浙江省高等教育重点建设教材

概率论与生物统计

倪海儿 钱国英 主编

ZHEJIANG UNIVERSITY PRESS
浙江大學出版社

浙江省高等教育重点建设教材

概率论与生物统计

倪海儿 钱国英 主编

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

图书在版编目（CIP）数据

概率论与生物统计 / 倪海儿，钱国英主编．—杭州：浙江大学出版社，2006.8
ISBN 7-308-04858-6

Ⅰ．概… Ⅱ．①倪…②钱… Ⅲ．①概率论—高等学校—教材②生物统计—高等学校—教材
Ⅳ．O211②Q-332

中国版本图书馆 CIP 数据核字（2006）第 090494 号

概率论与生物统计
倪海儿 钱国英 主编

责任编辑 周卫群
出版发行 浙江大学出版社
（杭州浙大路 38 号 邮政编码 310027）
（网址：http://www.zjupress.com）
（E-mail：zupress@mail.hz.zj.cn）
排　　版 浙江大学出版社电脑排版中心
印　　刷 浙江省良渚印刷厂
开　　本 787mm×960mm 1/16
印　　张 18.75
字　　数 350 千
版、印次 2006 年 8 月第 1 版 2006 年 8 月第 1 次印刷
书　　号 ISBN 7-308-04858-6/O·347
定　　价 28.00 元

浙江大学出版社发行部邮购电话 （0571)88072522

前 言

随着生物科学的进展，在生物学的各个分支里，古老的、描述性的研究方法已逐步地为精确的、更有把握的定量的研究方法所取代。在各种定量的研究方法中，统计方法是被广泛应用的方法之一。到二十世纪初，已形成了一门由生物学和概率统计有机结合而成的边缘学科——生物统计学。它是用概率论与数理统计的原理和方法对生物学科领域中的随机现象进行分析推断的学科。

生物统计学研究的对象是生物科学领域中的各种随机现象，而概率论是研究随机现象的工具，因此它是生物统计学的基础。在生物统计学的学习中只会机械地套用生物统计的公式是不够的，只有掌握了统计分析的原理，才能对生物科学中的各种问题用正确的统计分析方法进行分析，对计算的结果进行合理的解释与推断。因此，本书首先叙述了概率论的基本原理，在此基础上，尽量清楚地阐述生物统计的基本概念及基本原理。为体现统计方法在生物科学领域中应用的广泛性，从文献和著作中选取了一些实例，但是因为讲授上的原因，对部分实例作了一些修改，因此在此不再注明它的出处。另外，本教材的例子或习题中的结论，也不能作为生物学上结论看待。

本教材参考学时为50～70学时，其中多元统计分析的内容可根据各专业和课时进行选择。结合本教材的内容，教师可介绍一些常用的统计分析软件，如SAS、SPSS等，这些统计软件的应用已有许多专门的教材和参考书，限于篇幅，本书中不作介绍。

编 者

2006年8月

目　　录

第1章 事件与概率

§1.1 事件与概率

一、随机现象及随机事件

我们所常见的现象可分为两类:必然现象与随机现象.

如果在一定的条件下其出现的结果是肯定的、可以预言的,这一类现象叫做确定性现象或必然现象. 例如,在标准大气压下,把水加热到100℃(条件组),水必然沸腾(结果),这种在一定的条件组下必然会发生的结果称为必然事件;反之,在一定的条件组下肯定不会发生的结果叫做不可能事件.例如,“在某一池塘中,有1～4龄的鱼,从中任抽一条,抽到这条鱼的年龄为6龄”是不可能事件;必然事件和不可能事件都是必然现象.除必然现象外,在生物界中还广泛存在着与这类现象有着本质区别的另一类现象——随机现象,下面举几个例子.

例1.1.1 任取一尾鲤鱼,数它的椎骨,所得的结果可能是36枚、37枚或38枚. 这就是说:尽管条件组(任取一尾鲤鱼,数它的椎骨)是确定的,但所得的结果却是不确定的.

例1.1.2 任取一尾白鲢,测定其卵巢中DNA含量,如果这一操作进行多次,可以发现每次测定值不会完全相同. 造成这一测定值变化的因素是大量的,如被测个体之间的差异,测试设备及试剂在测试过程中的微小变化等等.这些因素的变化,导致了测定结果的不同.

例1.1.3 在同一地区,种植同一品种的作物,产量会不同.

这一类现象的共同点是:在条件组重复实现时,所出现的结果常常是不同的. 这种现象称为随机现象. 生物学中所遇到的现象大多数属于这一类型. 生物统计方法处理的对象就是生物学中的随机现象.

我们把一个条件组的实现叫做试验.随机现象就是在重复进行试验时,可能会出现不同结果的现象. 对应于随机现象的试验叫做随机试验. 在本课程中我

们只讨论随机试验，并把它简称为试验(Experiment，记为 E).

在生物学中，一次试验会产生什么结果，除了受条件组的影响外，还会受到条件组以外的许多因素的影响. 这种除条件组以外的其他作用于受试对象的各种因素，统称为随机因素. 在试验中，随机因素是我们未加控制或无法控制的因素. 正因为随机因素的影响，导致了试验的随机性，即试验结果的不确定性.

随机试验的结果叫做随机事件，简称为事件. 在一定的条件组下必然发生的事件叫做必然事件. 反之，在一定的条件组下必然不会发生的事件，叫做不可能事件. 必然事件和不可能事件都不具有不确定性，但为了今后讨论的方便，我们把它们当成随机事件的两种极端情况.

二、频数、频率及概率

随机试验的特点之一是：就个别试验而言，它可以时而出现这种结果，时而出现那种结果，即某事件是否出现就一次试验来说，具有不确定性，但是，如果试验重复的次数足够多，我们便会发现它具有某种规律性.

例 1.1.4 抛一枚硬币，每次抛掷后可能“出现正面”，也可能“出现反面”，历史上有人对此作过试验，得出下列结果：

实验者	抛掷次数 n	出现正面的次数 k	k/n
Buffon	4040	2048	0.5069
K. Pearson	12000	6019	0.5016
K. Pearson	24000	12012	0.5005

这一事实表明，随机现象虽有其偶然性的一面，但也有其必然性的一面. 如果抛一枚硬币的次数足够多，我们可以预期出现正面的次数大约是总抛掷次数的一半.

在 n 次重复试验中，某一事件 A 出现的次数 k 叫做该事件的频数，比值 k/n 叫做事件的相对频率，简称为频率. 当试验次数 n 足够大时，“抛出正面”这一事件 A 的频率将会在 0.5 附近徘徊. 这个规律性，即某个试验重复许多次以后，一个事件 A 出现的频率徘徊在某个数 p 附近，是由随机事件本身的内在属性所决定的. 常数 p 可以用来表示在一次试验中事件 A 发生的可能性的大小，把它叫做事件的概率，记为 $P(A) = p$.

例 1.1.5 根据对长江草鱼产卵鱼群的调查，得到下列结果：

鉴定鱼数 n	10	20	50	100	500	1000
雌鱼数 k	5	8	15	24	110	232
雌鱼频率 k/n	0.50	0.40	0.30	0.24	0.22	0.23

虽然整个产卵鱼群的性比是未知的，但我们可以近似地用 0.23 作为在该鱼群中任取一尾鱼恰好取到雌鱼的概率，用 A 表示这一事件，则 $P(A) = 0.23$.

有了概率的概念后，我们可以较详细地说明什么叫随机试验. 一个随机试验是指具备下列几个特点的试验：

(i) 可以重复进行；

(ii) 虽然不能断定某特定的结果将在某一次试验中出现，但能对试验的一切可能结果作出描述；

(iii) 每一结果（即每一事件）都对应着一个确定的实数，它是这个事件的概率，当试验次数足够大时，该事件的频率将稳定于它的概率.

因此，当我们要描述一个随机试验时，除了试验的条件组外，重要的两件事是"事件"和"事件的概率".

§1.2　事件的运算

一、基本事件与复合事件

在一定的研究范围内，不能再"分解"的事件叫做基本事件，由基本事件"复合"而成的事件叫做复合事件.

例 1.2.1　某水库中有年龄从 1 到 8 龄的鲫鱼，任取一尾鱼观察其年龄. 如果以 e_i 表示"取到的是 $i(i = 1,2,\cdots,8)$ 龄鱼"这一事件，则 $e_1,e_2,\cdots,e_8$ 都是基本事件；又"取到的鱼的年龄不小于 6 龄"也是一个随机事件，它是由 e_6,e_7,e_8"复合"而成，这个事件便是一个复合事件.

例 1.2.2　在池中任取一尾鱼，测量其体长，对于每一个特殊的数值 x，即"测得体长为 x"是一个基本事件，"体长 π 在 a 与 b 之间"则是一个复合事件.

一个随机试验 E 的所有基本事件所形成的集合，叫做这个试验的样本空间，记为 S. S 中的元素就是试验 E 的基本事件，基本事件也叫做样本点. 而复合事件便可理解为样本点所组成的集合，即 S 的子集.

容易看到，样本空间对应于必然事件，S 中的空集对应于不可能事件. 以后我们把必然事件记为 U，不可能事件记为 $\varnothing$.

二、事件关系及其运算

用 e 表示随机试验的基本事件，则任一事件 A 可表示成 $A=\{e|e\in A\}$. 这里 $e\in A$，表示 e 是组成事件 A 的基本事件. 如在例 1.2.1 中，用 A 表示“抽到的鱼的年龄大于 6 龄”这个事件，则有 $A=\{e_7,e_8\}$.

两个事件 A,B，当且仅当它们由相同的样本点所组成时，才说它们是相等的，记为 $A=B$.

如果两事件 A 与 B 没有公共的样本点，则称这两个事件是互斥的(或互不相容的).

可以借助图形来说明事件的互斥，在图 1.2.1 中，正方形表示样本空间 S，两个圆分别表示事件 A 及 B，A 与 B 互斥意味着这两个圆不相交.

下面我们引进事件的主要运算：

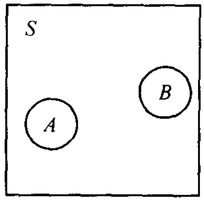

图 1.2.1

1. 事件 A 与 B 中至少有一个发生所构成的事件(“A 或 B”)称为事件 A 与 B 的和，记为 $A\cup B$. 显然，$A\cup B$ 由所有属于 A 或 B 的基本事件所构成，见图 1.2.2a.

一般地，用 $\bigcup\limits_{i=1}^{n}A_i$ 表示 n 个事件 $A_1,A_2,\cdots,A_n$ 之和.

2. 由事件 A 与 B 同时发生所构成的事件称为事件 A 与 B 的积(或交)，记为 $A\cap B$ 或 AB. AB 是由所有既属于 A 又属于 B 的基本事件所构成，如图 1.2.2b 所示.

易见，A 与 B 互斥的充要条件是 $AB=\varnothing$. 当 A 与 B 互斥时，通常把 $A\cup B$ 记作 $A+B$.

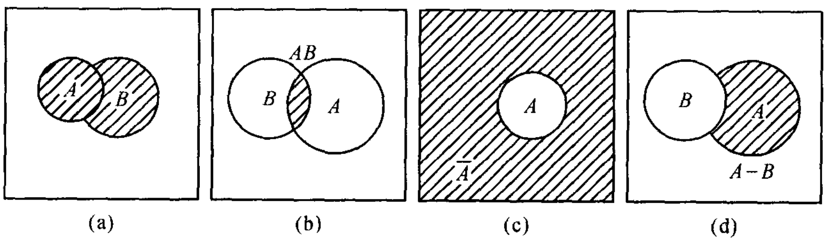

图 1.2.2

3. 若事件 A 与 B 满足以下两条件：

$$A\cup B=S;\ AB=\varnothing$$

则事件 B 是 A 的对立事件，记为 $B=\overline{A}$，显然，此时也有 $A=\overline{B}$，见图 1.2.2c. 当 A 与 B 互为对立事件时，在一次试验中，两者中必出现一个且只出现一个.

4. 由事件 A 发生而事件 B 不发生所构成的事件叫做 A 与 B 之差，记为 $A-B$. $A-B$ 是由所有属于 A 但不属于 B 的基本事件所构成，见图 1.2.2d. A 的对立事件即 $S-A$.

§1.3　古典概型

一、古典概型

在本节中，我们将讨论一类最简单的随机现象，其特征是：

(1) 试验的样本空间只含有有限个样本点，记为 $e_1,e_2,\cdots,e_n$；

(2) $e_1,e_2,\cdots,e_n$ 在试验中出现的可能性相同.

这种概率模型称为古典概型. 在这个的模型中，定义事件 A 的概率为

$$P(A)=k/n \tag{1.3.1}$$

上式中 k 是组成 A 的基本事件个数. 从以上的概率定义可以看到它有下列性质：

(1) 对于任一事件 A，$0\leqslant P(A)\leqslant 1$；　(1.3.2)

(2) 必然事件的概率为 1，不可能事件的概率为 0，即

$$P(U)=1;\ P(\varnothing)=0 \tag{1.3.3}$$

(3) 设事件 $A_1,A_2,\cdots,A_m$ 两两互不相容（即 $A_iA_j=\varnothing,i\neq j$），则

$$P\Big(\sum_{i=1}^{m}A_i\Big)=\sum_{i=1}^{m}P(A_i) \tag{1.3.4}$$

从性质 (3) 可以推得：如果 $\overline{A}$ 是 A 的对立事件，则

$$P(\overline{A})=1-P(A) \tag{1.3.5}$$

事实上，由于 $A+\overline{A}=S$，且 $A\cap\overline{A}=\varnothing$，所以

$$1=P(S)=P(A+\overline{A})=P(A)+P(\overline{A})$$

亦即

$$P(\overline{A})=1-P(A)$$

例 1.3.1　设一口塘中共有 $m+n$ 尾鱼，其中 m 尾是雌鱼，n 尾是雄鱼，从中任取 r 尾 ($r<n$)，问对于下面的两种不同的取法，取到的 r 尾鱼均为雄鱼的概率是多少？

(1) 每次取出一尾检查后放回去，再任意地取下一尾（这种做法叫做返回抽样或回置抽样）.

(2) 每次取出检查后的鱼不再放回去（这种做法叫做不返回抽样或不回置

抽样).

解:设 A 是"取出的 r 尾均为雄鱼"这一事件.

(1) 由于每次取出检查的鱼都放回去,故每次都可以取到 $m+n$ 尾鱼中的任一尾,于是,基本事件的总数可以用计算重复排列数的公式算出,它等于 $(m+n)^r$.

满足事件 A 所描述的基本事件只能从 n 尾雄鱼中取得,同样的道理,属于事件 A 的基本事件数为 n^r.

如果抽取的过程是随机的,由实际知识可以判断,所有 $(m+n)^r$ 个基本事件是等可能的,因此所求的概率为

$$P(A)=\frac{n^r}{(m+n)^r}=\left(\frac{n}{m+n}\right)^r$$

(2) 由于鱼在取出检查后不再放回,且我们把取出 r 尾鱼作一组而不考虑它们之间的排列次序,因此可以用组合公式计算基本事件的总数,即

总的基本事件数 $=C_{m+n}^r$

同样,属于 A 的基本事件数 $=C_n^r$

故要求的概率为

$$P(A)=\frac{C_n^r}{C_{m+n}^r}=\frac{n(n-1)\cdot\cdots\cdot(n-r+1)}{(m+n)(m+n-1)\cdot\cdots\cdot(m+n-r+1)}$$
$$=\frac{n}{m+n}\cdot\frac{n-1}{m+n-1}\cdot\cdots\cdot\frac{n-r+1}{m+n-r+1}$$

可以看出,取样方法不同时,所算得的概率也不同,即

$$\left(\frac{n}{m+n}\right)^r\neq\frac{n}{m+n}\cdot\frac{n-1}{m+n-1}\cdot\cdots\cdot\frac{n-r+1}{m+n-r+1}$$

但若 $m+n$ 很大,r 相对于 $m+n$ 又很小,则

$$\frac{n-k}{m+n-k}\approx\frac{n}{m+n},\ k=1,2,\cdots,r-1$$

此时,近似地有

$$\left(\frac{n}{m+n}\right)^r=\frac{n}{m+n}\cdot\frac{n-1}{m+n-1}\cdot\cdots\cdot\frac{n-r+1}{m+n-r+1}$$

也就是说,此时回置抽样和不回置抽样在本问题中可以近似地看成是一样的.

二、概率的一般定义

古典概型有很大的局限性,它要求基本事件的总数是有限的,而且每个基本事件都是等可能的. 但是,在生物学中所遇到的随机现象却往往并非如此,例如,初生婴儿的性别是个随机现象,它的样本空间只含有两个基本事件:"男"、"女".

但据各国人口统计资料表明，男婴的频率稳定于 22/43 左右，即，我们可以认为“婴儿是男的”这一基本事件的概率接近于 22/43，而不是 1/2. 换言之，这两个基本事件不是等可能的. 又如对某种生物的定量性状进行测定时，所得到的样本点的数目，通常（就理论上说）是无穷的，同样无法把这种情况纳入古典概型. 所以，我们有必要根据古典概型所提供的有关事件及概率的一些特征，并联系到频率与概率之间的关系，加以抽象，概括出一般的概率的定义.

定义 1.3.1 设随机试验 E 的样本空间为 S，对于 E 的每一事件 A（即 S 的每一个子集），赋予一个实数，记为 $P(A)$，如果 $P(A)$ 具有下列性质：

1° 对于任一事件 A，有 $0 \leqslant P(A) \leqslant 1$；

2° $P(S) = 1$；

3° 对于两两不相容的事件 $A_k (k = 1, 2, \cdots)$，有

$$P\left(\sum_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \tag{1.3.6}$$

则称 $P(A)$ 是 A 的概率.

从以上概率的定义我们可以导出下列结论（推导略）：

1° 不可能事件的概率为 0，即 $P(\varnothing) = 0$；

2° 概率的有限可加性：如果事件 $A_1, A_2, \cdots, A_n$ 两两互不相容，则

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i);$$

3° 对于任两事件 A 及 B，有

$$P(A \cup B) = P(A) + P(B) - P(AB) \tag{1.3.7}$$

4° $P(\overline{A}) = 1 - P(A)$

三、小概率原理

我们已经知道，概率 $P(A)$ 是用来衡量一次试验中事件 A 出现的可能性大小的数量指标. 我们还知道，不可能事件的概率为 0，容易设想如果一个事件的概率接近于 0（这种事件，称之为小概率事件），那么，我们认为它在一次试验中实际上是不可能出现的，这就是在统计推断中起着重要作用的“小概率原理”.

一个事件的概率小到怎样的程度才能算作小概率事件，并没有绝对的标准，要根据实际情况和需要决定. 国际上通常取 5% 或 1%（有时也取 10%）作为小概率的上界，即只要一个事件的概率小于 0.05 或 0.01，我们便可认为这个事件在一次试验中实际上是不可能发生（或出现）的.

§ 1.4 事件的独立性

一、条件概率

先看一个简单的例子，设某池中有 10 尾鱼，其中成鱼及幼鱼各 5 尾，成鱼中有 1 尾患病，幼鱼中有 3 尾患病. 现在从该池中任取 1 尾进行检查，以 A 表示“抽到幼鱼”，以 B 表示“抽到病鱼”，则显然有

$$P(A)=\frac{5}{10},\ P(B)=\frac{4}{10}$$

现在，我们来考虑事件“在幼鱼中抽到病鱼”. 因为共有 5 条幼鱼而其中有 3 条是病鱼，因此，这个事件的概率应当是 3/5.

让我们进一步来剖析这个例子. 当我们说：“在幼鱼中任取一条发现是病鱼”时，我们实际上已改变了随机试验的条件组，详细地说，我们是在原有的随机试验的条件组（从有 10 尾具有上述特征的鱼中任取 1 条）中，添加了一个新的条件“在幼鱼中”，由于这个新的条件的添加，原来的样本空间被改变了，从而，得出了另一个概率，由于后者是在原有的试验中添加条件后所获得的，所以我们称之为条件概率，确切地说，是在事件 A 出现的条件下，事件 B 出现的概率，称作事件 B 对于事件 A 的条件概率，记为 $P(B/A)$.

在这个例子中，$P(B/A)=\frac{3}{5}$.

我们再来计算 $P(AB)$，此处，事件 AB 的含义是：抽到的一尾鱼既是幼鱼又是病鱼，所以 $P(AB)=\frac{3}{10}$，容易看出

$$P(B/A)=\frac{P(AB)}{P(A)}$$

这个结果在古典概型中是普遍成立的. 事实上，设样本空间共有 n 个样本点，其中属于 A 的有 m 个；属于 B 的有 k 个，属于 AB 的有 r 个（见图 1.4.1），则

$$P(AB)=\frac{r}{n},\ P(A)=\frac{m}{n}$$

当在原有的条件组中添加了“事件 A 已经发生”这个条件后，基本事件只剩 m 个，在这 m 个基本事件中，属于 B 的有 r 个，因此

$$P(B/A)=\frac{r}{m}=\frac{r/n}{m/n}=\frac{P(AB)}{P(A)}$$

注意以上结果是在古典概型的前提下导出的，对于一般的情况，我们给出下

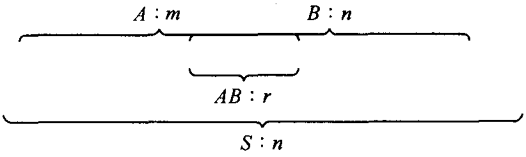

图 1.4.1

列定义：

定义 1.4.1　设 $P(A)>0$，则称分数

$$\frac{P(AB)}{P(A)}$$

为事件 B 对于事件 A 的条件概率，记为 $P(B/A)$，即

$$P(B/A)=\frac{P(AB)}{P(A)} \tag{1.4.1}$$

上式移项即得

$$P(AB)=P(B/A)P(A) \tag{1.4.2}$$

这个公式叫做概率的乘法公式.

二、事件的独立性

设考虑两个事件 A 与 B，在前面已经说过，A 的发生对于 B 的发生可能提供某种信息，但也可能会有这种情况，A 的发生对于 B 的发生不提供任何信息，此时，应当有

$$P(B/A)=P(B)$$

则(1.4.2)式的乘法公式成为

$$P(AB)=P(A)P(B)$$

定义 1.4.2　对于事件 A,B，如果等式

$$P(AB)=P(A)P(B) \tag{1.4.3}$$

成立，则说事件 A 与事件 B 独立.

可以证明，如果事件 A 与 B 独立，则事件 A 与 $\overline{B}$，$\overline{A}$ 与 B，$\overline{A}$ 与 $\overline{B}$ 也都相互独立.

例 1.4.1　罐子里有红球 r 个，黑球 b 个，混合均匀后，任取一个. 然后采用放回和不放回方式，再取第二个，求当第一次取到红球(事件 A)时，第二次也取到红球(事件 B)的概率及两次都取到红球的概率.

解：这里有两种抽样方式，回置抽样和不回置抽样，以下分别对这两种抽样方式进行讨论.

(1) 回置抽样　此时，第一次取球的结果对第二次取到红球的概率不影

响，即 A,B 两事件是独立的，于是有

$$P(B/A)=P(B)=P(A)=\frac{r}{r+b}$$

$$P(AB)=P(A)P(B)=\left(\frac{r}{r+b}\right)^2$$

(2) 不回置抽样　因为第一次取出的红球不放回，当第二次抽取时罐中只有 $r+b-1$ 个球，其中红球有 $r-1$ 个，第二次取到红球的概率是

$$P(B/A)=\frac{r-1}{r+b-1}$$

两次都取到红球的概率是

$$P(AB)=P(A)P(B/A)=\frac{r}{r+b}\cdot\frac{r-1}{r+b-1}$$

本例实际上是例 1.3.1 的特殊情况.

以上讲的是两个事件的独立性，对于 n 个事件，则可定义它们的独立性如下：

定义 1.4.3　设有 n 个事件 $A_1,A_2,\cdots,A_n$，如果对任何 $k=2,3,\cdots,n$ 均有

$$P(A_{i_1}\cdots A_{i_k})=P(A_{i_1})\cdots P(A_{i_k})$$

则称 $A_1,A_2,\cdots,A_n$ 相互独立.

显然，n 个相互独立的事件一定是两两独立的，而两两独立的事件不一定是相互独立的.

例 1.4.2　对菌种进行诱变处理时，优良菌株出现的概率一般很低，在工作中我们怎样把这些优良菌株筛选出来呢？如果对成千上万个变异个体一一进行鉴定，显然是不可能的，客观上只允许取其中一小部分进行鉴定，现在的问题是怎样以较大的把握，较小的工作量找出优良菌株. 如果经过诱变处理后，优良菌株的突变率为 $p=0.05$，现欲以 95% 的把握至少得到一个优良突变菌株，应该取几个菌株来培养鉴定呢？

挑菌株属不回置抽样，但因菌株数目甚大，故可看作回置抽样（参看例 1.3.1），设应该取 n 个菌株才有 95% 的把握至少找到一个优良突变菌株，以 A 表示“n 个中至少有一个是优良菌株”这一事件，则

$$\begin{aligned}P(A)&=1-P(\overline{A})=1-P(\overline{A}_1\overline{A}_2\cdots\overline{A}_n)\\&=1-P(\overline{A}_1)P(\overline{A}_2)\cdots P(\overline{A}_n)\end{aligned}$$

因为 $P(\overline{A}_i)=1-P(A_i)=1-0.05$，又，要求 $P(A)=0.95$，所以 $0.95=1-(1-0.05)^n$，故

$$0.95^n=0.05$$

两边取对数得