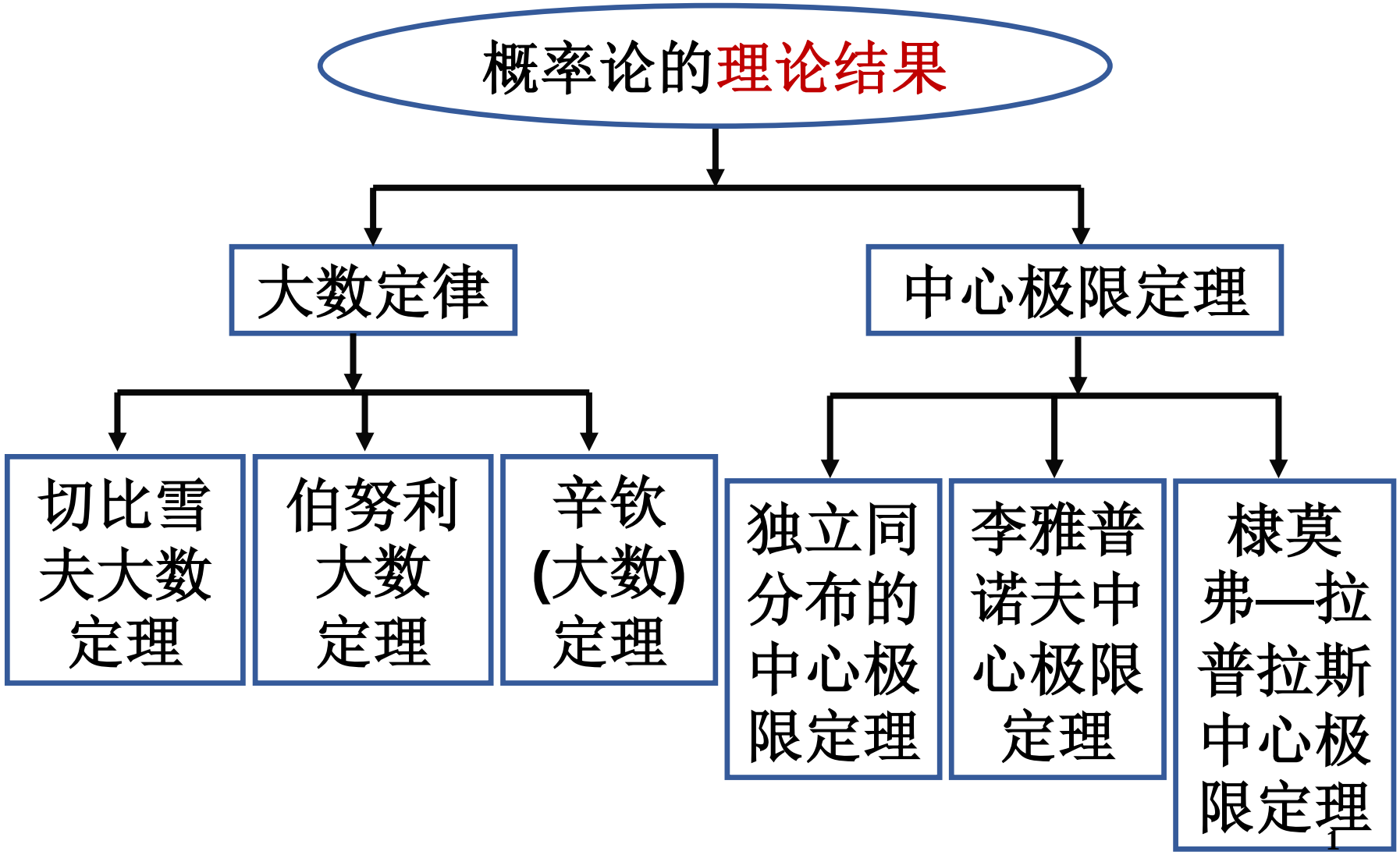
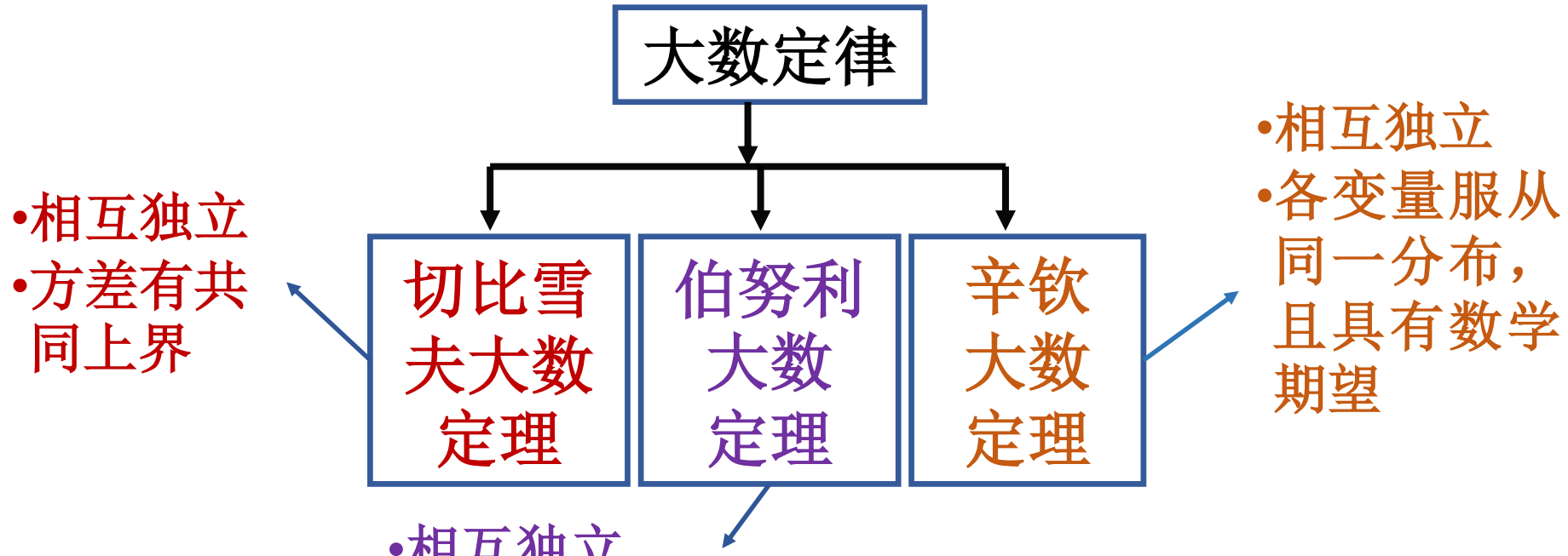
第五章 知识回顾


概率论的理论结果
大数定律
中心极限定理
切比雪夫大数定理
伯努利大数定理
辛钦(大数)定理
独立同分布的中心极限定理
李雅普诺夫中心极限定理
棣莫弗—拉普拉斯中心极限定理

第五章 知识回顾

➢在一定条件下,一列随机变量的算术平均值（按某种意义）收敛于这些项的均值的定理。（为用平均值估计期望提供了理论依据）

$$\lim_{n\to\infty} P\{|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E(X_i)|<\varepsilon\}=1$$



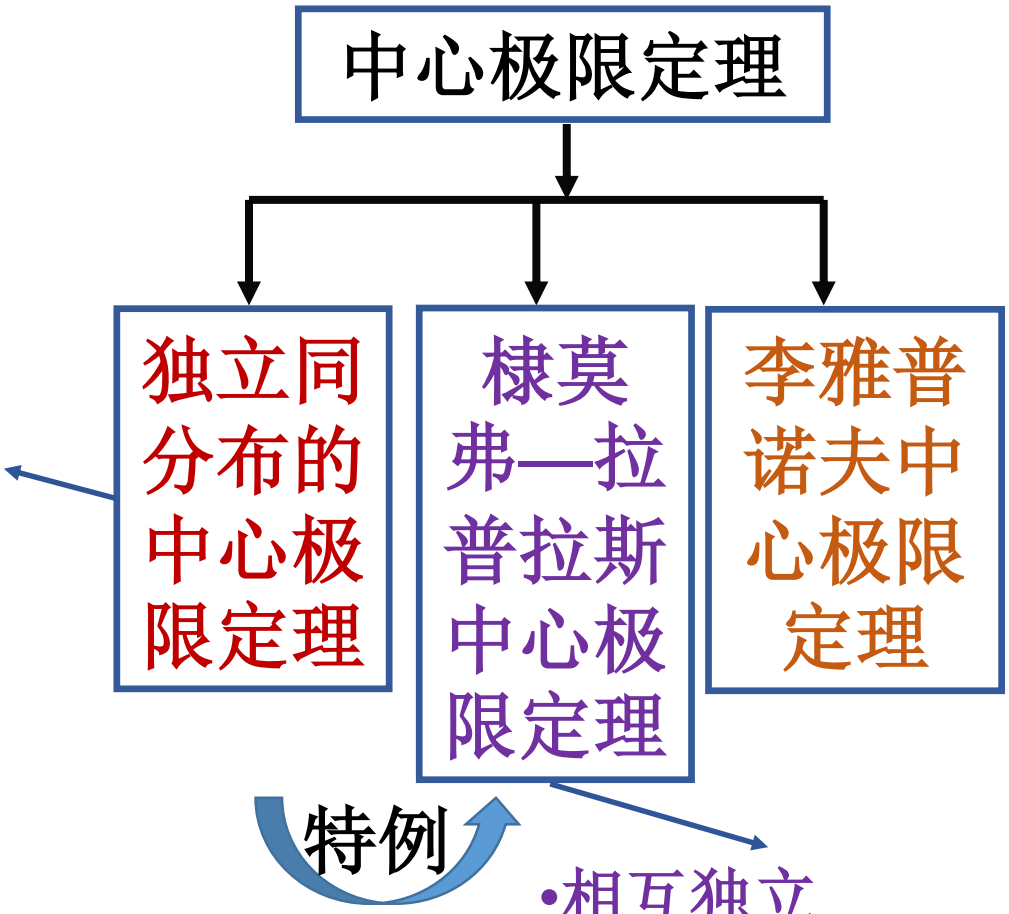
第五章 知识回顾

- 背景：如果一个量是由大量相互独立的随机因素的影响所造成，而每一个别因素在总影响中所起的作用不大， 则这种量一般都近似服从正态分布。

- 在一定条件下，大量相互独立的随机变量之和的概率分布近似于正态分布的定理—中心极限定理。

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P\left\{ \frac{\sum_{k=1}^{n} X_k - E(\sum_{k=1}^{n} X_k)}{\sqrt{D(\sum_{k=1}^{n} X_k)}} \le x \right\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

第五章 知识回顾



- 相互独立
- 同分布
- 有限数学期望和方差

- 相互独立
- 服从二项分布

- “由大量微小的、独立的随机因素”（不要求同分布）累积成的变量（即，满足具体条件）
- 随机因素个数趋于无穷时，以正态分布为极限。

$$\lim_{n\to\infty} P\left\{\frac{\eta_n - np}{\sqrt{np(1-p)}} \le x\right\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

第五章 知识回顾

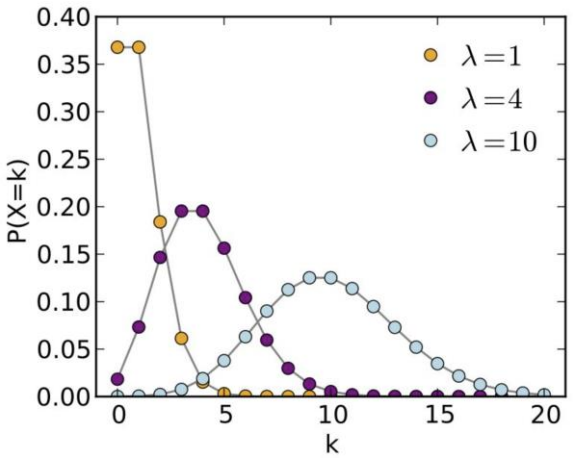
▲在第二章中已介绍当 $n \to \infty$ 时，二项分布以泊松分布为极限分布；而在本章中二项分布又以正态分布为极限分布。在实际计算中：

▲如果 n 很大，但 np或 nq 不大 (即 p 很小或 $q=1-p$ 很小)，那么应该用泊松定理去近似；

第五章 知识回顾

▲如果 n，np 或 nq 都较大，那么应该用中心极限定理去近似。

▲ 泊松分布 $X \sim P(\lambda)$ 的图形特点：



泊松分布的图形由λ决定：

（1）当λ较小时，泊松分布是偏峰的；

（2）随着λ增大，泊松分布逐渐趋于对称，接近正态分布。

总之，当随机变量数很大时，一般考虑中心极限定理的特点，用正态分布（连续）近似。

第六章
样本及抽样分布

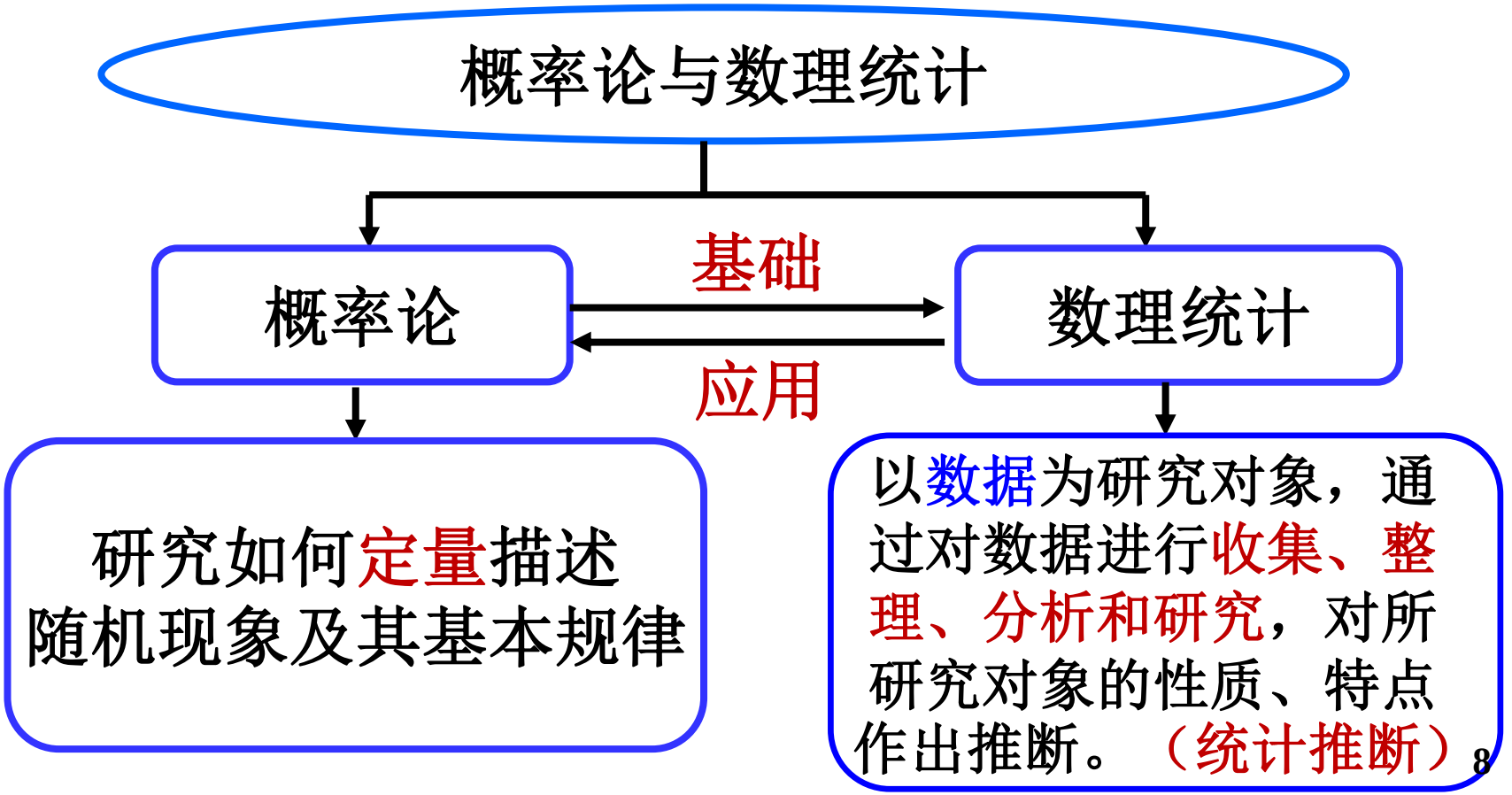
王笑尘

北京邮电大学网络空间安全学院

wxiaochen@bupt.edu.cn

什么是概率论与数理统计？

- 概率论与数理统计是研究和揭示随机现象统计规律性的一门数学学科。



什么是概率论与数理统计？

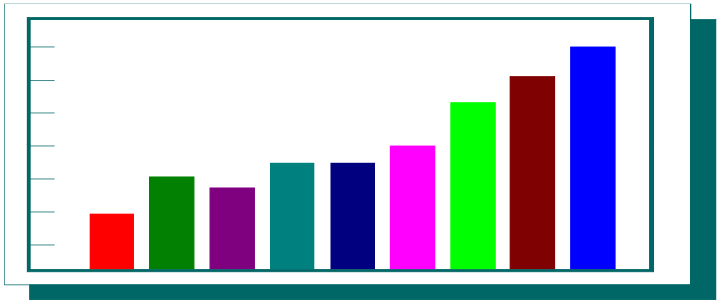
- 在概率论中所研究和讨论的随机变量，它的分布都是已知的，在这前提下去进一步的研究它的性质、特点和规律性。

- 而在数理统计中所研究和讨论的随机变量，它的分布是未知的或不完全知道的。于是就必须通过对所研究和讨论的随机变量进行重复独立的观察和试验，得到许多观察值(数据)，对这些数据进行分析后才能对其分布作出种种判断。

数理统计的客观背景

- 从历史的典籍中，人们不难发现许多关于钱粮、户口、地震、水灾等等的记载，这说明人们很早就开始了统计的工作。
- 但是当时的统计，只是对有关事实的简单记录和整理，而没有在一定理论的指导下，作出超越这些数据范围之外的推断。

- 到了十九世纪末二十世纪初，随着近代数学和概率论的发展，才真正诞生了数理统计学这门学科。
- 同时随着计算机的诞生与发展，为数据处理提供了强有力的技术支持，这就导致了数理统计与计算机结合的必然的发展趋势。
- 目前国内外著名的统计软件包：***R***，***SAS***，***SPSS***，***STAT*** 等，都提供了快速、简便地进行数据处理和分析的方法与工具。



数理统计研究的对象 --- 带有随机性的数据

数理统计的任务

数理统计学是一门应用性很强的学科，它是研究：

- 怎样以有效的方式收集、整理有限的数据资料；
- 如何对所得的数据资料进行分析、研究；
- 从而对所研究的对象的性质、特点作出推断。

直至为采取一定的决策和行动提供依据和建议。

数理统计的特征

--- 数理统计方法具有“部分推断整体”的特征

▲ 在数理统计中，不是对所研究的对象全体（称为总体）进行观察，而是抽取其中的部分（称为样本)进行观察获得数据（抽样），并通过这些数据对总体进行推断。

▲ 由于在数理统计中是从一小部分样本观察值去推断该全体对象（总体）情况，即由部分推断全体。所以这里使用的推理方法是“归纳推理”。

“部分推断整体”—要较好地反映所研究和讨论的随机变量整体的特性，就必须研究：

(1) 如何抽样，抽多少，怎么抽。

抽样方法问题

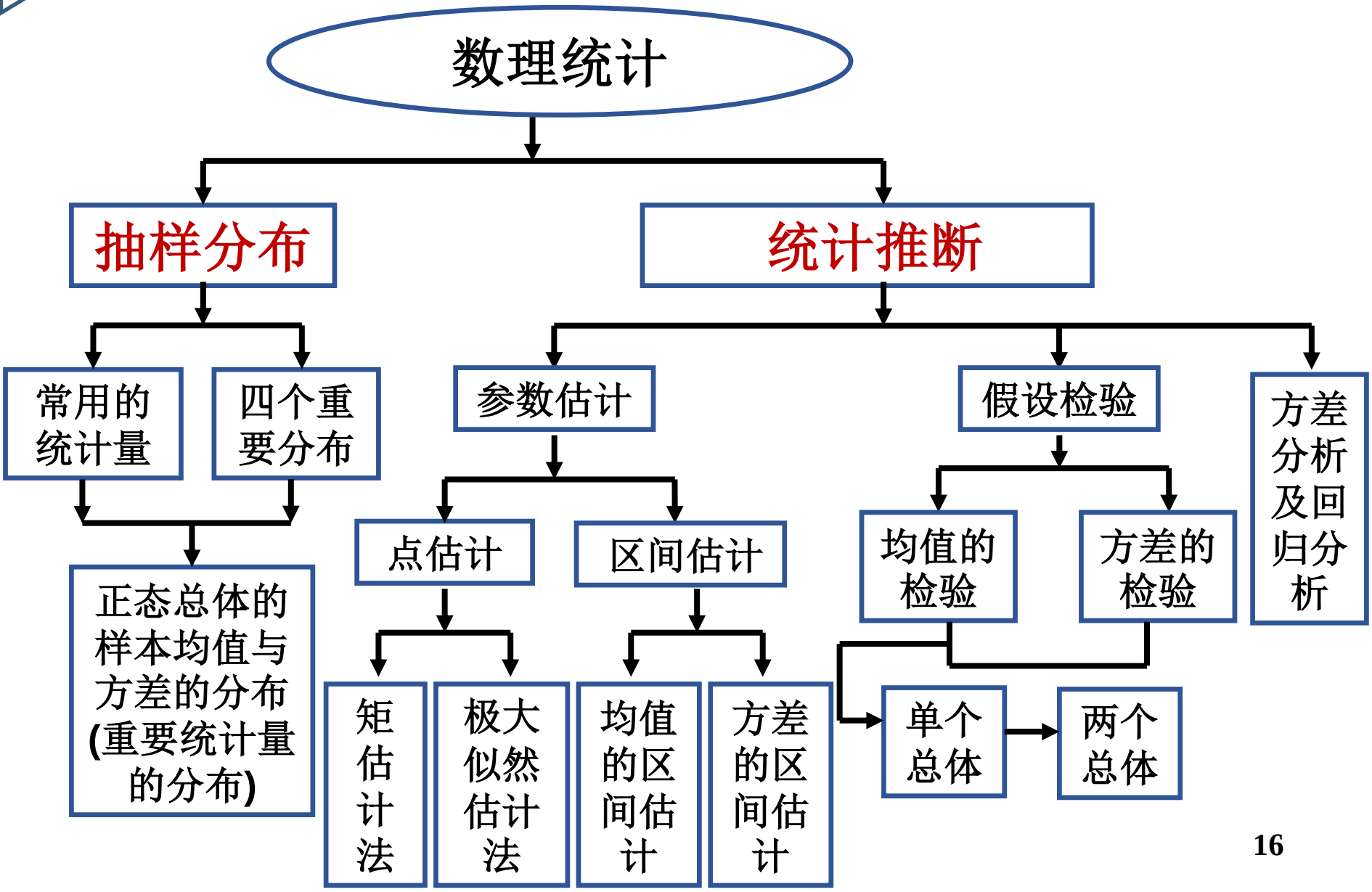
(2) 如何对抽样的结果进行合理分析，作出科学的判断。

统计推断问题

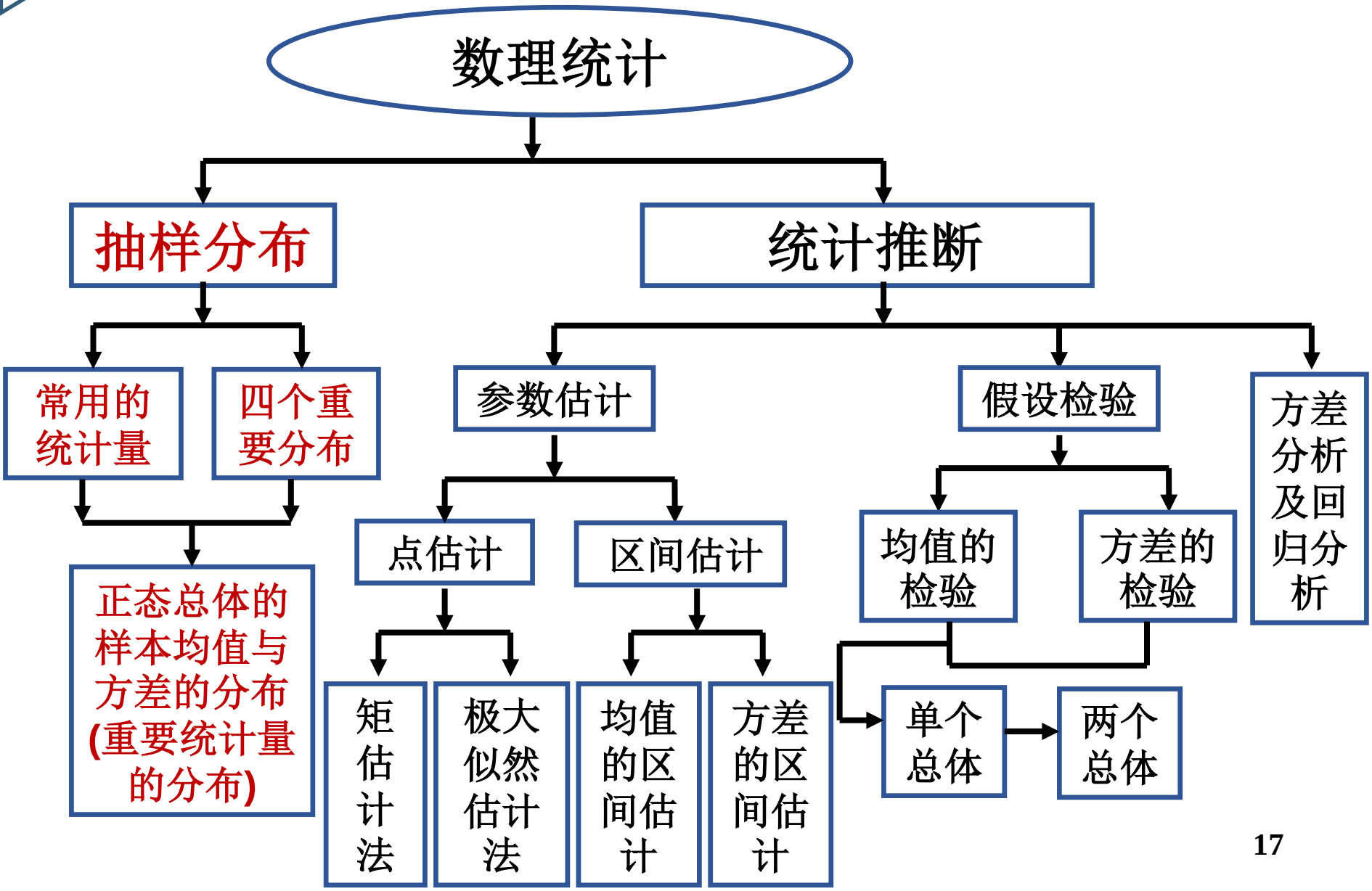
后面所讨论的统计问题主要属于下面这种类型：

1. 从所研究的随机变量的某个集合中抽取一部分元素；
2. 对这部分元素的某些数量指标进行试验与观察；
3. 根据试验与观察获得的数据来推断这集合中全体元素的数量指标的分布情况或数字特征。

数理统计部分知识结构图



数理统计部分知识结构图



第一节 随机样本

一. 总体和个体

定义 将研究对象的某项数量指标的值的全体称为总体(母体)；将总体中的每个元素称为个体。

例1. （1）当研究某地区职工月收入平均水平时，这地区所有职工的月收入组成了总体；而每个职工月收入就是个体。

（2）研究某批灯泡的质量，则该批灯泡寿命的全体就组成了总体；而每个灯泡的寿命就是个体。

…

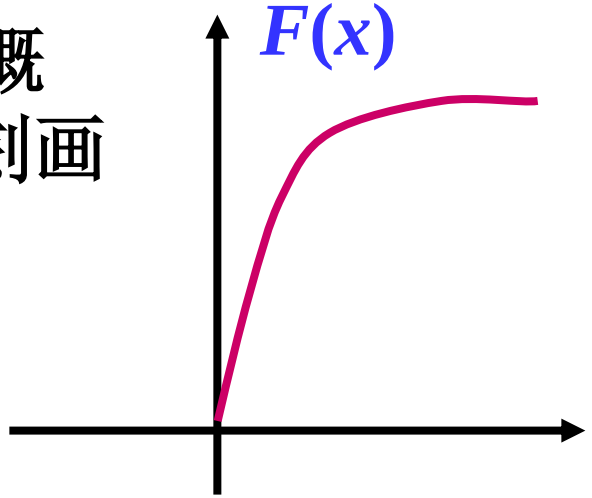
注：

- 由于每个个体（如，每个灯泡）的出现是随机的，所以相应的数量指标（如，灯泡寿命）的出现也带有随机性；
- 研究对象的某项数量指标可用一个随机变量X表示；
- 那么，一个总体（某项数量指标的全体）就对应于一个随机变量X；
- 因此，对总体的研究也就是对随机变量X的研究；
- 于是，X的分布函数和数字特征就称为总体的分布函数和数字特征。

总体—可以用一个随机变量及其分布来描述

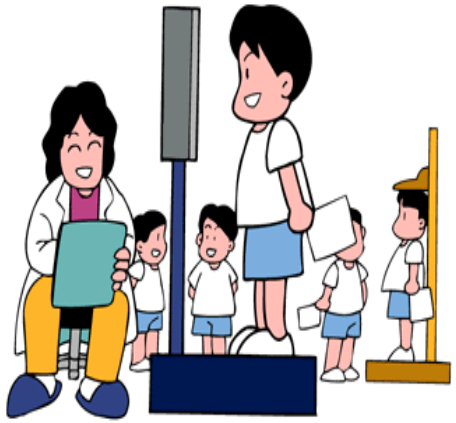
例2. (1) 研究某批灯泡的寿命时，关心的数量指标就是寿命，那么，此总体就可以用一维随机变量 X 表示，或用其分布函数 $F(x)$ 表示。

寿命 X 可用一概率分布函数来刻画



鉴于此，常用随机变量的记号或用其分布函数表示总体，如总体 X 或 总体 $F(x)$。

(2) 在研究某地区中学生的营养状况时，若关心的数量指标是身高和体重，现用 X 和 Y 分别表示身高和体重，则此总体可用二维随机变量 (X, Y) 或其联合分布函数 $F(x, y)$ 来表示。



注：▲ 在数理化统计中，总体这个概念的要旨是：

总体对应随机变量，其就是一个概率分布。

- 总体依其包含的个体总数分为有限总体(个体的个数是有限) 和 无限总体(个体的个数是无限的)。

- 但当有限总体它所含的个体的数很大时也可视其为无限总体。

二. 抽样和样本

抽样

- 为推断总体分布及各种特征，按一定规则从总体中抽取若干个体进行观察试验，以获得有关总体的信息，这一抽取过程称为“抽样”；
- 所抽取的部分个体称为 样本，样本中所包含的个体数目称为 样本容量。

例如:

- 从某批国产轿车中抽5 辆进行耗油量试验。这一过程即为“抽样”。
- 这 5 辆轿车为一个样本，其样本容量为 5。

定义1. 从总体中抽取一部分个体进行观察，被抽出的部分个体称为总体的一个样本。

注: ▲ 为了了解总体的分布，我们从总体中随机地抽取n个个体，记其指标值为 $x_1, x_2, \cdots x_n$ 则 $x_1, x_2, \cdots x_n$ 称为总体的一个样本/样本值。

▲ 由于每一次观测所取得的观测值 $x_1, x_2, \cdots x_n$ 具有随机性。因此，从另一个角度来讲，样本是一个随机变量 $(X_1, X_2, \cdots X_n)$。

▲ 即样本具有双重性：

(1) 是一个n维随机变量 $(X_1, X_2, \cdots X_n)$；

(2) 是n个具体的观察数值 $x_1, x_2, \cdots x_n$。

- 通常，$X_1, X_2, \cdots X_n$ 是相互独立的并与总体 X 具有相同的分布。一般称其为来自总体 X 的一个简单随机样本。
- 对于有限总体和无限总体都可以通过放回抽样的方式得到简单随机样本。
- 当个体的总数 N比要得到的样本容量n大得多时，可将不放回抽样近似地当作放回抽样来处理。

三. 简单随机样本

定义2 设 X 是具有分布函数 F 的随机变量，若 $X_1, X_2 \cdots X_n$ 是具有同一分布函数 F 的、相互独立的随机变量，则 称 $X_1, X_2 \cdots X_n$ 为从总体 X (或从总体 F 或从分布函数 F) 得到的容量为 n 的简单随机样本简称样本。

▲ 它们的观察值 $x_1, x_2, \cdots x_n$ 为样本值，又称为 X 的 n 个独立的观察值。

▲ 可视样本为一个随机向量，记为 $(X_1, X_2, \cdots X_n)$

从而，容量为 n 的样本可以看作 n 维随机变量。

▲ 若 $X_1, X_2, \cdots X_n$ 为总体 X 的一个样本，X 的分布函数为 $F(x)$，概率密度为 $f(x)$，则：

$X_1, X_2, \cdots X_n$ 联合分布函数为：

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

$X_1, X_2, \cdots X_n$ 联合概率密度为：

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

第二节 抽样分布

问题的提出

- 在上节所介绍内容中已经知道：样本是进行统计推断的依据。
- 但在实际应用时，往往不是直接使用样本本身，而是针对不同的问题构造样本的适当函数，利用这些样本的函数进行统计推断。
- 亦即用样本去推断总体情况，需要对样本进行一定的“加工”，这就要构造一些样本的适当函数，它把样本中所含的（某一方面）的信息集中起来。
- 这种不含任何未知参数的样本的函数称为统计量。它是完全由样本决定的量。

一. 统计量的定义

1. 定义 设 $X_1, X_2 \cdots X_n$ 是来自总体 X 的一个样本，

$g(X_1, X_2 \cdots X_n)$ 是 $X_1, X_2 \cdots X_n$ 的函数。

若 g 是连续函数且 g 中不含任何未知参数，

则称 $g(X_1, X_2 \cdots X_n)$ 是一个统计量。

注: ▲ 统计量是完全由样本确定的量。

▲ 统计量是样本的函数，所以也具有双重性。
(1) 当样本是随机变量时，统计量也是随机变量；(2) 当样本为观测值时，统计量是一个具体数值。

▲ 统计量的构造总是有目的的。

2. 几个常用的统计量

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

它反映了总体均值的信息

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

它反映了总体方差的信息，是总体方差的无偏估计(P145)

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

(4). 样本 k 阶原点矩： $A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$ $k=1, 2,\ldots$

它反映了总体k 阶矩的信息

(5). 样本 k 阶中心矩： $B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$

它反映了总体k 阶中心矩的信息

注：▲ (1) ~ (5)均是随机变量，实际上它们是样本的数字特征；它们的观察值是具体的实数值，仍称为样本均值、样本方差、样本 k 阶原点矩与样本 k 阶中心矩。

▲ 若总体 X 的 k 阶原点矩 $E(X^k)=\mu_k$ 存在，

则当 $n\to\infty$ 时有：$A_k \xrightarrow{p} \mu_k$，　$k=1,2,\cdots$

证明见后

这个结论表明：样本的 k 阶矩A_k依概率收敛到总体的 k 阶矩μ_k。这也是参数估计中的矩估计法的理论根据。

▲ 若总体 X 的 k 阶原点矩 $E(X^k)=\mu_k$ 存在，

则当 $n\to\infty$ 时有：$A_k \xrightarrow{p} \mu_k$，　$k=1,2,\cdots$

证明　因为$X_1, X_2,\ldots, X_n$独立且与X同分布，所以$X_1^k, X_2^k,\ldots, X_n^k$独立且与$X^k$同分布。故有$E(X_1^k)=E(X_2^k)=\cdots=E(X_n^k)=E(X^k)=\mu_k$。根据辛钦大数定律（独立同分布、且具有数学期望）：当 $n\to\infty$

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k \xrightarrow{p} E(X^k)=\mu_k$$

▲ 设总体X的均值为μ，方差为σ^2，$X_1, X_2, \ldots, X_n$为取自总体X的样本，则 (教材第5版P145)

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

3. 抽样分布

- 样本的函数为统计量；
- 统计量作为随机变量，因而就有一定的分布；
- 统计量的分布称为抽样分布。

4. 顺序统计量

定义： 设$(X_1,X_2,\ldots,X_n)$是从总体X中抽取的一个样本，$(x_1,x_2,\ldots,x_n)$是其中一个观测值，将观测值按从小到大的次序重新排列为：

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$$

定义$X_{(k)}$取值为$x_{(k)}(k=1,2,\ldots,n)$，由此得到

$$\left(X_{(1)}, X_{(2)}, \ldots, X_{(n)}\right)$$

称其为样本$(X_1,X_2,\ldots,X_n)$的顺序统计量，对应的$\left(x_{(1)}, x_{(2)}, \ldots, x_{(n)}\right)$称为其观测值。

$X_{(k)}$ 称为第 k 个顺序统计量(即它的每次取值总是取每次样本观测值由小到大排序后的第 k 个值).

特别的

$X_{(1)} = \min\limits_{1\le i\le n} X_i$ 为最小顺序统计量

$X_{(n)} = \max\limits_{1\le i\le n} X_i$ 为最大顺序统计量

说明

$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 也都是随机变量并且它们一般不相互独立，也不同分布。

例设总体X的分布为仅取 0, 1, 2 的离散均匀分布，

设总体 X 的分布如下：

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

现抽取容量为 3 的样本, 共有 27 种可能取值, 列表如下

x_1	x_2	x_3	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	x_1	x_2	x_3	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	x_1	x_2	x_3	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$
0	0	0	0	0	0	1	1	0	0	1	1	2	2	0	0	2	2
0	0	1	0	0	1	0	1	2	0	1	2	1	1	2	1	1	2
0	1	0	0	0	1	0	2	1	0	1	2	1	2	1	1	1	2
1	0	0	0	0	1	1	0	2	0	1	2	2	1	1	1	1	2
0	0	2	0	0	2	2	0	1	0	1	2	1	2	2	1	2	2
0	2	0	0	0	2	1	2	0	0	1	2	2	1	2	1	2	2
2	0	0	0	0	2	2	1	0	0	1	2	2	2	1	1	2	2
0	1	1	0	1	1	0	2	2	0	2	2	1	1	1	1	1	1
1	0	1	0	1	1	2	0	2	0	2	2	2	2	2	2	2	2

由此可得 $X_{(1)}$，$X_{(2)}$，$X_{(3)}$ 的概率分布如下：

$X_{(1)}$	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

其分布
各不相同

进而可得 $X_{(1)}$与 $X_{(2)}$ 的联合分布如下：

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

$X_{(1)}$与$X_{(2)}$
并不独立

$$P(X_{(1)}=0)P(X_{(2)}=0)=\frac{19}{27}\cdot\frac{7}{27}\text{，而}P(X_{(1)}=0,X_{(2)}=0)=\frac{7}{27}$$

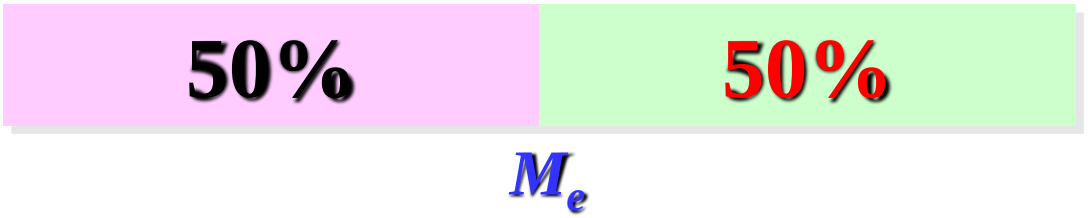
注: 在一个样本中, $X_1, X_2, \cdots\cdots, X_n$ 是独立同分布的, 而顺序统计量 $X_{(1)}, X_{(2)} \cdots\cdots, X_{(n)}$ 则可能既不独立, 分布也不相同。

常用顺序统计量

1、极差 $R = X_{(n)} - X_{(1)}$为样本极差

极差反映了随机变量X取值的分散程度。

2、中位数 排序后处于中间位置上的值



数值确定：$M_e = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & n\text{为奇数} \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right), & n\text{为偶数} \end{cases}$

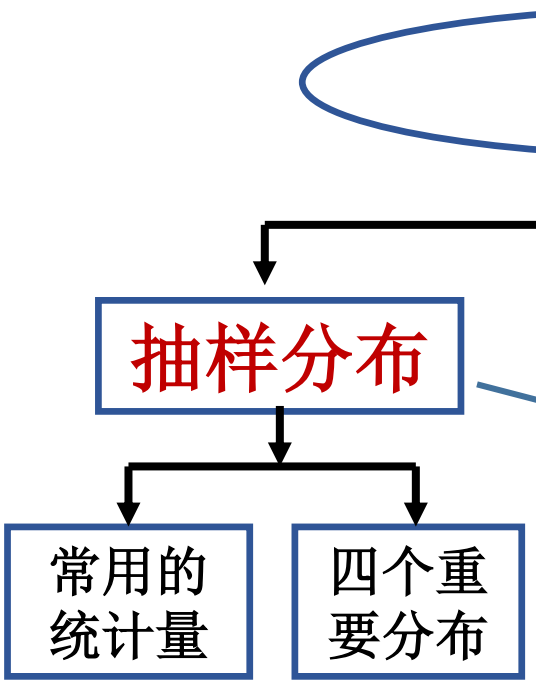
常用顺序统计量

3、分位数

设$X_{(1)} \le X_{(1)} \le \cdots \le X_{(n)}$为取自总体 **X** 的次序统计量，称 M_p为**p**分位数。

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

第六章 样本及抽样分布



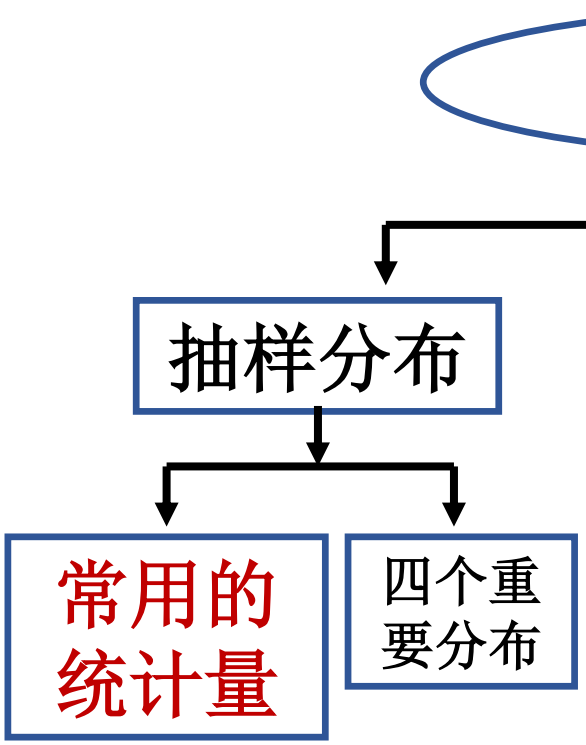
基本概念：

1、个体和总体（对应一个随机变量，即总体X）；

2、 抽样和样本（具有双重性：变量/观察数值；用于推断总体特征）；

3、简单随机样本：相互独立，且与总体同分布的n维随机变量。

第六章 样本及抽样分布



1、针对不同的问题/研究目的，由样本构成的函数；

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

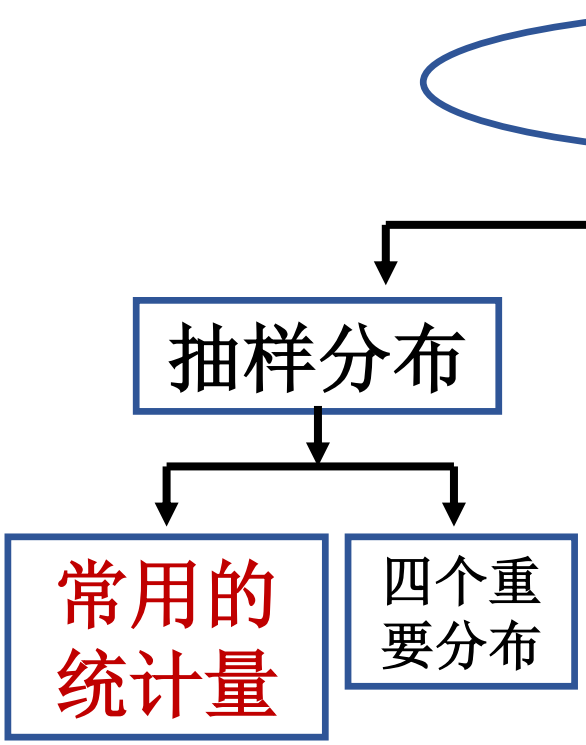
3、样本方差

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

4、样本k阶原点矩$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$；

5、样本k阶中心矩$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$。

第六章 样本及抽样分布



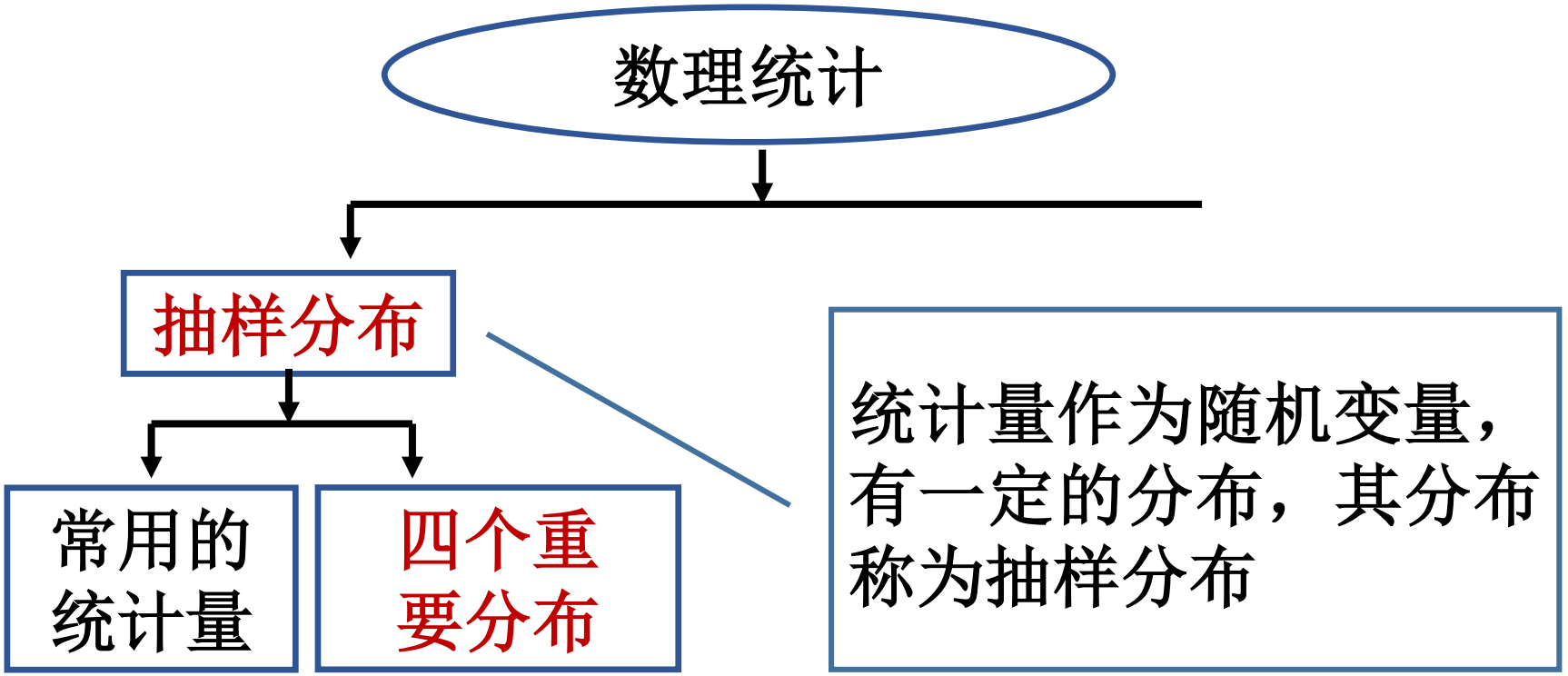
▲ 若总体 X 的 k 阶原点矩 $E(X^k)=\mu_k$ 存在，

则当 $n\to\infty$ 时有：$A_k \xrightarrow{p} \mu_k$，　$k=1,2,\cdots$

▲ 设总体X的均值为μ，方差为σ^2，$X_1, X_2, \ldots, X_n$ 为取自总体X的样本，则 (教材第5版P145)

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

第六章 样本及抽样分布



二. 几个重要的分布

1. χ^2 分布

定义. 设 $X_1, X_2, \cdots, X_n$ 是来自正态分布 $N(0, 1)$ 的样本，则称统计量：

$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$ 为服从自由度为 n 的 χ^2 分布。记为：$\chi^2 \sim \chi^2(n)$

注: ▲ 自由度 n 是指 χ^2 中所包含独立变量的个数。

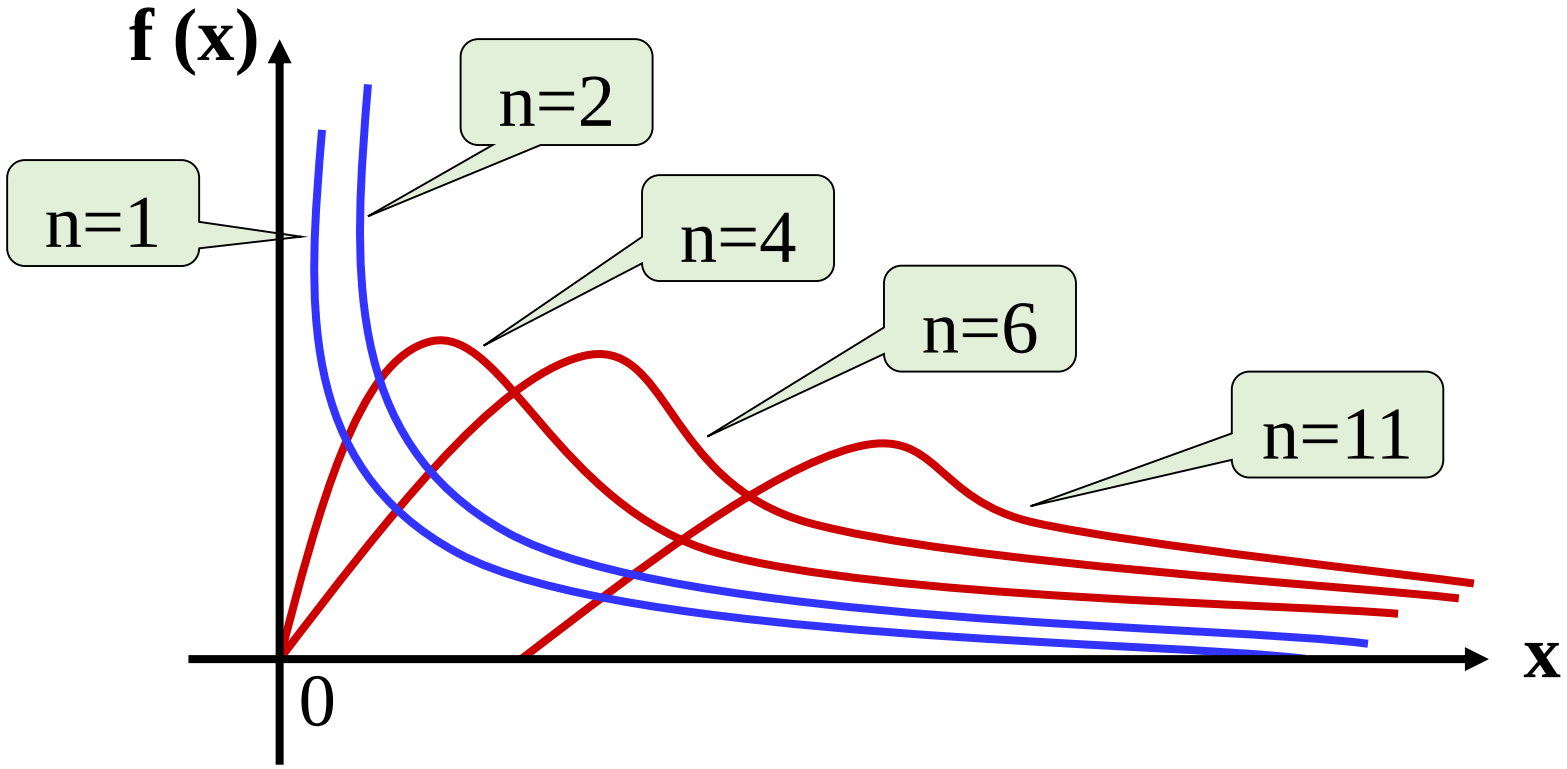
▲ χ^2分布的概率密度函数为：

$$f(x;n)=\begin{cases}\dfrac{1}{2^{n/2}\Gamma(n/2)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}} & x\geq 0\\ 0 & x<0\end{cases}$$

其中：伽玛函数 $\Gamma(x)$ 通过积分：

$$\Gamma(x)=\int_0^{\infty}e^{-t}t^{x-1}dt,\quad x>0$$ 来定义。

其图形如下：



（参见教材 （第五版）P142 图 6-7）

▲若 $\chi^2 \sim \chi^2(n)$，则 $E(\chi^2)=n$, $D(\chi^2)=2n$

▲ χ^2 分布的可加性：

若 $\chi_1^2 \sim \chi^2(n_1)$, $\chi_2^2 \sim \chi^2(n_2)$ 且

χ_1^2, χ_2^2 相互独立，则 $\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$

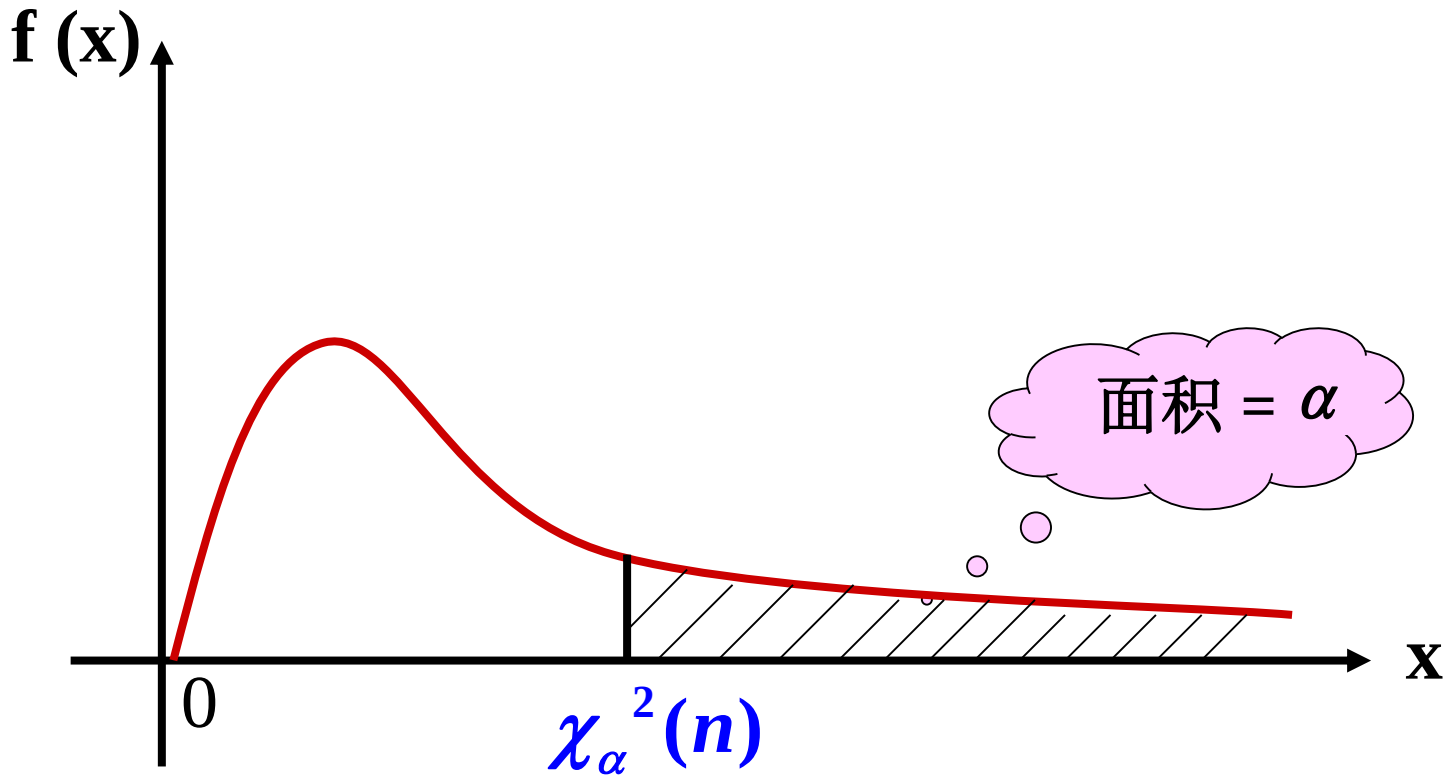
▲ χ^2 分布的上 α 分位点：

对于给定的 α $(0 < \alpha < 1)$, 称满足：

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

为 χ^2 分布的上 α 分位点。

其图形如下：



对于不同的α 与 n , $\chi^2(n)$有表可查（见教材P400附表5）。

一般: (a) 当 $n \leq 40$时可直接查表

(b) 当 $n > 40$时可用近似公式：

$$\chi^2_\alpha(n) \approx \frac{1}{2}(z_\alpha + \sqrt{2n-1})^2$$

费歇R.AFisher证明

z_α 是正态分布的上 α 分位点

例如：

$$\chi^2_{0.1}(25) = 34.382 \Longleftrightarrow P(\chi^2(25) > 34.382) = 0.1$$

$$\chi^2_{0.95}(40) = 26.509 \Longleftrightarrow P(\chi^2(40) > 26.509) = 0.95$$

$$\chi^2_{0.05}(50) \approx \frac{1}{2}(z_{0.05} + \sqrt{2\times 50 - 1})^2$$

$$= \frac{1}{2}(1.645 + \sqrt{99})^2 = 67.221$$

2. t 分布

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

$$T = \frac{X}{\sqrt{Y/n}}$$

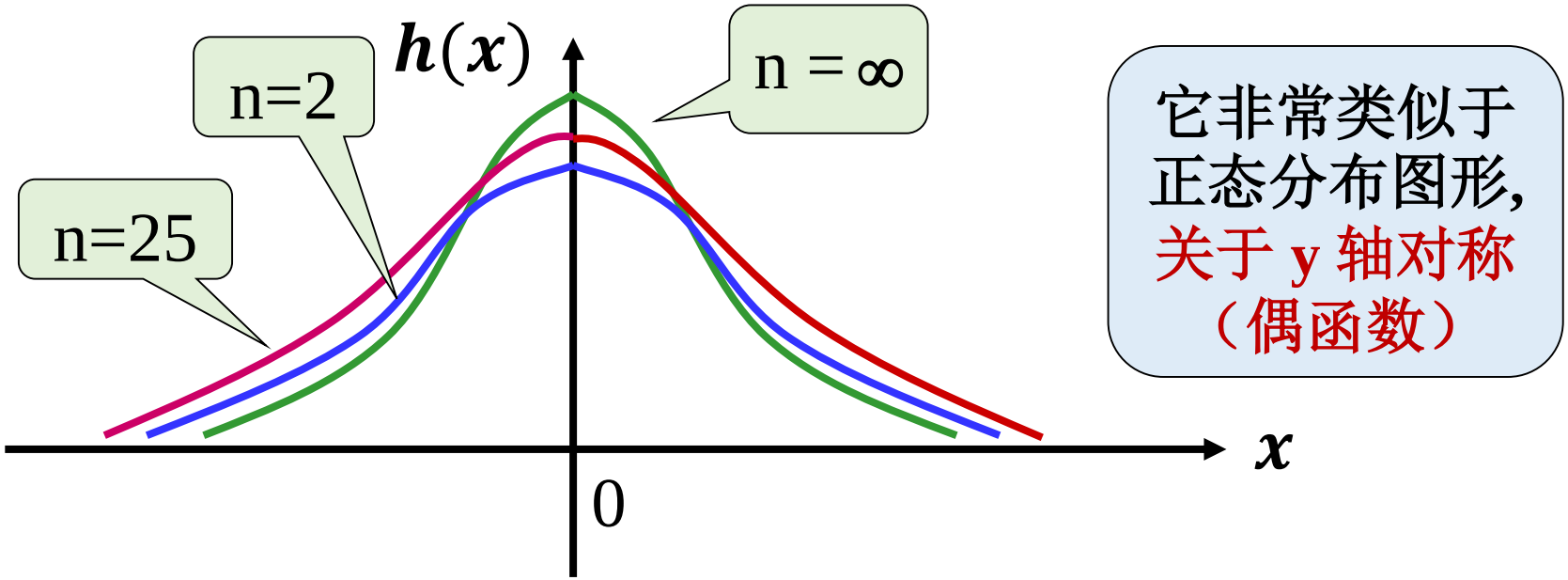
为服从自由度为 n 的 t 分布。记为 $T \sim t(n)$

注: ▲ t 分布是英国统计学家哥塞特（Gosset）首先发现的，并以学生(student)的笔名在英国的《Biometrike》杂志上发表的一篇文章中提出了他的研究结果，故 t 分布也称为学生分布。

▲ t 分布的概率密度函数为：

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

其图形如下：



它非常类似于正态分布图形，关于 y 轴对称（偶函数）

（参见教材(第五版) P143 图 6-9）

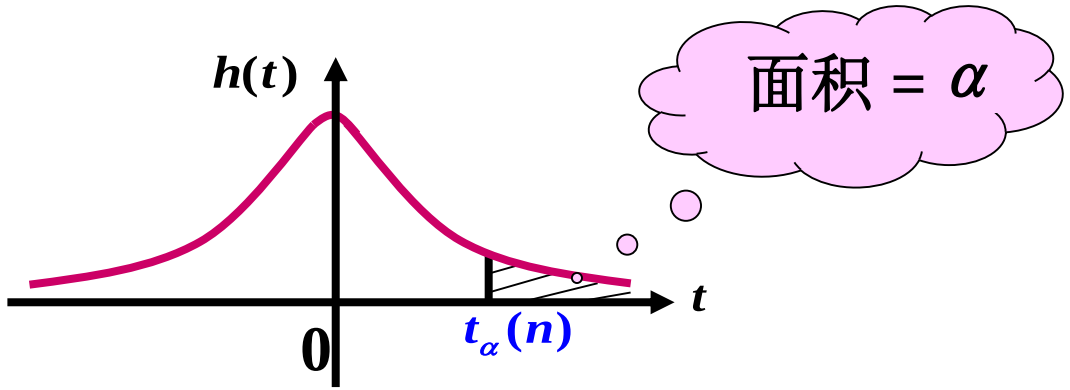
▲ 当 n 充分大时，$\lim\limits_{n\to\infty} h(x;\, n) = \dfrac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}$

即当 n 充分大时，t 分布可以近似看作是标准正态分布；但当 n 较小时， t 分布与正态分布的差异是不能忽略的。

▲ T 分布的上 α 分位点：对于给定的 α ，$(0<\alpha<1)$

称满足条件：$P(t>t_\alpha(n))=\int_{t_\alpha(n)}^{\infty}h(t)\,dt=\alpha$

的点 $t_\alpha(n)$ 为 t 分布的上 α 分位点。



对于不同的 α 与 n，$t_\alpha(n)$ 有表可查（见教材 P399附表4）

一般：

(a) 当 $n\le 45$ 时可直接查表

(b) 当 $n>45$ 时可用近似公式：

$$t_\alpha(n)\approx z_\alpha$$ (用正态分布近似)

例如：

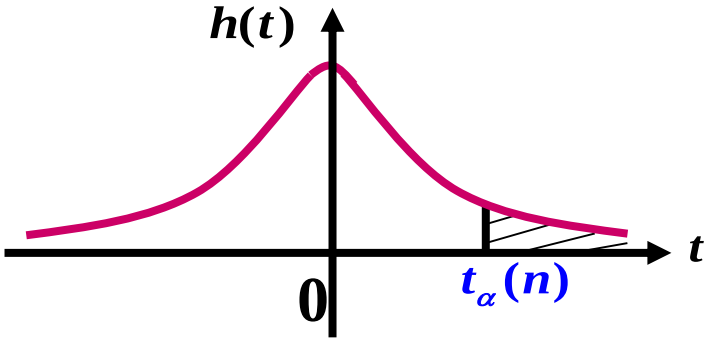
$$t_{0.05}(30)=1.6973 \Longleftrightarrow P(t(30)>1.6973)=0.05$$

$$t_{0.01}(35)=2.4377 \Longleftrightarrow P(t(35)>2.4377)=0.01$$

$$t_{0.05}(50)\approx z_{0.05}=z(1-0.05)=z(0.95)=1.645$$

▲ 由上α 分位点定义及 $h(t)$ 对称性得：

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



3. F分布

定义. 设 $X \sim \chi^2(n_1)$, $Y \sim \chi^2(n_2)$, X 与 Y 相互独立，则称统计量：

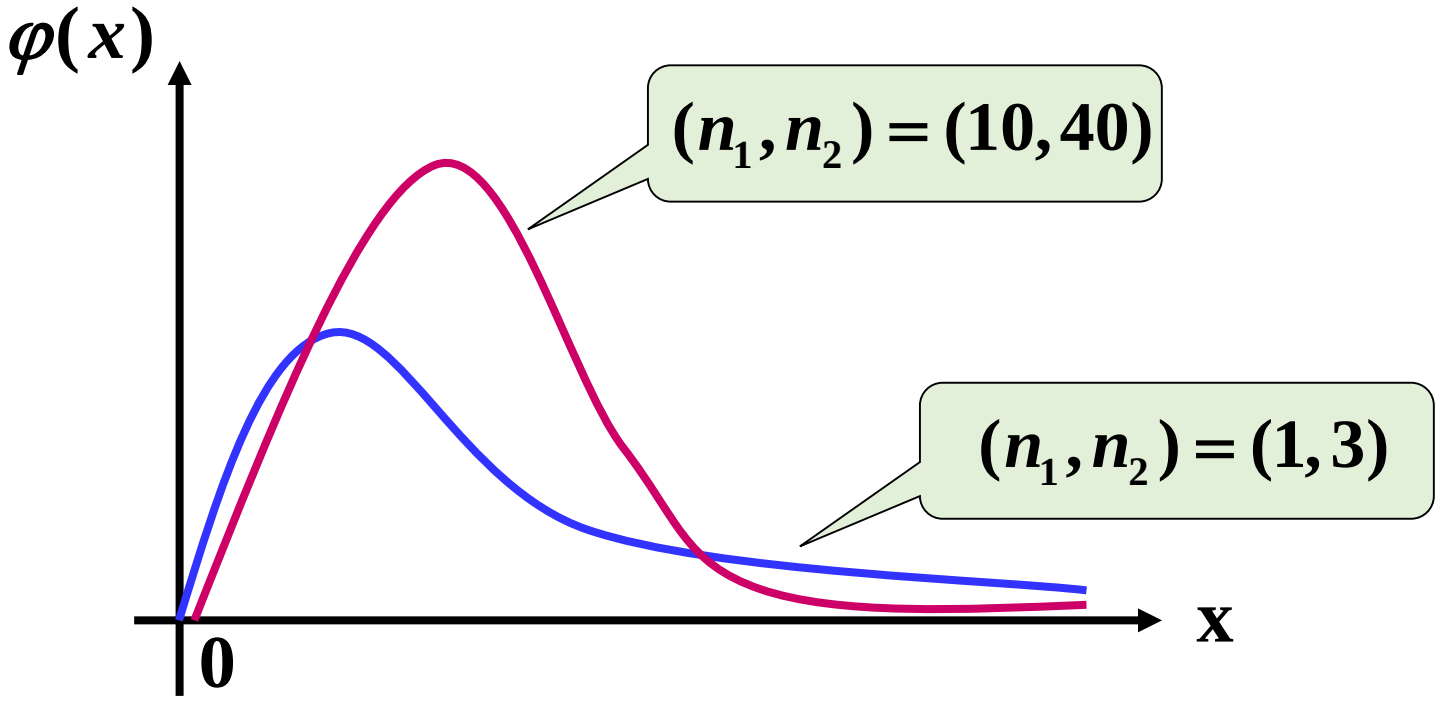
$$F = \frac{X/n_1}{Y/n_2}$$

为服从自由度 n_1 及 n_2 的 F 分布，记作：$F \sim F(n_1, n_2)$

注: ▲ 若 $F \sim F(n_1, n_2)$，则 F 的概率密度为：

$$\varphi(x; n_1, n_2) = \begin{cases} \dfrac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} (\frac{n_1}{n_2})(\frac{n_1}{n_2}x)^{\frac{n_1}{2}-1} \left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

其图形如下：



（参见教材(第五版) P144 图 6-11）

F分布是为纪念英国著名统计学家费歇尔（R.A.Fisher，1890~1962）而命名的。是数理统计的重要分布之一。

▲ 若 $F \sim F(n_1, n_2)$ 则 $\dfrac{1}{F} \sim F(n_2, n_1)$

▲ 若 $X \sim t(n)$ 则: $X^2 \sim F(1, n)$

证明: ∵ $X \sim t(n)$, 所以由 t 分布的定义, 即:

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

$$\therefore \ X^2 = \left(\frac{\mu}{\sqrt{\chi^2(n)/n}}\right)^2 = \frac{\mu^2}{\chi^2(n)/n} = \frac{\chi^2(1)/1}{\chi^2(n)/n}$$

由 F 分布定义得:

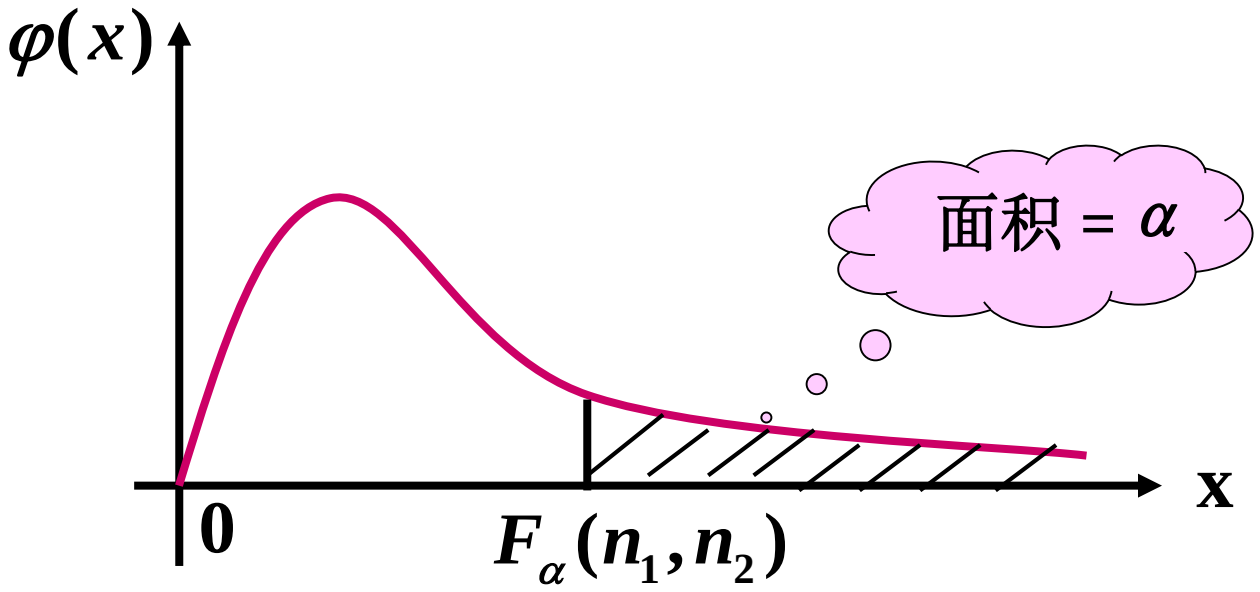
$$X^2 \sim F(1, n)$$

▲ F 分布的上 α 分位点：对于给定的 α ，$(0<\alpha<1)$

称满足条件：

$$P(F > F_\alpha(n_1,n_2)) = \int_{F_\alpha(n_1,n_2)}^{\infty} \varphi(x)dx = \alpha$$

的点 $F_\alpha(n_1,n_2)$ 为 F 分布的上 α 分位点。



对于不同的 α 与 n ，$F_\alpha(n_1,n_2)$ 有表可查（见教材P401附表6）

▲ F 分布的上 α 分位的性质(推导见P145)：

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

例如：

$$F_{0.05}(15,12)=2.62 \Longleftrightarrow P(F>2.62)=0.05$$

$$F_{0.95}(15,12)=F_{1-0.05}(15,12)$$

$$=\frac{1}{F_{0.05}(12,15)}=\frac{1}{2.48}\approx 0.04$$

$$\Longleftrightarrow \quad P(F>0.04)=0.95$$

三. 正态分布的样本均值与样本方差的分布

4. 正态分布—复习

(1). 正态分布的定义

若随机变量 X 的概率密度为：

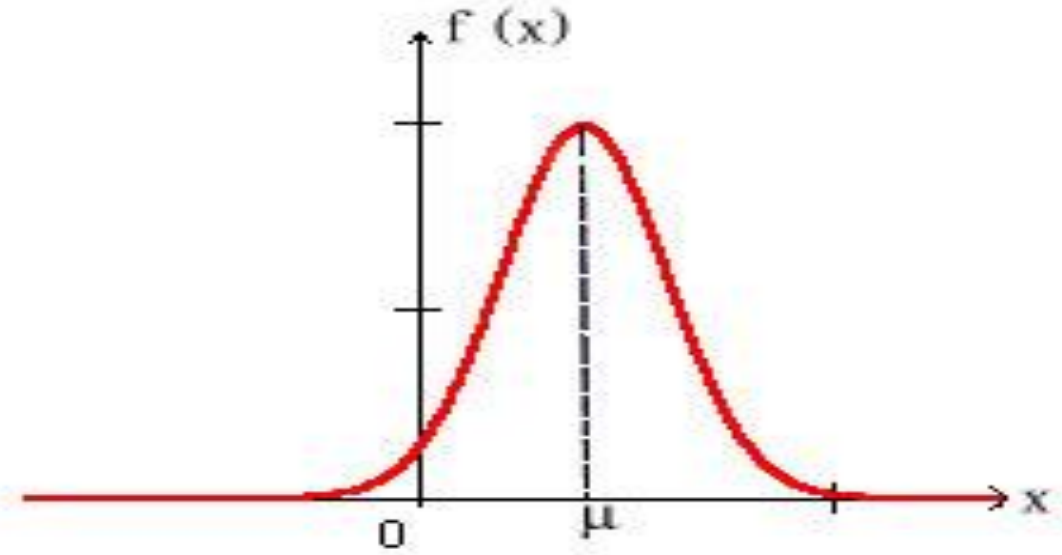
$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\quad -\infty<x<\infty$$

其中：μ和σ^2都是常数，μ任意，$\sigma>0$，
则称X服从参数为μ和σ^2的正态分布。记作：

$$X\sim N(\mu,\sigma^2)$$

$f(x)$ 所确定的曲线叫作正态曲线。

(2). 正态分布$N(\mu,\sigma^2)$ 的图形特点



正态分布的密度曲线是一条关于μ对称的钟形曲线，特点是“两头小，中间大，左右对称”。

三. 正态分布的样本均值与样本方差的分布

定理 1 (样本均值和样本方差的分布)

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

$\overline{X}, S^2$ 是其样本均值和样本方差

则

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

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

(3) $\overline{X}$ 和 S^2 相互独立

(4) $E(S^2) = \sigma^2$

只证（1），（2）与（3）的证明见教材（第五版）P146

证明：(1) 由已知 $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$,

$\therefore\ X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

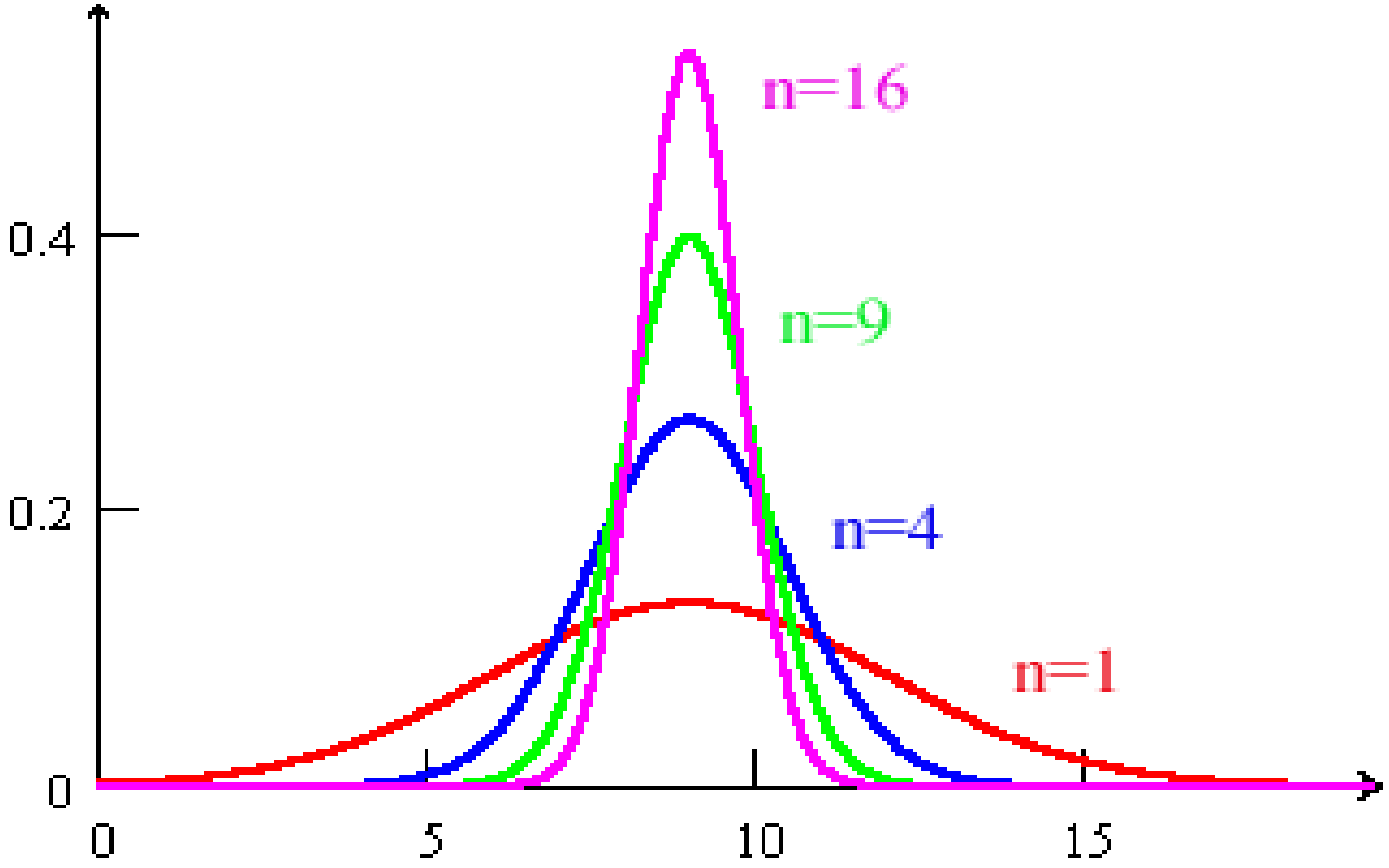
正态分布的可加性

若 $X_i \sim N(\mu, \sigma^2)$,

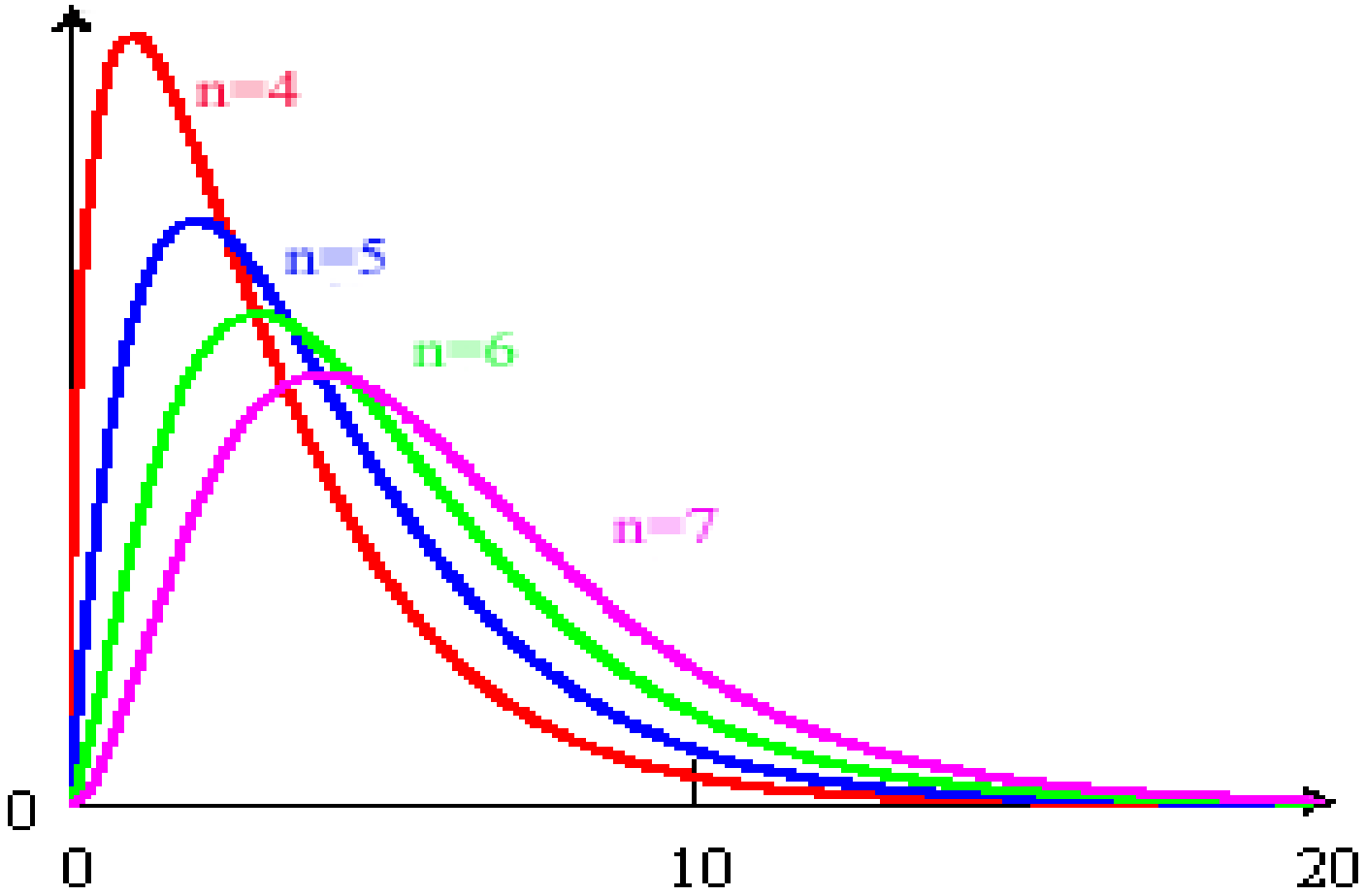
则 $\sum_{i=1}^{n} X_i \sim N(n\mu,\ n\sigma^2)$

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

则：$\frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu\ , \frac{\sigma^2}{n})$ 即 $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$



n 取不同值时样本均值 $\overline{X}$ 的分布



n 取不同值时 $\dfrac{(n-1)S^2}{\sigma^2}$ 的分布

推论. 设 $X_1, X_2 \cdots X_n$ 是总体 $N(\mu, \sigma^2)$ 的一个样本，

则 $\dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

注: ▲ 推论的实质是把服从一般正态分布的随机变量 $\overline{X}$ 化为标准正态分布的一个方法。它类似于把一个随机变量 $X \sim N(\mu, \sigma^2)$ 经线性变换 $Z = \dfrac{X - \mu}{\sigma}$ 化为服从标准正态分布。

▲ 对于一般的有：

$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ 是 $\overline{X}$ 的线性函数

$$P(a<\overline{X}\le b)=P\left(\frac{a-\mu}{\sigma/\sqrt{n}}<\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\le\frac{b-\mu}{\sigma/\sqrt{n}}\right)$$

由推论

$$=\Phi\left(\frac{b-\mu}{\sigma/\sqrt{n}}\right)-\Phi\left(\frac{a-\mu}{\sigma/\sqrt{n}}\right)$$

定理 2. 设 $X_1, X_2, \ldots, X_n$ 是取自正态总体 $N(\mu, \sigma^2)$

的样本, $\bar{X}$ 和 S^2 分别为样本均值和样本方差,

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

证明:

定理 2. 设 $X_1, X_2, \ldots, X_n$ 是取自正态总体 $N(\mu, \sigma^2)$

的样本, $\bar{X}$ 和 S^2 分别为样本均值和样本方差,

则有: $\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

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

并且两者相互独立

$\therefore$ 由 t 分布的定义得:

由定理1 的结论与 推论

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

定理 3. $X \sim N(\mu_1, \sigma_1^2)$，$Y \sim N(\mu_2, \sigma_2^2)$，且 X 与 Y 相互独立，$X_1, X_2, \cdots, X_{n_1}$ 是取自 X 的样本，$Y_1, Y_2, \cdots, Y_{n_2}$ 是取自 Y 的样本。$\overline{X}$ 和 $\overline{Y}$ 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) $$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1);$$

(2) 当 $\sigma_1^2=\sigma_2^2=\sigma^2$ 时，

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{s_W\cdot\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

其中：$s_W^2=\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$

证明见后

证明：（1）

$$\because \ \frac{(n_1-1)S_1^{\ 2}}{\sigma^2} \sim \chi^2(n_1-1)$$

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

由假设 S_1^2, S_2^2 相互独立，则由F分布的定义知

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

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

（2） 当 $\sigma_1^2=\sigma_2^2=\sigma^2$ 时，

$$\because \quad \overline{X}\sim N(\mu_1,\frac{\sigma^2}{n_1}),\ \ \overline{Y}\sim N(\mu_2,\frac{\sigma^2}{n_2})$$

而 $(-\overline{Y})\sim N(-\mu_2,\frac{\sigma^2}{n_2})$

$$\therefore\ \overline{X}-\overline{Y}\sim N(\mu_1-\mu_2,\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2})$$

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

由定理1推论

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

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

由χ^2分布的可加性

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

则由 t 分布定义得：

$$\frac{U}{\sqrt{V/(n_1+n_2-2)}} = \frac{\dfrac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}}{\sqrt{\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2}\Big/(n_1+n_2-2)}}$$

$\sim N(0,1)$

$\sim \chi^2(n_1+n_2-2)$

$$= \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{s_w\cdot\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \sim t(n_1+n_2-2)$$

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

例1. 在总体 $N(52, 6.3^2)$ 中随机抽取一容量为 36 的样本,

求：样本均值 $\overline{X}$ 落在 50. 8 到 53. 8 之间的概率

解:

例1. 在总体 $N(52, 6.3^2)$ 中随机抽取一容量为 36 的样本,

求：样本均值 $\overline{X}$ 落在 50. 8 到 53. 8 之间的概率

解: ∵ 样本的容量为 36

$$\therefore \text{样本均值 } \overline{X} \sim N(52, \frac{6.3^2}{36}) = N(52, 1.05^2)$$

从而：

$$P(50.8 < \overline{X} < 53.8) = P\{\frac{50.8-52}{1.05} < \frac{\overline{X}-52}{1.05} < \frac{53.8-52}{1.05}\}$$

$$= \Phi(1.71) - \Phi(-1.14)$$

$$= \Phi(1.71) + \Phi(1.14) - 1$$

$$= 0.9564 + 0.8729 - 1 = 0.8293$$

第六章作业（教材第五版）：

P149：2、4、5、6、7、8、9

注：作业不得抄袭；写上姓名、班级、学号和页码（如1/5），待第七章讲授结束，与第五章和第七章作业一起提交至教学云平台，标明题目属于第五章/第六章/第七章。