详细解答

STEP I 1996，第 12 题

概率与统计 • 随机样本的最大值

STEP I 1996，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/96-s1-q12

关键思路

阅卷员的选择，归根究底就是从 $\{1,2,\ldots,m\}$ 中选出一个由 n 个不同分数组成的集合，而每一个这样的集合出现的机会均等。因此，概率（probability）就是集合的数目除以 $\binom{m}{n}$。说最高分数是 k，即是固定了其中一个分数，余下的 $n-1$ 个则要从比它小的分数中选出。恒等式（identity）随之唾手可得，因为各概率之和为 1。至于期望值（expectation），窍门是把 $k\binom{k-1}{n-1}$ 还原成单独一个二项式系数（binomial coefficient），这样便可以再次使用同一条恒等式。

解答

K 的分布

阅卷员给出 n 个互不相同的分数，而根据对称性，$\{1,\ldots,m\}$ 中每一个大小为 n 的子集（subset），成为所用分数集合的机会均等。（哪一份答卷得到哪一个分数，对 K 并无影响。）这样的子集共有 $\binom{m}{n}$ 个。

事件 $K=k$ 发生，当且仅当该集合包含 k，而其余 $n-1$ 个分数全部落在 $\{1,\ldots,k-1\}$ 之内。选出这 $n-1$ 个分数的方法有 $\binom{k-1}{n-1}$ 种。这需要 $k-1\geq n-1$，即 $k\geq n$：n 个互不相同的正整数之中，最大的一个至少是 n。此外，当然还有 $k\leq m$。



（图中所示为 $m=10$、$n=4$、$k=7$ 的情况。）因此

$$\mathrm{P}(K=k)=\binom{k-1}{n-1}\Big/\binom{m}{n},\qquad n\le k\le m.$$

恒等式

K 必定恰好取 $n, n+1, \ldots, m$ 其中一个值，所以这些概率之和为 1：

$$\sum_{k=n}^{m}\binom{k-1}{n-1}\Big/\binom{m}{n}=1\quad\Longrightarrow\quad\sum_{k=n}^{m}\binom{k-1}{n-1}=\binom{m}{n}.$$

这对所有满足 $1\le n\le m$ 的整数都成立，下一步正要用到这一点。

期望值

根据定义，

$$\mathrm{E}(K)=\frac{1}{\binom{m}{n}}\sum_{k=n}^{m}k\binom{k-1}{n-1}.$$

> **想一想** 怎样才能消去因子 k，使恒等式可以再次派上用场？

把它并入二项式系数之中：

$$k\binom{k-1}{n-1}=\frac{k\,(k-1)!}{(n-1)!\,(k-n)!}=n\cdot\frac{k!}{n!\,(k-n)!}=n\binom{k}{n}.$$

现在 $\binom{k}{n}=\binom{(k+1)-1}{(n+1)-1}$。当 k 由 n 取到 m 时，$j=k+1$ 由 $n+1$ 取到 $m+1$。于是，以 $n+1$ 和 $m+1$ 代替 n 和 m 来使用恒等式（这是容许的，因为 $1\le n+1\le m+1$），便得

$$\sum_{k=n}^{m}\binom{k}{n}=\sum_{j=n+1}^{m+1}\binom{j-1}{(n+1)-1}=\binom{m+1}{n+1}.$$

所以

$$\begin{aligned}\mathrm{E}(K)&=\frac{n\binom{m+1}{n+1}}{\binom{m}{n}}=n\cdot\frac{(m+1)!}{(n+1)!\,(m-n)!}\cdot\frac{n!\,(m-n)!}{m!}\\&=\frac{n(m+1)}{n+1}.\end{aligned}$$

验算

- $n=m$：每个分数都用上了，所以必定有 $K=m$，而确实 $\frac{m(m+1)}{m+1}=m$。
- $n=1$：K 是单独一个分数，在 $\{1,\ldots,m\}$ 上均匀分布，平均值为 $\frac{m+1}{2}$，与公式吻合。

- $m=6$、$n=2$: 列出 $\{1,\ldots,6\}$ 中全部 15 对, 得 $\mathrm{E}(K)=\frac{70}{15}=\frac{14}{3}$, 而公式给出 $\frac{2\cdot 7}{3}=\frac{14}{3}$。我们已就所有 $1\le n\le m\le 9$ 逐一列出每个子集，验证了分布、恒等式和平均值。
- 有一个漂亮的方法可以看出答案：选出的 n 个分数把余下的 $m-n$ 个分数分成 $n+1$ 个间隔，各间隔的期望大小相等，均为 $\frac{m-n}{n+1}$。于是 $\mathrm{E}(K)=m-\frac{m-n}{n+1}=\frac{n(m+1)}{n+1}$。

常见错误

- 分子数的是把分数分配给答卷的有序方式，分母数的却是无序的集合（或者反过来）。只要前后一致，两种数法都可行，因为因子 $n!$ 会互相抵消。
- 用了 $\binom{k}{n}$ 而不是 $\binom{k-1}{n-1}$。前者数的是所有分数都*不超过* k 的集合，而不是最高分数*等于* k 的集合。

STEP 之外

这条恒等式就是帕斯卡三角形（Pascal's triangle）中的“曲棍球棒”恒等式（hockey-stick identity），而我们在这里是用两种方法数同一样东西来证明它，这是组合数学中的标准技巧。K 是不放回抽取的随机样本中的最大值，属于*顺序统计量*（order statistic）。“德国坦克问题”（German tank problem）把本题倒过来，由观察所得的 K 去估计 m：从 $\mathrm{E}(K)=\frac{n(m+1)}{n+1}$ 解出 m，便得到无偏估计量（unbiased estimator）$\frac{n+1}{n}K-1$。想进一步了解顺序统计量和计数论证，可参阅 Blitzstein 与 Hwang 的 *Introduction to Probability*（概率论入门教材）。

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。