# 走进概率的世界

## —— 信息学竞赛中概率问题求解初探

安徽省合肥一中 梅诗珂

# 引言

- 算法设计中很多问题的解决都用到了概率分析
  - 一个大家熟知的例子是，快速排序中通过随机选择划分点而使极端情况出现的概率大大减小
- 在信息学竞赛中，与概率有关的问题占据着相当的分量
  - 在 **05**， **06**， **08** 年的 **NOI** 中都出现了与概率有关的试题

# 全文总览

样本空间

↓

随机变量

- 连续型随机变量
  - SPOJ RNG
  - SGU Random Shooting
- 离散型随机变量
  - SRM 349 LastMarble
  - UVA Randomness

# 要用到的定义

- 连续型随机变量的概率分布
  - 设有随机变量 **X**，称 ***F(x) =P(X≤x)*** 为 ***X*** 的概率分布函数，如果有非负可积函数 ***f(x)*** 使

$$F(x)=\int_{-\infty}^{x} f(t)dt$$

  成立，则称 ***f(x)*** 是 ***X*** 的概率密度函数
- 均匀分布
  - 若随机变量 ***X*** 在 **[*a*,*b*]** 上等概率地取每个值，称 ***X*** 在 **[*a*,*b*]** 上均匀分布，由概率密度的定义知

$$f(x)=\frac{1}{b-a} \quad (a\le x\le b)$$

# SPOJ RNG

- 题目大意
  - 有 $N$ 个随机数生成器，第 $i$ 个等概率地返回 $[0,R_i]$ 中的一个实数 $(1 \le i \le N)$
  - 问所有随机生成数的和小于等于 $b$ 的概率是多少
- 约束条件
  - $N$ 、$R_i$ 都是范围在 1 到 10 内的正整数 $(1 \le i \le N)$

# 题意分析

- 第 $i$ 个随机数生成器返回的值是一个在 $[0,R_i]$ 中均匀分布的连续型随机变量
  - 不妨设为 $X_i$，显然这 $N$ 个随机变量是互相独立的 $(1 \le i \le N)$
  - 它们的和，即 $X_1+X_2+\cdots+X_N$，也是一个随机变量，不妨设为 $S$
  - 那么 $S \le b$ 的概率就是我们要求的，记为 $P(S \le b)$

# 方法一

当 $N=2$ 时
（$N=1$ 略）


- $(X_1,X_2)$ 的取值范围可看成平面直角坐标系的一个矩形
- $S=X_1+X_2 \le b$ 可以看成半平面
- $P(S \le b)$ 就是它们的公共部分的面积与矩形 $(R_1 \times R_2)$ 的比值

# 方法一

当 $N=3$ 时

- 与 $N=2$ 的情况类似
  - $(X_1,X_2,X_3)$ 的取值范围可看成空间直角坐标系的长方体
  - $S=X_1+X_2+X_3 \le b$ 可以看成半空间
  - $P(S \le b)$ 就是它们的公共部分的体积，与长方体的体积 $(R_1 \times R_2 \times R_3)$ 的比值

# 方法一

## 当 $N$ 为任意值时

- $(X_1,X_2,\cdots,X_N)$ 的取值范围就是 $N$ 维空间中的一个区域，$S \le b$ 是一个半 $N$ 维空间，它们公共部分的体积与 $(R_1 \times R_2 \times \cdots \times R_N)$ 的比值就是 $P(S \le b)$

  - 不妨记为 $V([0, R_1],[0, R_2],\ldots,[0, R_N],b)$

  - 补集转化 $[0,R_i] = [0,+\infty) - (R_i ,+\infty)$

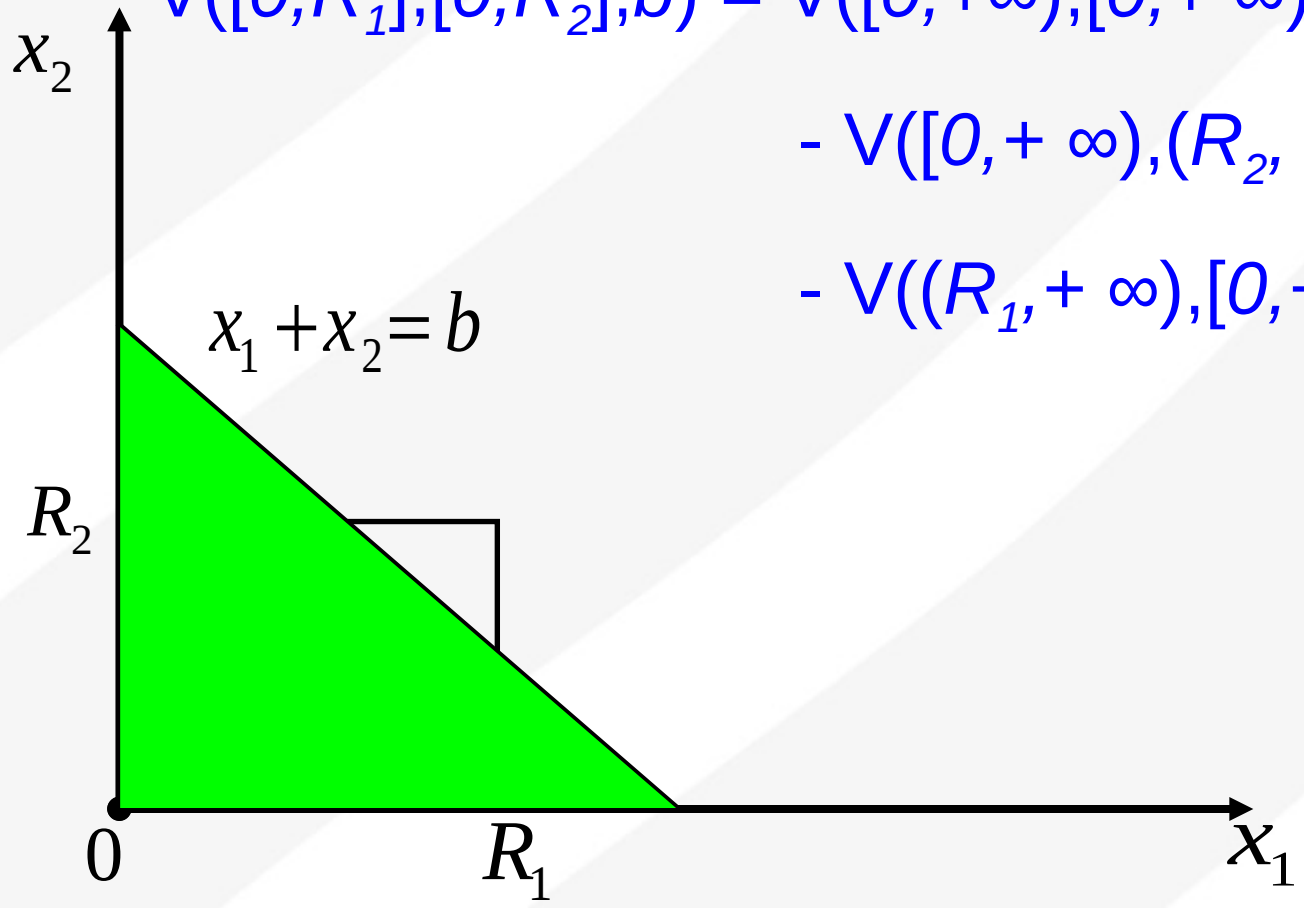
# 方法一

- 以 ***N=2*** 为例

$V([0,R_1],[0,R_2],b) = V([0,+\infty),[0,+\infty),b)$

$- V([0,+\infty),(R_2,+\infty),b)$

$- V((R_1,+\infty),[0,+\infty),b)$



# 方法一

- 当 $X_i$的取值范围为 $[0,+\infty)$ 或 $(R_i,+\infty)$ 时，怎样求 $V$
- 当 $X_i$的取值范围为 $(R_i,+\infty)$ 时，定义 $X'_i=X_i-R_i$
- 用 $X'_i$替换 $X_i$，同时把 $b$ 减去 $R_i$，问题等价。
- 求 $V([0,+\infty),\ldots,[0,+\infty),b)$

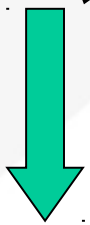
N=1 时 $V((0,+\infty),b) = b$

N=2 时 $V((0,+\infty),(0,+\infty),b)=b^2/2!$

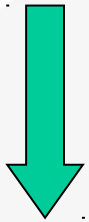
N=3 时 $V((0,+\infty),(0,+\infty),(0,+\infty),b)= b^3/6$

# 方法一总结

从 *N*=2、 3 开始分析



求区域体积



问题解决

有限区间

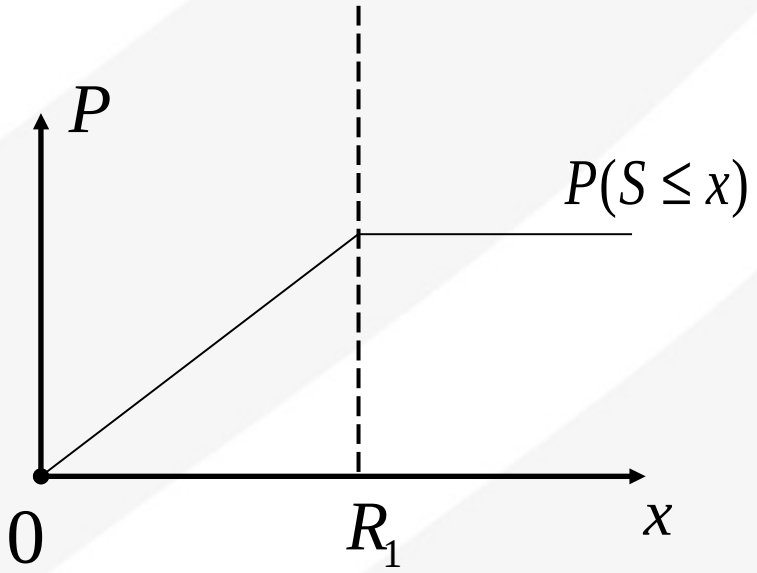
补集转化

无限区间

# 方法二

- 当 *N=1* 时（先说明 *X*）

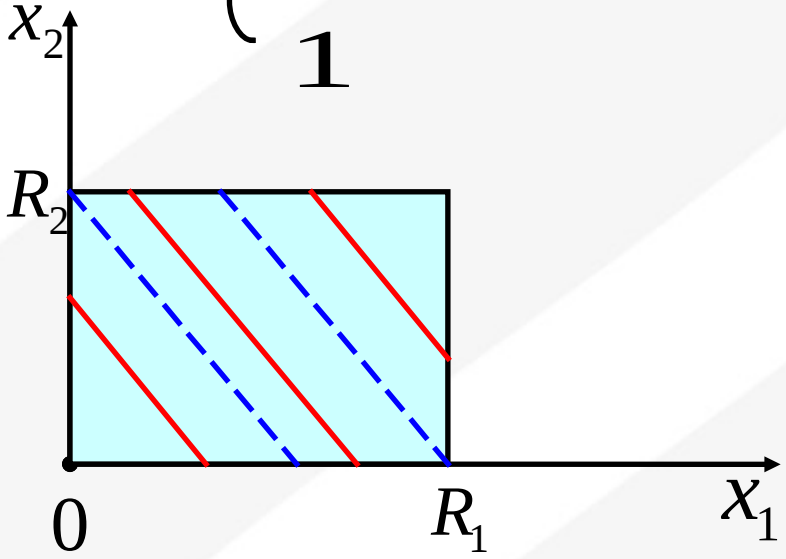
$$P(S \le x) = \begin{cases} x / R_1 & (0 \le x \le R_1) \\ 1 & x > R_1 \end{cases}$$



# 方法二

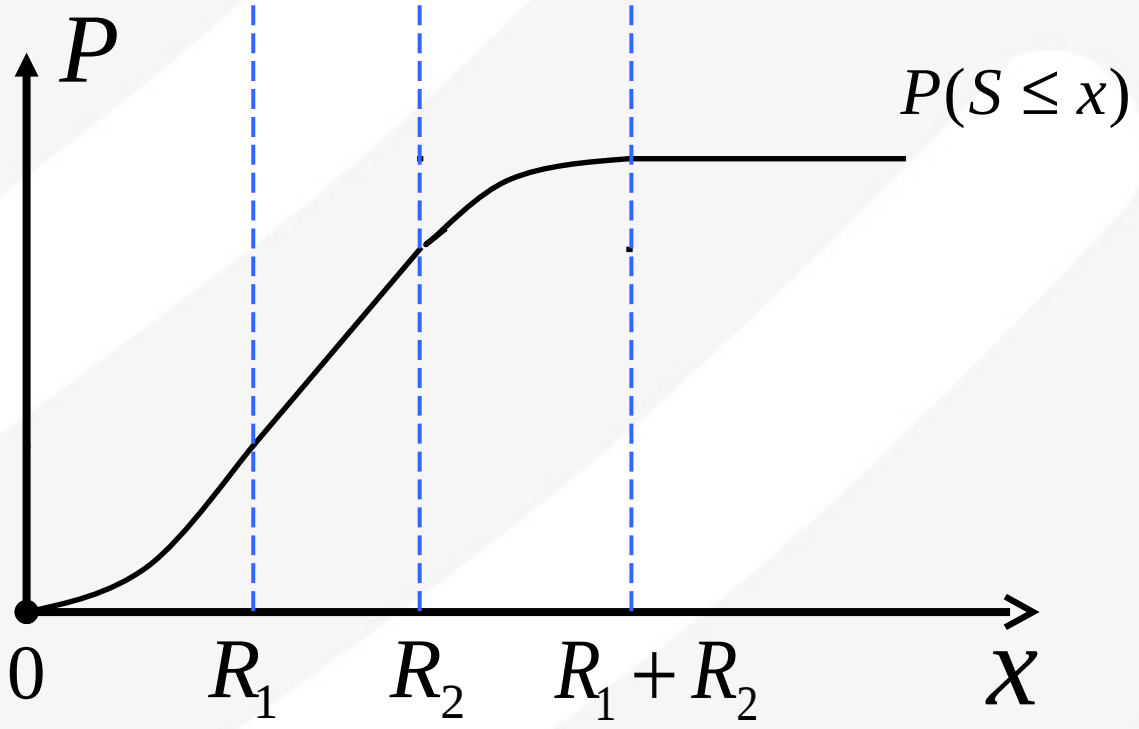
- 当 $N=2$ 时

$$P(S \le x) = \begin{cases} \dfrac{x^2}{2R_1R_2} & (0 \le x \le R2) \quad (1) \\ \dfrac{2x - R_2}{2R_1} & (R2 < x \le R1) \quad (2) \\ \dfrac{-x^2 + 2(R_1 + R_2)x - (R_1^2 + R_2^2)}{2R_1R_2} & (R1 < x \le R1 + R2) \quad (3) \\ 1 & (R1 + R2 < x) \quad (4) \end{cases}$$



# 方法二

- 画出函数图像



# 方法二

- $N$ 更大时
  - 回顾 $N=1,N=2$ 时 $P(S \le x)$ 的求解过程，我们发现可以划分若干区间，使每个区间的 $P(S \le x)$ 都可以表示成多项式
  - 如何划分区间
  - 以全体整数为划分点

# 推想

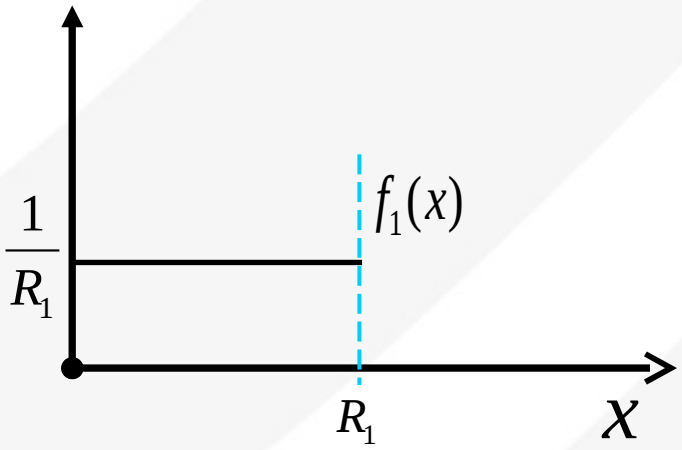
- 对任意的 $N$, 函数 $P(S \leq x)$ 在任意相邻整数区间内都可表示成多项式
- 推想正确吗?

# 证明思路

- 要证明 $P(S \le x)$ 在相邻整数区间能用多项式表示
  - 只需证明 $S$ 的概率密度函数在相邻整数区间能用多项式表示。
  - 用归纳法，设 $f_i(x)$ 为随机变量 $X_1, X_2, \cdots, X_i$ 的和 ( 一个随机变量 ) 的概率密度函数

**当 $i=1$ 时**

$$f_1(x) = \begin{cases} 1/R_1 & (0 \le x \le R_1) \\ 0 & (x > R_1) \end{cases}$$



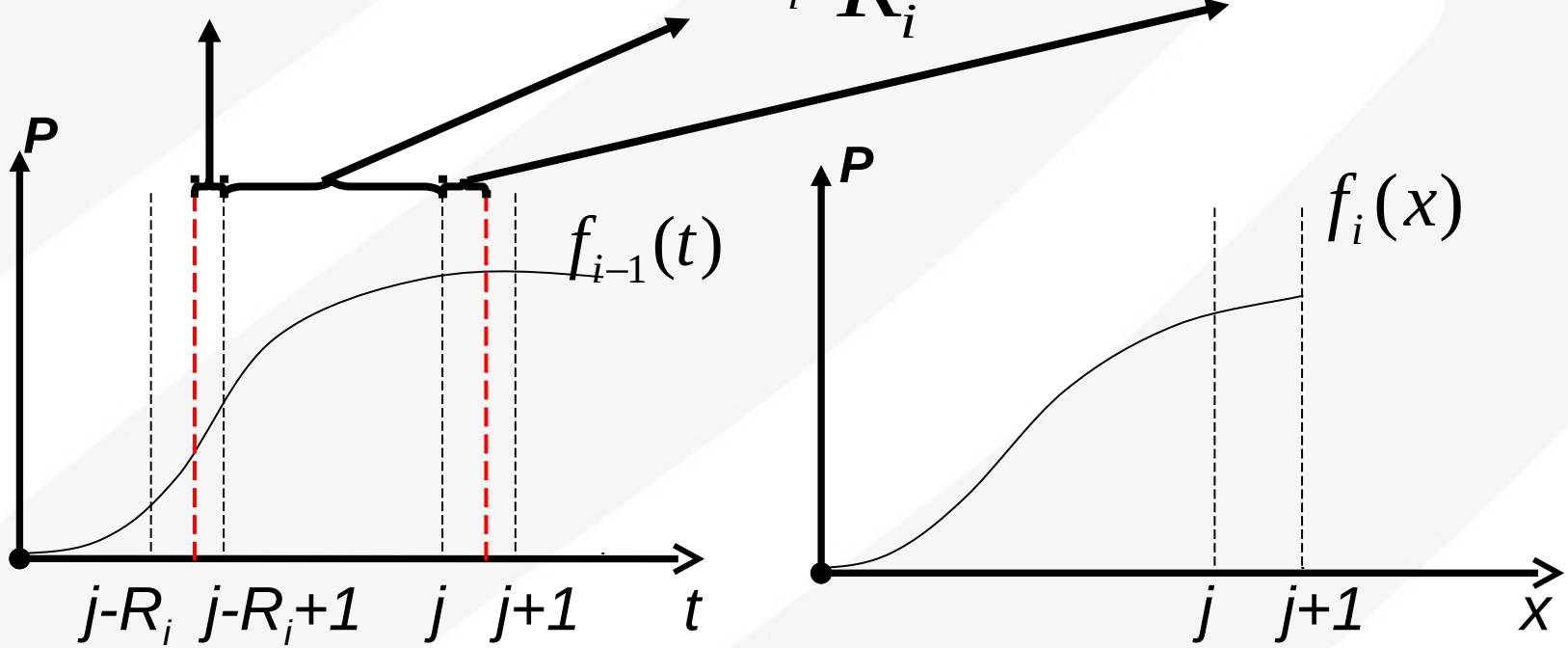
# 证明思路

- 设 ***i=N-1*** 时结论成立，证明 ***i=N*** 时成立。
- 由于前 ***(i-1)*** 个数的和与 $X_i$ 是互相独立的，有

$$f_i(x) = \frac{1}{R_i}\int_{x-R_i}^{x} f_{i-1}(t)\,dt$$

[x-Ri, j-R+1]

[j,x]

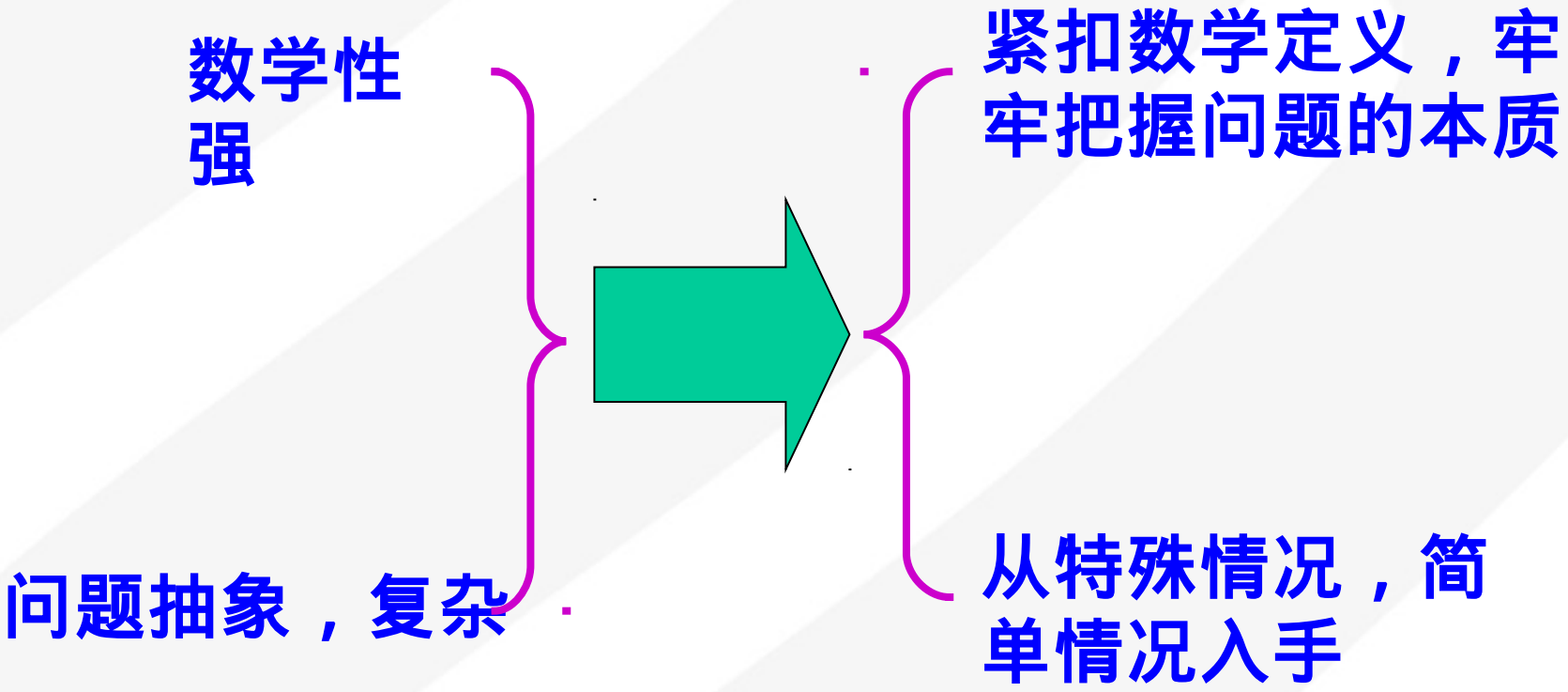


# 本题总结

- 在解决本题的过程中，我们遇到了这样的困难：N 个随机变量代表着 N 维空间，较为抽象
- 两个解法都从 N 较小的情况开始分析，发现规律
  - 比较两种解法，第二种比第一种思考与编程复杂度都大一些
  - 但是第二种解法可推广性强，在 N 个随机变量不为均匀分布时依然适用
  - 随机变量均匀分布是能用第一种方法解题的根本原因

# 总结

- 通过对上面例题的分析，我们发现概率问题有如下特点



# 谢谢大家

# 欢迎提问

# V([0,R1],[0,R2],…,[0,RN],x) 的表示。

- **V([0,R1],[0,R2],…[0,RN],x)**
  **=V([0,+∞)-(R1,+∞),[0,+∞)-(R2,+∞),…,**
  **[0,+∞)-(R2,+∞),x)**
- 设 **(1≤i≤N)**，设集合为所有满足下标集合，那么

$$\mathbf{V([0,R1],[0,R2],[0,RN],x)=}\sum_{i=0}^{N}(-1)^{i}\sum_{|Q|=i}V((D_1,+\infty),(D_2,+\infty),...(D_{N,}+\infty),x)$$

- 该式与容斥原理的公式相近。

- 用归纳法，当 **N=1** 时，**V((0,+∞),x) = x** 显然符合结论。
- 设当 **N=2,3,…,k-1** 时都有结论成立，那么 **N=k** 时，**V([0,+∞), [0,+∞),…, [0,+∞),x)** 就是一个 **k** 维锥体的 **k** 维体积，椎体的底面面积是 $\frac{x^{k-1}}{(k-1)!}$
- 同时我们知道体积就是截面面积的积分值，而对与锥体的顶点距离为 **h** 的截面而言，其截面面积为 $\frac{x^{k-1}}{(k-1)!}(\frac{h}{x})^{k-1}$，所以
- **V([0,+∞), [0,+∞),…, [0,+∞),x) =** $\int_0^x \frac{x^{k-1}}{(k-1)!}(\frac{h}{x})^{k-1}dh = \frac{1}{(k-1)!}(\frac{x^k}{k}-0) = \frac{x^k}{k!}$

■

- 方法二的可推广性强。这是因为每个随机变量都是均匀分布这个条件对方法二中的证明来说可以减弱：只要每个随机变量的概率密度函数是多项式即可。而方法一中之所以能把概率看成 **N** 维体积的比值，其根本原因就是每个随机变量都是均匀分布的。
- 举个简单的例子：如果要求的是 **N** 个随机数的平方和小于等于 **b** 的概率，那么方法一将无能为力，而方法二只要简单套用即可。
- 两种方法都从分析简单情况着手，但第二种方法对题目数学本质把握更透彻，这使方法二在一定程度上成为解决连续型随机变量问题的一般性方法。

# 对 $F(x)=\int_{-\infty}^{x} f(t)dt$ 的解释

一般的积分式的积分区间是一个有限闭区间，但是一些情况下它不能满足需要，于是有了无穷积分，也即积分区间为无穷区间的积分。

如果可积函数 **f(t)** 在 **(-∞,b)** 上都有定义，并且如果极限 $\lim_{a\to-\infty}\int_a^b f(t)dt$ 存在并有限，则把该极限记为 $\int_{-\infty}^{b} f(t)dt$