详细解答

STEP III 2001，第 12 题

概率与统计 · 以指示变量求期望值

STEP III 2001，第 12 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

step.maths.org/questions/01-s3-q12

关键思路

在第 (i) 部分，不要尝试求出黑球数目的分布：把这个数目拆开，每个黑球对应一个指示变量（indicator），然后运用期望值的线性性质（linearity of expectation）。不放回地抽球直至袋空，其实就是把所有球随机排成一列，所以每个指示变量都归结为一个对称性论证（symmetry argument）。在第 (ii) 部分，两次抽到白球之间袋中的球并无改变，所以等候每个白球出现的次数服从几何分布（geometric distribution）；而题目中的“hence”，则来自这个观察：取走一个白球之后，我们回到同样的处境，只是白球少了一个。

解答

第 (i) 部分

设想一直抽下去，直至袋空为止。由于每次都是随机抽取，而且抽出的球不放回，这 $b+w$ 个球被抽出的次序，就是它们的一个均匀随机排列（uniformly random ordering）。到最后一个白球被取走时已抽出的黑球，正好就是在这个排列中位于最后一个白球之前的黑球。

把黑球编号为 $1, 2, \ldots, b$；若黑球 j 在最后一个白球之前被抽出，令 $I_j = 1$，否则令 $I_j = 0$。设 N 为最后一个白球被取走时已抽出的黑球数目，则

$$N = I_1 + I_2 + \cdots + I_b, \qquad \text{所以} \qquad \mathrm{E}(N) = \sum_{j=1}^{b} \mathrm{P}(I_j = 1)$$

这是根据期望值的线性性质（它并不需要各个 I_j 之间互相独立）。

想一想 黑球 j 排在所有白球之后的概率是多少?

只看由黑球 j 和 w 个白球组成的这 $w+1$ 个球。在所有球的均匀随机排列中，这 $w+1$ 个球的相对次序同样是均匀随机的，所以黑球 j 处于它们当中 $w+1$ 个位置的任何一个，机会均等。它排在这组球最后的概率是 $\frac{1}{w+1}$，而这正是事件 $I_j = 0$。因此

$$\mathrm{P}(I_j = 1) = 1 - \frac{1}{w+1} = \frac{w}{w+1},$$

从而

$$\boxed{\mathrm{E}(N) = \frac{bw}{w+1}.}$$

第 (ii) 部分

现在黑球会放回袋中，所以在抽到白球之前，袋中始终有 b 个黑球和 w 个白球。因此每次抽到黑球的概率是 $p = \frac{b}{b+w}$，抽到白球的概率是 $q = \frac{w}{b+w}$，而且与之前各次抽取独立，因为每次袋中的球都相同。

设 X 为抽到第一个白球之前抽到黑球的次数。那么 $X = n$ 表示先抽到 n 次黑球，然后抽到一个白球，所以

$$\mathrm{P}(X = n) = p^n q, \qquad n = 0, 1, 2, \ldots$$

若 $b = 0$，则 $X = 0$，结果显然成立，故取 $b \geq 1$；这时 $0 < p < 1$，利用当 $|x| < 1$ 时 $\sum_{n\geq 1} n x^{n-1} = \frac{1}{(1-x)^2}$（把等比级数（geometric series）求导即得），

$$\mathrm{E}(X) = \sum_{n=0}^{\infty} n p^n q = pq \sum_{n=1}^{\infty} n p^{n-1} = \frac{pq}{(1-p)^2} = \frac{p}{q} = \frac{b/(b+w)}{w/(b+w)} = \frac{b}{w},$$

因为 $1 - p = q$。这就是所要证明的结果。

至于最后一个白球，把抽到黑球的总次数写成 $X_w + X_{w-1} + \cdots + X_1$，其中 X_k 是袋中恰好有 k 个白球期间抽到黑球的次数。当袋中尚余 k 个白球时，袋中有 b 个黑球和 k 个白球，所以这个阶段正是把 w 换成 k 的“第一个白球”问题，而刚才证明的结果给出 $\mathrm{E}(X_k) = \frac{b}{k}$。根据期望值的线性性质，最后一个白球被取走时抽到黑球的期望次数是

$$\boxed{\sum_{k=1}^{w} \frac{b}{k} = b\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{w}\right).}$$

验算

当 $w = 1$ 时，两部分的结果都很容易直接看出。在 (i) 中，每一个黑球排在白球之前的概率都是 $\frac{1}{2}$，得出 $\frac{b}{2}$，与 $\frac{bw}{w+1}$ 相符。在 (ii) 中，只有一个阶段，而总和就是 $\frac{b}{1} = b$。在 (i) 中当 $w \to \infty$ 时，几乎每个黑球都在最后一个白球之前被抽出，而确实有 $\frac{bw}{w+1} \to b$。

我们对所有 $b \le 4$、$w \le 4$ 的情况逐一列出每个排列来验算 (i)；当 $b = 3$、$w = 4$ 时，结果恰好是 $\frac{12}{5}$。至于 (ii)，以 $b = 3$、$w = 4$ 进行 2×10^5 次模拟，所得平均值约为 6.245，而 $3(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}) = \frac{25}{4} = 6.25$。

常见错误

- 在 (i) 中，假设各个黑球是否被抽出是互相独立的。期望值的线性性质并不需要独立性，这正是指示变量如此好用的原因。
- 在 (ii) 中，采用了把抽到白球那一次也计算在内的几何分布，结果得出 $\frac{1}{q} = \frac{b+w}{w}$，比答案多了一。
- 在最后一部分的每个阶段都以 $b + w$ 作分母：袋中白球的数目每过一个阶段就减少一个。

STEP 之外

第二部分其实是乔装了的优惠券收集问题（coupon collector's problem）：把各个阶段的几何等候时间加起来，便得到一个调和和（harmonic sum），其增长速度与 $\ln w$ 相若。像第 (i) 部分那样把一个计数写成指示变量之和，是概率论中最强而有力的技巧之一；日后计算随机置换（random permutation）中不动点（fixed point）、纪录（record）和循环（cycle）的期望数目时，你会再次遇上它。Blitzstein 与 Hwang 合著的 *Introduction to Probability*（概率论入门）中，有许多关于这两个想法的例子。

Beyond STEP 详细解答：https://beyondstep.keithmcnulty.org/zh-hans/。解答文字采用 CC0 1.0；题目本身版权属 Cambridge University Press & Assessment 所有（©），本文件不予转载。由机器从英文翻译而成；尚未经审阅。