高校应用数学学报
2024, 39(1): 41-50

高放废物缓冲材料的均匀性分析 –基于密度比模型下的Gini系数

廖文琛[1], 谭 煜[2], 庄玮玮[1*]

(1. 中国科学技术大学 管理学院, 安徽合肥 230041;
2. 兰州大学 土木工程与力学学院, 甘肃兰州 730000)

摘 要: 提出用密度比模型估计缓冲材料含水率和干密度的Gini系数, 并在此基础上进行假设检验, 以解决核安全领域中如何科学分析缓冲材料均匀性的难点. 对假设检验所涉及的渐近理论进行了探讨. 提出在进行Wald-型检验的同时可借助自助法来提高检验功效. 基于上述方法对覆膜储存法试验所得的实际数据进行相应的估计和假设检验, 所得的结果验证了覆膜储存法对缓冲材料均匀性的保护是有效的.

关键词: Gini系数; 密度比模型; 自助法检验; 核安全
中图分类号: O213.2
文献标识码: A **文章编号**: 1000-4424(2024)01-0041-10

§1 引 言

随着能源危机愈演愈烈, 核能源是我国能源转型的重要方向之一, 而安全处置核废料则是核电产业可持续发展的必要条件[1]. 在国际通用的核废料深地质处置模型中, 缓冲屏障的性能对保障核安全至关重要. 缓冲屏障是由预制的压实膨润土砌块填充而成. 要有效阻止核素向周围环境中迁移, 缓冲屏障须具有极低的渗透性，较强的吸附性，较高的膨胀性能和一定的热传导性能, 这些性能提出的基本要求之一是缓冲屏障具有足够均匀的密度分布[2]. 目前的技术能够保障缓冲材料在安装后保持原有的均匀性, 但砌块在制好后往往不能立刻安装, 若在存储过程中不能得到良好的养护, 缓冲材料则极可能受到环境影响, 出现明显的密度差异, 从而导致核素沿着低密度的方向向外迁移. 比如, 在低湿环境中, 砌块会干燥失水, 致使干密度增加, 产生干燥裂纹, 形成潜在的渗透通道[3-6]. 工程试验中, 通常需要分析存储前后砌块的均匀性变化情况来判断养护手段是否有效. 其中, 最主要的两个指标是含水率和干密度的分布均匀情况, 两者分布均匀则表示缓冲材料在均匀性这方面的性能优良[7]. 如何更科学地分析这两个指标分布的变化情况, 是一个亟需解决的工程难题.

收稿日期: 2023-03-10 修回日期: 2023-06-15
*通讯作者, Email: weizh@ustc.edu.cn
基金项目: 国家自然科学基金(71971204); 安徽省杰青项目(2208085J43)

目前, 工程学中通常用离散系数来判断含水率和干密度的均匀度[8]. 而离散系数仅仅包含离散信息, 并不能全面反映砌块的均匀性, 且对异常值极为敏感. 因此, 本文提出借助更科学的统计学工具–Gini系数来分析砌块性能. Gini系数是由Gini[9]首次提出的, 在经济学中被广泛用于衡量国民收入或社会福利公平性. 它不仅同时包含了总体的离散信息和集中信息, 还不易受随机误差的干扰, 作为衡量均匀性和公平性的优良指标, 早已被推广到经济学之外的诸多领域, 在社会学[10-11], 数据科学[12-13], 医药学[14], 生态学[15], 运筹管理[16]等学科中都有重要应用.

由于目前砌块制备和测试的成本较高($\geq$ 10000元每块), 实验中的砌块样本量相对有限, 用常用的经验法来进行估计和检验时, 可能会受抽样误差影响较大. 为此, 本文引入了密度比模型. 密度比模型作为半参数模型的一种, 兼具非参数模型和参数模型的优点, 可以有效提升估计准确率和检验结果可靠性. 密度比模型的雏形可以追溯到文[17]. 随后, Anderson[18]假设样本所服从分布的密度函数之比是可估参数, 正式提出了密度比模型的概念. Owen[19], Qin和Zhang[20], Keziou和Leoni-Aubin[21], Chen和Liu[22]等对如何运用密度比模型进行估计及相关的渐近性质进行了大量讨论. Yuan等人[23]基于密度比模型对两个半连续总体的Gini系数进行了估计, 并证明了估计是渐近正态的.

本文基于密度比模型, 对兰州大学使用覆膜存储法进行储存前后的缓冲砌块数据的Gini系数进行了估计, 并通过假设检验分析了Gini系数是如何变化的, 以借此判断缓冲材料的均匀性变化情况. 首先基于Yuan等人[23]的理论进行了Wald-型检验. 考虑自助法在原始样本质量可靠的情况下能够扩大样本容量, 从而使推断结果更加稳定[24], 故进一步使用自助法来进行检验. 在使用自助法前, 本文对所构造统计量的渐近性进行了进一步的理论分析. 本文的主要结构为: §2简单介绍了本文所涉及的Gini系数和密度比模型的相关基本概念, 给出了运用密度比模型估计Gini系数的方法; §3根据实际背景提出了假设检验, 构造出相应的检验统计量, 并介绍所采用的两种检验方法的理论依据和具体步骤; §4对要分析的数据进行了描述, 并使用前文所述的方法进行假设检验; §5对本文的主要工作和结论进行了整体性的梳理.

§2 基于密度比模型的Gini系数

本章介绍Gini系数和密度比模型的一些基本概念, 并给出基于密度比模型估计多个总体的Gini系数的方法.

2.1 Gini系数的含义

Gini系数是基于Lorenz[25]提出的Lorenz曲线构造的. 假设总体X的分布函数为$F(x)$, p分位点为$\xi_{F,p}=\xi_F(p)=F^{-1}(p)$, $p\in[0,1]$, 则Lorenz曲线的函数形式为

$$L_F(p)=\frac{\int_0^p \xi_F(t)\mathrm{d}t}{\mu},\quad p\in[0,1],$$

其中μ是该分布的期望. Lorenz曲线反映了总体的分布均匀情况, 曲率越小表示分布越均匀. 当Lorenz曲线与x轴成45度时, 为绝对平等线, 代表绝对均匀的状态. Gini系数与Lorenz曲线的关系如图1所示, 从其端点向x轴引垂线形成一个三角形, 设某总体的Lorenz曲线将这个三角形分割成两个几何图形, 记Lorenz曲线之上的几何图形面积为S_1, 其下部分的面积为S_2, 则Gini系数为$\mathcal{G}=S_1/(S_1+S_2)$. 显然, $\mathcal{G}\in[0,1]$, 且$\mathcal{G}$越接近于0 表示总体越均匀. 上述面积法通常不便于

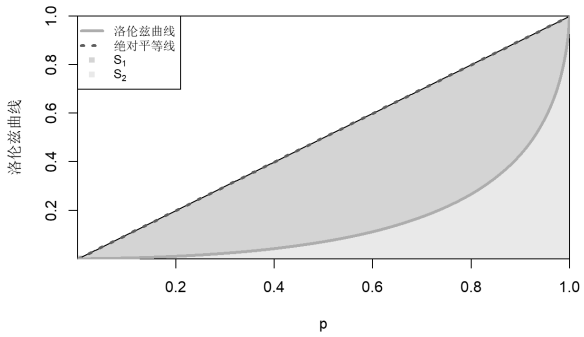

图 1 Gini系数和Lorenz曲线的关系

计算, 为了解决这个难点, David[26]用收入差绝对值的期望与均值之比构造出另一种Gini系数的等价形式.

定义2.1[26] 记X_1和X_2是从分布为$F(x)$的总体中随机抽取的两个样本点，E表示期望，μ为总体均值，则总体的Gini系数为

$$\mathcal{G}=\frac{\mathrm{E}|X_1-X_2|}{2\mu}=\frac{\int_{-\infty}^{\infty}[2xF(x)-x]\,\mathrm{d}F(x)}{\mu}=\frac{\int_{-\infty}^{\infty}2xF(x)\mathrm{d}F(x)}{\mu}-1.$$

2.2 密度比模型下的估计

假设有$m+1$个总体, 它们的累积分布函数和概率密度函数分别为$F_k(x)$和$f_k(x)$, 从中分别独立抽取样本, 观测的样本值为$\{x_{kj}\}$, 其中$k=0,1,\cdots,m$, $j=1,2,\cdots,n_k$, n_k为样本量. 密度比模型对这些分布进行了假设

$$\mathrm{d}F_k(x)=\exp\{\boldsymbol{\theta}_k^{\top}\boldsymbol{q}(x)\}\mathrm{d}F_0(x),\quad k=1,2,\cdots,m, \tag{1}$$

其中$\boldsymbol{\theta}=(\boldsymbol{\theta}_1^{\top},\boldsymbol{\theta}_2^{\top},\cdots,\boldsymbol{\theta}_m^{\top})$ 是未知的参数向量, $\boldsymbol{q}(x)$是维数为d的基函数. 为了简化符号, 通常取$\boldsymbol{\theta}_0=\mathbf{0}$. 式(1)中的$F_0$是无需指定的. 基函数$\boldsymbol{q}(x)$是线性的, 同时第一个元素通常取1以保证$\boldsymbol{\theta}_k$正则化. 非参数部分使得密度比模型的应用范围十分广泛. 通过调节基函数可以使密度比模型对对应的分布进行良好的拟合. 譬如, 总体服从正态分布族时可选取$\boldsymbol{q}(x)=(1,x,x^2)^{\top}$; 分析对数正态族时可选用$\boldsymbol{q}(x)=(1,\log x,\log^2 x)^{\top}$; 用$\boldsymbol{q}(x)=(1,x,\log x)^{\top}$可以对Gamma函数族进行较为精确的拟合. $\boldsymbol{q}(x)=(1,x,x^2,\log|x|,\log^2|x|)^{\top}$ 已涵盖了大部分分布族的需求. 此外, Zhang等人[27]还提出了自适应的方法来为未知的分布选取适合的基函数.

估计模型中参数部分的常用方法是经验似然法[27-28]. 已知每个样本$\{x_{k,j}\}$的样本量分别为n_k, 则总样本量为$n=n_0+n_1+\cdots+n_m$, 每个样本对总样本的占比为$\rho_k=n_k/n, k=0,1,\cdots,m$. 显然, $x_{k,1},x_{k,2},\cdots,x_{k,n_k}$是独立同分布的. 记$p_{k,j}=\mathrm{d}F_0(x_{k,j})$ 于是可以得到经验似然函数

$$\ell_n(\boldsymbol{\theta},p_{k,j})=\sum_{k,j}\log(p_{k,j})+\sum_{k,j}\boldsymbol{\theta}_k^{\top}\boldsymbol{q}(x_{kj}),\ k=0,1,\cdots,m,\ j=1,2,\cdots,n_k. \tag{2}$$

模型还假设存在约束条件

$$\int\exp\{\boldsymbol{\theta}_k^{\top}\boldsymbol{q}(x_{kj})\}\mathrm{d}G_0=1. \tag{3}$$

在约束式(3)下, 使似然函数最大的点就是所求的对$\boldsymbol{\theta}$和F_0的估计. 除了最大似然法, 对偶经验似然也是一种常用的估计方法. 对偶似然法通常是将约束等式直接加入原始函数中, 把原始问题转换为对偶问题, 在满足一定条件时, 对偶问题与原始问题等价. 在估计$\boldsymbol{\theta}$和F_0时, 把约束式(3)引入式(2)得到的对偶似然函数与经验法求得的结果相同, 并且对偶似然相对更加易于计算[21, 29]. 在对偶似然的方法下, $\widehat{\boldsymbol{\theta}}$是对偶似然函数式(4)的极大值点.

$$\widetilde{l}_n(\boldsymbol{\theta}, F_0) = \sum_{k,j} \log\left\{\sum_{r=0}^{m} \rho_r \left[\exp\left\{\boldsymbol{\theta}_r^{\top}\boldsymbol{q}(x_{kj})\right\}\right]\right\} + \sum_{k,j}\left\{\boldsymbol{\theta}_k^{\top}\boldsymbol{q}(x_{kj})\right\}. \tag{4}$$

求出$\hat{\boldsymbol{\theta}}$后, 便可以得到$\widehat{p}_{kj} = \left\{nh\left(x_{kj}; \widehat{\boldsymbol{\theta}}\right)\right\}^{-1}$, 其中$h(x;\boldsymbol{\theta}) = \sum_{k=0}^{m}\rho_k \exp\left\{\theta_k^{\top}\boldsymbol{q}(x)\right\}$. 记$\omega_k(x) = \exp\left\{\boldsymbol{\theta}_k^{\top}\boldsymbol{q}(x)\right\}$, 其估计为$\widehat{\omega}_k(x) = \exp\left\{\widehat{\boldsymbol{\theta}}_k^{\top}\boldsymbol{q}(x)\right\}$, 得到密度比模型下对$F_k$的估计是

$$\widehat{F}_k(x) = \sum_{r,j} \widehat{p}_{rj}\widehat{\omega}_k(x_{rj})\mathbf{I}(x_{rj} \le x), \tag{5}$$

其中$\mathbf{I}(A)$表示事件A的示性函数, $r = 0, 1, \cdots, m$, $j = 1, \cdots, n_r$.

根据定义1.1很容易推导出, 密度比模型下Gini系数的估计为

$$\widehat{\mathcal{G}}_k = \frac{\int_{-\infty}^{\infty} 2x\widehat{F}(x)\mathrm{d}\widehat{F}(x)}{\mu} - 1 = \frac{2\sum_{r,j}\widehat{p}_{rj}\widehat{\omega}_k(x_{rj})x_{rj}\left[\sum_{l,s}\widehat{p}_{ls}\widehat{\omega}_k(x_{ls})\mathbf{I}(x_{ls} \le x_{rj})\right]}{\sum_{r,j}\widehat{p}_{rj}\widehat{\omega}_k(x_{rj})x_{rj}} - 1,$$

其中$l = 0, 1, \cdots, m$, $s = 1, 2, \cdots, n_l$. 值得注意的是, 估计基准函数时, 根据密度比模型的结构, 式(5), (6)中的$\omega_0(x) = 1$.

§3 假设检验

在实际检验中, 本文关注的是高放废物砌块在存储前后均匀性的变化情况. 用F_0和F_1分别代表存储前后相关指标的累积分布函数, 并记它们的概率密度函数为f_0 和f_1, 相应的Gini系数为$\mathcal{G}_0$和$\mathcal{G}_1$. 工程中, 砌块越均匀表示其性质越稳定, 通常期望的结果是存储后均匀性不会显著变差, 即$\mathcal{G}_1 \le \mathcal{G}_0$. 据此本文提出原假设和备择假设

$$H_0 : \mathcal{G}_1 \le \mathcal{G}_0 \quad vs. \quad H_1 : \mathcal{G}_1 > \mathcal{G}_0.$$

根据提出的假设检验, 可以构造检验统计量

$$S = \sqrt{n}\left(\mathcal{G}_1 - \mathcal{G}_0\right) = \sqrt{n}\left[\frac{\int_{-\infty}^{\infty} 2xF_1(x)\mathrm{d}F_1(x)}{\mu_1} - \frac{\int_{-\infty}^{\infty} 2xF_0(x)\mathrm{d}F_0(x)}{\mu_0}\right].$$

为了简化, 记$\delta = \mathcal{G}_1 - \mathcal{G}_0$, 上述假设检验可转换为

$$H_0' : \delta \le 0 \quad vs. \quad H_1' : \delta > 0,$$

且有$S = \sqrt{n}\delta$.

3.3 基本条件及渐近性质

记实际观测到的样本值为$\{x_{0j}\}_{j=1}^{n_0}$和$\{x_{1j}\}_{j=1}^{n_1}$, 则两个样本各自对应的样本量为$n_r (r = 0, 1)$. 要使用密度比模型对数据进行估计, 对统计量的性质进行分析, 应满足以下基本的假设条件.

假设3.1 对$r = 0, 1$, 假设

1. F_0和F_1满足式(1)的结构;

2. 当样本总量$n \longrightarrow \infty$, $\rho_r = n_r/n$趋近于$(0,1)$内的某个常数;

3. 参数向量$\boldsymbol{q}(x)$线性独立, 且第一个元素为1;

4. $\int x^2 \mathrm{d}F_0 < \infty$, $\int \exp\left\{\boldsymbol{\theta}_0^{\mathrm{T}}\boldsymbol{q}(x)\right\}\mathrm{d}F_0 < \infty$, 且$\int x^2 \exp\left\{\boldsymbol{\theta}_0^{\mathrm{T}}\boldsymbol{q}(x)\right\}\mathrm{d}F_0 < \infty$, 在参数$\boldsymbol{\theta}$的真值$\boldsymbol{\theta}_0$的邻域内.

第二个条件要求样本容量n_0和n_1以相同的速率趋向无穷, 以防样本量过大时模型失效. 第三个条件可以确保模型是可识别的. 四个条件共同保障了密度比模型下对两总体的Gini系数的线性近似的可靠性. 在满足这些基本假设条件下, 可以通过第2.2节所述的方法, 用密度比模型分别估计出两种分布的Gini系数和统计量

$$\widehat{\mathcal{G}}_0 = \frac{2\sum_{r,j}\widehat{p}_{rj}x_{rj}\left[\sum_{l,s}\widehat{p}_{ls}\mathbf{I}(x_{ls}\le x_{rj})\right]}{\sum_{r,j}\widehat{p}_{rj}x_{rj}} - 1,$$

$$\widehat{\mathcal{G}}_1 = \frac{2\sum_{r,j}\widehat{p}_{rj}\widehat{\omega}(x)x_{rj}\left[\sum_{l,s}\widehat{p}_{ls}\widehat{\omega}(x)\mathbf{I}(x_{ls}\le x_{rj})\right]}{\sum_{r,j}\widehat{p}_{rj}\widehat{\omega}(x)x_{rj}} - 1,\ \widehat{S} = \sqrt{n\widehat{\delta}} = \sqrt{n}\left(\widehat{\mathcal{G}}_1 - \widehat{\mathcal{G}}_0\right),$$

其中$r = 0,1$, $j = 1,2,\cdots,n_r$, $l = 0,1$, $s = 1,2,\cdots,n_l$,

$$\widehat{\omega}(x) = \exp\left\{\widehat{\boldsymbol{\theta}}_1^{\top}\boldsymbol{q}(x)\right\},\ \widehat{p}_{kj} = \left\{nh\left(x_{kj};\widehat{\boldsymbol{\theta}}\right)\right\}^{-1}.$$

关于基于密度比模型的统计量, Yuan等人[23]讨论了分布半连续时的情形, 本文分析的连续分布情形为半连续的特殊情况. 根据Yuan等人[23]的定理4, 可知两分布密度比模型下的Gini系数之差是渐近正态的. 首先引入一些标记. 记

$$\omega(x) = \exp\left\{\boldsymbol{\theta}_1^{\top}\boldsymbol{q}(x)\right\},\quad h(x) = \rho_1\omega(x) - \rho_0,\quad h_1(x) = \rho_1\omega(x)/h(x),$$

$$u_0(x) = 2\left[xF_0(x) + \int_x^{\infty} y\mathrm{d}F_0(y) - \int_0^{\infty} xF_0(x)\mathrm{d}F_0(x)\right] - x,$$

$$u_1(x) = 2\left[xF_1(x) + \int_x^{\infty} y\mathrm{d}F_1(y) - \int_0^{\infty} xF_1(x)\mathrm{d}F_1(x)\right] - x,$$

$$\boldsymbol{J} = \begin{pmatrix} -\frac{\mathcal{G}_0}{\mu_0} & \frac{1}{\mu_0} & 0 & 0 \\ 0 & 0 & -\frac{\mathcal{G}_1}{\mu_1} & \frac{1}{\mu_1} \end{pmatrix},\quad \boldsymbol{A}_{\boldsymbol{\theta}} = \rho_0\mathrm{E}\left[h_1(X)\boldsymbol{q}(X)\boldsymbol{q}(X)^{\top}\middle|F_0\right],$$

其中$\mathrm{E}(\cdot|F_0)$表示分布F_0下的期望, X表示服从分布F_0的随机变量. 用$\xrightarrow{d}$表示依分布收敛, 则两分布密度比模型下的Gini系数之差的渐近性质如下.

引理3.1[23] 在假设3.1的条件下, 当$n \to \infty$时, $\sqrt{n}\left(\widehat{\delta} - \delta\right) \xrightarrow{d} N\left(0,\sigma^2\right)$, 其中

$$\sigma^2 = (-1,1)\,\boldsymbol{\Sigma}\,(-1,1)^{\top},\quad \boldsymbol{\Sigma} = \boldsymbol{J}\left[\mathrm{E}\left(\frac{\boldsymbol{u}(X)\boldsymbol{u}(X)^{\top}}{h(X)}\middle|F_0\right) + \frac{1}{\rho_1^2}\boldsymbol{B}\right]\boldsymbol{J}^{\mathrm{T}},$$

$$\boldsymbol{u}(x) = (x, u_0(x), \omega(x), \omega(x)u_0(x))^{\top},\quad \tilde{\boldsymbol{u}}_0(x) = -\rho_1(x, u_0(x))^{\top},$$

$$\tilde{\boldsymbol{u}}_1(x) = \rho_0(x, u_1(x))^{\top},\quad \tilde{\boldsymbol{u}}_(x) = (\tilde{\boldsymbol{u}}_0(x)^{\top}, \tilde{\boldsymbol{u}}_1(x)^{\top})^{\top},$$

$$\boldsymbol{B} = \mathrm{E}\left[h_1(X)\tilde{\boldsymbol{u}}_(X)\boldsymbol{q}(X)^{\top}\middle|F_0\right]\boldsymbol{A}_{\boldsymbol{\theta}}^{-1}\mathrm{E}\left[h_1(X)\boldsymbol{q}(X)\tilde{\boldsymbol{u}}_(X)^{\top}\middle|F_0\right].$$

显然引理3.1即密度比模型下统计量的渐近性质, $\widehat{S} - S = \sqrt{n}\left(\widehat{\delta} - \delta\right) \xrightarrow{d} N\left(0,\sigma^2\right)$. 据此可进一步推断出检验统计量分别在原假设为真和备择假设为真的条件下的渐近性.

定理3.1 在假设3.1的条件下, 当$n \to \infty$时,

1. 若$H_0^{'}$为真, 则$\widehat{S} \leq \widehat{S} - S$, 而$\widehat{S} - S \xrightarrow{d} N\left(0, \sigma^2\right)$, 其中$\sigma^2$如引理3.1所述, 该分布在上$\alpha$分位点处为有限正常数, $0 < \alpha < 1/2$;

2. 若$H_1^{'}$为真, 则$\widehat{S} \to \infty$.

证 1. 在$H_0^{'}$为真的情况下, $\delta \leq 0$, 故$S = \sqrt{n}\delta \leq 0$, 从而易得$\widehat{S} \leq \widehat{S} - S$. 根据引理3.1, 当$n \to \infty$时, $\widehat{S} - S \xrightarrow{d} N\left(0, \sigma^2\right)$. 根据正态分布的性质, 易知$N\left(0, \sigma^2\right)$在$0 < \alpha < 1/2$的分位点上为有限正常数. 2. 在$H_1^{'}$为真的情况下, $\widehat{\delta} \xrightarrow{d} \delta > 0$, 故$n \to \infty$时, $\sqrt{n}\widehat{\delta} \to \infty$.

3.4 检验步骤

3.4.1 Wald-型检验

Wald-型检验方法是已知统计量的抽样分布的情况下最常见的检验方法之一, 原理直观易于理解, Yuan等人[23]在数据模拟中使用此法进行了双边检验. 它是根据统计量服从的正态分布确定的临界值[31]. 由引理3.1, 可以直接得到

$$(\widehat{S} - S)/\sigma \xrightarrow{d} N(0,1).$$

本文要分析的是一个典型的右侧检验问题, 则所求的p值为$p = 1 - \Phi[(\widehat{S} - S)/\sigma]$. $\Phi(x)$表示标准正态的分布函数. 在给定置信水平α的情况下, 倘若$p < \alpha$, 则拒绝原假设. 由于σ中含有未知参数和统计量, 因此需要进行估计. 用密度比模型所得的相应估计值替代σ中的未知参数, 可以得到σ的一致估计$\widehat{\sigma}$, 且$(\widehat{S} - S)/\widehat{\sigma} \xrightarrow{d} N(0,1)$. 则实际检验中, p值的估计为

$$\widehat{p} = 1 - \Phi[(\widehat{S} - S)/\widehat{\sigma}].$$

若$\widehat{p} < \alpha$, 则拒绝原假设.

3.4.2 自助法检验

自助法检验是一种结合重抽样进行假设检验的方法, 常用于对原始样本进行扩容以及抽样分布未知的情形, 当原始样本质量较好时, 相较于直接使用原始样本的检验方法, 可以提升功效，并更具稳健性[24]. 因此本文考虑使用此法进一步保障检验结果的可靠性. 将从原始样本$\{x_{0j}\}_{j=1}^{n_0}$ 和$\{x_{1j}\}_{j=1}^{n_1}$ 中分别有放回地独立随机抽取n_0, n_1 个样本, 记为$\{x_{0j}^*\}_{j=1}^{n_0}$和$\{x_{1j}^*\}_{j=1}^{n_1}$. 则每一次重抽样后, 根据前面2.2节的方法, 用密度比模型估计出Gini系数和统计量为

$$\widehat{\mathcal{G}}_0^* = \frac{2\sum_{r,j}\widehat{p}_{rj}^* x_{rj}^* \left[\sum_{l,s}\widehat{p}_{ls}^* \mathbf{I}(x_{ls}^* \leq x_{rj}^*)\right]}{\sum_{r,j}\widehat{p}_{rj}^* x_{rj}^*} - 1,$$

$$\widehat{\mathcal{G}}_1^* = \frac{2\sum_{r,j}\widehat{p}_{rj}^*\widehat{\omega}(x)^* x_{rj}^* \left[\sum_{l,s}\widehat{p}_{ls}^*\widehat{\omega}(x)^* \mathbf{I}(x_{ls}^* \leq x_{rj}^*)\right]}{\sum_{r,j}\widehat{p}_{rj}^*\widehat{\omega}(x)^* x_{rj}^*} - 1,$$

$$\widehat{S}^* = \sqrt{n}\left(\widehat{\mathcal{G}}_1^* - \widehat{\mathcal{G}}_0^*\right),$$

$\widehat{\cdot}^*$表示相应统计量在重抽样数据下通过密度比模型得到的估计值.

参考Zhuang等人[32]的定理2, 可以直接得到下述结论.

定理3.2 在假设3.1的条件下, 当$n \to \infty$时,

$$\sup_x |\mathbf{P}^*\{\widehat{S}^* - \widehat{S} \le x\} - \mathbf{P}\{\widehat{S} - S \le x\}| = \mathbf{o}_p(1),$$

其中$\mathbf{P}^*$表示给定样本下的条件概论.

再结合定理3.1, 可以估计p值

$$\widehat{p} = P\left\{\widehat{S}^* - \widehat{S} \ge \widehat{S} \middle| \{x_{0j}\}_{j=1}^{n_0}, \{x_{1j}\}_{j=1}^{n_1}\right\}.$$

由于真实的概率无法计算, 因此实际中借助Monte Carlo方法来对p值进行逼近, 将重抽样重复B次, 记每次重抽样估计的统计量为$\widehat{S}_b^*(b-1, 2, \cdots, B)$, 拟合的结果为

$$\widehat{p} \simeq \frac{1}{B}\sum_{b=1}^{B} \mathbf{I}(\widehat{S}_b^* - \widehat{S} \ge \widehat{S}).$$

在给定置信水平α的情况下, 若$\widehat{p} < \alpha$, 则拒绝原假设.

§4 结果分析

本文的原始数据由兰州大学环境岩土工程团队提供. 该团队提出了用覆膜储存法来解决储存过程中缓冲砌块质量劣化的难点, 并进行了为期5个月的存储试验. 试验将缓冲材料划分出64个面积相等的区域并进行编号, 再分别独立测量出64块砌块的含水率和干密度. 砌块的含水率和干密度的主要信息如表1所示(保留四位小数).

表 1 存储前后的含水率(%)和干密度(g/cm^3)

	存储前			存储后		
	均值	标准差	离散系数	均值	标准差	离散系数
含水率	16.5800	0.8946	5.3957	15.8800	0.7677	4.8342
干密度	1.7134	0.0317	1.8473	1.6946	0.0386	2.2753

本文分别计算出覆膜储存试验前后含水率和干密度的Gini系数, 先按照第3.4.1节的wald-型检验法进行初步的分析, 再借助第3.4.2节的自助法进一步保障检验结果的可靠性, 通过含水率和干密度的变化情况来判断缓冲材料的均匀性是否显著变差, 从而判断覆膜储存法是否能有效保护高放废物缓冲材料使环境影响降到可以接受的范围.

根据经验, 本文在使用密度比模型进行估计时采用的基函数是高斯分布所对应的$\boldsymbol{q}(x) = (1, x, x^2)^\top$. 在两种检验中, 置信水平$\alpha$ 都设为工程学中常用的0.05. 在进行自助法检验时, 设定重抽样次数为$B = 1000$. 经过一系列运算, 将得到的结果保留四位小数后, 整理制成表2.

表 2 存储前后的Gini系数和假设检验的p值

	基尼系数		p值	
	试验前	试验后	Wald-型检验	自助法检验
含水率	0.0180	0.0198	0.3270	0.1190
干密度	0.0179	0.0198	0.3150	0.0890

表1中砌块在存储后的含水率的标准差和离散系数分别减小了14.19%和10.41%, 看似变得均匀许多. 然而, 表2中根据原始数据得到的含水率在存储前后的Gini系数分别为0.0180

和0.0198, 存储后的数据增加了10.00%, 说明存储后的含水率没有存储前均匀. 指标间的矛盾之处在于标准差可能随着含水率整体降低而减小, 离散系数对最大值和最小值的影响赋予较大的权重, 而Gini系数是着眼于整体的, 由此可知, 基尼系数的引入有助于排查出某些容易被标准差和离散系数所忽视的恶化情况. 从存储前后含水率的Gini系数的绝对值来看, 两者都非常接近于0, 说明存储后含水率的分布仍然保持在一个较为均匀的水平, 但要确定降低是否显著, 即存储后的含水率是否没有变得显著不均, 还需要进一步进行假设检验. 干密度在存储后的标准差增加了21.81%, 离散系数增大了23.16%, 这两个指标表明存储后的干密度不均程度变大了. 再看表2, 存储前后的Gini系数分别为0.0179 和0.0198, 存储后的数据增加了10.61%, 同样说明存储后的干密度没有存储前均匀. 与含水率的情况类似, 从存储前后干密度的基尼系数的绝对值来看, 两者都非常接近于0, 说明存储后干密度的分布仍然保持在一个较为均匀的水平, 而要确定降低是否显著, 需要进一步进行假设检验.

先看对含水率样本进行假设检验的结果: Wald-型检验的p值为0.3270, 大于给定的置信水平$\alpha = 0.05$; 自助法得到的p值为0.1190, 同样大于给定的置信水平$\alpha = 0.05$. 干密度的检验结果类似, Wald-型检验的p值为0.3150, 自助法得到的p值为0.0890, 均大于给定的置信水平$\alpha = 0.05$. 比较两种检验结果的数据, 无论是对含水率还是对干密度进行检验, 重抽样方法下的p值都减小许多. 这是因为原始样本容量只有64的情况下, Wald-型检验有一定的纳伪风险, 而自助法能够提高检验功效, 这对高度注重安全性的核工程是十分重要的. 从检验结果来看, 虽然两种检验方法得到的p值在数值上有差异, 但都小于给定的置信水平, 因此得到的检验结果一致, 都表明没有足够的证据可以推翻对含水率和对干密度提出得原假设, 即相较存储前, 存储后的含水率和干密度的不均程度被控制在工程学认定的合理范围内. 进一步可知, 覆膜储存法使缓冲材料的均匀性保持了良好的稳定性.

综上所述, 在经历5个月的存储后, 缓冲材料始终保持着含水率和干密度的基尼系数都十分接近0的状态, 没有显著变得更加不均匀. 可知覆膜储存法对高放废物缓冲材料可以起到一定的保护作用, 降低环境对缓冲材料均匀性的影响程度, 使其发生的变化保持在工程学的可接受范围内.

§5 结论

本文考虑了在核安全工程中使用一种基于半参数模型的均匀性指标来分析高放废物缓冲材料的均匀性变化情况, 给出了密度比模型下的Gini系数的通用表达式; 构造出相应的检验统计量, 并讨论了统计量的渐近性质, 为实际应用提供理论支撑; 用密度比模型估计了实际试验中的缓冲砌块样本的系数, 通过比较样本的标准差, 离散系数和Gini系数所得到的结论, 论证了引入Gini系数的必要性和重要性; 最后分别通过Wald-型检验和自助法对实际试验样本进行了假设检验, 结果表明缓冲材料在经过兰州大学提出的覆膜储存法养护存储后, 含水率和干密度的Gini系数没有显著变得不均, 可知覆膜储存法对高放废物缓冲材料可以起到一定的保护作用, 使缓冲材料保持了较为良好的均匀性.

致谢 感谢国家自然科学基金(No. 71971204)和安徽省杰青项目(No. 2208085J43)对本文提供的支持, 感谢兰州大学环境岩土工程团队提供的原始数据.

参考文献:

[1] 王驹. 中国高放废物地质处置21世纪进展[J]. 原子能科学技术, 2019, 53(10): 2072-2082.

[2] 张虎元, 张国超, 余荣光, 等. 混合型缓冲砌块膨胀性的空间分布及其各向异性[J]. 岩石力学与工程学报, 2019, 38(增(2)): 3469-3480.

[3] Zhang Huyuan, Tan Yu, Zhu Fei, et al. Shrinkage property of bentonite-sand mixtures as influenced by sand content and water salinity[J]. Construction and Building Materials, 2019, 224(C): 78-88.

[4] Tan Yu, Zhang Huyuan, He Dongjin, et al. Deterioration of exposed buffer block: desiccation shrinkage and cracking[J]. Bulletin of Engineering Geology and the Environment, 2019, 78(7): 5431-5444.

[5] Tan Yu, Zhang Huyuan, Wang Ying. Evaporation and shrinkage processes of compacted bentonite-sand mixtures[J]. Soils and Foundations, 2020, 60(2): 505-519.

[6] Tan Yu, Zhou Guangping, Ding Zhinan, et al. Effect of desiccation on the hydraulic conductivity of compacted bentonite–sand blocks[J]. IOP Conference Series: Earth and Environmental Science, 2021, 861(4): 042122. doi: 10.1088/1755-1315/861/4/042122.

[7] Tan Yu, Zhang Huyuan, Zhang Tongwei, et al. Anisotropic hydro-mechanical behavior of full-scale compacted bentonite-sand blocks[J]. Engineering Geology, 2021, 287: 106093. doi: 10.1016/j.enggeo.2021.106093.

[8] Tan Yu, Zhang Huyuan, Ding Zhinan, et al. Particle Size Effects on the Volumetric Shrinkage of Bentonite–Sand Mixtures[J]. International Journal of Geomechanics, 2022, 22(8): 06022019. doi: 10.1061/(ASCE)GM.1943-5622.0002447.

[9] Gini C. Measurement of inequality of incomes[J]. The Economic Journal, 1912, 31(124): 124-126.

[10] MacDonald J, Mohler G, Brantingham P J. Association between race, shooting hot spots, and the surge in gun violence during the COVID-19 pandemic in Philadelphia, New York and Los Angeles[J]. Preventive Medicine, 2022, 165: 107241. doi: 10.1016/j.ypmed.2022.107241.

[11] 曾静, 沙治慧. 我国生育保障水平测度及其区域差异研究–基于31省(自治区, 直辖市)数据的实证分析[J]. 中国卫生政策研究, 2022, 15(4): 17-23.

[12] Algehyne E A, Jibril M L, Algehainy N A, et al. Fuzzy neural network expert system with an improved Gini index random forest-based feature importance measure algorithm for early diagnosis of breast cancer in Saudi Arabia[J]. Big Data and Cognitive Computing, 2022, 6(1): 13. doi: 10.3390/bdcc6010013.

[13] Miao Yonghao, Wang Jingjing, Zhang Boyao, et al. Practical framework of Gini index in the application of machinery fault feature extraction[J]. Mechanical Systems and Signal Processing, 2022, 165: 108333. doi: 10.1016/j.ymssp.2021.108333.

[14] Xiong Qiangqiang, Liu Yaolin, Xing Lijun, et al. Measuring spatio-temporal disparity of location-based accessibility to emergency medical services[J]. Health & Place, 2022, 74: 102766. doi: 10.1016/j.healthplace.2022.102766.

[15] 黄晓芬, 白鸥. 浙江省森林乡村空间分布特征及其影响因素[J]. 浙江农林大学学报, 2022, 39(4): 884-893.

[16] 李敏, 唐杰, 吴中明. 交通拥堵管理中可交易电子券方案的公平性分析[J]. 运筹与管理, 2022, 31(5): 86-92.

[17] Anderson J A. Separate sample logistic discrimination[J]. Biometrika, 1972, 59(1): 19-35.

[18] Anderson J A. Multivariate logistic compounds[J]. Biometrika, 1979, 66(1): 17-26.

[19] Owen A B. Empirical likelihood ratio confidence intervals for a single functional[J]. Biometrika, 1988, 75(2): 237-249.
[20] Qin Jing, Zhang Biao. A goodness-of-fit test for logistic regression models based on case-control data[J]. Biometrika, 1997, 84(3): 609-618.
[21] Keziou A, Leoni-Aubin S. On empirical likelihood for semiparametric two-sample density ratio models[J]. Journal of Statistical Planning and Inference, 2008, 138(4): 915-928.
[22] Chen Jiahua, Liu Yukun. Quantile and quantile-function estimations under density ratio model[J]. The Annals of Statistics, 2013, 41(3): 1669-1692.
[23] Yuan Meng, Li Pengfei, Wu Changbao. Semi-parametric inference on Gini indices of two semi-continuous populations under density ratio models[J]. The Econometrics Journal, 2023, 26(2): 174-188.
[24] Shao Jun, Tu Dongsheng. The jackknife and bootstrap[M]. Berlin: Springer Science & Business Media, 2012.
[25] Lorenz M O. Methods of measuring the concentration of wealth[J]. Publications of the American statistical association, 1905, 9(70): 209-219.
[26] David H A. Miscellanea: Gini's mean difference rediscovered[J]. Biometrika, 1968, 55(3): 573-575.
[27] Zhang Archer Gong, Chen Jiahua. Density ratio model with data-adaptive basis function[J]. Journal of Multivariate Analysis, 2022, 191(C): 105043. doi: 10.1016/j.jmva.2022.105043.
[28] Fokianos K, Kedem B, Qin Jing, et al. A semiparametric approach to the one-way layout[J]. Technometrics, 2001, 43(1): 56-65.
[29] Owen A B. Empirical Likelihood[M]. London: Chapman and Hall/CRC, 2001.
[30] Li Huapeng, Liu Yang, Liu Yukun, et al. Comparison of empirical likelihood and its dual likelihood under density ratio model[J]. Journal of Nonparametric Statistics, 2018, 30(3): 581-597.
[31] Qin Yongsong, Rao J N K, Wu Changbao. Empirical likelihood confidence intervals for the Gini measure of income inequality[J]. Economic Modelling, 2010, 27(6): 1429-1435.
[32] Zhuang Weiwei, Hu Boyi, Chen Jiahua. Semiparametric inference for the dominance index under the density ratio model[J]. Biometrika, 2019, 106(1): 229-241.

Homogeneity analysis of buffer barriers of high-level radioactive waste repositories by using the Gini coefficient under the density ratio model

LIAO Wen-chen[1], TAN Yu[2], ZHUANG Wei-wei[1]

(1. School of Management, University of Science and Technology of China, Hefei 230041, China;
2. College of Civil Engineering and Mechanics, Lanzhou University, Lanzhou 730000, China)

Abstract: The density ratio model based estimator of the Gini coefficient and hypothesis testing on this basis are introduced to nuclear safety field to measure the homogeneity of buffer barriers. The asymptotic theory involved in hypothesis testing is discussed. In addition to analysis by the Wald-type test, a valid Bootstrap process is designed to improve power. The above mentioned methods are applied to real data of the curing method called film covering storage method. The results suggest that film covering storage method can effectively protect the buffer barrier from becoming more inhomogeneous.

Keywords: Gini coefficient; density ratio model; bootstrap test; nuclear safety

MR Subject Classification: 62P12; 62P30