

# Probability and Statistics - P12

方差

(variance) 对于随机变量  $X$ , 有期望  $E(X)$

如果  $E\{(X-E(X))^2\}$  存在, 则称为随机变量  $X$  的方差

记为  $D(X)$  或  $\text{Var}(X)$ . 即  $D(X) = \text{Var}(X) = E\{(X-E(X))^2\}$

同时引入  $\sqrt{D(X)}$ , 称为标准差 (standard deviation,  $sd$ ), 记为  $\sigma(X)$ , 或称均方差

方差 (标准差) 描述了随机变量  $X$  与其数学期望  $E(X)$  的偏离程度

对于离散随机变量  $X$ , 有分布函数  $P_{f_x}\{X=x_k\} = P_k$

$$\text{于是有 } D(X) = \sum_{k=1}^{\infty} [x_k - E(X)]^2 P_k = \sum_{k=1}^{\infty} x_k^2 P_k - 2 \sum_{k=1}^{\infty} x_k P_k \cdot E(X) + \sum_{k=1}^{\infty} P_k [E(X)]^2$$

$$= E(X^2) - [E(X)]^2$$

对于连续随机变量  $X$ , 有分布函数  $f(x)$

$$\text{于是有 } D(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - 2 \int_{-\infty}^{\infty} x f(x) E(X) dx + \int_{-\infty}^{\infty} [E(X)]^2 f(x) dx$$

$$= E(X^2) - [E(X)]^2$$

$$\text{即有 } D(X) = E(X^2) - [E(X)]^2$$

对于常数  $C$ , 则  $D(C) = 0$ .

$$\text{证明过程有, } D(C) = E\{(C-E(C))^2\} = E\{[C-C]^2\} = E(0) = 0$$

对于常数  $C$ , 随机变量  $X$ , 有  $D(X+C) = D(X)$ ,  $D(CX) = C^2 D(X)$

$$\text{证明过程有, } D(X+C) = E\{(X+C - E(X+C))^2\} = E\{(X+C - E(X)-C)^2\} = E\{(X-E(X))^2\} = D(X)$$

$$D(CX) = E\{(CX - E(CX))^2\} = E\{[C(X - E(X))]^2\} = C^2 E\{(X-E(X))^2\} = C^2 D(X)$$

对于随机变量  $X, Y$ , 有  $D(X+Y) = D(X) + D(Y) + 2E\{(X-E(X))(Y-E(Y))\}$

$$\text{证明过程有, } D(X+Y) = E\{(X+Y - E(X+Y))^2\} = E\{(X - (E(X)+E(Y)))^2\}$$

$$= E\{(X-E(X)) + (Y-E(Y))\}^2$$

$$= E\{(X-E(X))^2\} + 2E\{(X-E(X))(Y-E(Y))\} + E\{(Y-E(Y))^2\}$$

$$= D(X) + D(Y) + 2E\{(X-E(X))(Y-E(Y))\}$$

对于随机变量  $X, Y$ , 如果  $X, Y$  相互独立, 则有  $D(X+Y) = D(X) + D(Y)$

$$\text{证明过程有, } 2E\{(X-E(X))(Y-E(Y))\} = 2E\{XY - XE(Y) - YE(X) + E(X)E(Y)\}$$

$$= 2[E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y)]$$

$$= 2[E(XY) - E(X)E(Y)] = 0$$

$$\text{即有 } D(X+Y) = D(X) + D(Y)$$

# Probability and Statistics - P13

889 - sample variance

样本方差

(sample variance), 记为  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ , 即上面讲过的样本的方差  
但是特别注意, 样本方差公式中是  $\frac{1}{n-1}$  而方差中是  $\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$

这是由于一般希望样本方差是实际方差的无偏估计, 即  $E(S^2) = \sigma^2$

证明没有, 首先计算  $D(\bar{X}) = D\left(\frac{1}{n} \sum_{i=1}^n X_i\right)$

由于  $X_1, X_2, \dots, X_n$  是独立同分布的随机变量, 因有方差  $\sigma^2$ , 期望为  $\mu$

于是有  $D(\bar{X}) = D\left(\frac{1}{n} \sum_{i=1}^n X_i\right)$

$$= \frac{1}{n^2} D\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n D(X_i)$$

$$= \frac{1}{n} D(X_i) = \frac{1}{n} \sigma^2$$

假如直接取样本方差公式为  $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$

则有  $E(S^2) = E\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right]$

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

$$\approx \frac{1}{n} E\left[\sum_{i=1}^n (X_i - \mu)(\bar{X} - \mu)\right]$$

$$= (\bar{X} - \mu) E\left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)\right]$$

$$= E[(\bar{X} - \mu)^2]$$

$$\frac{1}{n} E\left[\sum_{i=1}^n (\bar{X} - \mu)^2\right] = E[(\bar{X} - \mu)^2]$$

$$\text{于是有 } E(S^2) = E\left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2\right] - 2(\bar{X} - \mu)^2 + (\bar{X} - \mu)^2$$

$$= E[\sigma^2 - (\bar{X} - \mu)^2], \text{ 且 } E(\bar{X}) = \mu$$

$$= \sigma^2 - D(\bar{X}) = \frac{n-1}{n} \sigma^2 \neq \sigma^2$$

于是考虑对  $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$  的公式进行调整

从系数  $\frac{n-1}{n}$  得出  $S^2 = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$

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

$$\text{则有 } E(S^2) = E\left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right]$$

$$= E\left[\frac{n}{n-1} \left[ \frac{1}{n} \sum_{i=1}^n [(X_i - \mu) - (\bar{X} - \mu)]^2 \right]\right]$$

$$= E\left[\frac{n}{n-1} \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^n (X_i - \mu)(\bar{X} - \mu) + \frac{1}{n} \sum_{i=1}^n (\bar{X} - \mu)^2 \right]\right]$$

$$= E\left[\frac{n}{n-1} [G^2 - (\bar{X} - \mu)^2]\right]$$

$$= \frac{n}{n-1} G^2 - \frac{n}{n-1} E[(\bar{X} - \mu)^2]$$

$$= \frac{n}{n-1} G^2 - \frac{n}{n-1} \cdot D(\bar{X})$$

$$= \frac{n}{n-1} G^2 - \frac{n}{n-1} \cdot \frac{1}{n} \sigma^2$$

$$= \sigma^2$$

特别地有, 当样本容量足够大时, 可以直接将  $S^2$  视为原分布的方差  $\sigma^2$

即当  $n \rightarrow +\infty$  时  $S^2 \sim \sigma^2$

所以  $S^2$  为样本方差, 且有  $E(S^2) = \sigma^2$

# Probability and

## Statistics - P14

多项式系数 (multinomial coefficient), 对于将  $n$  个可区分的物体放入  $k$  个可区分的盒子使得每个盒子分别有  $n_1, n_2, \dots, n_k$  个物体, 对于  $i=1, \dots, k$ ,  $0 \leq n_i \leq n$ ,  $\sum_{i=1}^k n_i = n$

有  $\frac{n!}{n_1! n_2! \dots n_k!}$  种不同的方式, 记为  $(n_1, n_2, \dots, n_k)$

$$\binom{n+1}{r+1} = \sum_{j=r}^n \binom{j}{r}$$

多项式系数也有类似于  $\binom{n}{r}$  的关系

$$\text{即 } (n_1, n_2, \dots, n_k) = (n_1-1, n_2, \dots, n_k) + (n_1, n_2-1, \dots, n_k) + \dots + (n_1, n_2, \dots, n_k-1)$$

组合证明过程有,  $(n_1, n_2, \dots, n_k)$  可以描述为将  $n$  个可区分的物体放入  $k$  个可区分的盒子且每个盒子分别有  $n_1, n_2, \dots, n_k$  个物体

则考虑编号为  $n$  的物体放入盒子  $i$  的情形, 其中  $1 \leq i \leq k$

则问题转换为, 将剩余  $n-1$  个物体放入  $k$  个盒子

使得每个盒子分别有  $n_1, n_2, \dots, n_{i-1}, \dots, n_k$  个物体

于是共有  $(n_1, n_2, \dots, n_{i-1}, \dots, n_k)$  种不同的方式

考虑  $i=1, 2, \dots, k$  的情形, 则合计有  $\sum_{i=1}^k (n_1, n_2, \dots, n_{i-1}, \dots, n_k)$  种不同方式

于是有  $(n_1, n_2, \dots, n_k) = \sum_{i=1}^k (n_1, n_2, \dots, n_{i-1}, \dots, n_k)$

$$= (n_1-1, n_2, \dots, n_k) + \dots + (n_1, n_2, \dots, n_k-1)$$

特别注意, 这里实际上不要求  $n_1, n_2, \dots, n_k$  为正整数

由于当  $n_i = 0$  时,  $(n_1, n_2, \dots, n_k) = \frac{n!}{n_1! n_2! \dots n_{i-1}! 0! n_{i+1}! \dots n_k!} = (n_1, n_2, \dots, n_{i-1}, n_{i+1}, \dots, n_k)$

而定义  $(n_1, n_2, \dots, n_{i-1}, \dots, n_k) = 0$ , 其余项可去除  $n_i$

于是  $(n_1, n_2, \dots, n_k) = \sum_{i=1}^k (n_1, n_2, \dots, n_{i-1}, \dots, n_k)$  依旧成立

对于方程  $x_1 + x_2 + \dots + x_r = n$ , 恰有  $\binom{r}{k} \binom{n-1}{r-k}$  个解, 使其中恰有  $k$  个  $x_i$  为 0

证明过程有, 对于  $0 \leq k < r$ , 如果恰有  $k$  个  $x_i$  为 0

则可以先从  $r$  个  $x_i$  中选择  $k$  个指定为 0, 则有  $\binom{r}{k}$  种不同选择方式

对于剩下的  $r-k$  个  $x_i$ , 可以改写方程为  $x_1 + x_2 + \dots + x_{r-k} = n$ , 并计算正整数向量解的个数

即有  $\binom{n-1}{r-k-1}$  种不同的解

于是合计有  $\binom{r}{k} \binom{n-1}{r-k-1} = \binom{r}{k} \binom{n-1}{n-r+k}$  个解, 使得其中恰有  $k$  个  $x_i$  为 0

考虑将  $n$  个不可区分的物体放入  $r$  个可区分的盒子里, 且每个盒子  $m_1, \dots, m_r$  个物体的方式数. 其中  $n \geq \sum_{i=1}^r m_i$

可分为两步, 首先从  $n$  个物体中取出  $m_1, \dots, m_r$  个分别放入每个盒子

由于物体不可区分, 所以只有 1 种选择方式

然后问题转换为将  $n - \sum_{i=1}^r m_i$  个物体放入  $r$  个盒子的形式

则共有  $\binom{n - \sum_{i=1}^r m_i + r - 1}{r - 1}$  种方式

于是合计有  $\binom{n - \sum_{i=1}^r m_i + r - 1}{r - 1}$  种不同方式

# Probability and

## Statistics - P15

对于正整数  $1 \leq i \leq j \leq n$ , 有  $\sum_{j=i}^n \binom{n}{j} \binom{j}{i} = \binom{n}{i} 2^{n-i}$

组合证明过程有, 对于  $\sum_{j=i}^n \binom{n}{j} \binom{j}{i}$  可以描述为这样一个过程

首先从包含  $n$  个元素的集合中选择  $i$  个元素。[ ] 的子集

再 [ ] 从  $i$  个元素的子集中选择  $j$  个元素的子集。

对于给定的正整数  $n$  和  $i$ , 有不同的选择方式数

注意对于给定的  $n, i \in \mathbb{Z}^+$ ,  $j$  的取值范围是  $i \leq j \leq n$

对于某个  $i \leq j \leq n$  的  $j$  值

第一步从  $n$  个元素中选择  $j$  个元素, 有  $\binom{n}{j}$  种不同的选择方式

第二步再从  $j$  个元素中选择  $i$  个元素, 有  $\binom{j}{i}$  种不同的选择方式

则对于某个确定的  $j$  值, 有  $\binom{n}{j} \binom{j}{i}$  种选择方式

于是合计有  $\sum_{j=i}^n \binom{n}{j} \binom{j}{i}$  种不同的选择方式

注意对于同样的过程, 也可以描述为另一种形式

首先从  $n$  个元素中选择  $i$  个元素, 则有  $\binom{n}{i}$  种不同选择方式

再对剩余  $n-i$  个元素分别决定是否放入 [ ] 有  $j$  个元素的集合

则有  $2^{n-i}$  种不同的选择方式

于是合计有  $\binom{n}{i} 2^{n-i}$  种不同的选择方式

注意到这两个过程的结果是一一对应的,

即有  $\sum_{j=i}^n \binom{n}{j} \binom{j}{i} = \binom{n}{i} 2^{n-i}$ , 其中  $0 \leq i \leq j \leq n$

特别地有, 当  $i=0$  时, 等式退化为  $\sum_{j=0}^n \binom{n}{j} = 2^n$

另外有  $\sum_{j=i}^n \binom{n}{j} \binom{j}{i} (-1)^{n-j} = 0$ , 其中  $0 \leq i < n$

$$\begin{aligned} \text{代数证明过程有 } \sum_{j=i}^n \binom{n}{j} \binom{j}{i} (-1)^{n-j} &= \sum_{j=i}^n \frac{n!}{(n-j)! j!} \cdot \frac{j!}{(j-i)! i!} \cdot (-1)^{n-j} \\ &= \sum_{j=i}^n \frac{n!}{(n-i)! i!} \cdot \frac{(n-i)!}{(n-j)! (j-i)!} \cdot (-1)^{n-j} \\ &= \binom{n}{i} \sum_{j=i}^n \binom{n-i}{j-i} (-1)^{n-j} |^{j=i}, \text{ 令 } k=j-i \\ &= \binom{n}{i} \sum_{k=0}^{n-i} \binom{n-i}{k} (-1)^{n-i-k} |^k = 0 \end{aligned}$$

对于正整数  $1 < k \leq n$ , 有  $\binom{n}{2} = \binom{k}{2} + k(n-k) + \binom{n-k}{2}$

组合证明过程有,  $\binom{n}{2}$  可描述为从  $n$  个元素的集合中选择 2 个元素的 [ ] 不同方式数

也可以描述为, 在  $n$  个元素中标记  $k$  个特殊元素, 则考虑特殊元素的情况

2 个元素都是特殊的 | 1 个特殊的和 1 个非特殊的 | 2 个都是非特殊的

$\binom{k}{2}$

$\binom{k}{1} \binom{n-k}{1}$

$\binom{n-k}{2}$

于是有  $\binom{n}{2} = \binom{k}{2} + k(n-k) + \binom{n-k}{2}$ ,  $0 \leq k \leq n$

特别地有, 当  $k=1$  时,  $\binom{k}{2} + k(n-k) + \binom{n-k}{2} = 0 + (n-1) + \frac{(n-1)(n-2)}{2} = \binom{n}{2}$

当  $k=0$  时,  $\binom{0}{2} + 0(n-0) + \binom{n-0}{2} = \binom{n}{2}$  平凡地为真

# Probability and

## Statistics - P16

令  $H_k(n)$  为向量  $(x_1, x_2, \dots, x_k)$  的个数，其中  $1 \leq x_1 \leq x_2 \leq \dots \leq x_k \leq n$

则有递归定义  $H_1(n) = n$ ,  $H_k(n) = \sum_{j=1}^n H_{k-1}(j)$  ( $k > 1$ )

证明过程有：当  $k=1$  时， $H_1(n)$  即为一维向量  $(x_1)$  的个数

而  $1 \leq x_1 \leq n$  有  $n$  种不同的取值方式

于是  $H_1(n) = n$

当  $k > 1$  时， $H_k(n)$  为  $k$  维向量  $(x_1, x_2, \dots, x_k)$  的个数

考虑第  $k$  维  $x_k$  的取值，可知  $1 \leq x_k \leq n$

对于给定的  $x_k$  的值  $j$ ，其中  $1 \leq j \leq n$

则剩余  $k-1$  维的取值服从  $1 \leq x_1 \leq x_2 \leq \dots \leq x_{k-1} \leq x_k = j$

于是有  $H_{k-1}(j)$  种不同的取值方式

即合计有  $\sum_{j=1}^n H_{k-1}(j)$  种不同的  $k$  维向量

可以递归地定义  $H_k(n)$ ，其中  $k, n \in \mathbb{N}^+$

$$H_k(n) = \begin{cases} n, & k=1 \\ \sum_{j=1}^n H_{k-1}(j), & k>1 \end{cases}$$

[1..n]

## 成绩排名

对于  $n$  个选手的成绩进行排名，并且允许选手排名相同。排名相同的选手不分先后

于是可以按成绩将选手分组，第一好成绩在第一组，成绩其次在第二组，以此类推

记  $N(n)$  表示不同结果的可能数，并令  $N(0) = 1$

则有  $N(n) = \sum_{i=1}^n \binom{n}{i} N(n-i)$

证明过程有，对于  $n$  个选手参加的比赛，其中  $n > 0$

则考虑并列最后一名的选手个数

令有  $i$  个人并列最后一名，其中  $1 \leq i \leq n$

注意当  $i=n$  时，全部选手并列，可以视其为全部并列最后一名

则从  $n$  个选手中选择  $i$  个有  $\binom{n}{i}$  种不同的方式

对于剩余  $n-i$  人，其情形等同于对  $n-i$  个选手的排名的方式，即  $N(n-i)$

即共同  $\binom{n}{i} N(n-i)$  种不同方式

于是合计有  $\sum_{i=1}^n \binom{n}{i} N(n-i)$  种不同排名

如果取  $j=n-i$ ，则有  $N(n) = \sum_{j=0}^{n-1} \binom{n}{j} N(j)$

可以递归地定义函数  $N(n)$ ，其中  $n \in \mathbb{N}$

$$N(n) = \begin{cases} 1, & n=0 \\ \sum_{i=1}^n \binom{n}{i} N(n-i), & n>0 \end{cases}$$

$scoreRank :: Int \rightarrow Int$

$scoreRank 0 = 1$

$scoreRank n =$

$let func x = (myCombination n x) * (scoreRank )$

$in sum (map func [1..n])$

注意这个序列在形式上与贝努利相似以

但不同的在于这个序列不仅考虑划分还要考虑划分间的排序

# Probability and

## Statistics - P17

相对频率 (relative frequency), 指对于一个具有样本空间  $S$  的试验，并可在相同条件下重复进行

对于样本空间中的事件  $E$ , 记  $n(E)$  为  $n$  次重复试验中事件  $E$  的发生次数

则定义事件  $E$  发生的概率  $P(E)$  为  $E$  的发生次数与试验总次数的比例

即  $E$  发生频率的极限, 有  $P(E) = \lim_{n \rightarrow \infty} \frac{n(E)}{n}$

但是这个对于概率的直观定义存在缺陷

1. 如何知道  $n(E)/n$  会收敛到一个固定的常数

2. 即使  $n(E)/n$  收敛到一个常数, 如何知道下一次重复试验后, 依旧收敛到相同常数  
一种观点是将  $\lim_{n \rightarrow \infty} n(E)/n$  的极限值存在当作一个对整个系统的假设

但是这个假设异常复杂, 无法作为一个最基本, 最简单的假设

于是假定对于样本空间  $S$  中的任一事件  $E$ , 都存在一个值  $P(E)$  表示事件  $E$  的概率  
并假定如此假设的概概率值符合一系列公理

### 概率论公理

一. 对于任意事件  $E$ , 有  $0 \leq P(E) \leq 1$ , 即任何事件  $E$  的概率在 0 到 1 之间  
直觉上即为, 事件  $E$  不可能以负数概率发生, 也不可能以大于 1 的概率发生

二. 对于样本空间  $S$ , 有  $P(S) = 1$

直觉上即为, 样本空间  $S$  包含了所有可能结果, 则样本空间  $S$  作为必然发生的事件

三. 对一系列互不相容的事件  $E_1, E_2, \dots$ , 有  $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$

直觉上即为, 对任意一组互不相容事件, 至少有一件事发生的概率等于各个事件概率之和

而满足以上三条公理的  $P(E)$  称为事件  $E$  的概率

不可能事件 (impossible event), 记为  $\emptyset$ , 指对于某一事件  $E$ , 不包含任何样本空间内的结果

则直觉上这样的事件不可能发生, 即称为不可能事件

互不相容 (mutually exclusive), 指对于事件  $E$  和  $F$ , 有  $E \cap F = \emptyset$ , 则称  $E, F$  是互不相容事件

假设有系列特殊的事件  $E_1, E_2, \dots$ , 其中  $E_1 = S$ ,  $E_2, E_3, \dots = \emptyset$ .

则有  $E_1, E_2, \dots$  互不相容, 且  $\bigcup_{i=1}^{\infty} E_i = S$

于是有  $P(S) = P(\bigcup_{i=1}^{\infty} E_i) = P(E_1) + \sum_{i=2}^{\infty} P(\emptyset) = P(S) + \sum_{i=2}^{\infty} P(\emptyset)$ , 即有  $P(\emptyset) = 0$

直觉上  $P(\emptyset) = 0$  表示不可能事件的发生概率率为 0

# Probability and Statistic - P18

49 - Second half

补

(complement). 对于任意事件  $E$ , 定义  $E$  的补, 记为  $E^c$  或  $\bar{E}$ .

指包含在样本空间中但不包含在  $E$  中的所有结果构成的事件.

即  $E^c$  发生当且仅当  $E$  不发生, 有  $E^c = S \setminus E$ .

对于事件  $E$  和  $E$  的补  $E^c$ , 有  $P(E^c) = 1 - P(E)$ .

通过公理可知, 且事件  $E$  和  $E^c$  是互不相容事件.

有  $1 = P(S) = P(E \cup E^c) = P(E) + P(E^c)$

即有  $P(E^c) = 1 - P(E)$ .

对于事件  $E$  和  $F$ , 如果  $E$  包含于  $F$ , 即有  $E \subset F$ , 则有  $P(E) \leq P(F)$ .

由于  $E \subset F$ , 则  $F$  可表示为  $F = E \cup E^c F$ .

且  $E$  和  $E^c F$  是互不相容的, 所以有  $P(E \cup E^c F) = P(E) + P(E^c F)$ .

又通过公理可知,  $P(E^c F) \geq 0$ .

于是有  $P(F) = P(E) + P(E^c F) \geq P(E)$ .

容斥恒等式 (inclusion-exclusion identity), 指对于事件  $E_1, E_2, \dots, E_n$

$$P(E_1 \cup E_2 \cup \dots \cup E_n) = \sum_{i=1}^n P(E_i) - \sum_{i_1 < i_2} P(E_{i_1}, E_{i_2}) + \dots +$$

$$(-1)^{r+1} \sum_{i_1 < i_2 < \dots < i_r} P(E_{i_1}, E_{i_2}, \dots, E_{i_r}) + \dots + (-1)^{n+1} P(E_1, E_2, \dots, E_n)$$

其中  $\sum_{i_1 < i_2 < \dots < i_r} P(E_{i_1}, E_{i_2}, \dots, E_{i_r})$  表示为一切下标集合  $\{i_1, \dots, i_r\}$  求和, 共有  $\binom{n}{r}$  项.

所以容斥恒等式有更简明的写法:

$$P(\bigcup_{i=1}^n E_i) = \sum_{r=1}^n (-1)^{r+1} \sum_{i_1 < \dots < i_r} P(E_{i_1}, \dots, E_{i_r})$$

上界与下界 在容斥恒等式中, 取前奇数项则有逐渐收紧的上界, 取前偶数项则有逐渐收紧的下界.

即  $P(\bigcup_{i=1}^n E_i) \leq \sum_{i=1}^n P(E_i)$ ,  $P(\bigcup_{i=1}^n E_i) \geq \sum_{i=1}^n P(E_i) - \sum_{j < i} P(E_j, E_i)$ , ...

证明过程有, 注意有  $\bigcup_{i=1}^n E_i = E_1 \cup E_2 \cup E_3 \cup \dots \cup E_{n-1} \cup E_n$ .

又  $E_1, E_1^c E_2, E_1^c E_2^c E_3, \dots, E_1^c E_2^c \dots E_{n-1}^c E_n$  是互不相容的事件.

即  $P(\bigcup_{i=1}^n E_i) = P(E_1) + \sum_{i=2}^n P(E_1^c \dots E_{i-1}^c | E_i)$ .

又  $E_1^c \dots E_{i-1}^c = (\bigcup_{j < i} E_j)^c$ , 则有  $P(E_1^c \dots E_{i-1}^c | E_i) = P(E_i) - P(E_i \cap (\bigcup_{j < i} E_j))$ .

于是  $P(\bigcup_{i=1}^n E_i) = \sum_{i=1}^n P(E_i) - \sum_{i=2}^n P(\bigcup_{j < i} E_j)$ .

即有  $P(\bigcup_{i=1}^n E_i) \leq \sum_{i=1}^n P(E_i)$ .

另外对于固定的  $i$ , 则  $P(\bigcup_{j < i} E_j | E_i) \leq \sum_{j < i} P(E_j | E_i)$ .

于是有  $P(\bigcup_{i=1}^n E_i) \geq \sum_{i=1}^n P(E_i) - \sum_{j < i} P(E_j | E_i)$ .

于是可以类推地得到容斥恒等式的前奇数项, 则有逐渐收紧的上界.

取前偶数项, 则有逐渐收紧的下界.

# Probability and Statistic - P19

等可能结果的样本空间，指对于样本空间为有限集  $S = \{s_1, s_2, \dots, s_n\}$  则令  $P\{s_i\} = P\{s_2\} = \dots = P\{s_n\}$  于是可知  $P\{\{s_i\}\} = \frac{1}{n}$ ,  $i = 1, 2, \dots, n$  则对于任何事件  $E \subseteq \{s_1, s_2, \dots, s_n\}$  有  $P(E) = E$  中的结果数 /  $S$  中的结果数 =  $\frac{|E|}{|S|}$  直觉上即为在一次所有结果均等可能地发生的试验中，任何事件发生的概率等于其包含结果数占样本空间中所有结果数的比例

## 错排

(derangement)，指对于一个包含  $n$  个元素的排列，如果该排列中所有元素均不在其原来的位置，则称为原排列的错排

$n$  个不同元素的排列的不同错排数记为  $D_n$

假设有  $n$  位男士将他们的帽子混在一起并随机发给每人 1 顶帽子

求没有一个人拿到自己的帽子的发放方式数有多少种

令  $E_i$  表示事件为第  $i$  位男士拿到了自己的帽子， $U$  表示事件空间

则没有人拿到自己的帽子为  $\prod_{i=1}^n \overline{E_i} = U - \bigcup_{i=1}^n E_i$

即有  $|\prod_{i=1}^n \overline{E_i}| = |U| - \sum_{i=1}^n |E_i| + \sum_{1 \leq i < j} |E_i E_j| + \dots +$

$$(-1)^k \sum_{1 \leq i_1 < \dots < i_k} |E_{i_1} E_{i_2} \dots E_{i_k}| + \dots + (-1)^n |E_1 E_2 \dots E_n|$$

注意  $E_{i_1} E_{i_2} \dots E_{i_k}$  表示第  $i_1, i_2, \dots, i_k$  位男士拿到了自己的帽子，对于  $1 \leq k \leq n$

即有  $|k \cdot (n-k)!|$  种不同的分配方式

又  $\sum_{1 \leq i_1 < \dots < i_k} |E_{i_1} E_{i_2} \dots E_{i_k}|$  中有  $\binom{n}{k}$  项

于是有  $(-1)^k \sum_{1 \leq i_1 < \dots < i_k} |E_{i_1} E_{i_2} \dots E_{i_k}| = \binom{n}{k} \cdot |k \cdot (n-k)!| \cdot (-1)^k$

$$= \frac{n!}{k!(n-k)!} \cdot (n-k)! \cdot (-1)^k = (-1)^k \frac{n!}{k!}$$

又  $U$  中为  $n$  顶帽子的全排列，则  $|U| = n!$

于是  $D_n = |\prod_{i=1}^n \overline{E_i}| = n! - \frac{n!}{1!} + \frac{n!}{2!} - \dots + (-1)^k \frac{n!}{k!} + \dots + (-1)^n \frac{n!}{n!}$

$$= n! \sum_{k=0}^n (-1)^k \frac{1}{k!} \quad \text{其中 } 0! = 1$$

由于排列为均匀随机排列，所以每个排列出现概率均为  $\frac{1}{n!}$

于是  $P(\prod_{i=1}^n \overline{E_i}) = \sum_{k=0}^n (-1)^k \frac{1}{k!}$

特别地有，当  $n \rightarrow +\infty$  时， $P(\prod_{i=1}^n \overline{E_i}) = \sum_{k=0}^n (-1)^k \frac{1}{k!} \sim e^{-1}$

于是，相比于直觉上当有无限多人时，每个人都拿错帽子应趋向于 1

实际上概率仅趋近于  $e^{-1} \approx 0.37$

# Probability and

## Statistic - P20

错排的递归形式，对于包含n个元素的排，其错排的数量为 $D_n$ 。

$$\text{则有 } D_n = (n-1)(D_{n-1} + D_{n-2}), n \geq 2$$

组合证明过程有，当 $n=0$ 时，定义 $D_0=1$ ，即0排列是其本身的错排；当 $n=1$ 时， $D_1=0$ 。

当 $n \geq 2$ 时，考虑第1个人拿到的帽子，有 $n-1$ 种不同选择。

令第*i*个人的帽子被第1个人拿走，考虑第*i*个人得到的帽子。如果第*i*个人拿到第1个人的帽子，则其余 $n-2$ 个人分配剩余 $n-2$ 顶帽子。

即有 $D_{n-2}$ 种不同错排方式。

如果第*i*个人没有拿到第1个人的帽子，则可以视为 $n-1$ 个人分配剩余 $n-1$ 顶帽子。

即有 $D_{n-1}$ 种不同错排方式。

于是合计有 $D_n = (n-1)(D_{n-2} + D_{n-1})$ 种不同错排方式。

生日悖论 (birthday paradox)，指使至少有两个人同一天生日的概率达到50%，至少需要多少人。

用整数1, 2, ..., k对屋子里的人进行编号，并假设每年均为 $n=365$ 天。

令 $b_i$ 表示第*i*个人的生日，则有 $1 \leq b_i \leq n$ ,  $i=1, 2, \dots, k$

并假设对每个 $b_i$ 都均匀地分布在 $[1, n]$ 上。

则令 $\Pr\{b_i=r\}$ 表示第*i*个人的生日是 $r$ 的概率， $r=1, 2, \dots, n$

于是有 $\Pr\{b_i=r\} = 1/n$ ,  $i=1, 2, \dots, k$ ,  $r=1, 2, \dots, n$

假设任意两个人的生日是相互独立的。

于是有对于指定的 $r$ ，有 $\Pr\{b_i=r \wedge b_j=r\} = 1/n^2$ ,  $1 \leq i < j \leq k$

于是 $\Pr\{b_i=b_j\} = \sum_{r=1}^n \Pr\{b_i=r \wedge b_j=r\} = 1/n$

令事件 $A_i$ 表示对所有 $j < i$ ,  $j$ 与*i*的生日不同， $i=1, 2, \dots, n$

事件 $B_j$ 表示前*j*个人的生日各不相同，

于是有 $B_j = \bigcap_{i=1}^j A_i$ ，即对于 $j > 1$ ，有 $B_j = A_j \cap B_{j-1}$

即 $\Pr\{B_j\} = \Pr\{B_{j-1}\} \Pr\{A_j | B_{j-1}\}$

给定 $B_{j-1}$ 发生， $A_j$ 的条件概率为 $\Pr\{A_j | B_{j-1}\} = (n-j+1)/n$

又有 $A_1 = 1$ 且 $B_1 = 1$

则 $\Pr\{B_k\} = \Pr\{B_1\} \Pr\{A_2 | B_1\} \Pr\{A_3 | B_2\} \cdots \Pr\{A_k | B_{k-1}\}$

$$= \frac{n}{n} \times \frac{n-1}{n} \times \frac{n-2}{n} \times \cdots \times \frac{n-k+1}{n} = 1 \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right)$$

又有 $1+x \leq e^x$ ，则 $\Pr\{B_k\} \leq e^{-1/n} e^{-2/n} \cdots e^{-(k-1)/n} = e^{-k(k-1)/2n}$

令 $e^{-k(k-1)/2n} \leq \frac{1}{2}$ ，则 $k(k-1) \geq 2n \ln^2$ ，即 $k \geq \sqrt{1 + \sqrt{1 + 8\ln^2 n}} / 2$

可知当 $n=365$ 时，有 $k \geq 23$

# Probability and

## Statistic - P21

Bonferroni 不等式，对于事件  $E$  和  $F$ ，有  $P(E \cup F) \geq P(E) + P(F) - 1$

证明过程有，根据容斥原理，有  $P(E \cup F) = P(E) + P(F) - P(E \cap F)$

$$\text{即 } P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

又根据概率论公理有  $P(E \cap F) \leq 1$

于是  $P(E \cup F) \geq P(E) + P(F) - 1$

扩展 Bonferroni 不等式到  $n$  个事件  $E_1, E_2, \dots, E_n$  的情形

即对于事件  $E_1, \dots, E_n$ ，有  $P(E_1 \cup E_2 \cup \dots \cup E_n) \geq P(E_1) + P(E_2) + \dots + P(E_n) - (n-1)$

证明过程有，基础步骤呢：当  $n=1$  时， $P(E_1) \geq P(E_1) - 0$  平凡地成立

当  $n=2$  时，根据 Bonferroni 不等式，有

$$P(E_1 \cup E_2) \geq P(E_1) + P(E_2) - 1$$

递归步骤呢：假设对于任意  $n \geq 2$ ，有  $P(E_1 \cup E_2 \cup \dots \cup E_n) \geq P(E_1) + P(E_2) + \dots + P(E_n) - (n-1)$

$$P(E_1 \cup E_2 \cup \dots \cup E_n \cup E_{n+1}) = P(E_1 \cup E_2 \cup \dots \cup E_n) + P(E_{n+1}) - P(E_1 \cup E_2 \cup \dots \cup E_n \cup E_{n+1})$$

$$\geq P(E_1 \cup E_2 \cup \dots \cup E_n) + P(E_{n+1}) - 1$$

$$(IH) \geq P(E_1) + \dots + P(E_n) - (n-1) + P(E_{n+1}) - 1$$

$$= P(E_1) + P(E_2) + \dots + P(E_n) + P(E_{n+1}) - n$$

根据数学归纳法，对于任意  $n \in \mathbb{Z}^+$ ，有事件  $E_1, E_2, \dots, E_n$  有

$$P(E_1 \cup E_2 \cup \dots \cup E_n) \geq P(E_1) + P(E_2) + \dots + P(E_n) - (n-1)$$

对于  $n \geq 0$ ，令  $f_n$  表示连续掷一枚均匀硬币  $n$  次且不出现连续正面的可能结果数

则有  $f_0 = 1, f_1 = 2, f_n = f_{n-1} + f_{n-2}, n \geq 2$

证明过程有，当  $n=0$  时，只有 1 种可能结果，即  $f_0 = 1$

当  $n=1$  时，可能出现正面或反面，即  $f_1 = 2$

当  $n \geq 2$  时，考虑第一次掷硬币的结果

当第一次为反面时，

则其余  $n-1$  次掷硬币不出现连续正面的不同结果数为  $f_{n-1}$

当第一次为正面时，则第二次必定是反面

则其余  $n-2$  次掷硬币不出现连续正面的不同结果数为  $f_{n-2}$

于是有  $f_n = f_{n-1} + f_{n-2}$

等价地有，可以描述为没有连续的 1 的  $n$  位二进制数

$n=0$  时，仅有空串入，即  $f_0 = 1$

$n=1$  时，有位串 0, 1，即  $f_1 = 2$

$n \geq 2$  时，或者是  $0 d_1 d_2 \dots d_{n-1}$ ，或者是  $1 0 d_1 d_2 \dots d_{n-2}$

于是有  $f_n = f_{n-1} + f_{n-2}$