敏感性问题中具有多种选择的一种随机化回答调查法*

孙 汉 杰 梁 小 筠
(华东师范大学统计系, 上海, 200062)

摘 要

Franklin,L.A.对于仅有两种选择的敏感性问题给出了一种以连续分布作为随机化分布的随机化回答调查法. 本文从Franklin的调查法出发,探讨了在具有多种选择敏感性问题调查中,以连续分布作为随机化分布的随机化回答调查法. 在这种调查法中,被调查者只需回答来自指定分布的随机数即可,不需要正面回答调查问题.

一、调查方法

一般情况下,假设调查总体可以分为互不相容的M类. 总体中的每个人属于且只属于其中的一类. 我们所要关心的是各类在总体中所占的比例: $\theta_1,\theta_2,\cdots,\theta_M$, 且 $\sum_{i=1}^{M}\theta_i=1$. 为估计 $\theta_1,\theta_2,\cdots,\theta_M$, 这里给出两种调查方法.

调查方法一:

从总体中独立地抽取M-1组容量分别为: $n_1,n_2,\cdots,n_{M-1}$ 的简单随机样本. 对第 $j(j=1,2,\cdots,M-1)$ 组的每个人分别进行 $k_j(k_j\geq 1)$ 次试验, 在第 $t(t=1,2,\cdots,k_j)$ 次试验中,个体 $i(i=1,2,\cdots,n_j)$ 的回答值为来自密度函数分别为: $g_{jt}^{(1)},g_{jt}^{(2)},\cdots,g_{jt}^{(M)}$ 的M个随机数中的一个. 本文中假设 $g_{jt}^{(s)}$ 为正态分布,其均值为 $\mu_{jt}^{(s)}$, 方差为 $\sigma_{jt}^{2(s)}$. 这里还要求 $|A|\neq 0$,其中

$$A=\begin{pmatrix} \sum_{t=1}^{k_1}\mu_{1t}^{(1)} & \sum_{t=1}^{k_1}\mu_{1t}^{(2)} & \cdots & \sum_{t=1}^{k_1}\mu_{1t}^{(M)} \\ \sum_{t=1}^{k_2}\mu_{2t}^{(1)} & \sum_{t=1}^{k_2}\mu_{2t}^{(2)} & \cdots & \sum_{t=1}^{k_2}\mu_{2t}^{(M)} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{t=1}^{k_{M-1}}\mu_{M-1t}^{(1)} & \sum_{t=1}^{k_{M-1}}\mu_{M-1t}^{(2)} & \cdots & \sum_{t=1}^{k_{M-1}}\mu_{M-1t}^{(M)} \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

(注:其实这里 $g_{jt}^{(s)}(j=1,2,\cdots,M-1,\ t=1,2,\cdots,k_j,\ s=1,2,\cdots,M)$ 的选取可以是任意的, 只要他们的均值及方差存在,并且均值满足 $|A|\neq 0$即可.)

在调查过程中,只有被调查者才能看见这M个随机数,并确切地知道这M个随机数所分别对应的密度函数, 而调查者仅仅知道这M个密度函数的具体形式. 调查时, 如果被调查者属于第 $s(s=1,2,\cdots,M)$ 类,则要求他回答来自密度函数为 $g_{jt}^{(s)}$ 的随机数. 并记他的回答值为 z_{jti},这里 $j=1,2,\cdots,M-1,\ t=1,2,\cdots,k_j,\ i=1,2,\cdots,n_j$.

调查方法二:

如果总体不大或由于其它原因导致样本量不能取得充分大时, 可考虑采用如下的抽样调查方法: 从总体中抽取样本量为n的简单随机样本,对每个样本分别进行k次试验. 在第 $t(t=1,2,\cdots,k)$ 次试验时,个体 $i(i=1,2,\cdots,n)$ 的回答值为如下的 M组随机数中的一组,

$$(z_{1t}^{(1)},z_{2t}^{(1)},\cdots,z_{M-1t}^{(1)}),(z_{1t}^{(2)},z_{2t}^{(2)},\cdots,z_{M-1t}^{(2)}),\ \cdots(z_{1t}^{(M)},z_{2t}^{(M)},\cdots,z_{M-1t}^{(M)})$$

*本文受到国家社科基金项目98BT3003的资助

这里 $z_{jt}^{(s)}$ 为取自密度函数为 $g_{jt}^{(s)}$ 的随机数, $(j=1,2,\cdots,M-1,\ \ s=1,2,\cdots,M)$. 并且这些随机数的产生是相互独立的.

对 $g_{jt}^{(s)}(t=1,2,\cdots,k.\ \ j=1,2,\cdots,M-1.\ \ s=1,2,\cdots,M.)$ 的要求类似于调查方法一.

调查过程中,只有被调查者才能看见这M组随机数,并确切地知道每组随机数所对应的密度函数组 $(g_{1t}^{(s)},\cdots,g_{M-1t}^{(s)})\ (s=1,2,\cdots,M)$,而调查者仅仅知道这些密度函数的确切形式. 调查时,如果被调查者属于第 $s(s=1,2,\cdots,M)$ 类, 则要求他回答 $(g_{1t}^{(s)},\cdots,g_{M-1t}^{(s)})$ 所对应的随机数组,记他的回答值为 $(z_{1ti},z_{2ti},\cdots,z_{M-1ti})$.

二　估计量及其方差

定理一　在使用调查方法一的情况下,

$$\underset{\sim}{\hat{\theta}}=(\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_{M-1},\hat{\theta}_M)'=A^{-1}(\sum_{t=1}^{k_1}\bar{z}_{1t\cdot},\sum_{t=1}^{k_2}\bar{z}_{2t\cdot},\cdots,\sum_{t=1}^{k_{M-1}}\bar{z}_{M-1t\cdot},1)' \tag{2.1}$$

为 $\underset{\sim}{\theta}=(\theta_1,\theta_2,\cdots,\theta_{M-1},\theta_M)'$ 的无偏估计. 其中, $\bar{z}_{jt\cdot}=\frac{1}{n_j}\sum_{i=1}^{n_j}z_{jti}\quad j=1,2,\cdots,M-1$.

证明:由于被调查者属于M类中的一类,第s类所对应的密度函数为 $g_{jt}^{(s)}(s=1,2,\cdots,M)$, 所以第j组第i个人第t次回答的答案 z_{jti} 所对应的密度函数为:

$$\theta_1 g_{jt}^{(1)}+\theta_2 g_{jt}^{(2)}+\cdots+\theta_M g_{jt}^{(M)} \tag{2.2}$$

并且 $z_{jt1},z_{jt2},\cdots,z_{jtn_j}$ 独立同分布,因而

$$\mathsf{E}\bar{z}_{jt\cdot}=\frac{1}{n_j}\sum_{i=1}^{n_j}\mathsf{E}z_{jti}=\mathsf{E}z_{jti}=\theta_1\mu_{jt}^{(1)}+\theta_2\mu_{jt}^{(2)}+\cdots+\theta_M\mu_{jt}^{(M)}$$

$$\mathsf{E}(\sum_{t=1}^{k_j}\bar{z}_{jt\cdot})=\theta_1\sum_{t=1}^{k_j}\mu_{jt}^{(1)}+\theta_2\sum_{t=1}^{k_j}\mu_{jt}^{(2)}+\cdots+\theta_M\sum_{t=1}^{k_j}\mu_{jt}^{(M)}$$

$$\mathsf{E}\begin{pmatrix}\sum_{t=1}^{k_1}\bar{z}_{1t\cdot}\\ \sum_{t=1}^{k_2}\bar{z}_{2t\cdot}\\ \vdots\\ \sum_{t=1}^{k_{M-1}}\bar{z}_{M-1t\cdot}\\ 1\end{pmatrix}=\begin{pmatrix}\sum_{t=1}^{k_1}\mu_{1t}^{(1)} & \sum_{t=1}^{k_1}\mu_{1t}^{(2)} & \cdots & \sum_{t=1}^{k_1}\mu_{1t}^{(M)}\\ \sum_{t=1}^{k_2}\mu_{2t}^{(1)} & \sum_{t=1}^{k_2}\mu_{2t}^{(2)} & \cdots & \sum_{t=1}^{k_2}\mu_{2t}^{(M)}\\ \cdots & \cdots & \cdots & \cdots\\ \sum_{t=1}^{k_{M-1}}\mu_{M-1t}^{(1)} & \sum_{t=1}^{k_{M-1}}\mu_{M-1t}^{(2)} & \cdots & \sum_{t=1}^{k_{M-1}}\mu_{M-1t}^{(M)}\\ 1 & 1 & \cdots & 1\end{pmatrix}\begin{pmatrix}\theta_1\\ \theta_2\\ \vdots\\ \theta_{M-1}\\ \theta_M\end{pmatrix}=A\underset{\sim}{\theta}$$

所以, $\mathsf{E}\underset{\sim}{\theta}=A^{-1}A\underset{\sim}{\theta}=\underset{\sim}{\theta}$.

引理:　在使用调查方法一的情况下,

$$Var(\sum_{t=1}^{k_j}\bar{z}_{jt\cdot})=\frac{1}{n_j}\{\sum_{s=1}^{M}\sum_{r>s}^{M}\theta_s\theta_r(\mu_j^{(s)}-\mu_j^{(r)})^2+\sum_{s=1}^{M}\theta_s\sigma_j^{2(s)}\} \tag{2.3}$$

其中,$\mu_j^{(s)}=\sum_{t=1}^{k_j}\mu_{jt}^{(s)}$、$\sigma_j^{2(s)}=\sum_{t=1}^{k_j}\sigma_{jt}^{2(s)}$、$j=1,2,\cdots,M-1$、$s=1,2,\cdots,M$. 并记 $\mu_{jt}^{2(s)}=(\mu_{jt}^{(s)})^2$、$j=1,2,\cdots,M-1$、$s=1,2,\cdots,M$.

证明: 由(2.2)式及$\sum_{s=1}^{M}\theta_s=1$,得

$$\begin{aligned}Var(z_{jt_l})&=\theta_1(\mu_{jt}^{2(1)}+\sigma_{jt}^{2(1)})+\theta_2(\mu_{jt}^{2(2)}+\sigma_{jt}^{2(2)})+\cdots+\theta_M(\mu_{jt}^{2(M)}+\sigma_{jt}^{2(M)})\\&\quad-[\theta_1\mu_{jt}^{(1)}+\theta_2\mu_{jt}^{(2)}+\cdots+\theta_M\mu_{jt}^{(M)}]^2\\&=\sum_{s=1}^{M}\sum_{r=1}^{M}\theta_s\theta_r[\mu_{jt}^{2(s)}-\mu_{jt}^{(s)}\mu_{jt}^{(r)}]+\sum_{s=1}^{M}\theta_s\sigma_{jt}^{2(s)}\\&=\sum_{s=1}^{M}\sum_{r>s}^{M}\theta_s\theta_r[\mu_{jt}^{(s)}-\mu_{jt}^{(r)}]^2+\sum_{s=1}^{M}\theta_s\sigma_{jt}^{2(s)}\end{aligned}\tag{2.4}$$

另外,当$1\le l<t\le k_j$时,对第j组提出的第l个和第t个问题的答案独立,所以(z_{jl_l},z_{jt_l})的联合密度函数为:

$\theta_1 g_{jl}^{(1)}g_{jt}^{(1)}+\theta_2 g_{jl}^{(2)}g_{jt}^{(2)}+\cdots+\theta_M g_{jl}^{(M)}g_{jt}^{(M)}$ $j=1,2,\cdots,M-1$. 故

$$\mathrm{E}(z_{jl_l}z_{jt_l})=\theta_1\mu_{jl}^{(1)}\mu_{jt}^{(1)}+\theta_2\mu_{jl}^{(2)}\mu_{jt}^{(2)}+\cdots+\theta_M\mu_{jl}^{(M)}\mu_{jt}^{(M)}=\sum_{s=1}^{M}\theta_s\mu_{jl}^{(s)}\mu_{jt}^{(s)}\tag{2.5}$$

$$\begin{aligned}\mathrm{E}z_{jl_l}\mathrm{E}z_{jt_l}&=[\theta_1\mu_{jl}^{(1)}+\theta_2\mu_{jl}^{(2)}+\cdots+\theta_M\mu_{jl}^{(M)}][\theta_1\mu_{jt}^{(1)}+\theta_2\mu_{jt}^{(2)}+\cdots+\theta_M\mu_{jt}^{(M)}]\\&=\sum_{s=1}^{M}\sum_{r=1}^{M}\theta_s\theta_r\mu_{jl}^{(s)}\mu_{jt}^{(r)}\end{aligned}\tag{2.6}$$

$$\begin{aligned}Cov(z_{jl_l},z_{jt_l})&=\sum_{s=1}^{M}\theta_s\mu_{jl}^{(s)}\mu_{jt}^{(s)}-\sum_{s=1}^{M}\sum_{r=1}^{M}\theta_s\theta_r\mu_{jl}^{(s)}\mu_{jt}^{(r)}=\sum_{s=1}^{M}\sum_{r=1}^{M}\theta_s\theta_r[\mu_{jl}^{(s)}\mu_{jt}^{(s)}-\mu_{jl}^{(s)}\mu_{jt}^{(r)}]\\&=\sum_{s=1}^{M}\sum_{r>s}^{M}\theta_s\theta_r(\mu_{jl}^{(s)}-\mu_{jl}^{(r)})(\mu_{jt}^{(s)}-\mu_{jt}^{(r)})\end{aligned}\tag{2.7}$$

根据(2.4)及(2.7)式及$z_{jt1},z_{jt2},\cdots,z_{jtn_j}$独立同分布可得到,

$$\begin{aligned}Var(\sum_{t=1}^{k_j}\bar{z}_{jt})&=\frac{1}{n_j}[\sum_{t=1}^{k_j}Var(z_{jt_l})+2\sum_{l<t}^{k_j}Cov(z_{jl_l},z_{jt_l})]\\&=\frac{1}{n_j}[\sum_{t=1}^{k_j}\sum_{s=1}^{M}\sum_{r>s}^{M}\theta_s\theta_r(\mu_{jt}^{(s)}-\mu_{jt}^{(r)})^2+\sum_{t=1}^{k_j}\sum_{s=1}^{M}\theta_s\sigma_{jt}^{2(s)}\\&\quad+2\sum_{l<t}^{k_j}\sum_{s=1}^{M}\sum_{r>s}^{M}\theta_s\theta_r(\mu_{jl}^{(s)}-\mu_{jl}^{(r)})(\mu_{jt}^{(s)}-\mu_{jt}^{(r)})]\\&=\frac{1}{n_j}\{\sum_{s=1}^{M}\sum_{r>s}^{M}\theta_s\theta_r[\sum_{t=1}^{k_j}(\mu_{jt}^{(s)}-\mu_{jt}^{(r)})]^2+\sum_{t=1}^{k_j}\sum_{s=1}^{M}\theta_s\sigma_{jt}^{2(s)}\}\\&=\frac{1}{n_j}[\sum_{s=1}^{M}\sum_{r>s}^{M}\theta_s\theta_r(\mu_j^{(s)}-\mu_j^{(r)})^2+\sum_{s=1}^{M}\theta_s\sigma_j^{2(s)}]\end{aligned}$$

定理二：在使用调查方法一的情况下，$\hat{\theta}$的协方差矩阵有如下形式，

$$V(\hat{\theta}) = A^{-1}\begin{pmatrix}\Sigma_1 & & & 0\\ & \Sigma_2 & & \\ & & \cdots & \\ 0 & & & \Sigma_{M-1}\end{pmatrix}(A^{-1})' \tag{2.8}$$

其中，$\Sigma_j = Var(\sum_{t=1}^{k_j}\varepsilon_{jt}) = \frac{1}{n_j}[\sum_{s=1}^{M}\sum_{r>s}^{M}\theta_s\theta_r(\mu_j^{(s)}-\mu_j^{(r)})^2 + \sum_{s=1}^{M}\theta_s\sigma_j^{2(s)}]$　$j = 1,2,\cdots,M-1$.

对于调查方法二可以得到类似的结论.

而 Franklin 所给出的当 $M=2$ 时的估计量，采用本文的记号可以表达为：

$$\hat{\theta}_1 = \frac{\sum_{t=1}^{k}\bar{z}_t - \sum_{t=1}^{k}\mu_t^{(2)}}{\sum_{t=1}^{k}\mu_t^{(1)} - \sum_{t=1}^{k}\mu_t^{(2)}},\qquad Var(\hat{\theta}_1) = \frac{\theta_1(1-\theta_1)}{n} + \frac{\theta_1\sum_{t=1}^{k}\sigma_t^{2(1)} + (1-\theta_1)\sum_{t=1}^{k}\sigma_t^{2(2)}}{n(\sum_{t=1}^{k}\mu_t^{(1)} - \sum_{t=1}^{k}\mu_t^{(2)})^2}$$

容易看出这是本文定理一及定理二在 $M=2$ 时的特例.

三　随机模拟

设总体中的元素分为四类，其中每类在总体中所占的比例为：$\theta_1 = 0.05, \theta_2 = 0.20, \theta_3 = 0.30, \theta_4 = 0.45$. 这里利用随机模拟的方法对$\theta_i, i = 1,2,3,4$. 进行估计，模拟时采用调查方法一. 假设从总体中抽取了三组"被调查者"，这里"被调查者"用来自均匀分布U(0,1)的随机数r表示，若$r < 0.05$则表示此"被调查者"来自第一类；若$0.05 \le r < 0.25$则表示此"被调查者"来自第二类；若$0.25 \le r < 0.55$则表示此"被调查者"来自第三类；若$r \ge 0.55$则表示此"被调查者"来自第四类. 为简单起见，对每组每个人只进行一次试验. 模拟总共进行了两次. 结果列表如下：

表一

	$\bar{\theta}_1$	$\bar{\theta}_2$	$\bar{\theta}_3$	$\bar{\theta}_4$
第一次	0.0901	0.1107	0.3720	0.4272
第二次	0.0582	0.2116	0.2621	0.4685

表二

		$j=1$	$j=2$	$j=3$	$j=4$
第一次	σ_j	0.8522	1.0082	1.6271	0.9079
	$\hat{\sigma}_j$	0.7647	1.0344	1.6517	0.8766
	M_j	0.7626	1.0340	1.6464	0.8733
第二次	σ_j	0.2580	1.3635	2.3874	1.1204
	$\hat{\sigma}_j$	0.2454	1.4243	2.5777	1.2227
	M_j	0.2445	1.4184	2.5672	1.2177

这里$\hat{\sigma}_j = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\hat{\theta}_{ji}-\bar{\theta}_j)^2}$，$M_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_{ji}-\theta_j)^2}$　$j = 1,2,3,4$. σ_t^2为利用定理二所计算出的$\hat{\theta}_i$的方差$(i = 1,2,3,4)$，其中N为模拟次数，这里两次模拟均取$N = 120$，$\hat{\theta}_{ji}$为每次模拟所得到的θ_j的估计值，$\bar{\theta}_j$为N次θ_j估计值的平均值.

参考文献

1 Frank,L.A. *A Comparison Of Estimators For Ranomized Response Sampling With Continuous Distributions From A Dichotomous Population*. Communications In Statistics Theory And Methods, 18(2), 489-505 (1989)