计算机问题求解 – 论题2-5

- 离散概率基础

2019年03月25日

Part I

“机会”数学

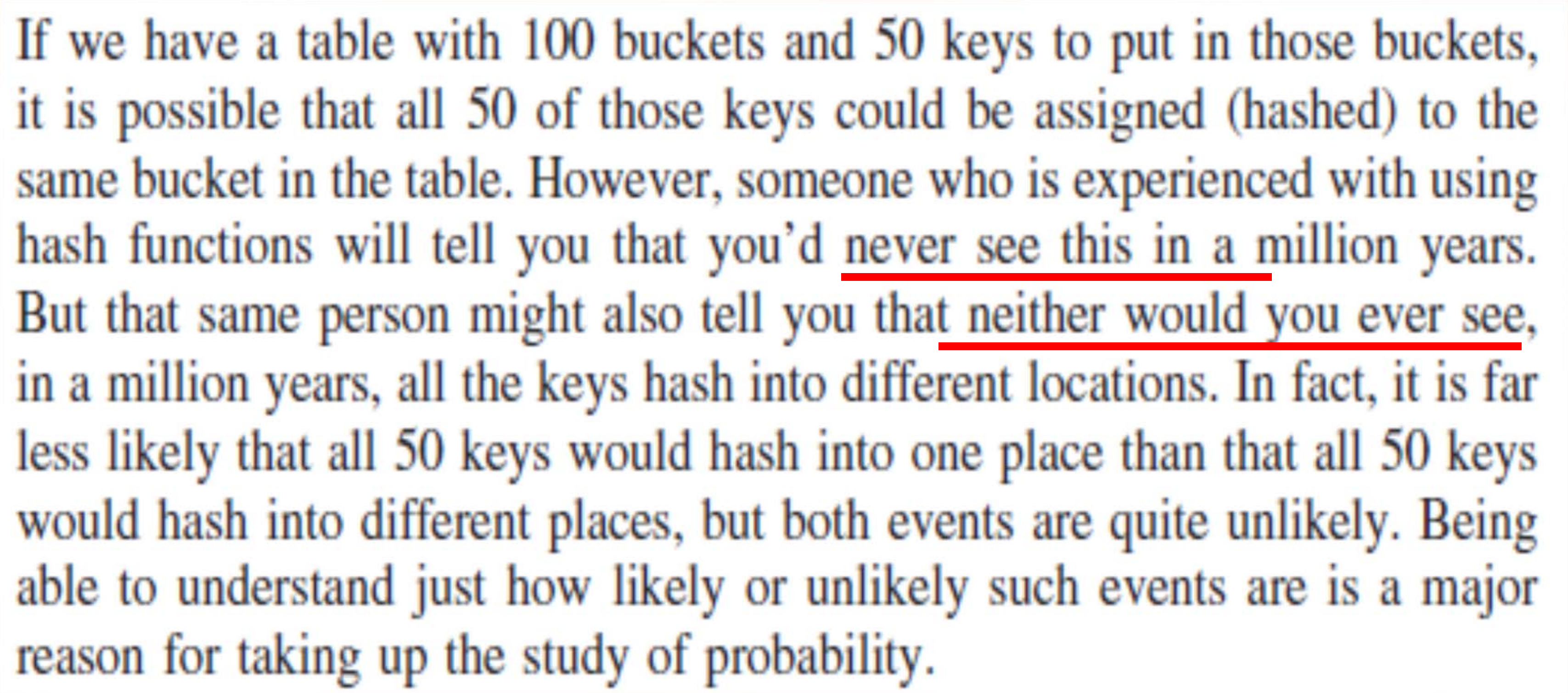
邮购商店和Hashing

将订单按照客户电话号码最后两位放入相应编号的格子中

计算机中的数据存取方法 - Hashing

问题1：

你能否类比左边的订单处理方式，解释一下什么是Hashing，特别是Hash函数的意义与作用？何为好坏？



If we have a table with 100 buckets and 50 keys to put in those buckets, it is possible that all 50 of those keys could be assigned (hashed) to the same bucket in the table. However, someone who is experienced with using hash functions will tell you that you'd never see this in a million years. But that same person might also tell you that neither would you ever see, in a million years, all the keys hash into different locations. In fact, it is far less likely that all 50 keys would hash into one place than that all 50 keys would hash into different places, but both events are quite unlikely. Being able to understand just how likely or unlikely such events are is a major reason for taking up the study of probability.

问题2：

这里的you'd never…和neither would you ever…有什么意义？

问题3：

真是百万年也碰不到一次吗？我们怎么能回答这个问题？

离散概率模型

问题4：

你能否以下面的过程为例解释离散概率模型中的主要概念？

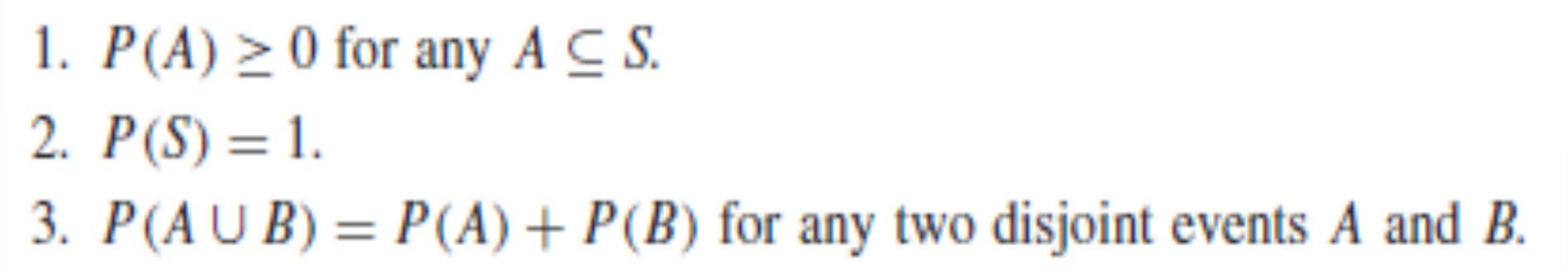
Sample space
Outcome and probability weight
Event and probability

$$P(E) = \sum_{x:x\in E} P(x).$$

A process: 掷两个骰子

Axioms for a probability space

满足下列性质的P称为一个probability distribution 或者一个probability measure。



1. $P(A) \geq 0$ for any $A \subseteq S$.
2. $P(S) = 1$.
3. $P(A \cup B) = P(A) + P(B)$ for any two disjoint events A and B.

记住：P是一个函数。

问题5：

为什么任何事件的概率值不会大于1？

Tree Diagrams: 结合计数与概率

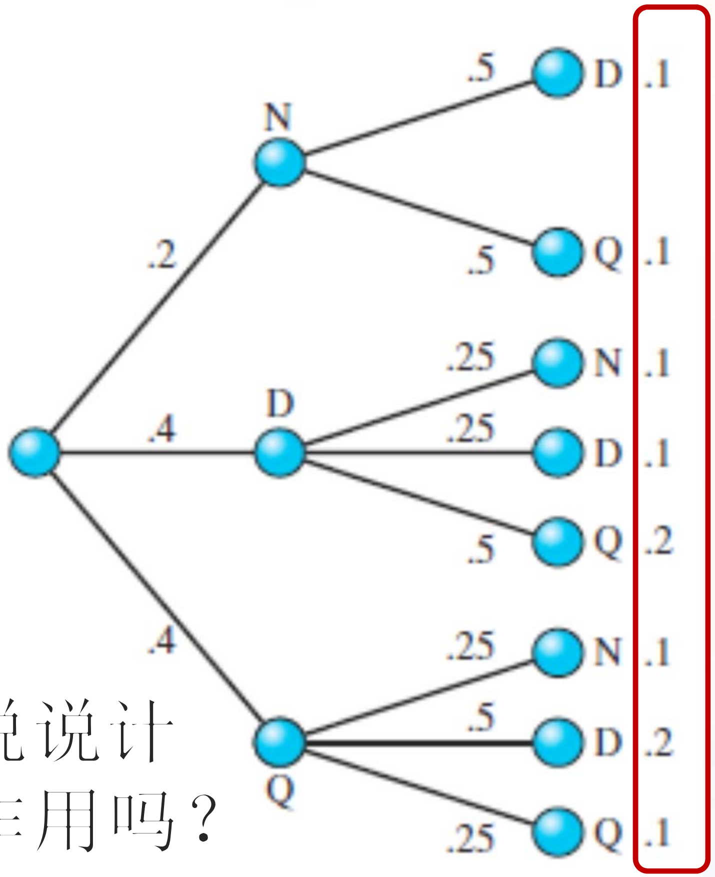
过程:

从下列硬币中依次取两枚(2-stage process):

nickel: 1
dime: 2
quarter: 2

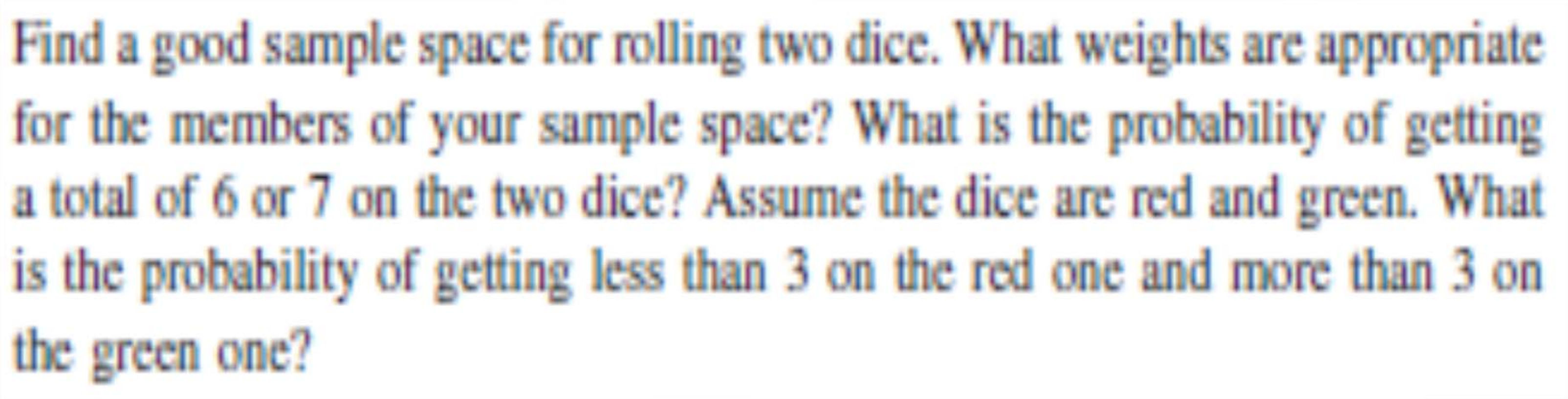
问题5:

结合这个例子，你能说说计数在计算概率时起的作用吗？这里隐含了什么假设？



为什么这个值等于该路径中前面结点上的值的乘积，将在后面解释。

有限样本空间上的离散概率计算

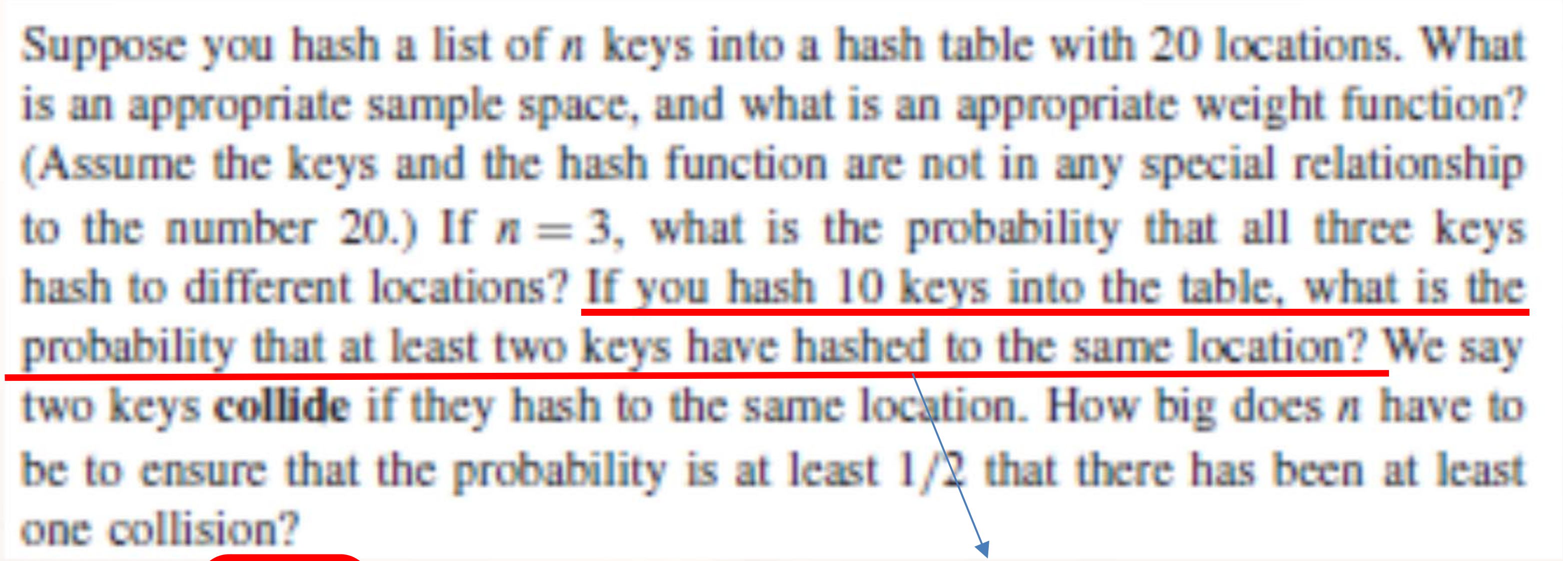


Find a good sample space for rolling two dice. What weights are appropriate for the members of your sample space? What is the probability of getting a total of 6 or 7 on the two dice? Assume the dice are red and green. What is the probability of getting less than 3 on the red one and more than 3 on the green one?

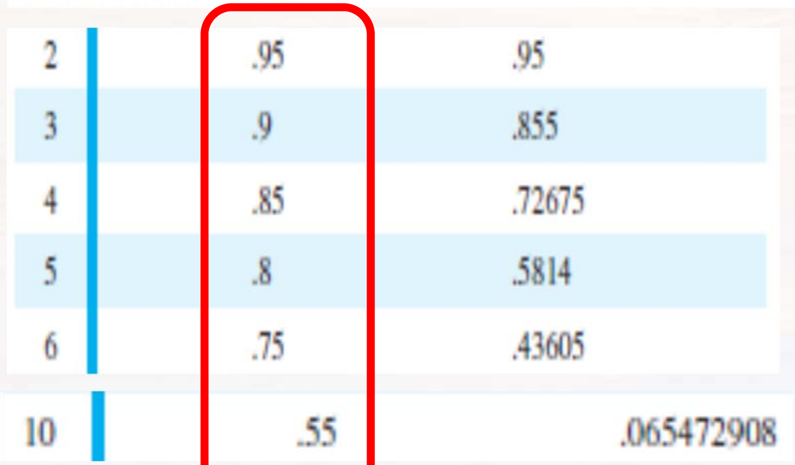
问题6：

“好的样本空间”是指什么？
带不带颜色有什么不同？

直接计数与间接计数



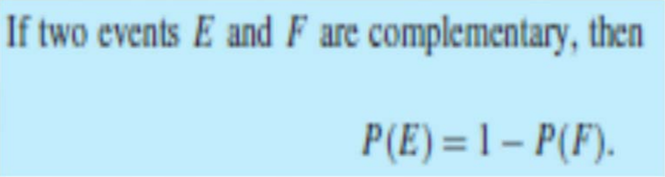
Suppose you hash a list of n keys into a hash table with 20 locations. What is an appropriate sample space, and what is an appropriate weight function? (Assume the keys and the hash function are not in any special relationship to the number 20.) If $n = 3$, what is the probability that all three keys hash to different locations? If you hash 10 keys into the table, what is the probability that at least two keys have hashed to the same location? We say two keys **collide** if they hash to the same location. How big does n have to be to ensure that the probability is at least 1/2 that there has been at least one collision?



2	.95	.95
3	.9	.855
4	.85	.72675
5	.8	.5814
6	.75	.43605
10	.55	.065472908

问题7：

这个你会算吗？

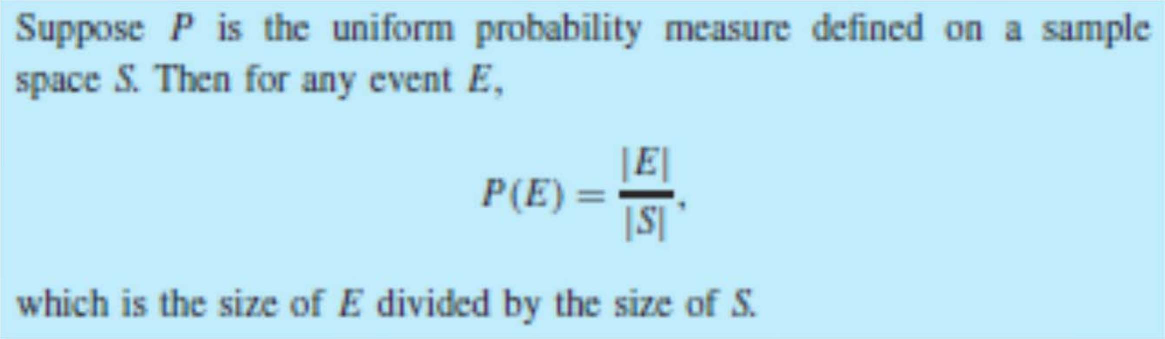


If two events E and F are complementary, then

$$P(E) = 1 - P(F).$$

顺便问一句，这是什么值？

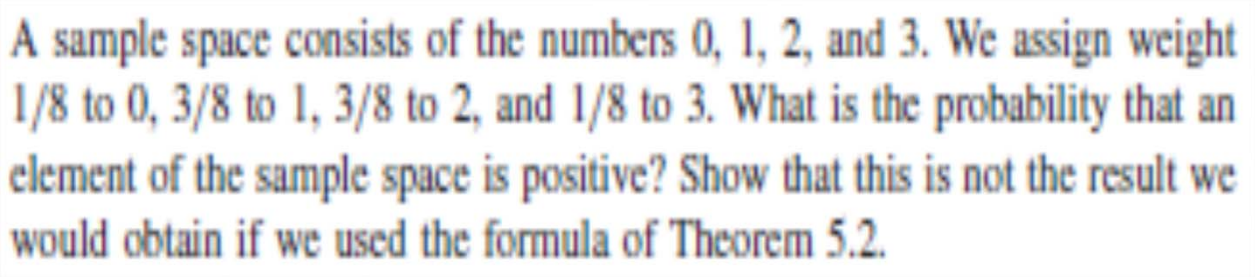
Uniform Probability Distribution



Suppose P is the uniform probability measure defined on a sample space S. Then for any event E,

$$P(E) = \frac{|E|}{|S|},$$

which is the size of E divided by the size of S.



A sample space consists of the numbers 0, 1, 2, and 3. We assign weight 1/8 to 0, 3/8 to 1, 3/8 to 2, and 1/8 to 3. What is the probability that an element of the sample space is positive? Show that this is not the result we would obtain if we used the formula of Theorem 5.2.

问题8：

造成差别的本质原因是什么？

Use the set {0, 1, 2, 3} as a sample space for the process of flipping a coin three times and counting the number of heads. Determine the appropriate probability weights $P(0)$, $P(1)$, $P(2)$, and $P(3)$.

$$P(1) = 3P(0)$$
$$P(2) = 3P(0)$$
$$P(3) = P(0)$$

$$P(0) + P(1) + P(2) + P(3) = 1.$$

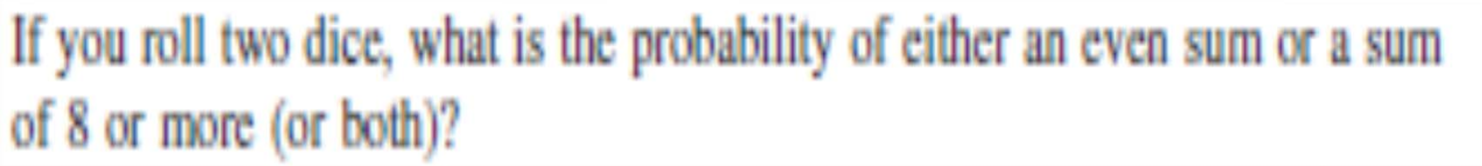
→

$$P(0) = \frac{1}{8}$$
$$P(1) = \frac{3}{8}$$
$$P(2) = \frac{3}{8}$$
$$P(3) = \frac{1}{8}.$$

问题9：

这注意到这个结果与二项式系数之间的关系吗？

事件的并集的概率


If you roll two dice, what is the probability of either an even sum or a sum of 8 or more (or both)?

E (偶数和): {(1,1), (1,3), (1,5), (2,2), (2,4), (2,6), (3,1), (3,3), (3,5), (4,2), (4,4), (4,6), (5,1), (5,3), (5,5), (6,2), (6,4), (6,6)}, 18个outcomes

F (和不小于8): {(2,6), (3,5), (3,6), (4,4), (4,5), (4,6), (5,3), (5,4), (5,5), (5,6), (6,2), (6,3), (6,4), (6,5), (6,6)}, 15个outcomes

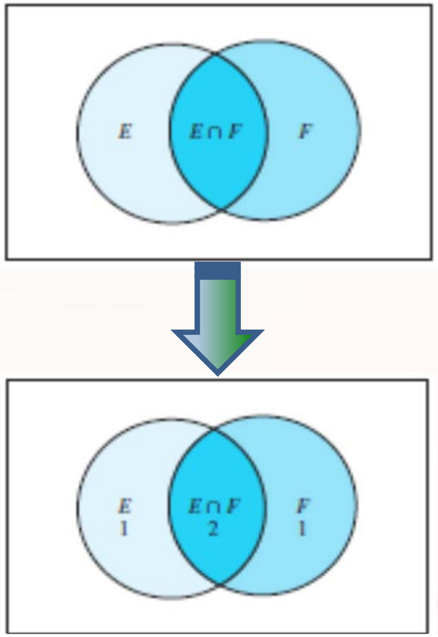
$$\text{显然}: p(E \cup F) = \frac{|E \cup F|}{|S|}$$

$$\text{其中}: |S| = 36, |E \cup F| = 18 + 15 - 9 = 24$$

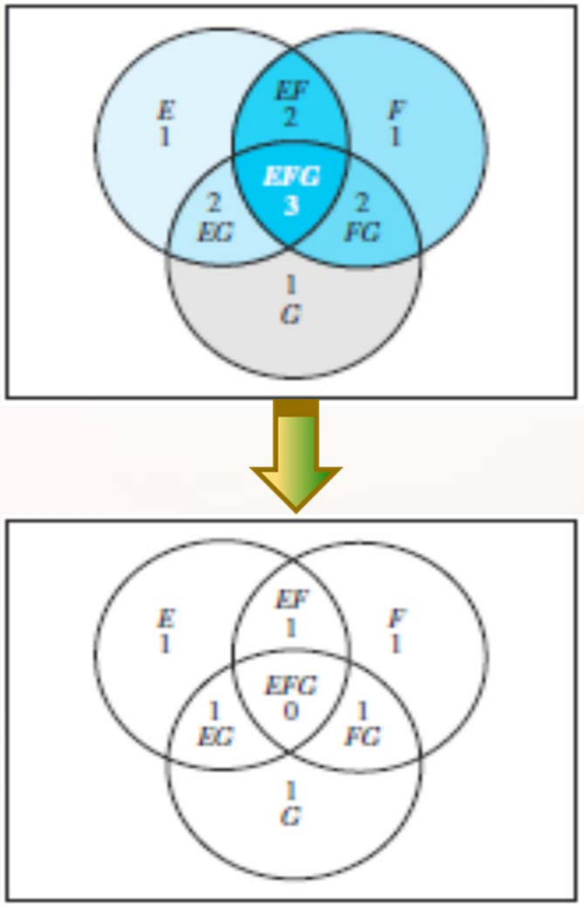
$$\therefore\ p(E \cup F) = \frac{2}{3}$$

为什么?

包含-排斥原理



$$P(E\cup F)=P(E)+P(F)-P(E\cap F).$$



$$P(E\cup F\cup G)=P(E)+P(F)+P(G)-P(E\cap F)-P(E\cap G)$$
$$-P(F\cap G)+P(E\cap F\cap G).$$

问题10：

你能解释这些图和公式吗？

和集合论中相关公式是什么关系？

一般形式及其缩略形式

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} P(E_i \cap E_j) + \sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\sum_{k=j+1}^{n} P(E_i \cap E_j \cap E_k) - \cdots$$

你应该能说处下一项该是什么?

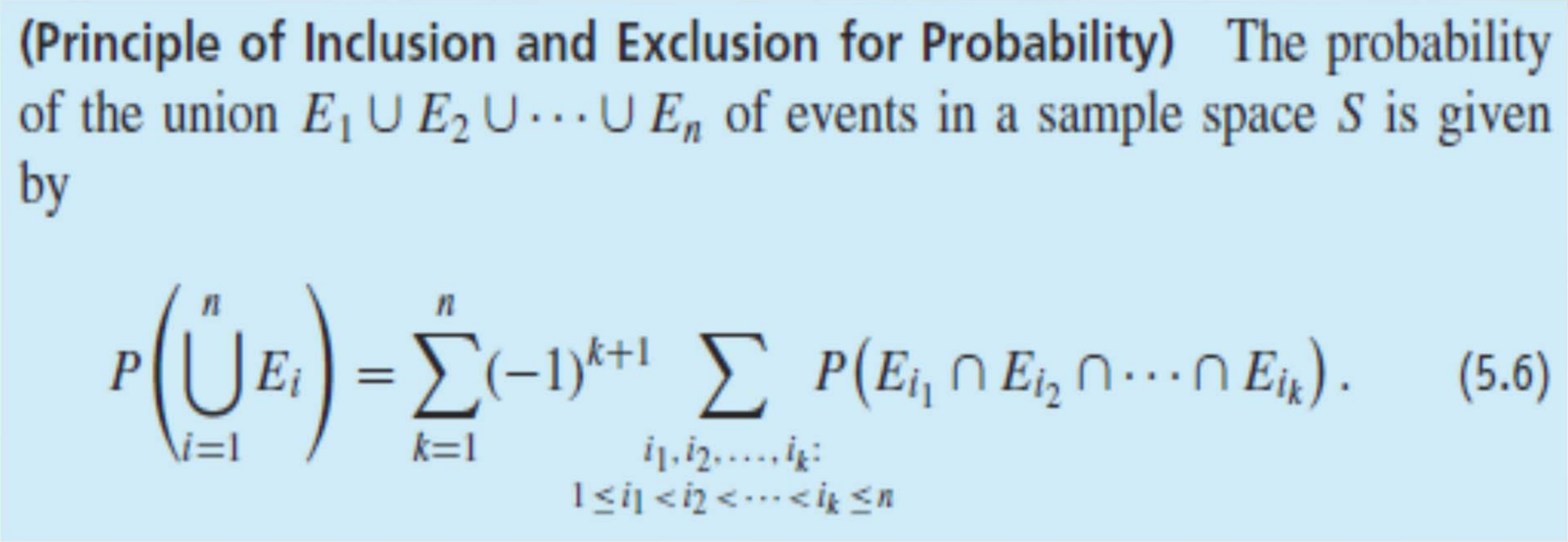
如何将多个Σ缩略成一个?

$$\boxed{\sum_{\substack{i_1, i_2, \ldots, i_k: \\ 1 \le i_1 < i_2 < \cdots < i_k \le n}} P(E_{i_1} \cap E_{i_2} \cap \cdots E_{i_k})}$$

我们可以将其记为 S_k

更可以推广到一般函数:

$$\sum_{\substack{i_1, i_2, \ldots, i_k: \\ 1 \le i_1 < i_2 < \cdots < i_k \le n}} f(i_1, i_2, \ldots, i_k)$$



(Principle of Inclusion and Exclusion for Probability) The probability of the union $E_1 \cup E_2 \cup \cdots \cup E_n$ of events in a sample space S is given by

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{\substack{i_1, i_2, \ldots, i_k: \\ 1 \le i_1 < i_2 < \cdots < i_k \le n}} P\left(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}\right). \qquad (5.6)$$

即：$p\left(\bigcup_{i=1}^{n} E_i\right) = S_1 - S_2 + S_3 - S_4 + \ldots + (-1)^{k+1} S_k$

问题11：

你能否说出这 *i* 个事件均不发生的概率是多少？

Hatcheck问题

- 大剧院衣帽间的员工太粗心，将*n*个客人的帽子上的标签搞乱了。他将*n*顶帽子随意地递交给每个客人。
  - 问题：“每个客人都拿错了帽子”的概率是多少？
- 数学模型：随机地排列自然数 1,2,3,…,*n*，生成一个序列：$i_1, i_2, i_3,\ldots,i_n$。出现下述情况的概率是多少： 对任意的 $k(1\le k\le n)$, $i_k \neq k$？
- 这样的序列称为 ***derangement***.

Derangement究竟有多少?

- 定义 “第*k*个人拿到自己的帽子”为事件E_k, 则derangement不满足这*n*个事件中的任何一个，所以：derangement的个数应该是：

$$n!-S_1+S_2-S_3+\ldots+(-1)^k S_k+\ldots+(-1)^n S_n$$

- 其实S_k就是恰好有*k*个人拿到自己的帽子，换句话说，满足这样的条件的置换中恰好有*k*个位置$i_k=k$。因此：

$$S_1=\binom{n}{1}(n-1)!;\ S_2=\binom{n}{2}(n-2)!;\ldots,\ S_k=\binom{n}{k}(n-k)!=\boxed{\frac{n!}{k!}}$$

结果末必如你的预想

$$n! - S_1 + S_2 - S_3 + \dots + (-1)^k S_k + \dots + (-1)^n S_n = n! \sum_{k=0}^{n} \frac{(-1)^k}{k!}$$

- 因此，我们需要的概率就是 $\underline{\displaystyle\sum_{k=0}^{n} \frac{(-1)^k}{k!}}$

这个式子眼熟吗?

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{k!} = e^{-1} \approx 0.368$$

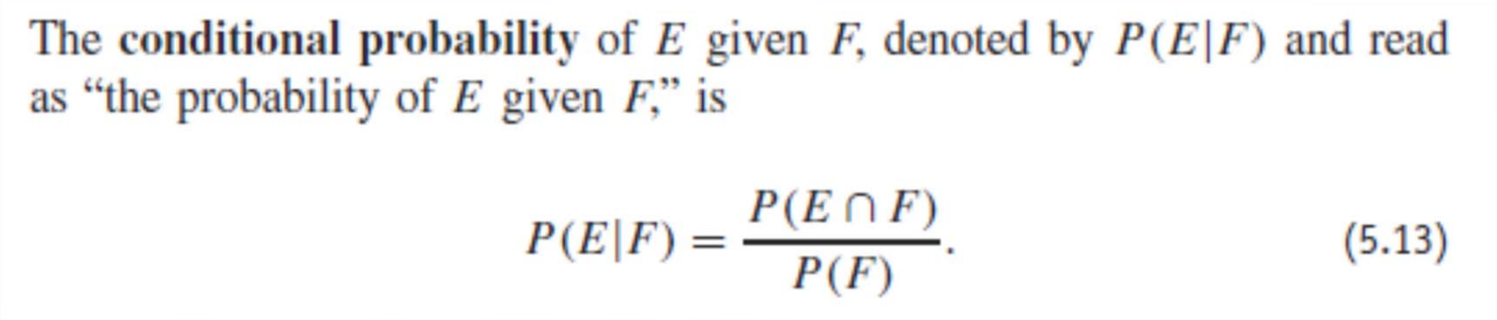
问题12：

这意味着什么？

Part II

条件概率

条件概率

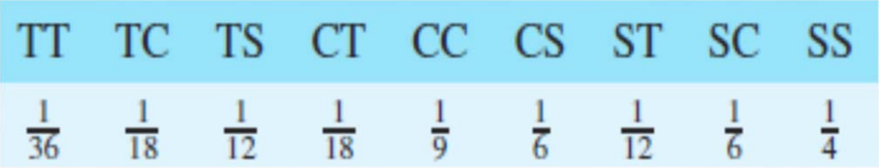


The **conditional probability** of E given F, denoted by $P(E|F)$ and read as "the probability of E given F," is

$$P(E|F) = \frac{P(E \cap F)}{P(F)}. \tag{5.13}$$

书上的例子：掷两个特殊的色子：出现三角形、圆、正方形的面数分别是1，2，3。事件E: 至少一个有圆的面朝上；事件F: 朝上的两个面图案相同。如果知道时间F已经发生，那么事件E的概率是多少？

- 按照常识来分析：可以理解为样本空间改变了。
- 利用上面的定义式来计算。



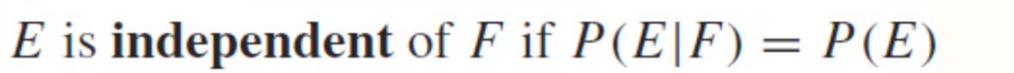
TT	TC	TS	CT	CC	CS	ST	SC	SS
$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{9}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{4}$

原来的样本空间大小是36，如果只考虑两面相同，样本空间大小缩小为14，其中出现圆图案的是4。
关键假设：三种图案间的比例不变！

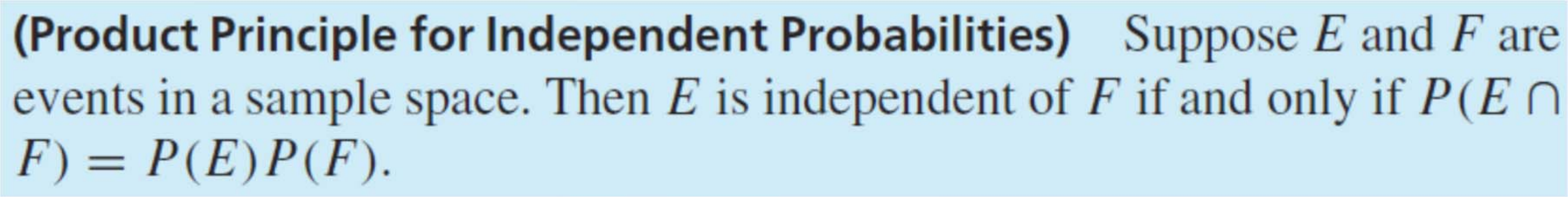
问题13：

条件概率那个式子是定义还是定理？

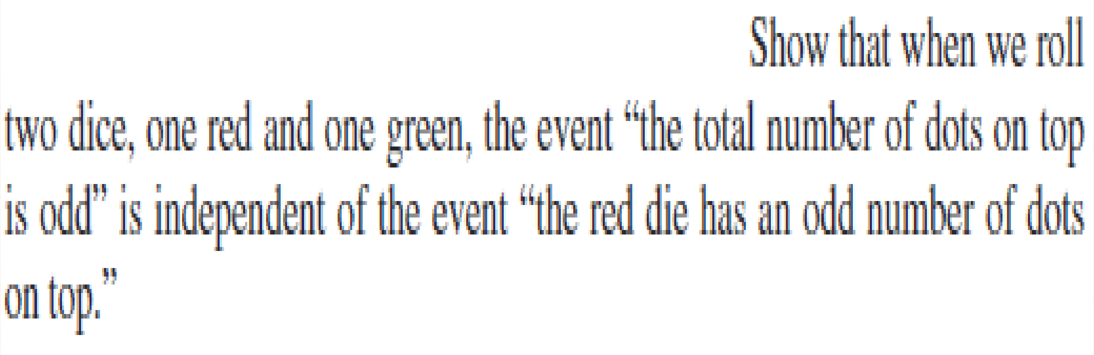
相互独立的事件



E is **independent** of F if $P(E|F) = P(E)$



(Product Principle for Independent Probabilities) Suppose E and F are events in a sample space. Then E is independent of F if and only if $P(E \cap F) = P(E)P(F)$.

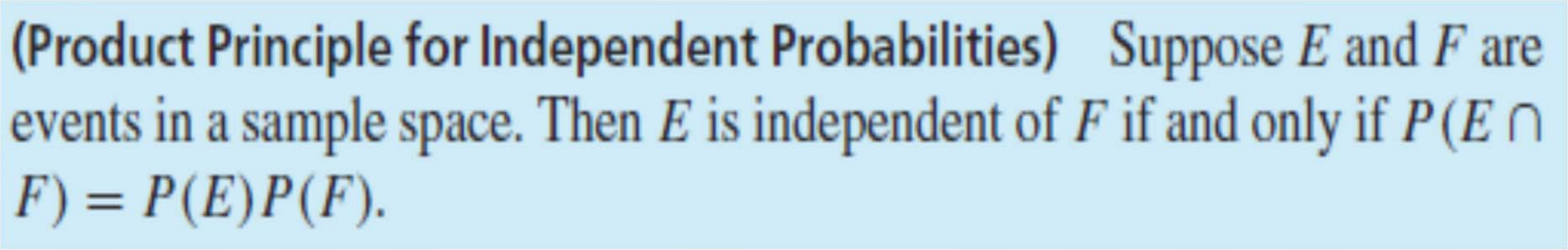


Show that when we roll two dice, one red and one green, the event "the total number of dots on top is odd" is independent of the event "the red die has an odd number of dots on top."

问题14：

你能解释吗? 并说说事件之间的相互“独立（independence)”与“互斥(disjoint)”有什么不同？

不计算条件概率也可判定独立



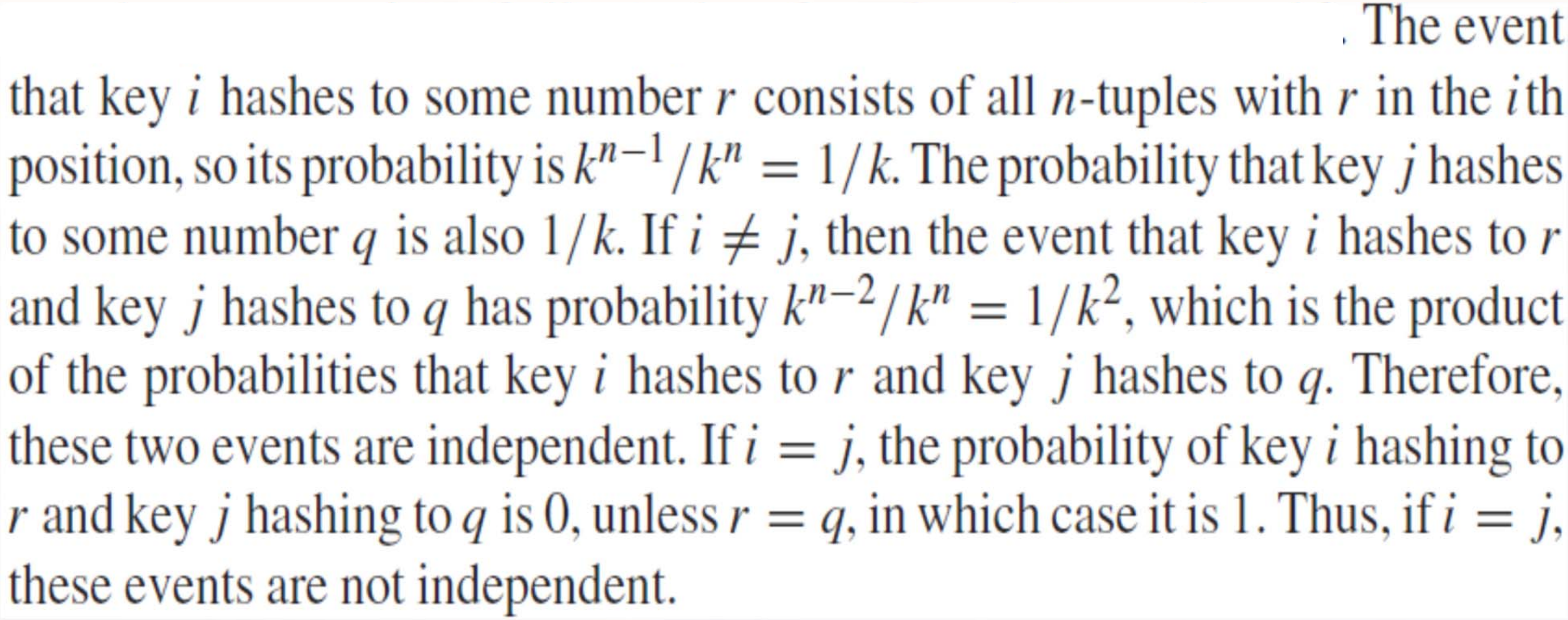
(Product Principle for Independent Probabilities) Suppose E and F are events in a sample space. Then E is independent of F if and only if $P(E \cap F) = P(E)P(F)$.

问题15:

连续两次掷骰子，直观上我们认为第2次的结果应该独立与第1次的结果，你能通过计算说明我们关于独立事件的定义确实符合直观吗？

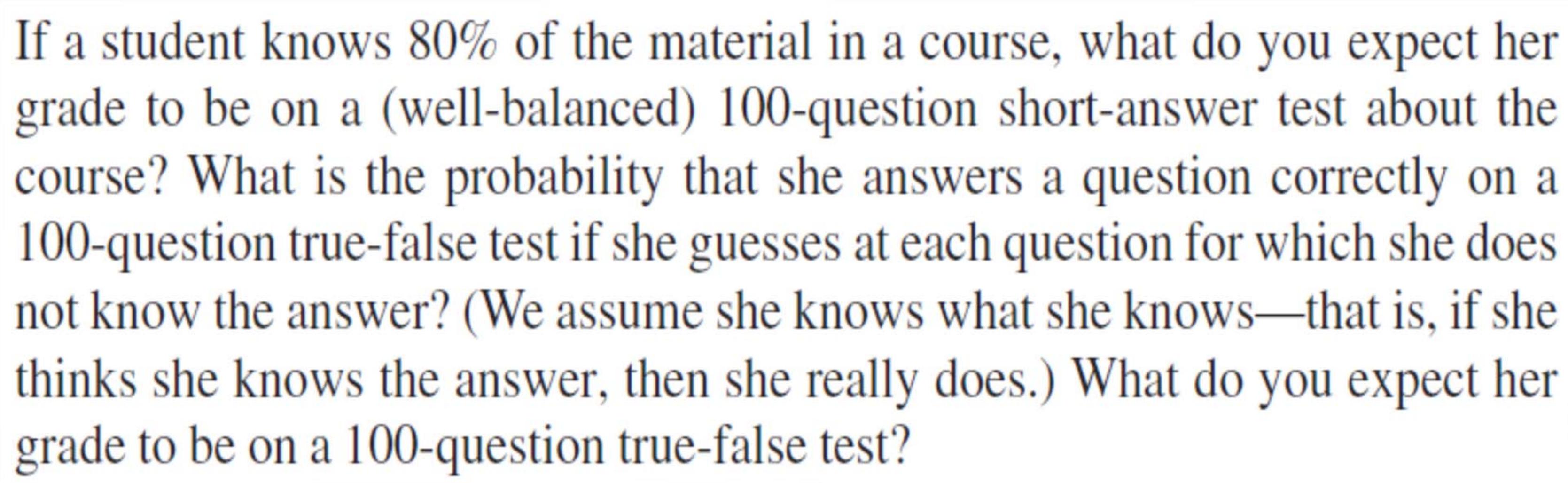
Hashing的样本空间和独立事件

如果将 n 个keys “哈希”到大小为 k 的表中，则样本空间包含长度为 n，元素为 $\{1,2,\ldots,k\}$ 中任意元素的序列。
考虑两个事件：“i 键哈希到位置r”和“j 键哈希到位置q”



. The event that key i hashes to some number r consists of all n-tuples with r in the ith position, so its probability is $k^{n-1}/k^n = 1/k$. The probability that key j hashes to some number q is also $1/k$. If $i \neq j$, then the event that key i hashes to r and key j hashes to q has probability $k^{n-2}/k^n = 1/k^2$, which is the product of the probabilities that key i hashes to r and key j hashes to q. Therefore, these two events are independent. If $i = j$, the probability of key i hashing to r and key j hashing to q is 0, unless $r = q$, in which case it is 1. Thus, if $i = j$, these events are not independent.

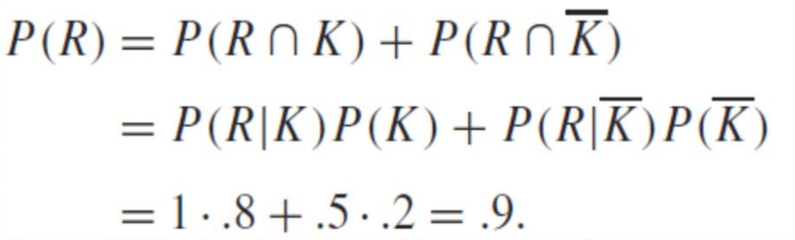
条件概率应用与概率分析–考试成绩



If a student knows 80% of the material in a course, what do you expect her grade to be on a (well-balanced) 100-question short-answer test about the course? What is the probability that she answers a question correctly on a 100-question true-false test if she guesses at each question for which she does not know the answer? (We assume she knows what she knows—that is, if she thinks she knows the answer, then she really does.) What do you expect her grade to be on a 100-question true-false test?

问题16:

你能否从“直观”上判断该期望多少分？



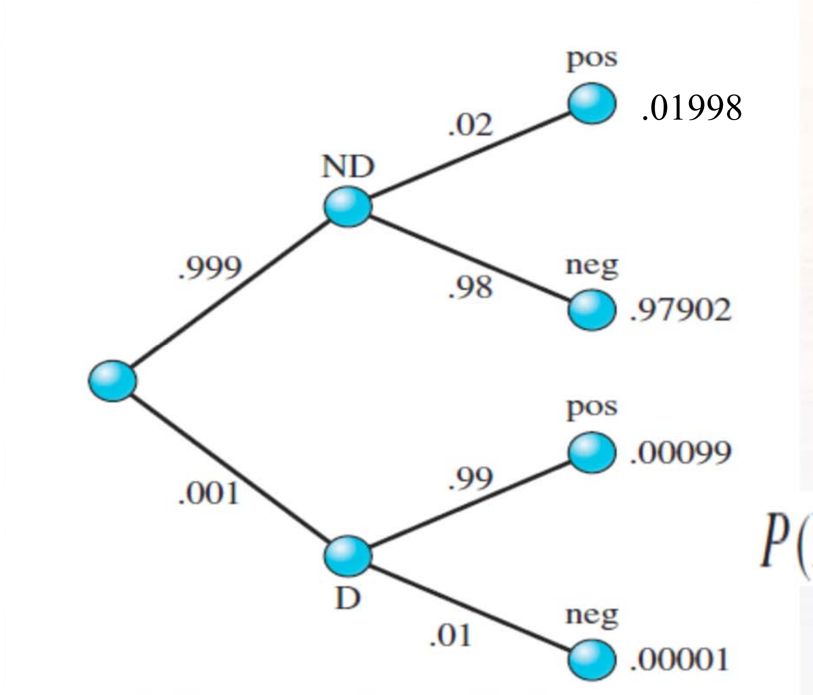
$$
\begin{aligned}
P(R) &= P(R \cap K) + P(R \cap \overline{K}) \\
&= P(R|K)P(K) + P(R|\overline{K})P(\overline{K}) \\
&= 1 \cdot .8 + .5 \cdot .2 = .9.
\end{aligned}
$$

附注：R：回答正确；K：知道正确答案

不那么“直观”的概率分析

两阶段过程：某种疾病检验，该病在全体人口中患病率为0.1%
1，任选一个人：患者(0.1%) 或非患者(99.9%)
2，检验：对患者正确率99%，对非患者出错率2%

问题：任选一人检验阳性的概率？检验阳性者实际患病的概率？



$$P(\mathrm{D}|\mathrm{pos}) = \frac{P(\mathrm{D} \cap \mathrm{pos})}{P(\mathrm{pos})}$$

$$P(\mathrm{D} \cap \mathrm{pos})/P(\mathrm{pos}) = .00099/.02097 = \underline{.0472}.$$

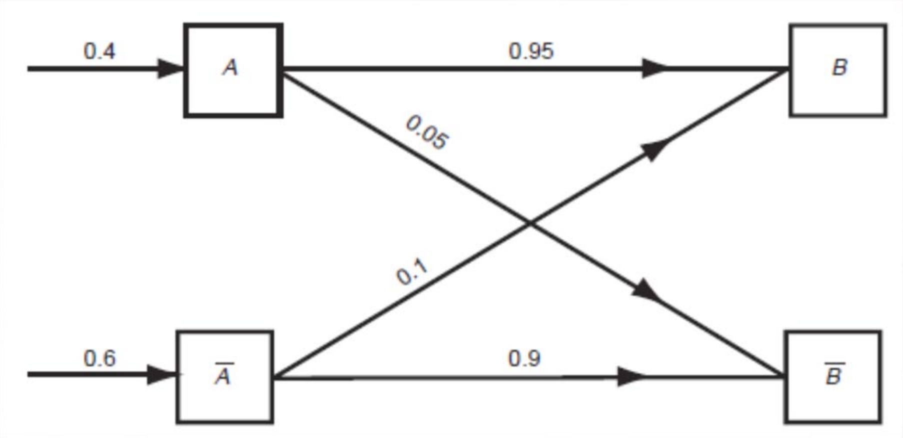
问题17：

Bayes定理是什么内容？它为什么成立？它有什么意义？

$$P(E|F)=\frac{P(F|E)P(E)}{P(F)}$$

网络通信的例子

Problem: a simple binary communication channel carries messages by using only two signals, say 0 and 1. We assume that, for a given binary channel, 40% of the time a 1 is transmitted; the probability that a transmitted 0 is correctly received is 0.90, and the probability that a transmitted 1 is correctly received is 0.95. Determine (a) the probability of a 1 being received, and (b) given a 1 is received, the probability that 1 was transmitted.



$$P(A) = 0.4, \qquad P(\overline{A}) = 0.6;$$
$$P(B|A) = 0.95, \qquad P(\overline{B}|A) = 0.05;$$
$$P(\overline{B}|\overline{A}) = 0.90, \qquad P(B|\overline{A}) = 0.10.$$

b $$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.95(0.4)}{0.44} = 0.863.$$

a $$P(B) = P(B|A)P(A) + P(B|\overline{A})P(\overline{A}) = 0.95(0.4) + 0.1(0.6) = 0.44$$

家庭作业

- CS pp.260-: 6, 10-13
- CS pp.274-: 2, 9, 10, 14, 15
- CS pp.290-: 3-4, 8, 11-13
- CS pp.307-: 5, 6, 8, 10, 17, 20, 21