§4.4 大数定律与中心极限定理

概率论与数理统计是研究随机现象统计规律性的学科. 随机现象的规律性只有在相同的条件下进行大量重复试验时才会呈现出来. 也就是说，要从随机现象中去寻求必然的法则，应该研究大量随机现象.

研究大量的随机现象，常常采用极限形式，由此导致对极限定理进行研究. 极限定理的内容很广泛，其中最重要的有两种:

大数定律
law of large numbers

与

中心极限定理
central limit theorem

1. 切比雪夫不等式Chebyshev's Inequality

重要

设随机变量 X 的期望$E(X)$与方差 $D(X)$ 存在，则对于任意实数 $\varepsilon > 0$,

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{D(X)}{\varepsilon^2}$$

或 $$P(|X - E(X)| < \varepsilon) \geq 1 - \frac{D(X)}{\varepsilon^2}$$

样本值围绕期望依方差波动

Chebyshev's Inequality补充

设随机变量 X 的期望$E(X)=\mu$与方差 $D(X)=\sigma^2$ 存在，则对于任意实数 $\varepsilon > 0$,

$$P(\left|\frac{X-\mu}{\sigma}\right| \geq \varepsilon) \leq \frac{1}{\varepsilon^2}$$

特别的:

$$P(|Z| > 2) \leq 1/4$$
$$P(|Z| > 3) \leq 1/9$$

证明 这里仅对X是连续型随机变量证明。

设X的概率密度函数为$f(x)$，对于任意$\varepsilon>0$，有

$$P\{|X-E(X)|\geq\varepsilon\}=\int_{|X-E(X)|\geq\varepsilon} f(x)dx$$

$$x-E(x)\geq\varepsilon \Rightarrow \leq \int_{|X-E(X)|\geq\varepsilon} \frac{[x-E(X)]^2}{\varepsilon^2} f(x)dx$$

$$\leq \frac{1}{\varepsilon^2}\int_{-\infty}^{+\infty}[x-E(X)]^2 f(x)dx$$

$$=\frac{D(X)}{\varepsilon^2}$$

$$P(|X-E(X)|\geq\varepsilon)\leq\frac{D(X)}{\varepsilon^2}$$

或 $$P(|X-E(X)|<\varepsilon)\geq 1-\frac{D(X)}{\varepsilon^2}$$

由切比雪夫不等式可看出：

- 当误差ε取定时，随着方差$D(X)$减小，X围绕$E(X)$取值的概率增大。
- 反之，随着方差$D(X)$增大，X围绕$E(X)$取值的概率减少。

进一步说明方差$D(X)$能描述X对其均值$E(X)$的偏离程度。

已知正常男性成人血液中，每一毫升白细胞数平均是7300，均方差是700. 利用切比雪夫不等式估计每毫升白细胞数在5200~9400之间的概率 .

标准差(Standard Deviation)，也称均方差(mean square error)

解：设每毫升白细胞数为X

依题意，$E(X)=7300, D(X)=700^2$

所求为 $P(5200 \leq X \leq 9400)$

$$P(5200 \le X \le 9400)$$

$$= P(5200-7300 \le X-7300 \le 9400-7300)$$

$$= P(-2100 \le X-E(X) \le 2100)$$

$$= P(|X-E(X)| \le 2100)$$

由切比雪夫不等式

$$P(|X-E(X)| \le 2100) \ge 1-\frac{D(X)}{(2100)^2}$$

$$=1-(\frac{700}{2100})^2=1-\frac{1}{9}=\frac{8}{9}$$

即估计每毫升白细胞数在5200~9400之间的概率不小于8/9 .

2. 大数定律law of large numbers

大量抛掷硬币正面出现频率

大数定律的客观背景：大量的随机现象中平均结果的稳定性

在一定条件下，多次重复进行某一试验，随机事件发生的频率随着次数的增多逐渐稳定在某一个常数附近，这一数值也就是随机事件的**概率**。

直观的经验表明，大量观测值的算术平均值the sample average也具有**稳定性**，即在相同条件下随着观测次数的增多，观测值的算术平均值逐渐稳定于某一常数附近，这一数值就是观测值(看作随机变量)的**数学期望**。

概率论中用来阐明大量随机现象平均结果的**稳定性**的定理统称为**大数定律**。

定理4.4.2 贝努里大数定律

重要

设 n_A 是 n 次独立重复试验中事件 A 发生的次数, p 是每次试验中 A 发生的概率, 则

$\forall \varepsilon > 0$ 有

$$\lim_{n\to\infty} P\left(\left|\frac{n_A}{n} - p\right| \geq \varepsilon\right) = 0$$

或

$$\lim_{n\to\infty} P\left(\left|\frac{n_A}{n} - p\right| < \varepsilon\right) = 1$$

贝努里大数定律的意义

在概率的统计定义中，事件 A 发生的频率“稳定于”事件A在一次试验中发生的概率，既任一随机事件的频率具有稳定性

在 n 足够大时，可以用频率近似代替 p。这种稳定称为依概率稳定.

n足够大时，频率收敛于概率

重要

定理4.4.3 切比雪夫大数定律(平均数法则)

设 r.v. 序列 $X_1, X_2, \cdots, X_n$ 相互独立，且具有相同的数学期望和方差

$$E(X_k)=\mu,\ D(X_k)=\sigma^2,\quad k=1,2,\cdots$$

则 $\forall \varepsilon>0$ 有

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k-\mu\right|\geq\varepsilon\right)=0$$

或

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k-\mu\right|<\varepsilon\right)=1$$

平均数法则的意义

具有相同数学期望和方差的独立 r.v.序列的算术平均值依概率收敛于数学期望，既算术平均值与数学期望有较大偏差的可能性很小。

当 n 足够大时, 算术平均值几乎是一常数.

数学期望 可被 算术均值 近似代替

n足够大时，同期望方差且独立，均值收敛于期望

辛钦大数定理

若$X_1, X_2,..,X_n$相互独立,服从同一分布,且具有相同的数学期望 $EX_k = \mu(k = 1,2,\cdots)$. 对任意正数$\varepsilon$, 有

$$\lim_{n\to+\infty} P\{\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu\right| < \varepsilon\} = 1.$$

$$\lim_{n\to+\infty} P\{\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu\right| \geq \varepsilon\} = 0.$$

辛钦大数定理的意义

- 当n很大时,独立同分布的随机变量的平均值$(\frac{1}{n}\sum_{k=1}^{n}X_k)$依概率收敛于它的数学期望$\mu$.

n足够大时，独立同分布，均值收敛于期望

3. 中心极限定理central limit theorem

中心极限定理的客观背景:

在实际问题中，常常需要考虑许多随机因素所产生总影响.

观察表明，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. 则这种量一般都服从或近似服从正态分布.

定理一 林德伯格-列维中心极限定理

[独立同分布的中心极限定理]

重要

设随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$

独立同分布, 且有相同期望和方差:

$$E(X_k) = \mu, \quad D(X_k) = \sigma^2 > 0, \quad k = 1, 2, \cdots$$

则对于任意实数 x,

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

注

重要

记 $Y_n = \dfrac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma}$

$$\lim_{n\to\infty} P(Y_n \le x) = \Phi(x)$$

即 n 足够大时，Y_n 的分布函数近似于标准正态随机变量的分布函数

$$Y_n \overset{\text{近似}}{\sim} N(0,1)$$

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \underset{\text{近似}}{\sim} N(\mu, \frac{\sigma^2}{n})$$

$$\sum_{k=1}^{n} X_k \underset{\text{近似}}{\sim} N(n\mu, n\sigma^2)$$

它表明:当n充分大时，n个具有期望和方差的独立同分布的$r.v$之和近似服从正态分布.

定理二

棣莫弗-拉普拉斯中心极限定理

[二项分布以正态分布为极限分布]

重要

设 $Y_n \sim B(n, p)$，$0 < p < 1$，$n = 1,2,\ldots$

则对任一实数 x，有

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

即 n 足够大时，

$Y_n \sim N(np, np(1-p))$ (近似)

独立同分布中心极限定理的特例，Yn=多个0-1分布随机变量和

例 某单位有200台电话分机，每台分机使用外线的概率为0.2, 假定每台分机是相互独立的，问要安装多少条外线，才能以95%以上的概率保证分机用外线时不等待？

解：设有X部分机同时使用外线，则有 $X \sim B(n, p)$,

$\mathrm{E}(X_n)=p$, $\mathrm{D}(X_n)=p(1-p)$

其中 $n = 200, p = 0.2, np = 40, np(1-p) = 32.$

设有N条外线。由题意有 $P(X \le N) \ge 0.95$

由棣莫佛-拉普拉斯定理有

$$P(X \le N) \approx \Phi\left(\frac{N-np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{N-40}{\sqrt{32}}\right).$$

查表得$\Phi(1.65) = 0.95.$

即 $N \ge 50$, 即至少要安装 50 条外线。

例 设有一批种子，其中良种占1/6. 试估计在任选的6000粒种子中，良种比例与 1/6 比较上下不超过1%的概率.

解 设 X 表示6000粒种子中的良种数，

则 $X \sim B(6000, 1/6)$

由德莫佛—拉普拉斯中心极限定理，

有 $X \overset{\text{近似}}{\sim} N\left(1000, \frac{5000}{6}\right)$

$$P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)=P\left(|X-1000|<60\right)$$

$$\approx \Phi\left(\frac{1060-1000}{\sqrt{5000/6}}\right)-\Phi\left(\frac{940-1000}{\sqrt{5000/6}}\right)$$

$$=\Phi\left(\frac{60}{\sqrt{5000/6}}\right)-\Phi\left(\frac{-60}{\sqrt{5000/6}}\right)$$

$$=2\Phi\left(\frac{60}{\sqrt{5000/6}}\right)-1\approx 0.9624$$

比较几个近似计算的结果

二项分布(精确结果) $P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)\approx 0.9590$

中心极限定理 $P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)\approx 0.9624$

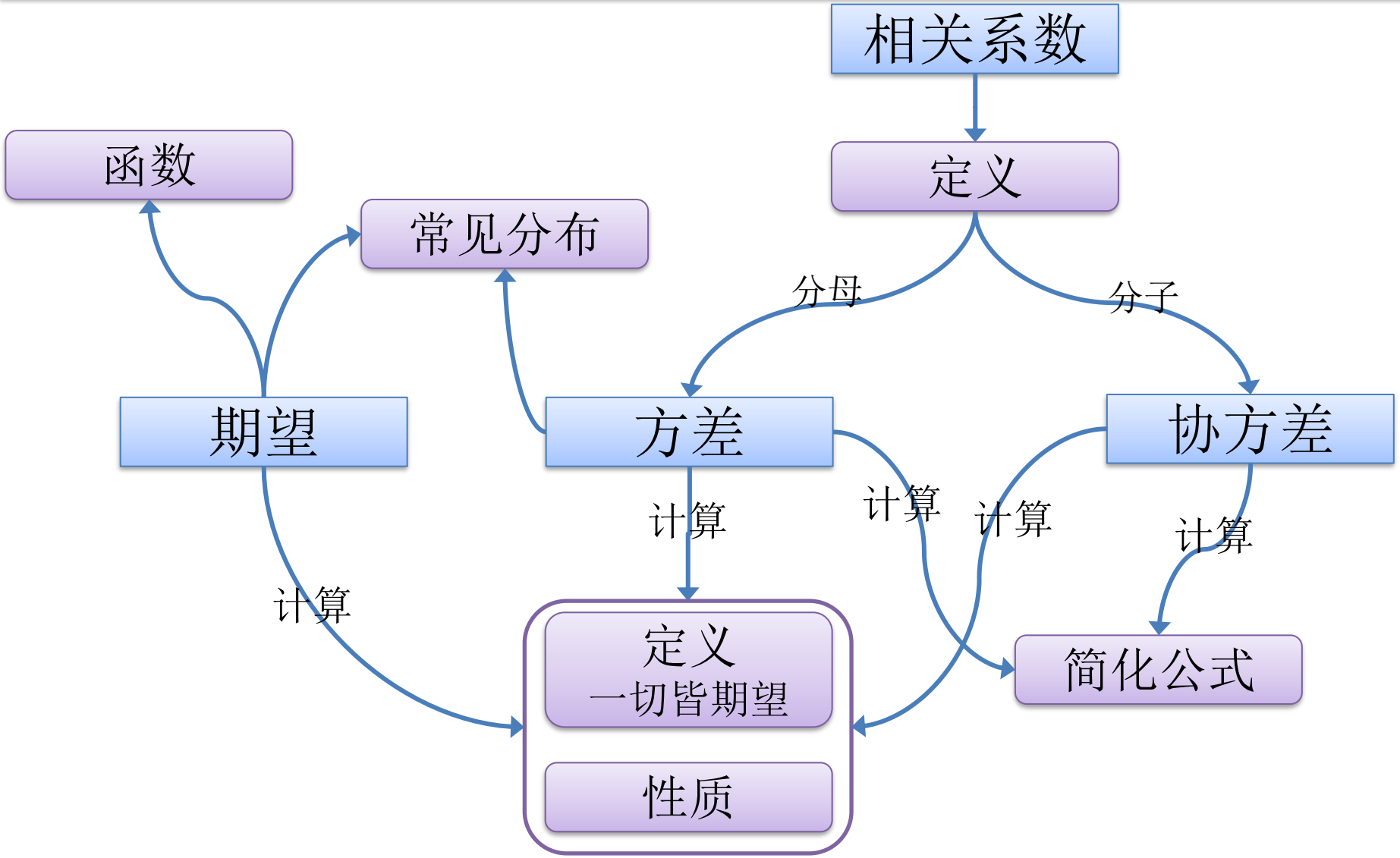
Poisson 分布 $P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)\approx 0.9379$

假设服从泊松分布，然后查表

Chebyshev 不等式 $P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)\geq 0.7685$

$$=P\left(\left|X-1000\right|<60\right)$$

Summary



第一章 随机事件及其概率

- 基本概念：是以后各章的基础知识
  - 事件之间的关系及公式表示
    - 给定一个事件与周围事件的关系，能用公式进行表示，并求出概率
  - 古典概率、几何概率、概率的公理化定义
  - 概率的性质及基本运算法则，如加法公式、减法公式等
  - 条件概率与乘法公式，全概率公式和贝叶斯公式
  - 互斥事件、对立事件、独立事件

第二章 Summary

2.1 r.v. X, $F(x)$, $f(x)$定义,关系,性质

2.1 离散型随机变量及其分布律

超几何分布
几何分布
两点分布
二项分布
泊松分布

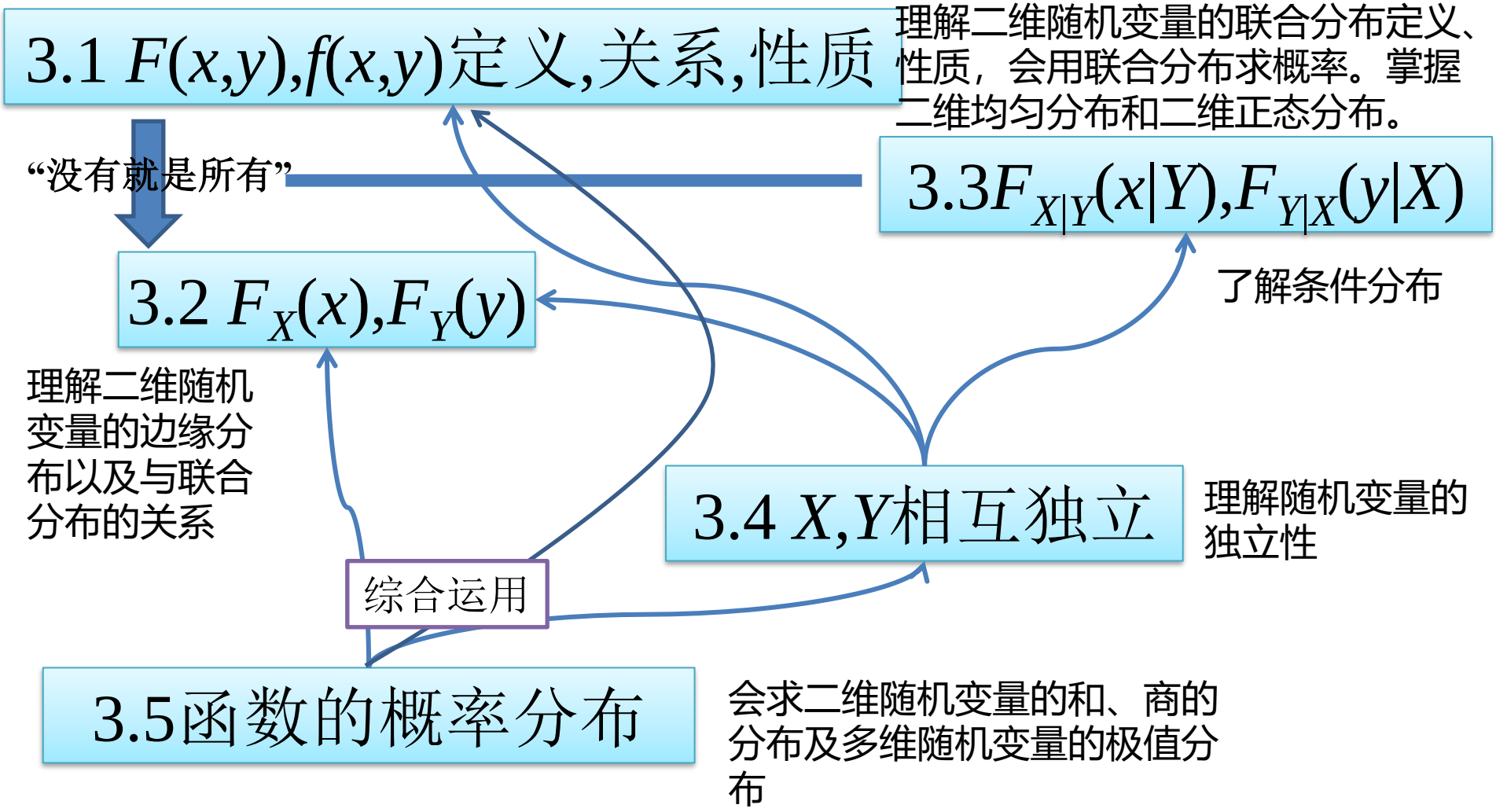
2.3连续型随机变量及其分布

均匀分布
指数分布
正态分布、标准正态分布

2.4 随机变量函数的分布

(1) 从分布函数出发
(2) 用公式直接求d.f.求反函数，代公式

第三章 Summary


3.1 $F(x,y),f(x,y)$定义,关系,性质
理解二维随机变量的联合分布定义、性质，会用联合分布求概率。掌握二维均匀分布和二维正态分布。
“没有就是所有”
3.3$F_{X|Y}(x|Y),F_{Y|X}(y|X)$
了解条件分布
3.2 $F_X(x),F_Y(y)$
理解二维随机变量的边缘分布以及与联合分布的关系
3.4 X,Y相互独立
理解随机变量的独立性
综合运用
3.5函数的概率分布
会求二维随机变量的和、商的分布及多维随机变量的极值分布