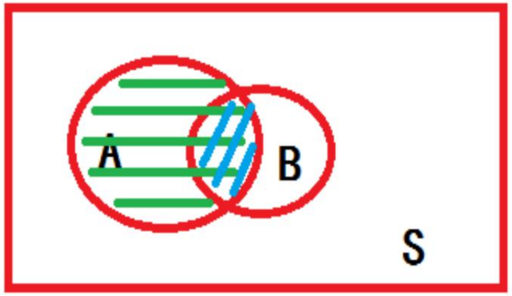
Ch.1 Basic Concepts of Probability

第一章 概率论的基本概念

上节回顾



- 条件概率 $P(B \mid A) = \dfrac{P(AB)}{P(A)}$
- 独立性 $P(AB) = P(A)P(B)$
- 加法原则

$$P(\bigcup_{i=1}^{n} A_k) = \sum P(A_i) - \sum P(A_iA_j) + \sum P(A_iA_jA_k) - \ldots + (-1)^{n-1}P(A_1A_2\ldots A_n)$$

- 乘法原则

$$P(A_1A_2\ldots A_n) = P(A_1)\,P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|A_1A_2\cdots A_{n-1})$$

多选题 1分

设置

下列说法中正确的是

- A A,B独立，则$P(B|A)=P(B)$
- B A,B独立，则A,B互斥
- C A,B互斥，则A,B独立
- D A,B独立，则$P(AB)=P(A)P(B)$
- E 如果A,B,C两两独立，则A、B、C相互独立
- F 如果A、B、C相互独立，则A，B，C两两独立
- G 如果$P(ABC)=P(A)P(B)P(C)$，则A、B、C相互独立
- H A与B独立，则$\bar{A}$与$\bar{B}$也独立

提交

1.4.3 独立事件

Remark 1： 区分互斥、对立、独立三个概念；

互斥： $A \cap B = \Phi$

对立： 互斥 $A \cap B = \Phi$，且 $A \cup B = S$

独立： $P(AB) = P(A)P(B)$

互斥必定不独立

Remark 2: 独立、互斥往往是根据实际意义去判断

1.4.3 独立事件

[定义1.3] 三个事件的相互独立，如果满足以下所有等式

$$\left.\begin{array}{l}\left.\begin{array}{l}P(AB)=P(A)P(B)\\ P(BC)=P(B)P(C)\\ P(AC)=P(A)P(C)\end{array}\right\}\text{两两独立}\\ P(ABC)=P(A)P(B)P(C)\end{array}\right\}\text{相互独立}$$

一般，设A_1，A_2，……，A_n是n个事件，如果其中的任意多个事件的积事件的概率都等于各个事件概率的积，则称A_1，A_2，……，A_n相互独立。

Remark 3：两两独立未必相互独立

例 四张卡片分别标以数字1,2,3,4，今任取一张。

A：取得的是1或2，$P(A)=1/2$

B：取得的是1或3，$P(B)=1/2$

C：取得的是1或4，$P(C)=1/2$

$P(AB)=1/4 \quad P(BC)=1/4 \quad P(CA)=1/4$

$P(AB)=P(A)P(B)$, $P(AC)=P(A)P(C)$, $P(BC)=P(B)P(C)$

但是:

$P(ABC)=1/4$, $\quad P(A)P(B)P(C)=1/8$

当5张数字时，事件不变，还两两独立吗?

填空题 1分

设置

若每个人的呼吸道中有感冒病毒的概率为0.005，在有1000人看比赛的球场中有感冒病毒的概率是[填空1]（结果请保留两位小数）

作答

若每个人的呼吸道中有感冒病毒的概率为0.005

在有1000人看比赛的球场中有感冒病毒的概率。

解：A_i: 事件“第i个人带有感冒病毒” （i＝1,2,...，1000）

假定每个人是否带有感冒病毒是相互独立的，则所求概率为

$$P\left(\bigcup_{i=1}^{1000} A_i\right) = 1 - P\left(\overline{A_1}\,\overline{A_2}\cdots\overline{A_{1000}}\right)$$

$$= 1 - P\left(\overline{A_1}\right)P\left(\overline{A_2}\right)\cdots P\left(\overline{A_{1000}}\right)$$

$$= 1 - (1 - 0.005)^{1000}$$

$$= 1 - 0.995^{1000}$$

$$\approx 0.99$$

一旦确定事件是相互独立的，在计算概率时，尽可能转化为事件的乘积进行计算

可见：虽然每人带感冒病毒的可能性很小，但许多人聚集在一起时，空气中含有感冒病毒的概率可能会很大

- 这种现象称为**小概率事件的聚众效应**
- 特殊时期，不聚众，戴口罩

Example 1.4.10 有一批产品是由甲、乙、丙三厂同时生产

	甲厂	乙厂	丙厂
产品百分比	15%	80%	5%
产品次品率	2%	1%	3%

如果从这批产品中随机抽取一件，试计算该产品是次品的概率多大?

解：设B_1、B_2、B_3分别表示抽得产品是甲厂、乙厂、丙厂生产的，A表示抽得产品为次品。

$$A=(A\cap B_1)\cup(A\cap B_2)\cup(A\cap B_3)$$

$$P(A)=P(A\cap B_1)+P(A\cap B_2)+P(A\cap B_3)$$

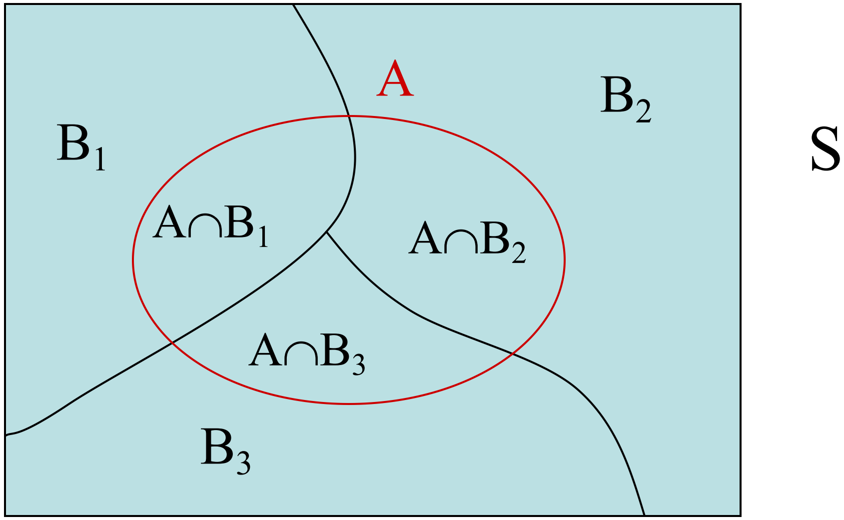
$$=P(A|B_1)P(B_1)+P(A|B_2)P(B_2)+P(A|B_3)P(B_3)$$

$$=0.02\times0.15+0.01\times0.8+0.03\times0.05$$

$$=0.0125$$

1.5　全概率公式及贝叶斯公式!

上面的例子，事实上是把事件 A 的概率计算转化为A在空间S上几个部分的概率计算。



空间S的划分：

设S为随机试验E的样本空间，$B_1, B_2, ---, B_n$为E的一组事件，如果

[1] $B_1, B_2, ---, B_n$是两两互斥的事件　不重

[2] $B_1\cup B_2\cup---\cup B_n = S$　不漏

则称$B_1, B_2, ---, B_n$称为样本空间S的一个划分

1.5 全概率公式及贝叶斯公式

[定理]：设S为随机试验E的样本空间，$B_1, B_2, ---, B_n$为S的一个划分，且 $P(B_i)>0$, $i=1,2,\ldots,n$。则E的任意事件A的概率为

$$P(A)=P(A|B_1)P(B_1)+P(A|B_2)P(B_2)+\ldots+P(A|B_n)P(B_n)$$

Remark：上述称为全概率公式，它事实上是先把在一个复杂空间S上事件A的概率，分解为在数个小的子空间上的概率之和

$$P(A)=P(AB_1)+P(AB_2)+\ldots+P(AB_n)$$

进一步地把在每个子空间上的概率分解为条件概率的积.

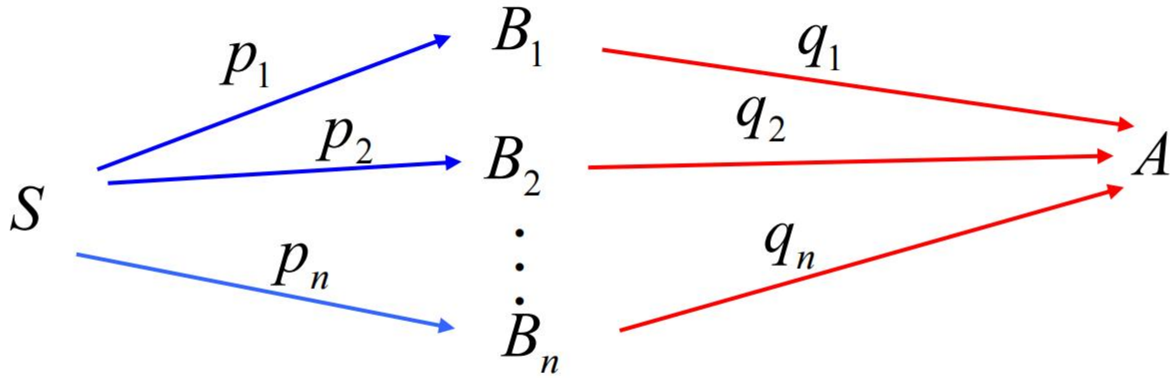
$$P(A)=P(A|B_1)P(B_1)+P(A|B_2)P(B_2)+\ldots+P(A|B_n)P(B_n)$$

现实意义：现实中A是一个复杂事件，如上例中的抽到次品，其原因众多；我们可以逐个分析造成A的原因，再将其合成起来。

全概率公式

设B_1,B_2, ---, B_n为样本空间S的一个划分

设$P(B_j)=p_j, P(A|B_j)=q_j, j=1,2,\ldots,n.$



则 $P(A)=\sum_{j=1}^{n} p_j q_j.$

注意：在运用全概率公式时,关键是构造合适的划分.

Example 1.5.1 甲乙丙分别操纵三门炮向一飞机射击。设他们的命中率分别为0.4、0.5、0.7；如只有一人射中，飞机坠毁的概率为0.2，如两个射中，飞机坠毁的概率为0.6，如三人都射中，则飞机必坠毁。

求：三人同时射击时飞机坠毁的概率?

解：A_0＝0人射中；

A_1＝有1人射中；

A_2＝有2人射中；

A_3＝有3人射中；

B＝飞机坠毁

$P(B|A_0)=0$; $P(B|A_1)=0.2$; $P(B|A_2)=0.6$; $P(B|A_3)=1$

C_1＝甲射中，C_2＝乙射中，C_3＝丙射中

解：A_0、A_1、A_2、A_3是S的一个划分。有

$P(B)=P(B|A_0)P(A_0)+P(B|A_1)P(A_1)+P(B|A_2)P(A_2)+P(B|A_3)P(A_3)$

现在要求出A_0、A_1、A_2、A_3的概率

$P(A_0)$ ＝P(有0个中) $=(1-0.4)(1-0.5)(1-0.7)=0.09$

$P(A_1)$＝P(有1个中)$=P(C_1\overline{C_2}\overline{C_3})+P(\overline{C_1}C_2\overline{C_3})+P(\overline{C_1}\overline{C_2}C_3)$

$=0.4(1-0.5)(1-0.7)+(1-0.4)0.5(1-0.7)+(1-0.4)(1-0.5)0.7$

$=0.36$

$P(A_2)$＝P(有2个中)$=P(C_1C_2\overline{C_3})+P(C_1\overline{C_2}C_3)+P(\overline{C_1}C_2C_3)$

$=0.4\times0.5(1-0.7)+0.4(1-0.5)0.7+(1-0.4)0.5\times0.7$

$=0.41$

$P(A_3)$＝P(3个全中)$=0.4\times0.5\times0.7=0.14$

$P(B)=0.458$

1.5.2 Bayes' Theorem！！！！

	甲厂	乙厂	丙厂
产品百分比	15%	80%	5%
产品次品率	2%	1%	3%

一类实际问题：“已知原因看结果”　全概率公式

已知导致结果A发生的各个原因，通过逐个分析原因，求出结果A发生的概率大小。例如，求抽到次品的概率。

另一类实际问题：“已知结果看原因”　贝叶斯公式

已知结果A发生的前提下，去分析导致它发生的原因，或者分析各种原因作用的大小。例如，已知抽到次品，求这个次品来自甲工厂的概率；最有可能来自哪个工厂。

1.5.2 Bayes’ Theorem

注意：事件A发生后，我们希望知道导致这个事件发生的各个原因的可能性大小。所有原因就是B_1、B_2、B_3，它们事实上是空间S的一个划分。

[Bayes定理]：设试验E的样本空间为S，$B_1, B_2, \ldots, B_n$为样本空间S的一个划分，A为E的事件，$P(A)>0$, $P(B_i)>0$ $(i=1,2,\ldots n)$，则：

$$P(B_i|A)=\frac{P(A\mid B_i)P(B_i)}{\sum_{j=1}^{n}P(A\mid B_j)P(B_j)}\quad i=1,2,\cdots,n$$

1.5.2 Bayes' Theorem

Proof: $P(B_i \mid A) = \dfrac{P(B_i A)}{P(A)} \qquad i = 1,2,\cdots,n$

注意到$B_1, B_2, \ldots, B_n$为样本空间S的一个划分

$$P(A) = \sum_{j=1}^{n} P(AB_j) = \sum_{j=1}^{n} P(A \mid B_j)P(B_j)$$ 全概率公式

同时 $P(B_i A) = P(A \mid B_i)\, P(B_i)$ 条件概率定义

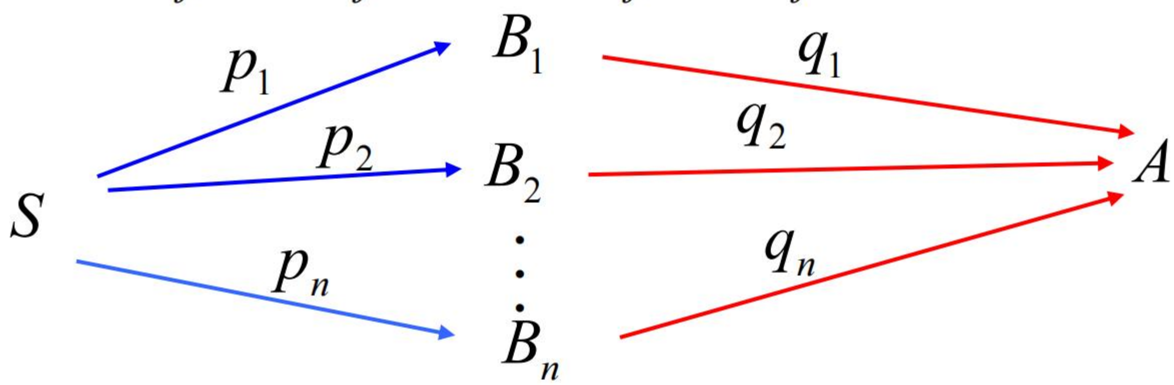
$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{n} P(A \mid B_j)P(B_j)} \qquad i = 1,2,\cdots,n$$

Remark 1：该公式于1763年由贝叶斯(Bayes)给出，故称Bayes公式。它是在观察到事件A已发生的条件下，寻找导致A发生的各个原因的概率。有时也称为逆概率公式。

贝叶斯公式

设B_1,B_2, ---, B_n为样本空间S的一个划分

$$\text{设} P(B_j)=p_j, P(A\mid B_j)=q_j, j=1,2,\dots,n.$$



$$P(B_i\mid A)=\frac{P(B_i)P(A\mid B_i)}{\sum_{j=1}^{n}P(B_j)P(A\mid B_j)}=\frac{p_i q_i}{\sum_{j=1}^{n}p_j q_j}$$

填空题 1分

设置

	甲厂	乙厂	丙厂
产品百分比	15%	80%	5%
产品次品率	2%	1%	3%

已知抽到一件次品，则该次品是甲工厂生产的概率为 [填空1]

作答

1.5.2 Bayes' Theorem

Remark 2： $P(B_i)$称为先验概率，$P(B_\mathrm{i} \mid A)$称为后验概率

($P(A|B_i)$叫做似然概率 *)

$P(B_i)$ ($i=1,2,\ldots,n$)是基于以往的统计，人们对诸事件发生可能性大小的认识，故称为先验概率。

当有了新的信息（知道A发生）后，人们对诸事件发生可能性大小$P(B_i \mid A)$有了新的估计。贝叶斯公式从数量上刻划了这种变化。

比较$P(B_1 \mid A)$、$P(B_2 \mid A)$、……、$P(B_n \mid A)$的大小，则知导致A发生的最可能的原因.

Remark 3： 注意Bayes公式对于任意事件B(不一定要是$\mathrm{B_i}$)成立

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{\sum_{j=1}^{n} P(A \mid B_j)P(B_j)}$$

1.5.3 贝叶斯定义的重要性

在现实的机器学习中，我们研究问题的通用的流程为:
我们认为所有观测到的数据都是p(observation|reason), reason不知道
而机器学习的第一个目的就是估计原因 $p(reason_i|observation)$;
第二个目标是通过全概率公式估算 p(observation)来预测 p(future)

例如研究股市涨跌时:

设企业利好的概率B_1, 银行降息的概率B_2，此处为先验概率$P(B_1)$;

其次我们观测统计市场（例如深股）的$P(A|B_i)$，即条件B_i下，事件A（例如股市上涨）发生的可能性，此处为似然概率;

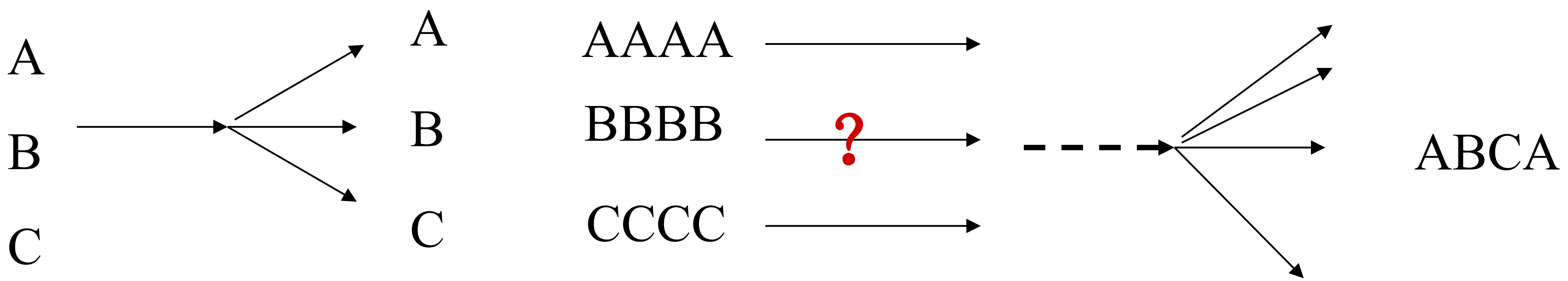
我们就可以通过贝叶斯公式研究$P(B_i|A)$，为当前股市上涨背后的原因，此处为后验概率。

Example 1.3.12（习题40 pp.29）将A,B,C三个字母之一输入信道，输出为原字母的概率为α，而输出为其他一字母的概率均为(1-α)/2. 今将字母串AAAA,BBBB,CCCC之一输入到信道，输入的概率分别为p_1, p_2, p_3 ($p_1+p_2+p_3=1$)。

假设：信道传输各个字母的工作是独立的

已知：输出为ABCA

问：输入的是AAAA的概率是多少?



$$P(\mathrm{AAAA}|\mathrm{ABCA}) = \frac{P(\mathrm{ABCA}|\mathrm{AAAA})P(\mathrm{AAAA})}{P(\mathrm{ABCA})}$$

$$P(\mathrm{ABCA}|\mathrm{AAAA})P(\mathrm{AAAA}) = \alpha \frac{(1-\alpha)}{2} \bullet \frac{(1-\alpha)}{2} \alpha p_1$$

$$= p_1 \alpha (1-\alpha)(1-\alpha)\alpha / 4$$

$$P(\mathrm{ABCA}) = P(\mathrm{ABCA}|\mathrm{AAAA})P(\mathrm{AAAA})$$

$$+ P(\mathrm{ABCA}|\mathrm{BBBB})P(\mathrm{BBBB})$$

$$+ P(\mathrm{ABCA}|\mathrm{CCCC})P(\mathrm{CCCC})$$

$$= p_1 \alpha (1-\alpha)(1-\alpha)\alpha/4$$

$$+ p_2 \alpha (1-\alpha)(1-\alpha)(1-\alpha)/8$$

$$+ p_3 \alpha (1-\alpha)(1-\alpha)(1-\alpha)/8$$

$$P(\mathrm{AAAA}|\mathrm{ABCA}) = 2p_1\alpha / \{2p_1\alpha + p_2(1-\alpha) + p_3(1-\alpha)\}$$

$$= 2p_1\alpha / \{(3\alpha - 1)p_1 + (1-\alpha)\}$$

思考：若输出ABCA，你认为输入的最有可能是AAAA,BBBB,CCCC中的哪个？

<u>假设α= 0.8，取不同的 p1,p2,p3，可以看到下面的结果：</u>

p1(AAAA)	p2(BBBB)	p3(CCCC)	α	(1-α)/2		ABCA|AAAA	ABCA|BBBB	ABCA|CCCC		AAAA|ABCA	BBBB|ABCA	CCCC|ABCA
0.3	0.4	0.3	0.8	0.1		0.0064	0.0008	0.0008		0.774194	0.129032	0.096774
0.1	0.8	0.1	0.8	0.1		0.0064	0.0008	0.0008		0.470588	0.470588	0.058824
0.05	0.9	0.05	0.8	0.1		0.0064	0.0008	0.0008		0.296296	0.666667	0.037037

从错误率来看，更像是AAAA->ABCA，

但第三行中，由于AAAA出现的概率太小，所以更有可能是由BBBB->ABCA

这就是先验概率对于整个估计的影响。

第一种直觉方法在机器/统计学习中被称为 最大似然概率，而后一种方法被称为最大后验方法

无数事实证明，最大后验方法比起最大似然方法预测效果很好，这也是贝叶斯概率理论压倒频率派的一个简单案例。

本周任务

- 完成第一章的作业（10月4日截至，在雨课堂中上传作业的照片或word、pdf文档）
- 画第一章的思维导图（10月9日截至，手画或使用xmind等软件，上传图片至雨课堂）
- 思考题

思考题

根据以往的临床记录，某种诊断癌症的试验具有5%的假阳性及3%的假阴性，即

若设A={试验反应是阳性}，C={被诊断者患有癌症}，则有

$$P(A\,|\,\bar{C}) = 5\%, P(\bar{A}\,|\,C) = 3\%,$$

现对自然人群进行普查，设患有癌症的概率为0.005,

P(C)=0.005，问这种方法能否用于普查?

提示：计算P(C|A)

上次课思考题：碰运气能否通过英语四级考试

每道题有4个选择，只有一个答案是对的

其选择可以视为Bernoulli试验，成功概率为1/4，失败概率为3/4；85道题的选择可以看为85重Bernoulli试验。

要及格，保证85道选择题能正确至少51道，即必须在85次中成功51次，其概率为：

$$= \sum_{k=51}^{85} C_{85}^{k} 0.25^{k} (1-0.25)^{85-k}$$

$$\approx 8.74 \times 10^{-12}$$

此概率非常之小，在1000亿碰运气的考生中，只有0.874个可能成功！！！

上次课思考题 甲乙两人进行乒乓球比赛，每局甲胜的概率为p，p≥1/2。问：对甲而言，采用三局两胜制有利，还是采用五局三胜制有利。**假设各局的胜负相互独立**。

解：采用三局两胜制，甲获胜的情况：甲甲，甲乙甲，乙甲甲，总的获胜概率是三种情况（互斥事件）概率的和

$$P_3 = P(A_1) + P(A_2) + P(A_3)$$
$$= p\times p + p(1-p)p + (1-p)p \times p$$
$$= 3p^2 - 2p^3$$

采用五局三胜制，如果要甲获胜，有以下几种可能：

甲前三盘连续获胜；

四盘（最后一盘必定是甲胜，前面三盘甲胜2盘，3选择2）；

五盘（最后一盘必定是甲胜，前面四盘甲胜2盘，4选择2）

总的获胜概率是三种情况概率的和（互斥事件）

$P_5 = P(B_1) + P(B_2) + P(B_3)$

$= p^3 + C_3^2 p^3 (1-p) + C_4^2 p^3 (1-p)^2$

$= 10p^3 - 15p^4 + 6p^5$

$p_5 - p_3 = 10p^3 - 15p^4 + 6p^5 - 3p^2 + 2p^3$

$= 12p^3 - 15p^4 + 6p^5 - 3p^2$

$= 3p^2(2p^3 - 5p^2 + 4p - 1)$

$= 3p^2(p-1)^2(2p-1)$

如果$p > 1/2$，$p_5 > p_3$，则$p_5 - p_3 > 0$，打五局比打三局要好

这就是增加了局数，偶然性就降低。