arXiv:2304.04249v1 [math.ST] 9 Apr 2023

存在观测值缺失时空间均值方差的收敛估计量

Ashwin K Seshadri

印度科学研究所大气与海洋科学中心及 Divecha 气候变化中心，印度班加罗尔 560012，电子邮箱：ashwins@iisc.ac.in

摘要

在地球科学中，一个反复出现的问题是使用点观测的加权平均值来估计物理场的空间均值。一个重要的变体是，个别观测值以小于一的某个概率被计入。这可能出现在不同的情境中：从数据缺失到估计各子样本中的统计量。在这种情- 况下，空间均值是随机变量的比值，其统计量涉及通过级数展开推导出的近似估计量。本文考察了存在缺失数据时空间均值方差的截断估-计量及其一般结构。在所有阶数下，方差估计量仅取决于基础场的一阶矩和二阶矩，而收敛要求这些矩为有限值。此外，当计入各个观测值的概率大于1/2 或点观测的数量很大时，都会发生收敛。在点观测采用统一加权的情况下，可以利用组合学轻松得到这些估计量，并且其中涉及第二类斯特林数。

1 引言

在地球科学中，我们经常寻求某些物理变量（如降雨量或温度）的空间平均值的估计（*Rodríguez-Iturbe* 和 *Mejía* （1974）；*Shen*等 （1994）；*Morrissey* 等（1999）；*Kundu* 和 *Siddani* （2007）；*Villarini* 等 （2008）；*Prakash*等（2019））。考察此类平均值的性质是一个有数十年历史的问题（$Kagan$ （1997））。一种有趣的变体是：单个观测值以小于一的已知概率被报告。在这种情况下，空间平均值的偏差和方差的估计量采取何种形式? $Seshadri$（2018）研究了这样一种情形，其中涉及位于 i （$1 \leq i \leq N$）处的点观测 r_i，每个观测值被赋予权重 β_i，使得 $\sum_i \beta_i = 1$。假设时间是离散的，并且在每个时刻并非所有观测值都会被报告。报告状态由布尔随机变量 s_i 表示：若该观测值在该时刻被报告，则 $s_i = 1$；否则为 $s_i = 0$。在我们的设定中，假设每个观测站点的报告状态与其他任何站点（包括邻近站点）是否被报告相互独立。这使得 s_i's 统计独立。假设每个 s_i 取一的概率为 α，我们定义不放回抽样样本上的空间平均值

$$r\left(t\right) \equiv \frac{R\left(t\right)}{S\left(t\right)} = \frac{\sum_i \beta_i s_i\left(t\right) r_i\left(t\right)}{\sum_i \beta_i s_i\left(t\right)}, \tag{1}$$

其中描述了仅对已报告的那些观测值计算的加权均值。权重β_i 在时间上是恒定的，但 s_i 和 r_i 会随时间变化。为使表达清晰，以下公式中省略时间依赖性，并将空间平均值记为 $r = f\left(R, S\right) = R/S$。一般而言，只要 $f\left(R, S\right)$ 在 $f\left(\mathbf{E}R, \mathbf{E}S\right)$, 的某个邻域内是解析的，其中 $\mathbf{E}$ 表示期望，就可以在该邻域内对 f 展开泰勒级数，并且该级数是收敛的。上述比值描述了基于可用观测值的空间均值。集合 $\{s_i$ 的不同实现：$i = 1, 2, \ldots N\}$ 会产生均值的不同估计，而对所有可能实现进行统计通常也很有意义。该问题的核心在于，s_i 是一个随机变量，而计数各个观测值时随机性的来源可能是多种多样的。一般而言，这种 s_i 的随机性会引发两类影响。各个观测值可能以非零概率缺失或未被报告，在这种情况下，每个时刻的空间平均值仅涉及

当时报告的那些观测值（*Vinnikov* 等人(2004)），人们可能会关注所报告空间平均值的统计特性，例如其时间方差（$Seshadri$ (2018)）。或者，随机变量 s_i 可能涉及有意省略观测值，以估计不放回子样本的性质（$Prakash\ et\ al.$(2019)）。在第一种情况下，方差估计量会将缺失数据对方差造成的额外影响计算在内。在第二种情况下，它会提供一个非参数估计，用以衡量由不同子样本组成的集合的离散程度。不过，从形式上看，这两个问题是相同的。

因此，尽管式（1）的公式很简单，甚至可能简单得出人意料，但其统计量却以多种形式反复出现。此外，如果将 s_i 固定为 1，标准的偏差和方差公式便会重新出现（$Gandin$ （1993）；$Kagan$ （1997））。然而，有趣的情况源于这些变量的随机性：此时分母也是一个随机变量，因此无法得到式（1）统计量的精确公式。正如 $Seshadri$ （2018）所示，沿袭更早且研究范围更广的文献（$Hartley\ and\ Ross$ （1954）；$Goodman\ and\ Hartley$ （1958）；$Hinkley$ （1969）；$Kendall\ et\ al.$ （1994）），仍然可以推导出实用的近似公式。尽管这些近似公式基于低阶截断，但与蒙特卡洛模拟的基准结果相比，其表现依然相当出色（$Seshadri$ （2018）；$Prakash\ et\ al.$ （2019））。因此，有必要考察由幂级数推导出的这些估计量的收敛性（$Oehlert$ （1992）；这正是本文的目标。

我们的基本方法在概念上很直接，即对函数的截断泰勒展开取期望（$Oehlert$ (1992)）。逐步提高近似阶数，就会得到相应的估计量。该方法的主要问题在于：随着逐步纳入泰勒级数中的更多项，这些估计量是否收敛，以及在什么条件下收敛。其基本理论众所周知，并且有两个要求。首先，泰勒级数本身必须收敛（$Bromwich\ and\ MacRobert$ (1991)）。其收敛可能是条件性的，即仅在距离展开中心较近时成立；也可能是无条件的，即对底层场的所有取值都成立。在更一般的问题中，采用这种方法推导近似估计量时，两种情况都有可能；相关例子不仅见于空间地球科学，也见于统计学和经济学的各个应用领域（$Loistl$ (1976)；$Hlawitschka$ (1994)；

Markowitz (2015))。一般而言，泰勒级数仅在有限的半径内收敛（*Bromwich andMacRobert* (1991)；*Gemignani* (1970)；*Hlawitschka* (1994)）。在此收敛半径内，还需进一步考察期望是否也收敛。这由控制收敛定理决定（*Weir* (1973)）。若泰勒级数绝对收敛，则满足控制收敛的条件，此时期望也收敛。这将是检验我们所构造估计量的相关标准，因为这些估计量涉及对描述式（1）中空间均值方差的级数取期望。

2 理论与推导

2.1 收敛

按照*Seshadri* (2018)，我们考虑空间变化场在由 i (1 ≤ i ≤ N) 编号的位置上的点观测 r_i。每个观测值以恒定概率 $1-\alpha$ 缺失，其中 $0<\alpha\leq 1$。每个观测值的状态由布尔随机变量 s_i 描述：如果观测值已报告，则 $s_i=1$ 否则 $s_i=0$。空间平均值定义于式 (1)，本文旨在推导其方差公式。由于分子和分母都是随机变量，因此不存在精确表达式，我们必须将 $f(R,S)\equiv R/S$ 在 (**E***R*,**E***S*) 处展开为泰勒级数，其中 **E** 表示期望。

$$f\left(R,S\right)\quad=\quad f\left(\mathbf{E}R,\mathbf{E}S\right)\quad+\quad\sum_{n=1}^{\infty}\frac{1}{n!}\sum_{k=0}^{n}\left(\begin{array}{c}n\\k\end{array}\right)\frac{\partial^{n}f}{\partial R^{n-k}\partial S^{k}}\left(R-\mathbf{E}R\right)^{n-k}\left(S-\mathbf{E}S\right)^{k}\quad(2)$$

其中 $\left(\begin{array}{c}n\\k\end{array}\right)=\frac{n}{n-k!\quad()!\,k!}$。这产生了级数展开式

$$f\left(R,S\right)=\sum_{n=0}^{\infty}\left(-1\right)^{n}\left(\frac{1}{\mathbf{E}S}\right)^{n+1}\left(R-\mathbf{E}R\right)\left(S-\mathbf{E}S\right)^{n}+\sum_{n=0}^{\infty}\left(-1\right)^{n}\frac{\mathbf{E}R}{\left(\mathbf{E}S\right)^{n+1}}\left(S-\mathbf{E}S\right)^{n}\quad(3)$$

如补充信息（SI）第 1 节所示。将此级数写作

$$f\left(R,S\right)=\sum_{n=0}^{\infty}a_n+\sum_{n=0}^{\infty}b_n\equiv\sum_{n=0}^{\infty}c_n, \tag{4}$$

它的平方是该级数与自身的柯西乘积。

$$f^2\left(R,S\right)=\left(\sum_{n=0}^{\infty}c_n\right)\left(\sum_{n=0}^{\infty}c_n\right). \tag{5}$$

上述关于 $f^2\left(R,S\right)$ 的级数何时收敛？根据默滕斯定理，如果各个级数都收敛，且至少有一个级数绝对收敛，则柯西乘积收敛（$Rudin$ （1976））。因此，只要 $\sum_{n=0}^{\infty}c_n$ 绝对收敛即可。根据三角不等式

$$\sum_{n=0}^{\infty}\left|c_n\right|=\sum_{n=0}^{\infty}\left|a_n+b_n\right|\leq\sum_{n=0}^{\infty}\left|a_n\right|+\sum_{n=0}^{\infty}\left|b_n\right| \tag{6}$$

只需 $\sum_{n=0}^{\infty}a_n$ 和 $\sum_{n=0}^{\infty}b_n$ 绝对收敛，即可保证 $\sum_{n=0}^{\infty}c_n$ 绝对收敛。我们来评估这两个级数各自的绝对收敛性。先看第一个级数。

$$\sum_{n=0}^{\infty}\left|a_n\right|=\left|R-\mathbf{E}R\right|\sum_{n=0}^{\infty}\left(\frac{1}{\mathbf{E}S}\right)^{n+1}\left|S-\mathbf{E}S\right|^n, \tag{7}$$

其中我们利用了以下事实：$\mathbf{E}S>0$ ，因为 $\alpha>0$。其收敛的一个充分条件由比值判别法给出。我们要求该比值

$$\frac{\left|a_{n+1}\right|}{\left|a_n\right|}=\frac{\left|S-\mathbf{E}S\right|}{\mathbf{E}S}<1 \tag{8}$$

作为 $n\to\infty$。由于公式 (1) 中的权重 β_i 之和为1，期望为 $\mathbf{E}S=\sum_{i=1}^{N}\beta_i\mathbf{E}s_i=\sum_{i=1}^{N}\beta_i\alpha=\alpha$。此外，随机变量 S 受 $0<S\leq1$限制，其中 $S=0$ 被排除，因为这将意味着不存在任何报告的观测值。该界限确保在

$$\frac{1-\alpha}{\alpha}<1, \tag{9}$$

或 $\alpha>1/2$。

对于 $\alpha < 1/2$，在 $S \geq 2\mathbf{E}S$ 的情况下，无法保证收敛。该事件发生的概率可通过 Hoeffding 不等式进行界定。我们使用这一结果的一个版本，适用于相互独立但不满足同分布条件的有界随机变量。我们可以写作 $S = \sum_i x_i$，其中$x_i = \beta_i s_i$。各个 x_i 的取值均介于 m_i 和 M_i之间，且相互独立，因为 s_i 相互独立。Hoeffding 的结果表明，事件 $\sum_i (x_i - \mathbf{E}x_i) \geq c$发生的概率小于或等于 $e^{-2c^2/\sum_{i(M_i-m_i)^2}}$ （*Vershynin* （2018））。由于 $0 \leq s_i \leq 1$、$m_i = 0$ 和 $M_i = \beta_i$。此外，$\mathbf{E}x_i = \alpha\beta_i$，因此 $\sum_i \mathbf{E}x_i = \alpha \sum_i \beta_i = \alpha$。这一结果意味着，$S - \alpha \geq \alpha$ 发生的概率，或等价地说，$S \geq 2\mathbf{E}S$ 发生的概率，小于

$$e^{-2\alpha^2/\sum_i \beta_i^2}, \tag{10}$$

如果权重的分布更加均匀，使得 $\sum_i \beta_i^2$ 更小，则可以得到更严格的界。在均匀权重的极限情形下，$\beta_i = 1/N$和 $\sum_{i=1}^{N} \beta_i^2 = 1/N$，$S > 2\mathbf{E}S$ 的概率上界为 $e^{-2N\alpha^2}$。实际上，Hoeffding不等式相当保守，尤其是当 α 较小时，还可以得到更强的界；不过，这一计算确实说明，较大的 N 有利于收敛。

对于均匀权重，可以更直接地求得界，因为 S 服从均值为 α 、标准差为 $\sqrt{\alpha_{(1-\alpha)\ N}}$ 的二项分布。因此，S 的取值必定与均值相差 $\sqrt{N \frac{\alpha}{1-\alpha}}$ 个标准差；当 N 较大时，这种情况极不可能发生。例如，即使概率 $\alpha = 0.1$ 很小，若 $N = 100$ ，则 S 必须与均值相差 3.3 个标准差。实际上，在地球物理问题中，观测值的报告概率要大得多。类似地，对于式（6）中第二个级数的绝对收敛

$$\sum_{n=0}^{\infty} |b_n| = |\mathbf{E}R| \sum_{n=0}^{\infty} \frac{1}{(\mathbf{E}S)^{n+1}} |S - \mathbf{E}S|^n , \tag{11}$$

该条件

$$\frac{|b_{n+1}|}{|b_n|} = \frac{|S - \mathbf{E}S|}{\mathbf{E}S} < 1 \tag{12}$$

是相同的。总之，$\sum_{n=0}^{\infty} |c_n|$ 的绝对收敛得益于较大的 α 和 N，并且是

无论 N 的值为何，均可确保 $\alpha > 1/2$ 。

当 $\sum_{n=0}^{\infty} |c_n|$ 收敛时，式（5）中的柯西乘积不仅收敛，而且绝对收敛。因此，根据控制收敛定理，我们可以对 f^2 取期望，即$\mathbf{E}f^2$，相应的$\mathbf{E}f^2$近似也收敛。最后，由于

$$\sigma_f^2 = \mathbf{E}\ f^2 - (\mathbf{E}f)^2 \leq = \mathbf{E}\ f^2 \tag{13}$$

若$\mathbf{E}f^2$ 的级数展开收敛，则 σ_f^2 的级数展开也收敛。

因此，方差级数 σ_f^2 收敛的一个充分条件是 $\frac{1-\alpha}{\alpha} < 1$。在实践中，要么是由于缺失数据概率较小，使得 $1-\alpha < 1/2$；要么在缺少这一条件时，潜在观测值 N 数量较大，足以弥补，因此方差级数能够收敛。因此，我们可以在地球物理学缺失数据问题中使用基于矩的实用估计量。本节余下部分将致力于推导空间均值方差的低阶估计量。

2.2 方差级数

在这里，我们将推导缺失数据情况下方差的低阶估计量，以研究其结构。将级数截断至 2^{nd} 阶

$$f(R,S) = f(\mathbf{E}R, \mathbf{E}S) + \ldots + \frac{1}{2}(R-\mathbf{E}R)^2 \frac{\partial^2 f}{\partial R^2} + \frac{1}{2}(S-\mathbf{E}S)^2 \frac{\partial^2 f}{\partial S^2} + (R-\mathbf{E}R)(S-\mathbf{E}S)\frac{\partial^2 f}{\partial R \partial S}, \tag{14}$$

将偏导数在（$\mathbf{E}R$,$\mathbf{E}S$）处计算，我们得到方差

$$\begin{aligned}\sigma_f^2 == \mathbf{E}\ f^2 - (\mathbf{E}f)^2 = &-6\frac{(\mathbf{E}R)^2}{(\mathbf{E}S)^2} + \frac{4}{(\mathbf{E}S)^2}\mathbf{E}R^2 + 10\frac{(\mathbf{E}R)^2}{(\mathbf{E}S)^4}\mathbf{E}S^2 - \frac{1}{(\mathbf{E}S)^4}(\mathbf{E}RS)^2 - \frac{(\mathbf{E}R)^2}{(\mathbf{E}S)^6}\left(\mathbf{E}S^2\right)^2 \\ &+ 2\frac{(\mathbf{E}R)}{(\mathbf{E}S)^4}\mathbf{E}RS^2 - \frac{4}{(\mathbf{E}S)^3}\mathbf{E}R^2S - 6\frac{(\mathbf{E}R)^2}{(\mathbf{E}S)^5}\mathbf{E}S^3 + \frac{(\mathbf{E}R)^2}{(\mathbf{E}S)^6}\mathbf{E}S^4 + \frac{1}{(\mathbf{E}S)^4}\mathbf{E}R^2S^2 \\ &+ 2\frac{(\mathbf{E}R)(\mathbf{E}RS)}{(\mathbf{E}S)^3}\left\{\frac{\mathbf{E}S^2}{(\mathbf{E}S)^2} - 1\right\}.\end{aligned} \tag{15}$$

这些计算详见补充信息第2节。对这些统计量的评估通常需要$\mathbf{E}R^2S^l$、$\mathbf{E}RS^l$和$\mathbf{E}S^l$的公式，其中 l 是非负整数。

2.3 矩的计算

这些矩$\mathbf{E}R^2S^l$、$\mathbf{E}RS^l$和$\mathbf{E}S^l$ ($l = 0, 1, 2, ..$)可以直接计算，但需要作出进一步假设。我们假设不同位置的观测值报告彼此独立，因此$\mathbf{E}s_is_j = \mathbf{E}s_i\mathbf{E}s_j$ if $i \neq j$。我们还假定可用性独立于所测量的场，因此$\mathbf{E}r_is_i = \mathbf{E}r_i\mathbf{E}s_i$。然后，利用多项式定理，可得矩，汇总于表 1 中。推导过程详见补充信息第 3 节。

根据多项式定理

$$\mathbf{E}S^l = \mathbf{E}\left(\sum_{i=1}^{N}\beta_i s_i\right)^l = \mathbf{E}\sum_{l_1+l_2+...l_N=l}\frac{l!}{l_1!l_2!\cdots l_N!}\beta_1^{l_1}\beta_2^{l_2}\cdots\beta_N^{l_N}s_1^{l_1}s_2^{l_2}\cdots s_N^{l_N}, \tag{16}$$

根据期望的线性性质

$$\mathbf{E}S^l = \sum_{l_1+l_2+...l_N=l}\frac{l!}{l_1!l_2!\cdots l_N!}\beta_1^{l_1}\beta_2^{l_2}\cdots\beta_N^{l_N}\mathbf{E}s_1^{l_1}s_2^{l_2}\cdots s_N^{l_N}, \tag{17}$$

根据 s_1、 s_2等的独立性

$$\mathbf{E}S^l = \sum_{l_1+l_2+...l_N=l}\frac{l!}{l_1!l_2!\cdots l_N!}\beta_1^{l_1}\beta_2^{l_2}\cdots\beta_N^{l_N}\left(\mathbf{E}s_1^{l_1}\mathbf{E}s_2^{l_2}\cdots\mathbf{E}s_N^{l_N}\right). \tag{18}$$

现在，$\mathbf{E}s_i^{l_i} = \alpha$ 若 $l_i > 0$ 则为 1，否则为 0。若乘积 $\mathbf{E}s_1^{l_1}\mathbf{E}s_2^{l_2}\cdots\mathbf{E}s_N^{l_N}$中有 m 个不同的项具有非零幂，则该乘积变为 α^m。类似地

$$\mathbf{E}RS^l = \mathbf{E}\sum_{i=1}^{N}\beta_i r_i s_i\sum_{l_1+l_2+...l_N=l}\frac{l!}{l_1!l_2!\cdots l_N!}\beta_1^{l_1}\beta_2^{l_2}\cdots\beta_N^{l_N}s_1^{l_1}s_2^{l_2}\cdots s_N^{l_N} \tag{19}$$

根据 $\mathbf{E}$ 的线性性质以及 r_i 与 s_i之间的独立性

$$\mathbf{E}RS^l = \sum_{i=1}^{N} \beta_i \mathbf{E} r_i \left\{ \sum_{l_1+l_2+\ldots l_N=l} \frac{l!}{l_1!l_2!\cdots l_N!} \beta_1^{l_1}\beta_2^{l_2}\cdots\beta_N^{l_N} \left(\mathbf{E}s_1^{l_1}\mathbf{E}s_2^{l_2}\cdots\mathbf{E}s_N^{l_N}\right)\mathbf{E}s_i \right\}, \tag{20}$$

其中，对于固定的 i，乘积 $\left(\mathbf{E}s_1^{l_1}\mathbf{E}s_2^{l_2}\cdots\mathbf{E}s_N^{l_N}\right)\mathbf{E}s_i$ 等于 α^m，前提是乘积中有 m 个不同的指标，且其中包括 i^{th} 指标。如果非零 l_k项中的某个指标等于 i，则乘积 $\mathbf{E}s_1^{l_1}\mathbf{E}s_2^{l_2}\cdots\mathbf{E}s_N^{l_N}$ 中必须有 m 个不同的项；否则有 $m-1$ 个不同的项。

同样，我们可以计算$\mathbf{E}R^2S^l$，详情见补充信息。例如，

$$\begin{aligned}\mathbf{E}R^2S^2 &= \alpha\sum\beta_i^4\mathbf{E}r_i^2 + \alpha^2\left\{\sum_{j\neq i}\beta_i^2\beta_j^2\mathbf{E}r_i^2 + \sum_{j\neq i}2\beta_i^3\beta_j\mathbf{E}r_ir_j + \sum_{j\neq i}2\beta_i^3\beta_j\mathbf{E}r_i^2 + \sum_{j\neq i}2\beta_i^2\beta_j^2\mathbf{E}r_ir_j\right\}\\ &+\alpha^3\left\{\sum_{i\neq j\neq k}\beta_i\beta_j\beta_k^2\mathbf{E}r_ir_j + \sum_{i\neq j\neq k}\beta_i^2\beta_j\beta_k\mathbf{E}r_i^2 + \sum_{i\neq j\neq k}4\beta_i^2\beta_j\beta_k\mathbf{E}r_ir_j\right\} + \alpha^4\sum_{i\neq j\neq k\neq m}\beta_i\beta_j\beta_k\beta_m\mathbf{E}r_ir_j.\end{aligned} \tag{21}$$

表 1：公式 (15) 中方差截断近似所涉及的矩。

矩	表达式
$\mathbf{E}S$	α
$\mathbf{E}S^2$	$\alpha\sum\beta_k^2 + \alpha^2\sum_{j\neq k}\beta_j\beta_k$
$\mathbf{E}S^3$	$\alpha\sum\beta_k^3 + \alpha^2\sum_{j\neq k}3\beta_j\beta_k^2 + \alpha^3\sum_{i\neq j\neq k}\beta_i\beta_j\beta_k$
$\mathbf{E}S^4$	$\alpha\sum\beta_k^4 + \alpha^2\left\{\sum_{j\neq k}4\beta_j\beta_k^3 + \sum_{j\neq k}3\beta_j^2\beta_k^2\right\} + \alpha^3\sum_{i\neq j\neq k}6\beta_i\beta_j\beta_k^2 + \alpha^4\sum_{i\neq j\neq k\neq m}\beta_i\beta_j\beta_k\beta_m$
$\mathbf{E}R$	$\alpha\sum_{i=1}^N\beta_i\mathbf{E}r_i$
$\mathbf{E}RS$	$\alpha\sum_{i=1}^N\beta_i^2\mathbf{E}r_i + \alpha^2\sum_{i\neq j}\beta_i\beta_j\mathbf{E}r_i$
$\mathbf{E}RS^2$	$\alpha\sum_{i=1}^N\beta_i^3\mathbf{E}r_i + \alpha^2\left\{\sum_{j\neq i}^N\beta_i\beta_j^2\mathbf{E}r_i + \sum_{j\neq i}2\beta_i^2\beta_j\mathbf{E}r_i\right\} + \alpha^3\sum_{i\neq j\neq k}2\beta_i\beta_j\beta_k\mathbf{E}r_i$
$\mathbf{E}R^2$	$\alpha\sum\beta_i^2\mathbf{E}r_i^2 + \alpha^2\sum_{i\neq j}\beta_i\beta_j\mathbf{E}r_ir_j$
$\mathbf{E}R^2S$	$\alpha\sum\beta_i^3\mathbf{E}r_i^2 + \alpha^2\left\{\sum_{i\neq j}\beta_i^2\beta_j\mathbf{E}r_i^2 + 2\sum_{i\neq j}\beta_i^2\beta_j\mathbf{E}r_ir_j\right\} + \alpha^3\sum_{i\neq j\neq k}\beta_i\beta_j\beta_k\mathbf{E}r_ir_j$

3 均匀权重

上述公式适用于总和为一的通用权重 β_i。对于均匀权重 $\beta_i = 1/N$，可以大幅简化重要性计算，例如计算算术平均值时。均匀权重的一种应用是无放回抽样，以推断样本平均值的方差。这会得到简化的矩公式。例如，根据表 1，对于均匀权重

$$\mathbf{E}S^2 = \alpha \sum \frac{1}{N^2} + \alpha^2 \sum_{j \neq k} \frac{1}{N^2} = \alpha \frac{1}{N} + \alpha^2 \frac{(N-1)}{N} \tag{22}$$

因为求和 分别在 N 个索引和 $N(N-1)$ 个索引上进行。类似地 y,

$$\mathbf{E}R^2S = \alpha \sum \frac{1}{N^3} \mathbf{E}r_i^2 + \alpha^2 \left\{ \sum_{i \neq j} \frac{1}{N^3} \mathbf{E}r_i^2 + \sum_{i \neq j} \frac{2}{N^3} \mathbf{E}r_i r_j \right\} + \alpha^3 \sum_{i \neq j \neq k} \frac{1}{N^3} \mathbf{E}r_i r_j, \tag{23}$$

简化为

$$\alpha \sum_i \frac{1}{N^3} \mathbf{E}r_i^2 + \alpha^2 \left\{ \sum_i \frac{N-1}{N^3} \mathbf{E}r_i^2 + \sum_{i \neq j} \frac{2}{N^3} \mathbf{E}r_i r_j \right\} + \alpha^3 \sum_{i \neq j} \frac{N-2}{N^3} \mathbf{E}r_i r_j. \tag{24}$$

表 2 总结了这些矩。相关简化详见补充信息第 4 节。与前文一样，除 α 外，这些量仅取决于底层场的 2^{nd} 阶矩。

Large N approx.
α
α^2
α^3
α^4
$\alpha \frac{1}{N} \sum_{i=1}^{N} \mathbf{E} r_i$
$\alpha^2 \frac{1}{N} \sum_{i=1}^{N} \mathbf{E} r_i$
$\alpha^3 \frac{1}{N} \sum_{i=1}^{N} \mathbf{E} r_i$
$\frac{}{2} \sum_i \mathbf{E} r_i^2 + \alpha^2 \frac{1}{N^2} \sum_{i \neq j} \mathbf{E} r_i r_j$
$\frac{}{2} \sum_i \mathbf{E} r_i^2 + \alpha^3 \frac{1}{N^2} \sum_{i \neq j} \mathbf{E} r_i r_j$
$\frac{}{2} \sum_i \mathbf{E} r_i^2 + \alpha^4 \frac{1}{N^2} \sum_{i \neq j} \mathbf{E} r_i r_j$

3.1 组合学推导

对于均匀权重的情况，这些公式可以通过组合学更简便地推导出来。对于均匀权重，公式 (18)

$$\mathbf{E}S^l = \sum_{l_1+l_2+\ldots l_N=l} \frac{l!}{l_1!l_2!\cdots l_N!}\left(\frac{1}{N}\right)^{l_1}\left(\frac{1}{N}\right)^{l_2}\cdots\left(\frac{1}{N}\right)^{l_N}\left(\mathbf{E}s_1^{l_1}\mathbf{E}s_2^{l_2}\cdots\mathbf{E}s_N^{l_N}\right) \tag{25}$$

变为

$$\frac{1}{N^l}\sum_{l_1+l_2+\ldots l_N=l} \frac{l!}{l_1!l_2!\cdots l_N!}\left(\mathbf{E}s_1^{l_1}\mathbf{E}s_2^{l_2}\cdots\mathbf{E}s_N^{l_N}\right). \tag{26}$$

考虑将 l 个可区分球分配到 N 个箱子中的问题。总共有 N^l 种不同的分配方式。这等于对 $l_1+l_2+\ldots l_N=l$ 求和所得的 $^l!_{l_1!l_2!\cdots l_N!}$。根据多项式定理，$^l!_{\,l_1!l_2!\cdots l_N!}$ 表示将 l_1 个球放入第 1^{st} 个箱子、将 l_2 个球放入第 2^{nd} 个箱子，依此类推，最终将 l 个球放入 N 个箱子中的分配方式数。现在考虑将 l 个可区分球分配到 N 个箱子中的问题，且恰有 m 个箱子被占用，这对应于$\mathbf{E}s_1^{l_1}\mathbf{E}s_2^{l_2}\cdots\mathbf{E}s_N^{l_N}$中具有非零幂次的 m 个不同项。

这对应于 N^l 乘以$\mathbf{E}S^l$中 α^m 的系数。该系数表示将 l 个可区分球分配到 N 个箱子中的方法数，且恰好占用 m 个箱子。计算分为三步：首先，将 l 个球的集合划分为恰好 m 个子集，此类划分的数量记为 $\left\{ \begin{array}{c} l \\ m \end{array} \right\}$；其次，从 N 个可用箱子中选出 m 个，共有 $\left(\begin{array}{c} N \\ m \end{array} \right)$ 种选法；最后，以 $m!$ 种不同方式排列这些箱子。因此，描述$\mathbf{E}s_1^{l_1}\mathbf{E}s_2^{l_2}\cdots\mathbf{E}s_N^{l_N}$中包含 m 个不同项的方法数的系数为

$$\left\{ \begin{array}{c} l \\ m \end{array} \right\}\left(\begin{array}{c} N \\ m \end{array} \right) m! = \left\{ \begin{array}{c} l \\ m \end{array} \right\}\frac{N!}{(N-m)!}, \tag{27}$$

其中，$\left\{ \begin{array}{c} l \\ m \end{array} \right\}$ 表示将 l 个可区分元素划分为恰好 m 个非空集合的划分数，可由第二类斯特林数给出，其公式为

$$\left\{ \begin{array}{c} l \\ m \end{array} \right\} = \frac{1}{m!} \sum_{q=0}^{m} (-1)^{m-q} \left(\begin{array}{c} m \\ q \end{array} \right) q^l. \tag{28}$$

最后我们得到

$$\mathbf{E}S^l = \frac{1}{N^l} \sum_{m=1}^{\min\{l,N\}} \left\{ \begin{array}{c} l \\ m \end{array} \right\} \frac{N!}{(N-m)!} \alpha^m \tag{29}$$

其中，斯特林数 $\left\{ \begin{array}{c} l \\ m \end{array} \right\}$ 可由式 (28) 中的公式计算。表 3 根据式 (29) 列出了 $1 \leq l \leq 4$ 时 $\mathbf{E}S^l$ 中的 α^m 系数，证实了表 2 中列出的独立推导公式。相应的第二类斯特林数列于表 7。

$m = 4$
$\frac{}{4)!} = \frac{(N-1)(N-2)(N-3)}{N^3}$

同样，对于均匀权重，公式（20）

$$\mathbf{E}RS^l = \sum_{i=1}^{N} \frac{1}{N}\mathbf{E}r_i \left\{ \sum_{l_1+l_2+\ldots l_N=l} \frac{l!}{l_1!l_2!\cdots l_N!} \left(\frac{1}{N}\right)^{l_1} \left(\frac{1}{N}\right)^{l_2} \cdots \left(\frac{1}{N}\right)^{l_N} \left(\mathbf{E}s_1^{l_1}\mathbf{E}s_2^{l_2}\cdots\mathbf{E}s_N^{l_N}\right)\mathbf{E}s_i \right\} \tag{30}$$

or

$$\mathbf{E}RS^l = \frac{1}{N^{l+1}} \sum_{i=1}^{N} \mathbf{E}r_i \left\{ \sum_{l_1+l_2+\ldots l_N=l} \frac{l!}{l_1!l_2!\cdots l_N!} \left(\mathbf{E}s_1^{l_1}\mathbf{E}s_2^{l_2}\cdots\mathbf{E}s_N^{l_N}\right)\mathbf{E}s_i \right\}, \tag{31}$$

这也是关于 α 的幂级数，其中（根据对称性） N 乘以 α^m 的系数为 $\frac{1}{N^{l+1}}\sum_{i=1}^{N}\mathbf{E}r_i$ 乘以将 $l+1$ 个可区分球分配到 N 个箱子中且恰好占用 m 个箱子的方式数。这样分配的方式数为 $\left\{ \begin{matrix} l+1 \\ m \end{matrix} \right\} \binom{N}{m} m! = \left\{ \begin{matrix} l+1 \\ m \end{matrix} \right\} \frac{N!}{(N-m)!}$，因此

最终

$$\mathbf{E}RS^l = \frac{1}{N^{l+1}} \sum_{i=1}^{N} \mathbf{E}r_i \left\{ \sum_{m=1}^{\min\{l+1,N\}} \frac{1}{N} \left\{ \begin{matrix} l+1 \\ m \end{matrix} \right\} \frac{N!}{(N-m)!}\alpha^m \right\}, \tag{32}$$

简化为

$$\mathbf{E}RS^l = \frac{1}{N^{l+1}} \sum_{i=1}^{N} \mathbf{E}r_i \left\{ \sum_{m=1}^{\min\{l+1,N\}} \left\{ \begin{matrix} l+1 \\ m \end{matrix} \right\} \frac{(N-1)!}{(N-m)!}\alpha^m \right\}. \tag{33}$$

$\mathbf{E}r_i\alpha^m$ 在$\mathbf{E}RS^l$ 中的系数显示在表 4 中，其结果与表 2 中的相应公式相同。

$m = 4$
$\frac{1)!}{4)!} = \frac{(N-1)(N-2)(N-3)}{N^4}$

最后，对于均匀权重，$\mathbf{E}R^2S^l$ 变为（根据补充信息中的公式 23）

$$\mathbf{E}R^2S^l = \sum_{i=1}^{N} \frac{1}{N^2}\mathbf{E}r_i^2 \left\{ \sum_{l_1+l_2+\ldots l_N=l} \frac{l!}{l_1!l_2!\cdots l_N!} \left(\frac{1}{N}\right)^{l_1} \left(\frac{1}{N}\right)^{l_2} \cdots \left(\frac{1}{N}\right)^{l_N} \left(\mathbf{E}s_1^{l_1}\mathbf{E}s_2^{l_2}\cdots\mathbf{E}s_N^{l_N}\right)\mathbf{E}s_i^2 \right\}$$
$$+\sum_{i\neq j}^{N} \frac{1}{N^2}\mathbf{E}r_ir_j \left\{ \sum_{l_1+l_2+\ldots l_N=l} \frac{l!}{l_1!l_2!\cdots l_N!} \left(\frac{1}{N}\right)^{l_1} \left(\frac{1}{N}\right)^{l_2} \cdots \left(\frac{1}{N}\right)^{l_N} \left(\mathbf{E}s_1^{l_1}\mathbf{E}s_2^{l_2}\cdots\mathbf{E}s_N^{l_N}\right)\mathbf{E}s_i\mathbf{E}s_j \right\}, \tag{34}$$

or

$$\mathbf{E}R^2S^l = \frac{1}{N^{l+2}}\sum_{i=1}^{N} \mathbf{E}r_i^2 \left\{ \sum_{l_1+l_2+\ldots l_N=l} \frac{l!}{l_1!l_2!\cdots l_N!} \left(\mathbf{E}s_1^{l_1}\mathbf{E}s_2^{l_2}\cdots\mathbf{E}s_N^{l_N}\right)\mathbf{E}s_i^2 \right\}$$
$$+ \frac{1}{N^{l+2}}\sum_{i\neq j}^{N} \mathbf{E}r_ir_j \left\{ \sum_{l_1+l_2+\ldots l_N=l} \frac{l!}{l_1!l_2!\cdots l_N!} \left(\mathbf{E}s_1^{l_1}\mathbf{E}s_2^{l_2}\cdots\mathbf{E}s_N^{l_N}\right)\mathbf{E}s_i\mathbf{E}s_j \right\}, \tag{35}$$

由于$\mathbf{E}s_i^2 = \mathbf{E}s_i$，我们可以像之前一样直接计算第一项。对于第二项，我们要找出含有$\mathbf{E}r_ir_j$ 的表达式在乘积 $\left(\mathbf{E}s_1^{l_1}\mathbf{E}s_2^{l_2}\cdots\mathbf{E}s_N^{l_N}\right)\mathbf{E}s_i\mathbf{E}s_j$中恰有 m 个不同项的不同方式数。这对应于将 $l+2$ 个可区分球分配到 N 个箱子中，且恰有 m 个箱子被占用。我们必须扣除已计数的分配，这些分配涉及$\mathbf{E}r_i^2$形式的项。每个$\mathbf{E}r_ir_j$、 $i \neq j$项都有 $2\binom{N}{2}$ 种选取方式，而 r_i中的项要么是$\mathbf{E}r_i^2$ 形式，要么是$\mathbf{E}r_ir_j$形式。因此，记 $\alpha^m\mathbf{E}r_ir_j$ 的系数为 $b_m^{(l)}$，记 $\alpha^m\mathbf{E}r_i^2$的系数为 $a_m^{(l)}$，我们得到

$$Na_m^{(l)} + 2\binom{N}{2} b_m^{(l)} = \frac{1}{N^{l+2}} \left\{ \begin{matrix} l+2 \\ m \end{matrix} \right\} \binom{N}{m} m!, \tag{36}$$

其中，按照上述论证，右边的项表示将 $l+2$ 个可区分球分配到从 N 个可能的箱子中选出的 m 个箱子里的方法数，并且 $Na_m^{(l)} = \frac{1}{N^{l+2}} \left\{ \begin{matrix} l+2 \\ m \end{matrix} \right\} \binom{N}{m} m!$。因此， $N(N-1)$

$$\overline{b_m^{(l)}} = \frac{1}{N^{l+2}} \left(\left\{ \begin{matrix} l+2 \\ m \end{matrix} \right\} - \left\{ \begin{matrix} l+1 \\ m \end{matrix} \right\} \right) \binom{N}{m} m! \qquad \overline{\qquad}$$

最终，我们得到

$$\mathbf{E}R^2S^l = \frac{1}{N^{l+2}}\sum_{i=1}\mathbf{E}r_i^2\sum_{m=1}^{\min\{l+1,N\}}\left\{\begin{array}{c} l+1 \\ m \end{array}\right\}\frac{(N-1)!}{(N-m)!}\alpha^m$$
$$+\frac{1}{N^{l+2}}\sum_{i\neq j}\mathbf{E}r_ir_j\sum_{m=1}^{\min\{l+2,N\}}\frac{(N-2)!}{(N-m)!}\left(\left\{\begin{array}{c} l+2 \\ m \end{array}\right\}-\left\{\begin{array}{c} l+1 \\ m \end{array}\right\}\right)\alpha^m. \quad (37)$$

这些系数见表5和表6，它们验证了表 2 中列出的表达式。总之，我们已经证明，对于均匀权重，方差公式中出现的矩可以通过组合学进行估计。然而，这并没有给出方差本身的公式。

$m = 4$
$\frac{1)!}{4)!} = \frac{(N-1)(N-2)(N-3)}{N^5}$

	$m=4$
$\left\{ \right] = \frac{N-2}{N^3}$	
$\left\{ \right] = 5\frac{N-2}{N^4}$	$\frac{1}{N^4}\frac{(N-2)!}{(N-4)!}\left[\left\{ \begin{matrix} 4 \\ 4 \end{matrix} \right\} - \left\{ \begin{matrix} 3 \\ 4 \end{matrix} \right\} \right] = \frac{(N-2)(N-3)}{N^4}$

表 7：第二类斯特林数 $\left\{ \begin{matrix} n \\ k \end{matrix} \right\}$，描述将 n 个元素划分为 k 个集合的划分数量，用于表 3-6 中使用的范围。

	k	1	2	3	4
n					
1		1	0	0	0
2		1	1	0	0
3		1	3	1	0
4		1	7	6	1

3.2 大 N 近似

如果该场从一个非常大的潜在位置集合中进行采样，这些公式会变得简化，因为对于固定的 α，我们只需考虑分子中 N 的最高次幂项。例如，在 $N \to \infty$的极限情况下，$\mathbf{E}S^4 \approx \alpha^{4\,(N-1)(N-2)(N-3)}_{N^3} \approx \alpha^4 = \mathbf{E}S^4$. 类似地，$\mathbf{E}R^2S^2 \approx \frac{\alpha^3}{N^2}\sum_i \mathbf{E}r_i^2 + \frac{\alpha^4}{N^2}\sum_{j\neq i}\mathbf{E}r_i r_j = \mathbf{E}R^2\mathbf{E}S^2$，依此类推。这些简化与完整公式一同列于表 2 中。在大 N 极限下，假设 $l \ll N$，我们得到

$$\mathbf{E}S^l \approx \frac{1}{N^l}\left\{ \begin{matrix} l \\ l \end{matrix} \right\} \frac{N!}{(N-l)!}\alpha^l = \frac{(N-1)\dots(N-l+1)}{N^l}\alpha^l \approx \alpha^l = (\mathbf{E}S)^l . \tag{38}$$

同样

$$\mathbf{E}RS^l \approx \frac{1}{N^{l+1}}\sum_{i=1}^{N}\mathbf{E}r_i \left\{ \begin{matrix} l+1 \\ l+1 \end{matrix} \right\} \frac{(N-1)!}{(N-l-1)!}\alpha^{l+1} \approx \frac{1}{N}\sum_{i=1}^{N}\mathbf{E}r_i\alpha^{l+1} = \mathbf{E}R\,(\mathbf{E}S)^l , \tag{39}$$

而

$$\mathbf{E}R^2S^l \approx \frac{1}{N^{l+2}}\sum_{i=1}\mathbf{E}r_i^2 \left\{ \begin{array}{c} l+1 \\ l+1 \end{array} \right\} \frac{(N-1)!}{(N-l-1)!}\alpha^{l+1} + \frac{1}{N^{l+2}}\sum_{i\neq j}\mathbf{E}r_ir_j\frac{(N-2)!}{(N-l-2)!}\left(\left\{ \begin{array}{c} l+2 \\ l+2 \end{array} \right\} - \left\{ \begin{array}{c} l+1 \\ l+2 \end{array} \right\} \right)$$
$$\approx \left\{ \frac{\alpha}{N^2}\sum_{i=1}\mathbf{E}r_i^2 + \frac{\alpha^2}{N^2}\sum_{i\neq j}\mathbf{E}r_ir_j \right\}\alpha^l = \mathbf{E}R^2\mathbf{E}S^l. \quad (40)$$

将这些结果代入式（15），方差公式变为

$$\sigma_f^2 = \frac{1}{\alpha^2}\left\{\mathbf{E}R^2 - (\mathbf{E}R)^2\right\} = \frac{1}{\alpha^2}\sigma_R^2, \quad (41)$$

当 R 和 S 的矩统计独立时，它也等于方程（12）中对方差的一阶近似。方程（1）中比值的方差，简单来说就是分子的方差因缺失观测而被 $1/\alpha^2$ 放大。

3.3 N 中等取值时的行为

如果 N 不一定很大，那么其行为会有趣得多。如果我们放宽关于$N \to \infty$的假设， R 和 S 的矩会协变，对于 $\beta_i = 1/N$的情况，我们必须参考表 2 的中间列。于是， 2^{nd} 阶方差近似可简化为

$$\sigma_f^2 = \frac{1}{\alpha}\left[1 + \frac{1}{N}\frac{\alpha-1}{\alpha} + \frac{1}{N^2}\frac{\alpha-1}{\alpha}\frac{2\alpha-1}{\alpha}\right]\frac{1}{N^2}\sum_i \mathbf{E}r_i^2$$
$$+ \left[1 + \frac{3}{N}\frac{\alpha-1}{\alpha} + \frac{1}{N^2}\frac{\alpha-1}{\alpha}\frac{6\alpha-4}{\alpha}\right]\frac{1}{N^2}\sum_{j\neq i}\mathbf{E}r_ir_j$$
$$- \left[1 + \frac{2}{N}\frac{\alpha-1}{\alpha} - \frac{1}{N^2}\left(\frac{\alpha-1}{\alpha}\right)\left(\frac{3\alpha-1}{\alpha}\right) + \frac{1}{N^3}\left\{6\left(\frac{\alpha-1}{\alpha}\right)^2 + \frac{1}{\alpha^2}\frac{\alpha-1}{\alpha}\right\}\right]\left\{\frac{1}{N}\sum_{i=1}^{N}\mathbf{E}r_i\right\}^2. \quad (42)$$

显然，每一项都比前一项小 $\frac{1}{N}\left|\frac{\alpha-1}{\alpha}\right|$倍。回想第 2.1 节，这个因子的倒数，即 $N\left|\frac{\alpha}{1-\alpha}\right|$，是其与期望值 $\mathbf{E}S$ 相差的标准差数的平方； S 必须达到这个数值，级数才不会收敛。这个量越大，收敛就越快。

发生。因此，当 α 接近于 **1** 时，我们可以预期级数会快速收敛；更一般地说，当 α 较小时，若 N 较大以抵消 $\left|{}^{\alpha-1}_{\alpha}\right| > 1$ 的影响，也可以预期级数会快速收敛。

简化

我们来算一下 c 考虑公式 **(42)** 中方差公式的以下简化情形：

- $\alpha = 1$: 在观测值以确定性报告的情况下，方差

$$\sigma_f^2 = \frac{1}{N^2}\sum_i \mathbf{E}r_i^2 + \frac{1}{N^2}\sum_{j\neq i}\mathbf{E}r_i r_j - \left\{\frac{1}{N}\sum_{i=1}^{N}\mathbf{E}r_i\right\}^2 \tag{43}$$

简化为

$$\sigma_f^2 = \frac{1}{N^2}\sum_i \sigma_i^2 + \frac{1}{N^2}\sum_{j\neq i}\sigma_{ij}, \tag{44}$$

其中 $\sigma_i^2 = \mathbf{E}r_i^2 - (\mathbf{E}r_i)^2$ 和 $\sigma_{ij} = \mathbf{E}r_i r_j - \mathbf{E}r_i\mathbf{E}r_j$ 分别是逐点方差和成对协方差。该公式等价地可写为 $\sigma_i^2 = \beta^T\Sigma\beta$，其中 Σ 是协方差矩阵，而 $\beta = \left\{\begin{array}{ccc}\beta_1 & \dots & \beta_n\end{array}\right\}$ 是等于 $\left\{\begin{array}{ccc}\frac{1}{N} & \dots & \frac{1}{N}\end{array}\right\}^T$ 的权重向量。

- $\alpha \to 1$: 当报告概率接近 **1** 时，我们可以只考虑每个级数中的第一项。

$$\sigma_f^2 = \frac{1}{\alpha}\frac{1}{N^2}\sum_i \mathbf{E}r_i^2 + \frac{1}{N^2}\sum_{j\neq i}\mathbf{E}r_i r_j - \left\{\frac{1}{N}\sum_{i=1}^{N}\mathbf{E}r_i\right\}^2 \tag{45}$$

由此变为

$$\sigma_f^2 = \frac{1}{\alpha}\frac{1}{N^2}\sum_i \sigma_i^2 + \frac{1}{N^2}\sum_{j\neq i}\sigma_{ij} + \frac{1-\alpha}{\alpha}\frac{1}{N}\sum_{i=1}^{N}(\mathbf{E}r_i)^2\,. \tag{46}$$

与前一种情况相比，逐点方差增加了 $1/\alpha$，并且对于逐点期望的平方，还需要额外的、与 ${}^{1-\alpha}_{\alpha}$ 成正比的修正。

- 单个时期：该公式可应用于单个时期，以描述在给定时间对潜在观测值进行无放回抽样。期望简化为变量

它们自身，因此 $\mathbf{E}r_i^2 = r_i^2$，$\mathbf{E}r_i r_j = r_i r_j$，以及 $\mathbf{E}r_i = r_i$。方差公式变为

$$\sigma_f^2 = \frac{1-\alpha}{\alpha}\frac{1}{N^2}\left[1 + \frac{1}{N}\frac{2\alpha-1}{\alpha} - \frac{1}{N^2}\frac{\alpha^2+\alpha-1}{\alpha^2} + \frac{1}{N^3}\frac{6\alpha^2-6\alpha+1}{\alpha^2}\right]\sum_i r_i^2 - \frac{1-\alpha}{\alpha}\frac{1}{N^3}\left[1 + \frac{1}{N}\frac{7\alpha-5}{\alpha} - \frac{1}{N^2}\frac{6\alpha^2-6\alpha+1}{\alpha^2}\right]\sum_{j\neq i} r_i r_j \quad (47)$$

当 N 很大，使得 $\frac{1}{N}\left|\frac{2\alpha-1}{\alpha}\right| \ll 1$ 和 $\frac{1}{N}\left|\frac{7\alpha-5}{\alpha}\right| \ll 1$ 均成立时，我们可以简化上述公式

$$\sigma_f^2 = \frac{1-\alpha}{\alpha}\frac{1}{N^2}\left\{\sum_i r_i^2 - \frac{1}{N}\sum_{j\neq i} r_i r_j\right\} \quad (48)$$

并且，对于较大的 N、$\sum_{j\neq i} r_i r_j \cong \left(\sum_i r_i\right)^2$，，因此定义该时期的空间方差

$$\sigma_s^2 = \frac{1}{N}\sum_i r_i^2 - \left(\frac{1}{N}\sum_i r_i\right)^2 \quad (49)$$

我们得到

$$\sigma_f^2 = \frac{1-\alpha}{\alpha}\frac{1}{N}\sigma_s^2 \quad (50)$$

如前所推导（*Seshadri* （2018）；*Prakash* 等 （2019））。因此，当观测值数量较大时，无放回抽样情况下的总体方差与潜在样本之间的方差成正比。然而，这一简单结果仅是一个特例；更一般地，需要使用公式（47）。

4 模拟

我们针对单个时期，对印度气象部门制作的网格降雨数据集进行了说明性模拟。该数据集记录了印度 357 个网格点的日降雨量；我们对 2001 年夏季季风月份的数据取一年平均，得到 357 个空间值。我们从这些网格值中随机选取一个子集（大小为 N，范围从 $10-300$），并固定 $N.$。对于每个 N，模拟一组伯努利随机变量 $\{s_i: \ 1 \leq i \leq N\}$。在每组中，每个 s_i 取值为 1 的概率为 α。这表示这些网格值中哪些会用于空间平均

计算。每个集合成员都有一个对应的空间平均值，并计算集合方差，再与式（50）中推导出的公式进行比较（图 1）。每个点表示 N 和 α. 的一种组合。左图显示，当方差较大时，即 $\frac{1-\alpha}{\alpha}\frac{1}{N}$ 较大时，模拟方差会偏离公式。右图显示相对误差的大小，以及相对误差为 0.1，的等值线，该等值线区分了近似效果良好和较差的区域。对于 $\alpha > 1/2$ ，较小的 N 就足够；而对于较小的 α，则需要大得多的数据集。 ——–

实际上，仅有 $\frac{1-\alpha}{\alpha}\frac{1}{N}$ 不足以确定收敛速率，但这一因子会在展开式的各项中反复出现。图 2 展示了式（42）中的两个修正项（截至 $1/N^2$）随 α 和 N变化的情况。由于这些项是对系数 1 的修正，因此只有在 α 或 N. 较大时才能忽略。第二个修正项通常可以忽略， α较小时除外。因此，当缺失数据概率 $1-\alpha$ 较小时，可以预期，方差估计量的低阶截断效果相当不错。

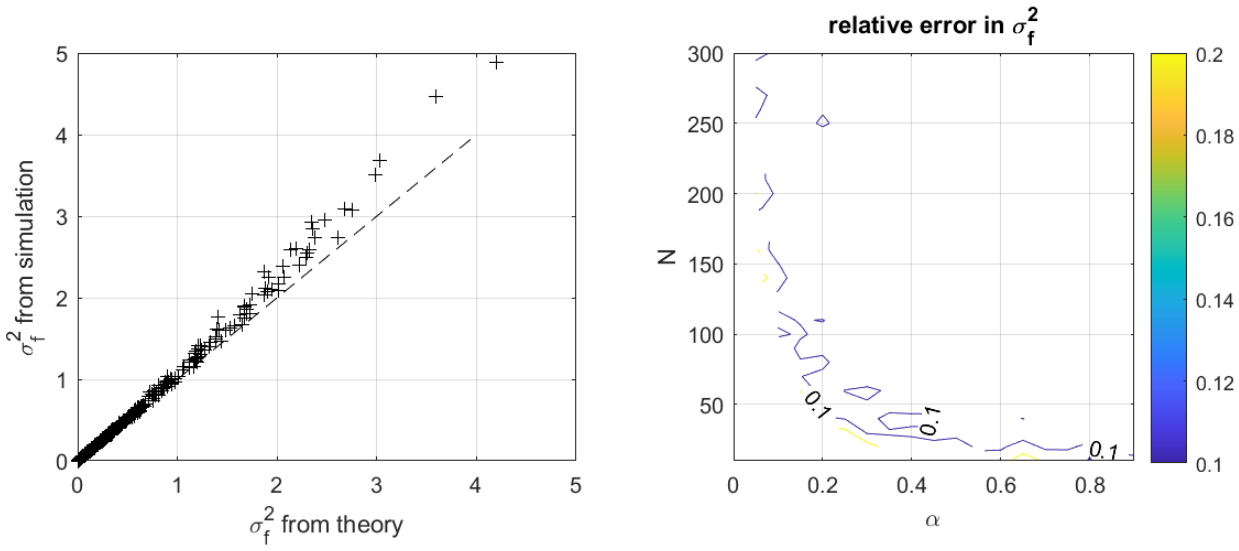


图1：蒙特卡洛模拟中单个时期的方差，与左图中的公式（50）推导结果进行比较，右图显示相对误差（绝对值）。标出了相对误差为 0.1 的等高线。

5 讨论

推导空间均值方差的收敛估计量，对地球科学具有广泛意义。这些方差估计量仅取决于物理场的一阶矩和二阶矩，其收敛要求这两阶矩有限。许多地球物理问题中的一阶矩和二阶矩都是有限的，因此可以针对广泛的应用场景推导收敛的方差估计量，涵盖从水文学到气候科学的领域。此外，只要单个观测值的报告概率足够高（即 $\alpha > 1/2$），就能保证收敛。我们通过描述方差级数展开中各连续项的行为，说明了这一收敛性。结果表明，即使数据集较小，只要缺失数据的概率不过高，我们仍能得到良好的方差估计量。反之，如果报告某个观测值的概率较低，那么数据集最好足够大（即 N）。

当权重均匀时，会出现一种重要的特殊情况，例如在求总体的算术平均值时。这是从潜在观测点总体中进行无放回抽样，以推断空间平均值的性质。与本研究一样，如果各个站点被观测到的概率保持不变，报告的观测值总数就会变化，并服从二项分布。虽然这与自助法不同。

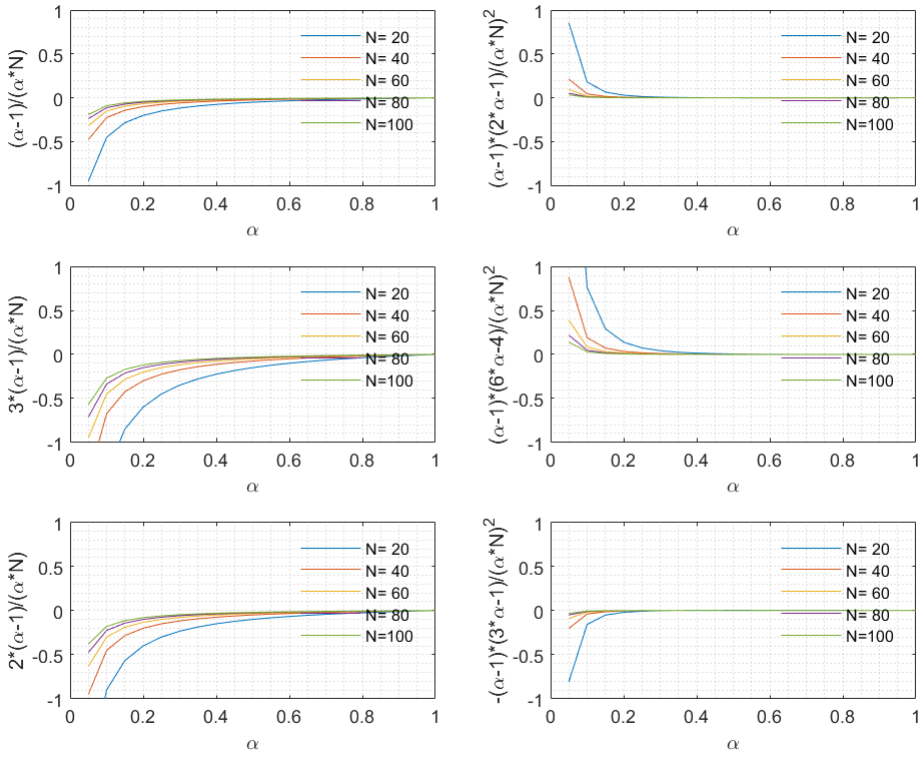


图 2：根据公式（42），$\sum_i \mathbf{E}r_i^2$（顶部图板）、$\sum_{j\neq i} \mathbf{E}r_i r_j$（中部图板）和 $\left\{\frac{1}{N}\sum_{i=1}^{N}\mathbf{E}r_i\right\}^2$（底部图板）的系数中的前两个修正项，作为 α 和 N 的函数。由于这些是对 1 的系数的修正，因此对于较大的 α 或 N.，它们通常很小

（*Diaconis* 和 *Efron* （1983））在其设计和无放回抽样中，仍能导出有用的非参数估计量公式，这些估计量出现在类似的情境中（*Hankin* 等人（2019）），并涉及总体子样本的统计量。由于其基础理论是非参数的，且具有普遍启示，因此即使这一情形的精确表述与更常见的处理方式不同，进一步研究仍有价值。

在本文的第二部分，我们对这一情形进行了详细分析。与一般问题相比，这个权重均匀的特殊情形，由于我们假设 s_i相互独立，具有额外的结构。因此，这种情形下的矩可以通过组合学直接估计。具体而言，估计量中出现的矩的所有系数都可以利用计数不同集合划分数量的第二类斯特林数轻松计算出来（$Aigner$ （2007））。由此，近似估计量的收敛问题也得到了明确阐述。将分析扩展到更高阶的方差近似或许会有帮助。

参考文献

Aigner, M. (2007),组合学教程, Springer.

Bromwich, T. J. I. 和 T. M. MacRobert（1991），无穷级数理论导论，纽约：Chelsea。

Diaconis, P.，和 B.Efron（1983），统计学中的计算机密集型方法，《科学美国人》，*248*, 116–131，doi:https://www.jstor.org/stable/24968902。

Gandin,L. S.（1993），气象场的最优平均，美国商务部办公室说明第397号。

Gemignani,M.C. (1970), 微积分与统计学, Dover。

Goodman, L. A. 和 H. O. Hartley（1958），《无偏比值型估计量的精度》，美国统计学会杂志，*53*，491–508，doi:10.1080/01621459.1958. 10501454。

Hankin, D. G.、M. S. Mohr 和 K. B. Newman（2019），抽样理论：生态学与自然资源科学，牛津大学出版社。

Hartley, H. O. 和 A. Ross（1954），无偏比值估计量，《自然》，*174*，270–271，doi:10.1038/ 174270a0。

Hinkley, D. V.（1969），关于两个相关正态随机变量的比值，《生物计量学》，*56*，635–639，doi:10.1093/biomet/56.3.635。

Hlawitschka, W.（1994），泰勒级数近似期望效用的经验特性，《美国经济评论》，*84*，713–719，doi:https://www.jstor.org/stable/2118079。

Kagan, R. L.（1997），《气象场的平均》，Kluwer Academic Publishers。

Kendall, M. G.、A. Stuart、J. Forster、K. Ord、S. F. Arnold 和 A. O'Hagan（1994），《*Kendall* 高级统计学理论》，Wiley。

Kundu, P. K. 和 R. K. Siddani（2007），用于描述区域平均降雨空间统计特征的一类新概率分布，*Journal of Geophysical Research: Atmospheres*，*112*，1–18，doi:10.1029/2006JD008042。

Loistl, O.（1976），通过泰勒级数展开对期望效用进行近似的错误：解析与计算结果，*The American Economic Review*，*66*，904–910，doi:https://www.jstor.org/stable/1827501。

Markowitz, H.（2015），期望效用的均值—方差近似，*European Journal ofOperations Research*，*234*，346–355，doi:10.1016/j.ejor.2012.08.023。

Morrissey, M. L.、J. A. Maliekal、J. S. Greene 和 J. Wang（1999），使用雨量计网络估计简单空间平均值的不确定性，*Water Resources Research*，*31*，2011–2017，doi:10.1029/95WR01232。

Oehlert, G. W.（1992），关于 delta 方法的一则说明，《美国统计学家》，*46*，27–29，doi：10.2307/2684406。

Prakash、S.、A. K. Seshadri、J. Srinivasan 和 D. S. Pai（2019），一种用于评估雨量计密度对印度月降雨量估算误差影响的新参数，水文气象学期刊，*20*，821–832，doi:10.1175/JHM-D-18-0161.1.

Rodríguez-Iturbe, I. 和 J. M. Mejía（1974），时空降雨网络的设计，水资源研究，*10*，713–728，doi:10.1029/WR010i004p00713.

Rudin, W.（1976），数学分析原理，McGraw Hill。

Seshadri, A. K.（2018），存在缺失数据时空间平均值的统计特性与最优平均，空间统计，*25*，1–21，doi:10.1016/j.spasta.2018.04.002。

Shen, S. S. P., G. R. North, 和 K.-Y. Kim（1994），全球平均温度最优估计的谱方法，气候学杂志，*7*，1999–2007，doi:10.1175/1520-0442(1994)007<1999:SATOEO>2.0.CO;2。

Vershynin, R.（2018），《高维概率：数据科学中的介绍与应用》，剑桥大学出版社。

Villarini, G.、P. V. Mandapaka、W. F. Krajewski 和 R. J. Moore（2008），降雨量与采样不确定性：雨量计视角，《地球物理研究杂志：大气》，*113*，1–12，doi:10.1029/2007JD009214。

Vinnikov, K. Y., A. Robock, N. C. Grody 和 A. Basist（2004），分析具有任意观测时间的气候记录中的昼夜和季节周期及趋势，地球物理研究快报，*31*，1–5，doi:10.1029/2003GL019196。

Weir, A. J. (1973), 勒贝格积分与测度, 剑桥大学出版社。