## **5.4 无替换抽样**

考虑一组标记为$1, 2 \ldots , n$的$n$个个体。 在没有替换的情况下随机抽取的结果是所有元素的随机排列。 当您尝试评估两个样本是否来自相同的基础分布时，您在 Data8 中使用了随机排列。

让我们来思考$(X\_1, X\_2, \ldots , X\_n)$的排列。对于任何 $i\_1, i\_2, \ldots , i\_n$ 的1到n的排列，$$ P(X\_1 = i\_1, X\_2 = i\_2, \ldots, X\_n = i\_n) = \frac{1}{n!} $$

注意，右边并不只是依赖于左边这种排列，参数$X\_1, X\_2, \ldots , X\_n$是可变的。

**相似性**

对每个固定的i，第i个参数Xi是一个1到n之间的整数。为了找到Xi的边缘分布，我们需要找到从1到n的每个k对应的P(Xi=k)值，由于所有排列具有相同的可能性，即

P(Xi=k)=(n−1)!n!=1n

使用一种现在很普遍的方法，把k放在i的位置上，让剩下n-1个元素随意排列，对于每个i而言，从1到n的Xi的分布都是一样的。

对于任意2个参数i和j而言，

P(Xi=k,Xj=l)=1n⋅1n−1,  1≤k≠l≤n

同样的，右边概率的取值并不需要左边的i和j取特定值。

我们早在匹配问题中就见过这种概率值的问题。在那个问题中，我们找到了匹配后的概率值，比如P(Xi=i,Xj=j)。但答案并不取决于i和j的取值，只有在考虑两个方向时他们的值才会起作用。同理，在我们当下研究的这个问题中也是一样的。

**范例：洗牌整齐的牌堆**

假设洗完了一副标准扑克牌，所有排列组合的可能性都是一样的。

问题一：第17张牌是A的可能性有多大？

回答一.根据我们上面的计算，第17张牌可能是52张牌中的任意一张。里面有四张A，所有概率值为4/52。

同理，第1张牌或者第32张牌是A的可能性都为4/52。根据相似性，所有这些非条件性的边缘分布的概率值都是相等的。如果这听起来很神秘的话，想象一下把发好的牌列首尾相连，你就挑不出哪张是第1张牌，哪张是第17张牌了。

问题二：如果第32张牌是A，那么第17张牌是A的可能性有多大？

回答2.根据我们上面计算的Xi和Xj的联合分布，答案是3/51。如果换成第2张牌是A，答案相同。

**简单随机样本**

简单随机样本是随机抽取的样本，不需要从有限总体中进行替换。样本是总体的一个随机子集，而不是总体的重新排布。如果你从52张标准扑克牌中抽取一份5张的随机样本，那么那么你拿到的就是子集的五张牌。这五张牌可以按任何顺序出现在你手里，但顺序无关紧要。重要的是这五张牌的集合。

为了找出你手上拿到特殊子集的五张牌的概率，首先得数出这五张牌如果不考虑顺序会有多少种组合数。

这五张牌总共有52×51×50×49×48种顺序。

为了得到这5张牌的图书组合，把其中一张放在位置1；第1张牌可以有5个选择。然后再挑1张放在位置4，以此类推。

所以出现特定牌组的概率是$$ \frac{5 \times 4 \times 3 \times 2 \times 1}{52 \times 51 \times 50 \times 49 \times 48} = \frac{5! 47!}{52!} = \frac{1}{\binom{52}{5}} $$。

这表明，在没有替换的情况下随机一张接一张地发5张牌， 相当于洗牌和抽出5张牌。

在scipy里的misc模块允许你计算出这些组合项。

from scipy import misc

misc.comb(52, 5)

2598960.0

有将近260万种组合。那是一个很大的数字。这有助于我们建立一个理论来理解这个问题以及其他的简单随机样本。在下个部分，我们会建立这个理论。在计算出从总体中抽出的简单随机样本的数量后，我们会结束这个部分。

假设有一个大小为$N$的总体（一个确定整数，而非一个随机变量），你想从中抽取一个大小为$n \le N$的简单随机样本。你能得到多少不同样本？

我们假设那个“样本”是$n$个个体的集合，可以以任何顺序出现。这就和扑克牌很像了。

**简单随机样本的数量**

一个类似的论点告诉我们简单随机样本的数量$$ \binom{N}{n} $$，他们出现的可能性相同。

**在一个简单随机样本中计算出好的元素**

如果一个整体由两类个体组成，这两类按照惯例会被称为“成功或失败”或者“好和坏”。在这里“好”几乎总是代表着你在计算的个体的种类。例如，如果你在尝试数出在一场选举中一个特定的候选人有多少支持者，无论你抱有何种政治观点，那一类的支持者都会被标记成“好”。

假设一个总体为$N$的个体包含$G$个好的个体，你从中抽取一个大小为$n$的简单随机样本。有多少样本会包含$g$个好的元素？

包含$g$个好个体的样本数量可以通过乘法法则计算得出：

在整体中，从$G$个好个体中选出$g$个好个体。你有$\binom{G}{g}$种做法。

对于这些有$g$个好个体的每个选项而言，你总共有$\binom{N-G}{n-g}$个坏个体的选项。

所以包含$g$个好个体的样本总数是$$ \binom{G}{g}\binom{N-G}{n-g} $$个。

这些被称为超几何分布，因为这些公式和数学中的超几何序列有关。我们在这个课程中不会解决相关序列的问题，但我仍然可以使用这个令人印象深刻的名词。我们会在稍后的课程中涉及更多的相关的概率的问题。

**技术笔记**：如果你很谨慎的话，你需要通过尝试找出哪些价值为$g$来开始，这是应该在这里被考虑到的事情。因为这是在样本中好的元素的个数，我们知道$g \le \min(n, G)$。通过考虑样本中坏的元素的个数，我们知道$n-g \le \min(n, N-G)$，因此$g \ge \max(0, n-N+G)$。

但你其实不用担心这些技术细节。在计算不可能选项时，仅仅只需定义$\binom{a}{b}$为0。比如$\binom{5}{10}$或者$\binom{6}{-4}$。如果在样本中得到$g$个好的元素是不可能的，那么用来计算概率为$g$的好的元素的超几何的算式会得到结果为0。