]=hist(x,k)</code>	频数表	x: 原始数据行向量 k: 等分区间数	n: 频数行向量 y: 区间中点行向量	<code>[n,y]=hist(x,k)</code> 中k取默认值10
<code>hist(x,k)</code>	直方图	x: 原始数据行向量 k: 等分区间数	直方图	<code>[n,y]=hist(x,k)</code> 中k取默认值10
<code>mean(x)</code>	均值	x: 原始数据行向量	均值 \bar{x}	
<code>median(x)</code>	中位数	x: 原始数据行向量	中位数	
<code>range(x)</code>	极差	x: 原始数据行向量	极差	
<code>std(x)</code>	标准差	x: 原始数据行向量	标准差 s	
<code>var(x)</code>	方差	x: 原始数据行向量	方差 s^2	
<code>skewness(x)</code>	偏度	x: 原始数据行向量	偏度 g_1	
<code>kurtosis(x)</code>	峰度	x: 原始数据行向量	峰度 g_2	15

§ 2 直方图和箱线图



一、直方图



matlab 命令

常用使用方法如下：

- **hist**函数用来作“**频数直方图**”，也可通过修改“频数直方图”的纵坐标将“频数直方图”转为“频率直方图”（结合使用**Bar**函数）。
- **ecdf**和**ecdfhist**函数用来作“**频率直方图**”。

`[fi,xi] = ecdf(x);`

`ecdfhist(fi,xi,10);`

其中x是样本数据，xi是不重复的从小到大排序的样本值，fi是与xi对应的经验分布函数值，10是直方图中柱子的数目。

§ 2 直方图和箱线图



二、箱线图



样本分位数

设有容量为 n 的样本观察值 x_1, x_2, \dots, x_n , 样本 p 分位数($0 < p < 1$)记为 x_p , 它具有以下性质:

- (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, \dots, x_n 按自小到大的次序排列成 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

§ 2 直方图和箱线图

二、箱线图



样本分位数

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： $n=12$ ， $p=0.9$ ， $np=10.8$ ， $n(1-p)=1.2$ ，故 x_p 应位于第11处

若 np 是整数，例如在 $n=20$ ， $p=0.95$ 时，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

$$\text{综上： } x_p = \begin{cases} x_{([np]+1)} & \text{当 } np \text{ 不是整数} \\ \frac{1}{2} [x_{(np)} + x_{(np+1)}] & \text{当 } np \text{ 是整数} \end{cases}$$

§ 2 直方图和箱线图

二、箱线图



样本分位数

当 $p=0.5$ 时，0.5分位 $x_{0.5}$ 也记为 Q_2 或 M 称为样本中位数，既有：

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

当 n 为奇数时中位数 $x_{0.5}$ 就是 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 这一数组中最中间的一个数；当 n 为偶数时中位数 $x_{0.5}$ 就是 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 这一数组中最中间的两个数的平均值

0.25分位 $x_{0.25}$ 称为第一四分位数，记为 Q_1 ；

0.75分位 $x_{0.75}$ 称为第三四分位数，记为 Q_3

§ 2 直方图和箱线图



二、箱线图



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。作法如下：

- (1) 画一水平数轴并标上Min, Q_1 , M , Q_3 , Max。画一个上、下侧平行于数轴的矩形箱子，箱子左右两侧分别位于 Q_1 、 Q_3 上方。在M点上方画一条垂直于箱内的线段。
- (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Max

§ 2 直方图和箱线图

二、箱线图



箱线图

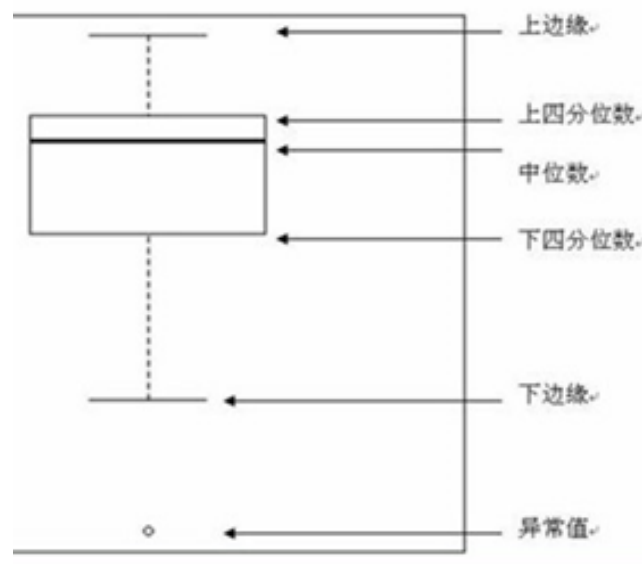
自箱线图可以形象地看出数据集的以下性质：

(1) 中心位置：中位数所在的位置

(2) 散布程度：数据落在 $[\text{Min}, \text{Max}]$ 内。

区间 $[\text{Min}, Q_1]$, $[Q_1, M]$, $[M, Q_3]$, $[Q_3, \text{Max}]$ 的数据个数各约占1/4。区间越短表示落在其中的点越集中，反之越分散

(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。若 Min 离 M 较近，则数据向左倾斜，反之右倾



§ 2 直方图和箱线图



二、箱线图



箱线图

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**。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。

Q_1 和 Q_3 之间的距离： $Q_3 - Q_1$ ，记为**IQR**，称为**四分位数间距**
若数据小于 $Q_1 - 1.5\text{IQR}$ 或大于 $Q_3 + 1.5\text{IQR}$ ，就认为它是疑似异常值

§ 2 直方图和箱线图



二、箱线图



箱线图

将箱线图的作法作如下改变：

(1') 同箱线图作法(1)

(2') 计算 $IQR = Q_3 - Q_1$ ，若一个数据小于 $Q_1 - 1.5IQR$ 或大于 $Q_3 + 1.5IQR$ ，就认为它是疑似异常值，画出疑似异常值，以*表示

(3') 自箱子左侧引一条水平线直至数据集中除去疑似异常值后的最小值；自箱子右侧引一条水平线直至数据集中除去疑似异常值后的最大值

最后作出的图形叫做**修正箱线图**

§ 2 直方图和箱线图

二、箱线图

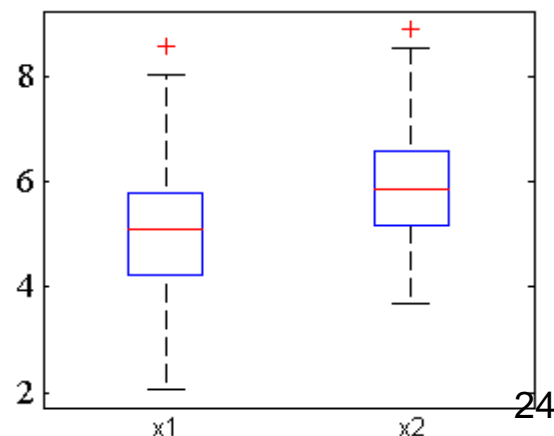


matlab画箱线图

`boxplot(X)`: 将X进行分组并画出每一组的箱线图，分组变量G值相同的数据为一组。

`boxplot(X,'Name',value)`: 设定参量值以调整箱线图的形状

例 `x1 = normrnd(5,1,100,1);`
`x2 = normrnd(6,1,100,1);`
`X=[x1,x2];`
`G=['x1';'x2']; % 定义分组变量`
`boxplot(X,G) % 画出箱线图`



§ 2 直方图和箱线图

三、绘制实例



数据说明

本次实验数据采用了300例新生儿出生体重及相关的四个参数：

- BPD(双顶径)
- HC(头围)
- AC(腹围)
- FL(股骨长度)



$$\begin{aligned} \text{Log}_{10} \text{ BW} = & 1.3596 + 0.0064(\text{HC}) + 0.0424(\text{AC}) \\ & + 0.174(\text{FL}) + 0.00061(\text{BPD})(\text{AC}) - 0.00386(\text{AC})(\text{FL}) \end{aligned}$$

§ 2 直方图和箱线图



三、绘制实例



数据读取

数据格式采用.xls格式（excel 03-07）。

1. 读取Excel文件：

(a) 命令方式xlsread:

读取命令：`[data,text] = xlsread(FileName, SheetName, Range);`

`data`保存的是数据单元格的值，`text`保存的是字符串单元格的内容。

例如：`[data,text] = xlsread('C:\Test\test.xls', 'testsheets', 'B2:D10');`

存储方式为矩阵，和Excel表格中的位置一致。

如果Sheet内都是数据，可直接使用`data = xlsread(文件名)`。

§ 2 直方图和箱线图

三、绘制实例



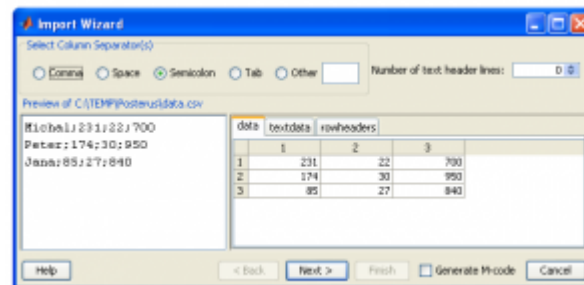
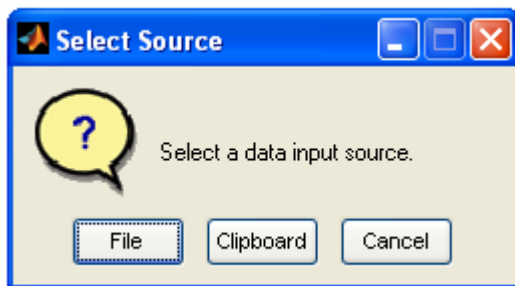
数据读取

数据格式采用.xls格式（excel 03-07）。

1. 读取Excel文件：

(b) 界面方式uiimport:

在Matlab中输入命令：uiimport，弹出如下窗口。不仅可以从文件导入，也可从剪贴板中导入。



§ 2 直方图和箱线图



三、绘制实例



补充：数据写入

数据格式采用.xls格式（excel 03-07）。

2. 将MATLAB数据写入到Excel

写入命令： `xlswrite(FileName, Output, SheetName, Range)`

其中Output为要写入的数据，可以是矩阵也可以是cell类型

例如： `xlswrite('C:\test\text.xls', eye(3), 'Sheet1', 'A1:C3')`

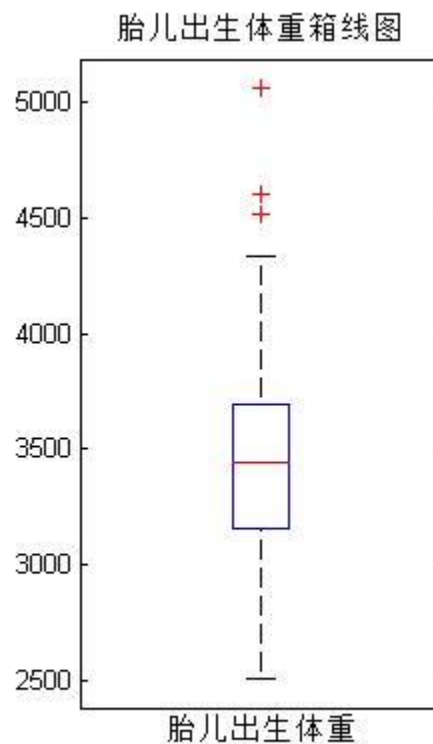
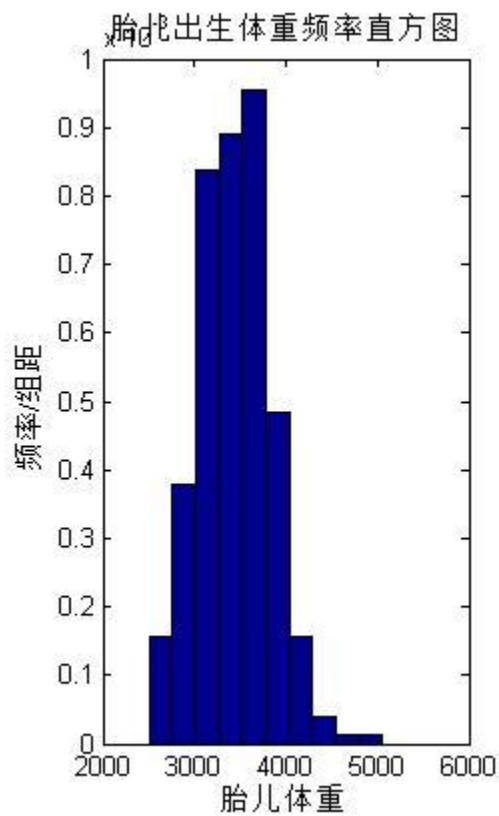


§ 2 直方图和箱线图

三、绘制实例



绘制结果



§ 3 抽样分布

一、统计量相关概念

样本是进行统计推断的依据，在应用中，往往不是直接使用样本本身，而是针对不同问题构造的适当函数，利用这些样本函数进行统计推断

定 义

设 (X_1, X_2, \dots, X_n) 是来自总体 X 的一个样本， $g(X_1, X_2, \dots, X_n)$ 是 X_1, X_2, \dots, X_n 的函数，若 g 是连续函数且 g 中不含任何未知数，则称 $g(X_1, X_2, \dots, X_n)$ 是一个**统计量**。

若 x_1, x_2, \dots, x_n 是 X_1, X_2, \dots, X_n 的**样本观察值**，则称 $g(x_1, x_2, \dots, x_n)$ 是 $g(X_1, X_2, \dots, X_n)$ 的**观察值(统计值)**

§ 3 抽样分布

一、统计量相关概念



示例

从正态总体 $N(\mu, \sigma^2)$ 中抽取样本 X_1, X_2, \dots, X_n ，其中 μ, σ^2 为未知参数。

则： $\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2, \frac{1}{\sigma^2} \sum_{i=1}^n X_i$ 不是统计量(因含有未知数 μ, σ^2)

$X = \frac{1}{n} \sum_{i=1}^n X_i$ 是统计量

当 μ 已知时， $\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$ 是统计量


已知总体 X 的分布，则统计量 $g(X_1, X_2, \dots, X_n)$ 应有确定的概率分布。这个分布被称为该**统计量的抽样分布**


§ 3 抽样分布

一、几种常用的统计量


设 X_1, X_2, \dots, X_n 是来自总体 X 的一个样本, x_1, x_2, \dots, x_n 是这一样本的观察值。

 **样本平均值** $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

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

 **样本标准差** $S = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$

 **样本k阶(原点)矩** $A_k = \frac{1}{n} \sum_{i=1}^n X_i^k, k = 1, 2, \dots$

 **样本k阶中心矩** $B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k, k = 2, 3, \dots$

§ 3 抽样分布



一、几种常用的统计量

它们的观察值分别为：

● **样本均值：** $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

样本标准差： $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$

样本方差：

● $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i^2 - n\bar{x}^2)$

样本k阶矩：

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

样本k阶矩：

$$b_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k, k = 2, 3, \dots$$

§ 3 抽样分布

一、几种常用的统计量

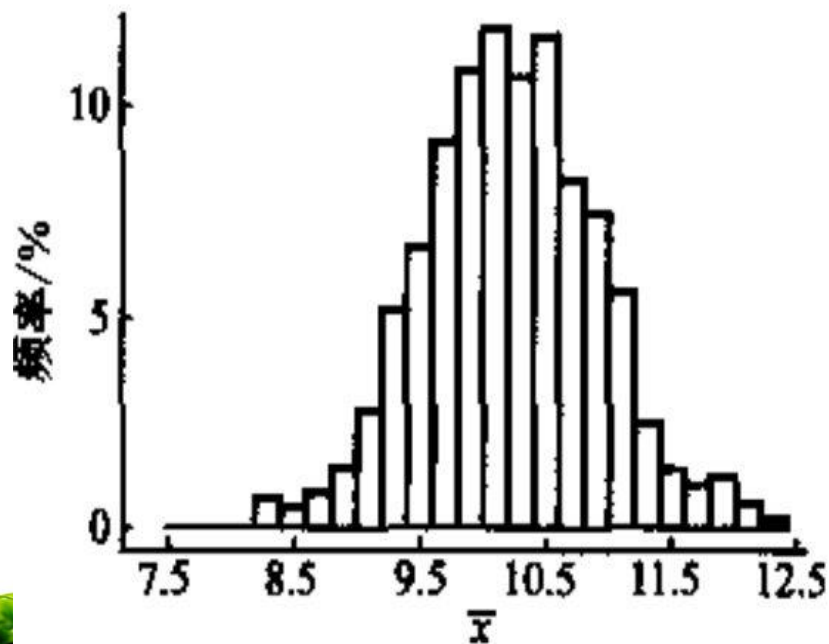
例 设有一个由20个数组成的总体，现从该总体同时取出容量为5的样本。抽取过程如下：

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

§ 3 抽样分布

一、几种常用的统计量

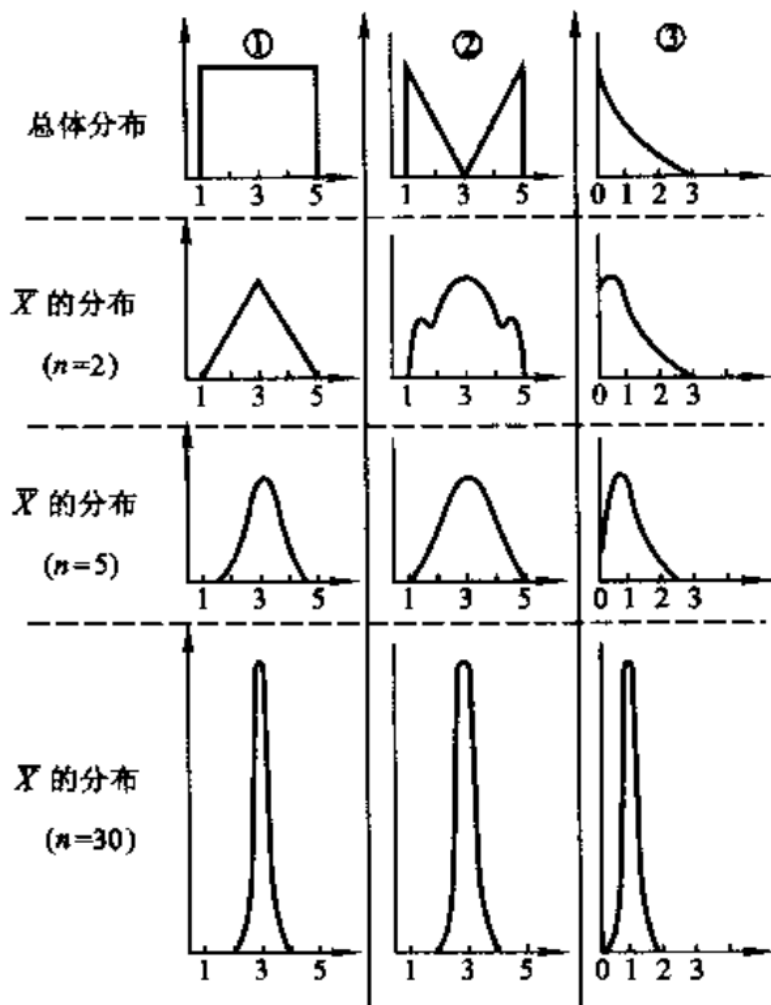
一共抽出4个样本，每个样本有5个观测值。由于样本的随机性，每个样本的样本均值 \bar{x} 都有差别。设想类似抽取无限制地进行，则可以得到大量的 \bar{x} 的值。用这样得到的前500个 \bar{x} 的值所形成的直方图为：



直方图反应了 \bar{x} 的抽样分布。
它的外形很像正态分布

§ 3 抽样分布

一、几种常用的统计量



例 左图给出三个不同总体样本均值的分布。三个总体分别是：①均匀分布②倒三角分布③指数分布。随着样本量的增加，样本均值 \bar{x} 的抽样分布**逐渐向正态分布逼近**。

它们的均值保持不变，而方差则缩小为原来的 $1/n$

§ 3 抽样分布



一、几种常用的统计量

具体说明

①的总体分布 $U(1,5)$ ，总体均值和方差分别为3和4/3，若从中抽取样本容量为30的样本，则样本均值的渐进分布为

$$\overline{X}_1 \sim N\left(3, \frac{4}{3 \times 30}\right) = N(3, 0.21^2)$$

②的总体分布的概率密度函数为

$$f(x) = \begin{cases} (3-x)/4, & 1 \leq x \leq 3 \\ (x-3)/4, & 3 \leq x \leq 5 \\ 0, & \text{else} \end{cases}$$

§ 3 抽样分布



一、几种常用的统计量

②是一个倒三角分布，总体均值和方差分别为3和2，若从中抽取样本容量为30的样本，则样本均值的渐进分布为

$$\overline{X}_2 \sim N\left(3, \frac{2}{30}\right) = N(3, 0.26^2)$$

③的总体分布为指数分布 $Exp(1)$ ，总体均值和方差都为1，若从中抽取样本容量为30的样本，则样本均值的渐进分布为

$$\overline{X}_3 \sim N\left(1, \frac{1}{30}\right) = N(1, 0.18^2)$$

三个总体都不是正态分布，但是其样本均值的分布都十分近似于正态分布，差别表现在均值和标准差上

§ 3 抽样分布

一、几种常用的统计量

定 理

设总体 X 具有二阶矩，即 $E(X)=\mu$ ， $Var(X)=\sigma^2<\infty$ ， x_1, x_2, \dots, x_n 为该总体的样本， \bar{X} 和 S^2 分别是样本均值和样本方差，则：

$$(1) \quad E(\bar{X}) = \mu, \quad Var(\bar{X}) = \sigma^2/n; \quad (2) \quad E(S^2) = \sigma^2$$

证明： $E(\bar{x}) = \frac{1}{n} E\left(\sum_{i=1}^n x_i\right) = \frac{n\mu}{n} = \mu;$ $E x_i^2 = (E x_i)^2 + Var(x_i) = \mu^2 + \sigma^2$

$$Var(\bar{x}) = \frac{1}{n^2} Var\left(\sum_{i=1}^n x_i\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$
$$E(\bar{x}^2) = (E\bar{x})^2 + Var(\bar{x}) = \mu^2 + \sigma^2/n$$

故(1)得正

$$\text{又 } \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2,$$

$$E\left(\sum_{i=1}^n (x_i - \bar{x})^2\right) = n(\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2/n) = (n-1)\sigma^2$$

两边同时除以 $n-1$ ，可得(2)

§ 3 抽样分布

一、几种常用的统计量

若总体 X 的 k 阶矩 $E(X^k)$ 记成 μ_k 存在, 则当 $n \rightarrow \infty$ 时,

$$A_k \xrightarrow{P} \mu_k, k = 1, 2, \dots$$

X_1, X_2, \dots, X_n 独立同分布, 则 $X_1^k, X_2^k, \dots, X_n^k$, 独立且与 X^k 同分布, 故有:

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

由辛钦大数定理知:

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

由依概率收敛的序列的性质知道:

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

矩估计法的理论根据

§ 3 抽样分布

二、经验分布函数

设 X_1, X_2, \dots, X_n 是总体 F 的一个样本，用 $S(x)$, $-\infty < x < +\infty$ 表示 X_1, X_2, \dots, 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 $F_n(x) = \frac{1}{n} S(x)$, $-\infty < x < +\infty$

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


$$F_3(x) = \begin{cases} 0 & x < 1 \\ \frac{1}{3} & 1 \leq x < 2 \\ \frac{2}{3} & 2 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$

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

$$F_3(x) = \begin{cases} 0 & x < 1 \\ \frac{2}{3} & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

§ 3 抽样分布

二、经验分布函数




设 x_1, x_2, \dots, x_n 是总体 F 的一个容量为 n 的样本值。将 x_1, x_2, \dots, x_n 按自小到大的次序排列，并重新编号。设为 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

则经验分布函数 $F_n(x)$ 的观察值为

$$F_n(x) = \begin{cases} 0 & x < x_{(1)} \\ \frac{k}{n} & x_{(k)} \leq x < x_{(k+1)}, k = 1, 2, \dots, n-1 \\ 1 & x \geq x_{(n)} \end{cases}$$

对于任一实数 x ，当 $n \rightarrow \infty$ 时 $F_n(x)$ 以概率1一致收敛于分布函数 $F(x)$ ，即 $P\left\{\lim_{n \rightarrow \infty} \sup_{-\infty < x < \infty} |F_n(x) - F(x)| = 0\right\} = 1$



当 n 充分大时 $F_n(x)$ 可以当作 $F(x)$ 用

§ 3 抽样分布

三、来自正态总体的几个常用统计量分布



χ^2 —分布

定 义

设 X_1, X_2, \dots, X_n 相互独立, 都服从正态分布 $N(0,1)$, 则称随机变量: $\chi^2 = X_1^2 + X_2^2 + \dots + X_n^2$ 服从的分布为**自由度为 n 的 χ^2 分布**

记为: $\chi^2 \sim \chi^2(n)$

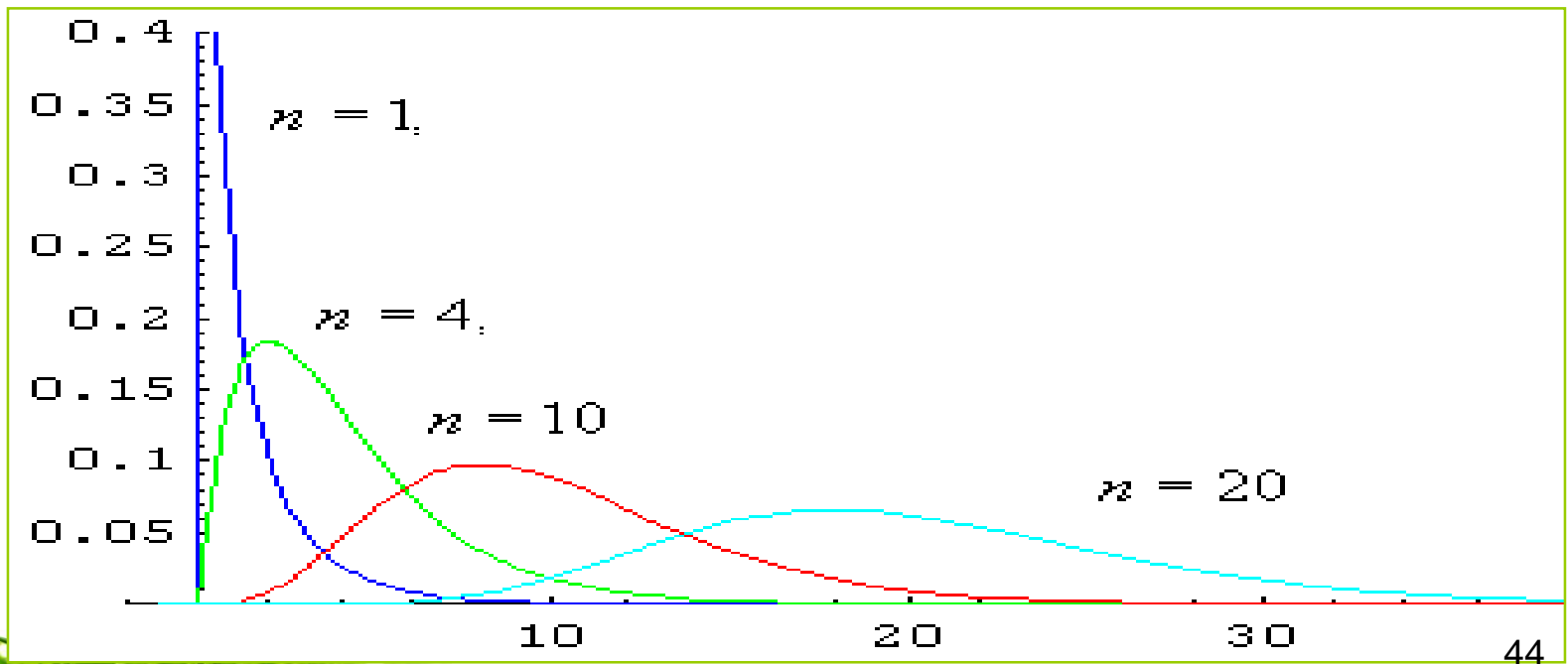
§ 3 抽样分布

三、来自正态总体的几个常用统计量分布



χ^2 —分布

$$\chi^2(n) \text{ 分布的概率密度为 } f(y) = \begin{cases} \frac{1}{2^{n/2} \Gamma(n/2)} y^{n/2-1} e^{-y/2} & y > 0 \\ 0 & \text{else} \end{cases}$$



§ 3 抽样分布

三、来自正态总体的几个常用统计量分布



χ^2 —分布

$\chi^2(n)$ 分布的性质

a. 设 X_1, X_2, \dots, X_n , 是来自正态总体 $N(\mu, \sigma^2)$ 的一个样本, 则 $Y = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \sim \chi^2(n)$

b. **分布可加性** 若 $X \sim \chi^2(n_1)$, $Y \sim \chi^2(n_2)$, X, Y 相互独立, 则 $X + Y \sim \chi^2(n_1 + n_2)$ 。

c. **期望与方差** 若 $X \sim \chi^2(n)$, 则 $E(X) = n$, $D(X) = 2n$ 。

d. 若 $X \sim \chi^2(n)$, 则当 n 充分大时, $\frac{X - n}{\sqrt{2n}}$ 的分布近似正态分布 $N(0, 1)$ 。

§ 3 抽样分布

三、来自正态总体的几个常用统计量分布



χ^2 —分布

$\chi^2(n)$ 分布的性质

χ^2 分布的上 α 分位点

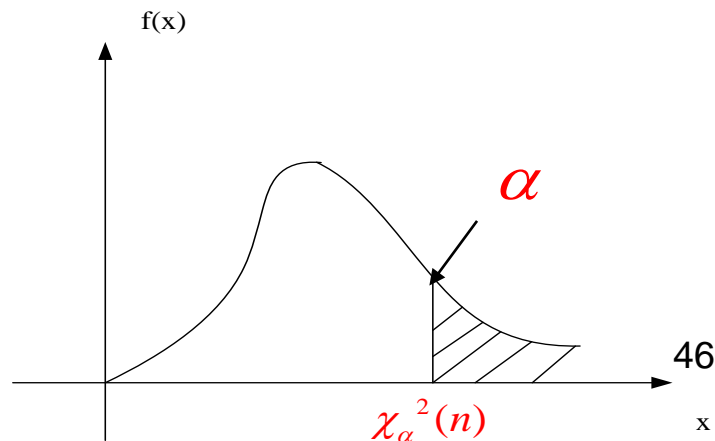
若对于给定的 $\alpha(0 < \alpha < 1)$, 存在 $\chi_\alpha^2(n)$ 使

$$P\{\chi^2 > \chi_\alpha^2(n)\} = \int_{\chi_\alpha^2(n)}^{\infty} f(x)dx = \alpha$$

则称点 $\chi_\alpha^2(n)$ 为 χ^2 分布的上 α 分位点

查卡方分布表

自由度



§ 3 抽样分布

三、来自正态总体的几个常用统计量分布



t —分布

定 义

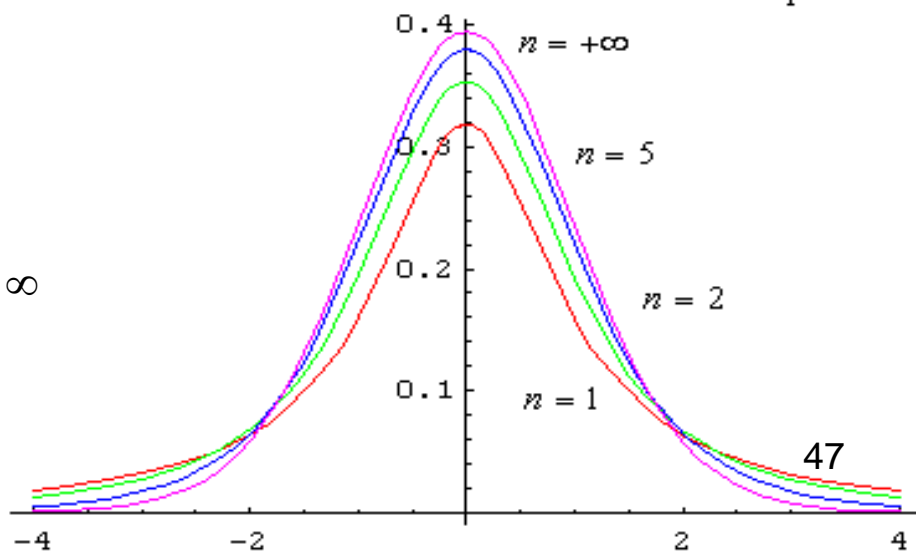
若 $X \sim N(0, 1)$, $Y \sim \chi^2(n)$, X 与 Y 独立, 则

$$t = \frac{X}{\sqrt{Y/n}} \sim t(n)$$

$t(n)$ 称为 **自由度为 n 的 t -分布**

$t(n)$ 分布的概率密度为

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



§ 3 抽样分布

三、来自正态总体的几个常用统计量分布



t —分布

1908以前 统计数据都是大量的，自然采集的，且服从正态分布 核心人物 K·皮尔逊

1899~1906 哥赛特发现 $t = \sqrt{n}(\bar{x} - \mu)/s$ 与传统 $N(0,1)$ 分布不同

$N(0,1)$ 和 $t(4)$ 的尾部概率 $P\{|X| \geq c\}$

	$c=2$	$c=2.5$	$c=3$	$c=3.5$
$X \sim N(0,1)$	0.0455	0.0124	0.0027	0.000465
$X \sim t(4)$	0.1161	0.0668	0.0399	0.0249

1906~1907 哥赛特向 K·皮尔逊学习统计学

1908 哥赛特在 Biometrics 上发表论文，提出 t 分布

1922 费希尔完整了 t 分布的证明并编制了 t 分布的分位数表

§ 3 抽样分布

三、来自正态总体的几个常用统计量分布



t —分布

例 设总体 $X \sim N(0,1)$, X_1, X_2, \dots, X_n 是 X 的一个样本,
求常数 C , 使统计量 $\frac{C(X_1 + X_2)}{\sqrt{X_3^2 + X_4^2 + X_5^2}}$ 服从 t 分布

解 由于 $X_i \sim N(0,1)$, $i=1,2,3,4,5$,
且相互独立, 故 $X_1 + X_2 \sim N(0,2)$,
 $X_3^2 + X_4^2 + X_5^2 \sim \chi^2(3)$

且两者相互独立, 由 t 分布的定义
可知, 要使

$$\frac{C(X_1 + X_2)}{\sqrt{X_3^2 + X_4^2 + X_5^2}} = \frac{\frac{C}{\sqrt{3}}(X_1 + X_2)}{\sqrt{(X_3^2 + X_4^2 + X_5^2)/3}}$$

服从 t 分布, 则 $\frac{C}{\sqrt{3}}(X_1 + X_2)$
必须服从 $N(0,1)$ 分布, 由
 $X_1 + X_2 \sim N(0,2)$ 得

$$\frac{C}{\sqrt{3}}(X_1 + X_2) \sim N\left(0, 2\left(\frac{C}{\sqrt{3}}\right)^2\right)$$

$$\Rightarrow \frac{2C^2}{3} = 1$$

故当 $C = \pm\sqrt{3/2}$ 时, 该统计
量服从自由度为3的 t 分布

§ 3 抽样分布

三、来自正态总体的几个常用统计量分布



t —分布

t 分布的性质

a. 具有自由度为 n 的 t 分布 $t \sim t(n)$, 其数学期望与方差为

$$E(t) = 0, D(t) = n/(n-2) \quad (n > 2)$$

b. t 分布的密度函数关于 $t=0$ 对称。当 n 充分大时, 其图形近似于标准正态分布概率密度的图形。

c. 当 $n \rightarrow \infty$ 时, 由 Γ 函数的性质有

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

即当 n 足够大时, $t \sim N(0,1)$; 当 n 较小时, t 分布 $N(0,1)$ 相差很大。

§ 3 抽样分布

三、来自正态总体的几个常用统计量分布



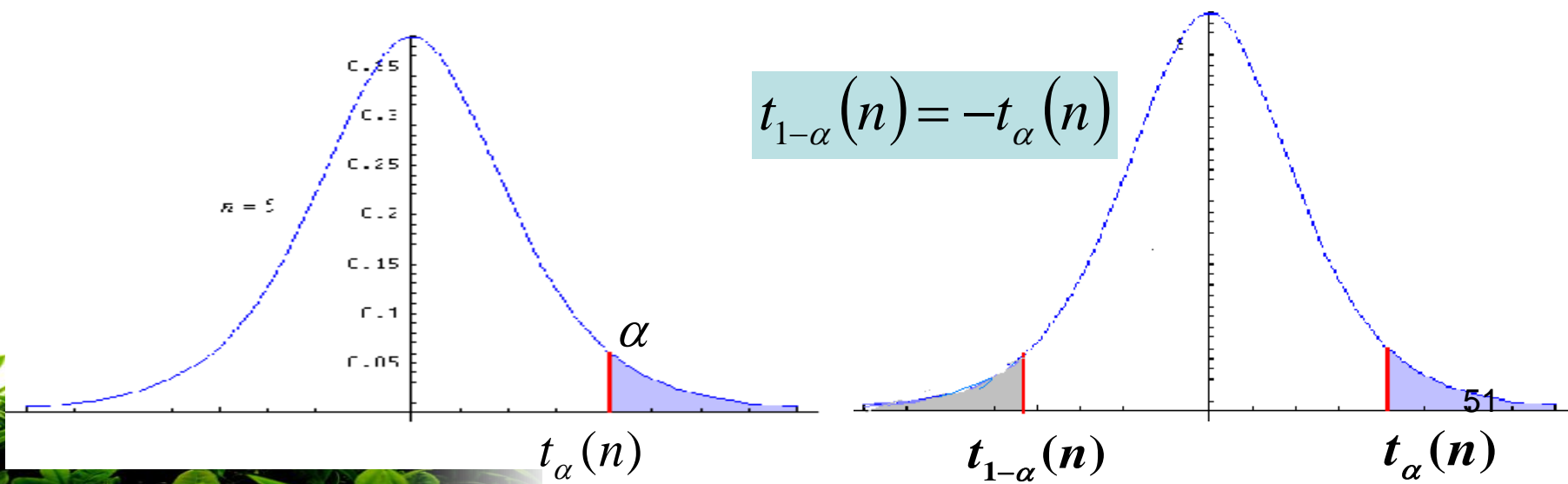
t —分布

t 分布的性质

t 分布的上 α 分位点

设 $t \sim t(n)$, 若对 $\alpha (0 < \alpha < 1)$, 存在 $t_\alpha(n) > 0$ 使 $P\{t > t_\alpha(n)\} = \alpha$
则称 $t_\alpha(n)$ 为 $t(n)$ 的上 α 分位点

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



§ 3 抽样分布

三、来自正态总体的几个常用统计量分布



F —分布

定 义

若 $U \sim \chi^2(n_1)$, $V \sim \chi^2(n_2)$, U 与 V 独立, 则

$$F = \frac{U/n_1}{V/n_2} \sim F(n_1, n_2)$$

称为第一自由度为 n_1 , 第二自由度为 n_2 的 F -分布

注: 若 $F \sim F(n_1, n_2)$, 则 $1/F \sim F(n_2, n_1)$ 。

§ 3 抽样分布

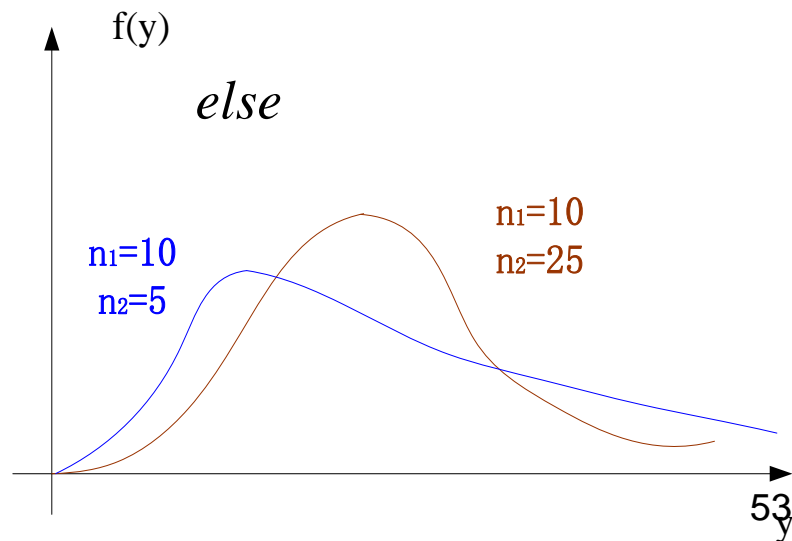
三、来自正态总体的几个常用统计量分布



F —分布

若 $F \sim F(n_1, n_2)$, F 分布的概率密度为

$$\psi(y) = \begin{cases} \frac{\Gamma\left(\frac{n_1 + n_2}{2}\right) (n_1/n_2)^{n_1/2} y^{n_1/2-1}}{\Gamma\left(\frac{n_1}{2}\right) \Gamma\left(\frac{n_2}{2}\right) \left(1 + \frac{n_1}{n_2} y\right)^{(n_1+n_2)/2}} & y > 0 \\ 0 & \text{else} \end{cases}$$



§ 3 抽样分布

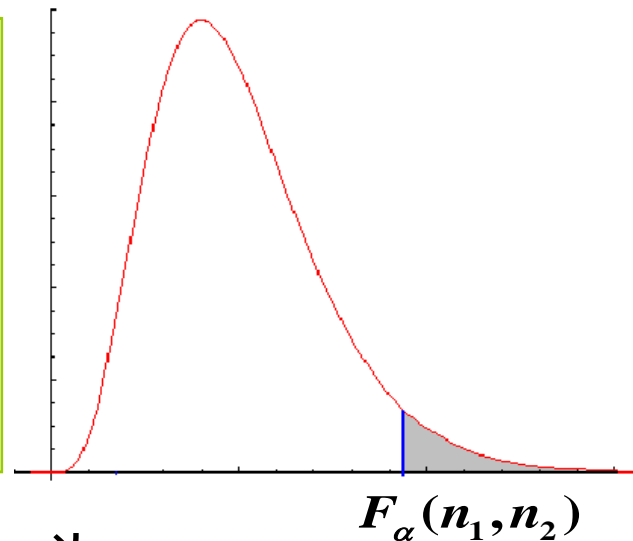
三、来自正态总体的几个常用统计量分布



F —分布

F 分布的上 α 分位点

对于 α : $0 < \alpha < 1$, 若存在 $F_{\alpha}(n_1, n_2) > 0$, 满足 $P\{F > F_{\alpha}(n_1, n_2)\} = \alpha$, 则称 $F_{\alpha}(n_1, n_2)$ 为 $F(n_1, n_2)$ 的上 α 分位点;



例 若 X 是服从 $F(6, 8)$ 的随机变量, 其中6为分子的自由度, 8为分母的自由度。试求:

- (1) 满足 $P\{X \geq A\} = 0.05$ 的 A ;
- (2) 满足 $P\{X < A\} = 0.05$ 的 B ;

§ 3 抽样分布

三、来自正态总体的几个常用统计量分布



F —分布

F 分布的上 α 分位点的性质

1. 当第一自由度 $n_1=1$ 时, $F(1,n)=[t(n)]$

证明 $t(n) = \frac{X}{\sqrt{Y/n}}, (X \sim N(0,1), Y \sim \chi^2(n)), X \text{与} Y \text{独立则}$

$$t^2(n) = \frac{X^2}{Y/n} = F(1,n)$$

§ 3 抽样分布

三、来自正态总体的几个常用统计量分布



F —分布

F 分布的上 α 分位点的性质

$$2. \quad F_{1-\alpha}(n_1, n_2) = \frac{1}{F_{\alpha}(n_2, n_1)}$$

证明： $F \sim F(n_1, n_2)$ 是已知条件


$$\begin{aligned} 1 - \alpha &= P\{F > F_{1-\alpha}(n_1, n_2)\} = P\left\{\frac{1}{F} < \frac{1}{F_{1-\alpha}(n_1, n_2)}\right\} \\ &= 1 - P\left\{\frac{1}{F} \geq \frac{1}{F_{1-\alpha}(n_1, n_2)}\right\} = 1 - P\left\{\frac{1}{F} > \frac{1}{F_{1-\alpha}(n_1, n_2)}\right\} \end{aligned} \quad \Rightarrow \quad P\left\{\frac{1}{F} > \frac{1}{F_{1-\alpha}(n_1, n_2)}\right\} = \alpha$$

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

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

§ 3 抽样分布

四、正态总体的几样本均值和样本方差的分布



设总体 X 的均值为 μ , 方差为 σ^2 , X_1, X_2, \dots, X_n 是来自总体 X 的一个样本, 则样本均值 \bar{X} 和样本方差 S^2 有下面结论成立:

$$E(\bar{X}) = \mu$$

$$D(\bar{X}) = \sigma^2/n$$

$$\begin{aligned} E(S^2) &= E\left[\frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right)\right] = \frac{1}{n-1} E\left[\frac{1}{n-1}\left(\sum_{i=1}^n (X_i^2) - n\sum_{i=1}^n (\bar{X}^2)\right)\right] \\ &= \frac{1}{n-1} E\left[\frac{1}{n-1}\left(\sum_{i=1}^n (\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2)\right)\right] = \sigma^2 \end{aligned}$$

§ 3 抽样分布

四、来自正态总体的几个常用统计量分布



相关定理

定理一(样本均值的分布)

设 X_1, X_2, \dots, X_n 是取自正态总体 $N(\mu, \sigma^2)$ 的样本, \bar{X} 是样本均值, 则有

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

例 从正态总体 $N(3.4, 6^2)$ 中抽取容量为 n 的样本, 如果要求其样本均值位于区间 $(1.4, 5.4)$ 内的概率不小于0.95, 问样本容量 n 至少应取多大

例 设总体 X 和 Y 相互独立, 且都服从正态分布 $N(30, 3^2)$, X_1, X_2, \dots, X_{20} 和 Y_1, Y_2, \dots, Y_{25} 是 X 和 Y 的样本, 求 $|\bar{X} - \bar{Y}| > 0.4$ 的概率

§ 3 抽样分布

四、来自正态总体的几个常用统计量分布



相关定理

定理二(样本方差的分布)

设 X_1, X_2, \dots, X_n 是取自正态总体 $N(\mu, \sigma^2)$ 的样本, \bar{X} , S^2 是样本均值和样本方差, 则有

$$(1) \bar{X} \text{ 和 } S^2 \text{ 相互独立} \quad (2) \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

$$(3) \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

证明: $U = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim N(0,1), V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

U 和 V 相互独立, 根据 t 分布定义有: $\frac{U}{\sqrt{V/(n-1)}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$

§ 3 抽样分布

四、来自正态总体的几个常用统计量分布



相关定理

定理二(样本方差的分布)

例 设在总体 $N(\mu, \sigma^2)$ 中抽取容量为16的样本，求：

$$(1) P\left\{\frac{S^2}{\sigma^2} \leq 1.6664\right\} \quad (2) D(S^2)$$

解 (1) 由定理二知 $\frac{(n-1)S^2}{\sigma^2} = \frac{15S^2}{\sigma^2} \sim \chi^2(15)$ 则有

$$P = P\{\chi^2(15) \leq 24.996\} = 0.95$$

$$(2) D(S^2) = D\left[\frac{\sigma^2}{n-1} \frac{(n-1)S^2}{\sigma^2}\right] = \left(\frac{\sigma^2}{n-1}\right)^2 D[\chi^2(n-1)] = \frac{2\sigma^4}{15}$$

§ 3 抽样分布

四、来自正态总体的几个常用统计量分布



相关定理

定理三(样本方差比的分布及样本均值差的分布)

设 X_1, X_2, \dots, X_{n_1} 与 Y_1, Y_2, \dots, 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。设 \bar{X} 和 \bar{Y} 分别是这两个样本的样本均值, S_1^2 和 S_2^2 分别是这两个样本的样本方差, 则有

$$(1) F = \frac{S_1^2 / S_2^2}{\sigma_1^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1) \quad \text{其中, } S_w = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

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

§ 3 抽样分布

四、来自正态总体的几个常用统计量分布



相关定理

(1)的证明: 由定理二知

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

S_1^2 和 S_2^2 相互独立, 则由F分布的定义知

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

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

§ 3 抽样分布

四、来自正态总体的几个常用统计量分布



相关定理

(2)的证明: 易知 $\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

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

由定理条件知 $\frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1)$, $\frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1)$

且它们相互独立, 故由 $V = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$

U 与 V 相互独立, 按 t 分布的定义知

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



例. 设 X_1, X_2, \dots, X_9 是来自正态总体 X 的简单随机样本,

$$Y_1 = \frac{1}{6}(X_1 + X_2 + \dots + X_6), \quad Y_2 = \frac{1}{3}(X_7 + X_8 + X_9),$$

$$S^2 = \frac{1}{2} \sum_{i=7}^9 (X_i - Y_2)^2, \quad Z = \frac{\sqrt{2}(Y_1 - Y_2)}{S}$$

证明: 统计量 Z 服从自由度为2的t分布。

例. 为研究甲、乙两地区学生的成绩, 随机抽取两个样本

甲地区: $n = 30, \quad \bar{x} = 86.5, \quad s_{1,n}^2 = 125.6,$

乙地区: $m = 20, \quad \bar{y} = 70.6, \quad s_{2,m}^2 = 110.4,$

若假定两地学生的成绩都服从正态分布, 您是否相信两地学生成绩的方差一样?

后话

大数据对统计学的冲击：

- 大数据是这个时代最常被引用的概念，可是其对于每个人的意义很不一样。对于一个统计学家而言，这是颠覆性的。
- 各种新奇的研究开始出现。



谢 谢!

