文章编号:1672-3961(2006)06-0051-06

敏感性问题中的均方误差与模型比较

王丽君[1], 黄奇成[2], 王兆旭[3]
(1. 曲阜师范大学 信息技术与传播学院, 山东 日照 276826; 2. 曲阜师范大学 运筹与管理学院, 山东 日照 276826;
3. 山东大学 基建处, 山东 济南 250061)

摘要:探讨了敏感性问题的几种模型,给出了它们相对应估计量的均方误差的计算公式.并结合实例对模型进行了比较.

关键词:敏感性问题;随机化回答模型;抽样调查;均方误差
中图分类号:C812 **文献标识码**:A

Mean square error and patterns comparison of sensitivity issue

WANG Li-jun[1], HUANG Qi-cheng[2], WANG Zhao-xu[3]
(1. College of Information Technology and Communication Science, Qufu Normal University, Rizhao 276826, China; 2. College of Operations Research and Management Science, Qufu Normal University, Rizhao 276826, China; 3. Office of Basic Construction, Shandong University, Jinan 250061, China)

Abstract: Several patterns of sensitivity issue are discussed. Calculation formulae about the Mean Square Error of the corresponding estimation are given. In addition, comparisons between the patterns are carried out combined with the actual sample.

Key words: sensitivity issue; romdomized response technique; sampling investigate; mean square error

0 敏感性问题的背景与概念

抽样调查的方法是从部分样本来推断或估计总体的某些特殊或数量指标,现在已经被广泛应用,它省时省力,能获得较为准确的结果,这一方面是由于方法本身的科学性;但另一方面也是很重要的一面是被调查者的回答必须都是真实的.

在当今的社会经济调查等各种统计调查中,经常遇到各种各样的敏感性问题.所谓敏感性问题,是指与个人或单位的隐私或私人的利益有关而不便向外界透露的问题.比如,个人或单位是否偷漏税及数额的多少;考生在考试中是否有作弊行为;吸毒、赌博;个人储蓄的多少;是否参加过走私货物的交易;是否有犯罪行为;各种类型的额外消费、公款吃喝;同性恋及类似的为社会所不赞成的各种行为.

对于这类敏感性问题,调查中若采用直接问答的方式,被调查者为了保护自己的隐私或出于其他目的,往往会拒绝回答.这样就破坏了我们收集的数据的真实性,而且破坏程度的大小我们也无法度量.

可以说,传统的调查方法在敏感性问题的面前无能为力,那就是调查者将难以控制样本信息,得不到可靠的样本数据.怎么办? 为了得到敏感性问题的可靠样本数据,我们就需要对此设计出一些好的调查方法,并且能对其合理的选用.

收稿日期:2006-07-22
作者简介:王丽君(1962-),女,副教授,学士,主要从事决策分析与计算机教育研究和教学.
E-mail: wanglijunsd@126.com

1 敏感性问题的提出和随机化回答技术(RRT)

总的说来,对于敏感性问题,若采用直接问答的方式,调查者将难以控制样本信息,得不到可靠的样本数据.从而调查的真实性就难以保证.因而为了得到敏感性问题的可靠的样本数据,有必要采用一种科学的可行的技术——随机化回答技术(Randomized Response Technique 简记为 RRT).

随机化回答,是指在调查中使用特定的随机化装置,使得被调查者以预定的概率 P 来回答敏感性问题,这一技术的宗旨就是最大限度地为被调查者保守秘密,从而取得被调查者的信任.比如在调查的学生考试作弊的问题中,设计外形完全一样的卡片 m 个,其中 m_1 个卡片上写上"你考试是否作弊?"另外 $m-m_1$ 个卡片上写上另外的问题.然后把这 m 个卡片折叠好,放在一盒子中.调查时,由被调查者从盒子里任抽一卡片,根据卡片上的问题做出回答,回答完毕再把卡片折叠好放回盒子.至于卡片上具体是什么问题,调查者无权过问.这样就起到了为被调查者保密的作用.因而相对于直接回答调查,易于得到被调查者的合作.

2 敏感性问题中的抽样调查方法

在选定了随机化回答技术之后,需要建立一定的模型对有关参数及总体特征进行估计.1965 年,沃纳首先提出了敏感性问题的调查方法.西蒙斯等人对这种方法进行了改进.对于不同的具体问题不同模型所得到的结果又有一定的异同和优劣.要根据具体问题予以讨论和取舍.

2.1 直接提出一个敏感性问题

从总体中抽取一个由 n 个被调查者组成的随机样本,现对这 n 个调查者直接提出敏感性问题,"你是 A 类中的成员"吗?设被调查者或者回答"是",或者回答"不是".

$$\text{设 } X_i=\begin{cases}1,\text{当被调查者回答问题用"是"时},\\0,\text{当被调查者回答问题用"不是"时},\end{cases}$$

如果在调查中属于"A"类的被调查者说真话的概率为 P_A,而属于"$\bar{A}$"类的成员中说真话的概率为 $P_{\bar{A}}$,

$$\text{设 } Y_i=\begin{cases}1,\text{被调查者是"}A\text{"类的成员},\\0,\text{被调查者是"}\bar{A}\text{"类的成员},\end{cases}$$

由全概率公式:

$$P(X_i=1)=P(Y_i=1)\cdot P(X_i=1\mid Y_i=1)+P(Y_i=0)\cdot P(X_i=1\mid Y_i=0)=$$
$$\pi P_A+(1-\pi)\cdot(1-P_A),$$
$$\therefore\ EX_i=\pi P_A+(1-\pi)\cdot(1-P_A),$$
$$Var(X_i)=[\pi P_A+(1-\pi)\cdot(1-P_{\bar{A}})]\cdot[1-\pi P_A-(1-\pi)\cdot(1-P_{\bar{A}})].$$

从而 $E\hat{\pi}=EX_i=\pi P_A+(1-\pi)\cdot(1-P_A)$,

$$Var(\hat{\pi})=\frac{n}{n^2}Var(X_i)=\frac{1}{n}\cdot[\pi P_A+(1-\pi)\cdot(1-P_A)]\cdot[1-\pi P_A-(1-\pi)\cdot(1-P_A)],$$

$$\therefore\ MSE(\hat{\pi})=E(\hat{\pi}-\pi)^2=Var(\hat{\pi})+(E\hat{\pi}-\pi)^2.$$

2.2 沃纳随机化回答模型(Warner model)

它开创了随机化回答的先河.其设计原则是根据敏感性特征设计两个相互对立的问题,让被调查者按预定的概率从中选一个回答,调查者无权过问被调查者究竟回答的是哪一个问题,从而起到了为被调查者保密的效果.

在调查中向被调查者提出两个陈述,每个陈述只要被调查者回答"是"或"不是",这两个陈述是"我是 A 类中的成员"(被提出的概率为 p),"我不是 A 类中的成员"设 m 为 n 个被调查者组成的随机样本中回答"是"的人数,则 π 的一个最大似然估计为:

$$\hat{\pi}=\frac{\varphi-(1-p)}{2p-1},\left(p\neq\frac{1}{2}\right),\text{其中 }\varphi=m/n,$$

易知:$E\hat{\pi}=\pi$.

则,$\hat{\pi}$是 π 的极大似然无偏估计.其方差为:

$$X_i=\begin{cases}1,若被调查者回答“是”,\\0,若被调查者回答“非”,\end{cases}$$

$$Var(\hat{\pi})=\frac{\pi(1-\pi)}{n}+\frac{p(1-p)}{(2p-1)^2n}=\frac{\phi(1-\phi)}{(2p-1)^2n},$$

其中 $\phi=p\pi+(1-p)(1-\pi)=P(X_i=1)$.

易知 $Var(\hat{\pi})$有一个无偏估计:

$$\hat{V}ar(\hat{\pi})=\frac{\hat{\pi}(1-\hat{\pi})}{n}+\frac{p(1-p)}{(2p-1)^2n}.$$

2.3　西蒙斯随机化回答模型(Simmons model)

其设计思想仍是基于沃纳的随机化回答思想,只是在设计中,用无关的问题 Y 代替了沃纳模型中的敏感性问题 A 的对立问题.其中以概率 p 提出第一个敏感性问题的陈述,以概率$(1-p)$提出第二个毫无敏感性的陈述,则总体中回答“是”的比例由全概率公式显然得:

$$X_i=\begin{cases}1,若被调查者回答“是”,\\0,若被调查者回答“非”,\end{cases}$$

$\varphi=p\pi+(1-p)(1-\pi_y)=P(X_i=1)$,(已知 π_y 是具有无关特性 Y 的所占比例),

则 π 的一个极大似然估计为:

$$\hat{\pi}=\frac{\hat{\varphi}-(1-p)\pi_y}{p}.$$

这里$\hat{\varphi}=m/n$,m 为总体 n 个被调查者中回答“是”的人数.

其方差为:

$$Var(\hat{\pi})=\frac{\varphi(1-\varphi)}{np^2},$$

它的一个无偏估计为:

$$\hat{V}ar(\hat{\pi})=\frac{\hat{\varphi}(1-\hat{\varphi})}{np^2}.$$

2.4　两个样本的无关回答模型

同样向被调查者提出两个陈述与上述方法不同的是,第一个非敏感性问题的比例虽然实际上已确定,但未知,为了确定两个比例 π 与π_y,我们有两个随机样本,它们的含量分别为 n_1,n_2,对于敏感性的问题有不同的比例 p_1 与 p_2,设 m_1,m_2 分别为样本含量为 n_1,n_2 时,随机样本中回答“是”的人数,设:

$$X_i=\begin{cases}1,第一组\ n_1\ 个样本中回答“是”,\\0,第一组\ n_1\ 个样本中回答“不是”,\end{cases}$$

$$Y_i=\begin{cases}1,第一组\ n_2\ 个样本中回答“是”,\\0,第一组\ n_2\ 个样本中回答“不是”.\end{cases}$$

则由全概率公式知:

$$\phi_1=p_1\pi+(1-p_1)\pi_y=P(X_i=1),$$

$$\phi_2=p_2\pi+(1-p_2)\pi_y=P(Y_i=1).$$

可以得到的 π 极大似然估计为:

$$\hat{\pi}=\frac{\frac{m_1}{n_1}(1-p_2)-\frac{m_2}{n_2}(1-p_1)}{p_1-p_2}.$$

易见:$E\hat{\pi}=\pi$,所以$\hat{\pi}$是 π 的无偏估计.

其方差为:

$$Var(\hat{\pi})=\frac{1}{(p_1-p_2)^2}\left[\frac{(1-p_2)^2\phi_1(1-\phi_1)}{n_1}+\frac{(1-p_1)^2\phi_2(1-\phi_2)}{n_2}\right].$$

上述误差随着 p_1,p_2,n_1,n_2 的不同选择而不同,当 $n_1+n_2=n$,采用最优的 n_1,n_2 时,即

$$\frac{n_1}{n_2}=\frac{(1-P_2)}{(1-P_1)}\sqrt{\frac{\phi_1(1-\phi_1)}{\phi_2(1-\phi_2)}}.$$

则$\hat{\pi}$的方差达到最小.此时有:

$$Var(\hat{\pi})_{\min}=\frac{1}{n(p_1-p_2)^2}[(1-p_2)\sqrt{\phi_1(1-\phi_1)}+(1-p_1)\sqrt{\phi_2(1-\phi_2)}]^2.$$

2.5 改进的随机化回答模型

由于在沃纳模型中存在不论抽中几号卡片都必须回答敏感性问题的缺点,消除不了被调查者的顾虑,在西蒙斯的随机化回答模型中要对调查者了解相关的人事资料具有一定的难操作性,而双样本模型的设计相对较为复杂.

因而提出了改进的随机化回答模型.在随机装置中按一定比例放入红、蓝、白三种颜色的球,各种球所占的比例及抽到哪种颜色的球后回答的问题都已经被事先约定.如果抽到红球则据实回答敏感性问题"是"或"不是",抽到蓝色球总是回答"是",但如果抽到白球总是回答"不是".由于调查者无法知道被调查者抽到何种颜色的球,也就无法知道被调查者回答的是哪一个问题,即回答"是"的可能是抽到敏感性问题的答案也可能是抽到蓝球的答案,故为其保密.

在一个随机装置中放入红、蓝、白三种颜色的球,其比例分别为 p_1,p_2,p_3.且 $p_1+p_2+p_3=1$.则

$$\phi=\frac{m}{n}=\pi p_1+p_2.$$

π 的极大似然估计量为:

$$\hat{\pi}=\frac{\phi-p_2}{p_1}.$$

其方差为:

$$Var(\hat{\pi})=\frac{\pi(1-\pi)}{n}+\frac{p_1(1-p_1)}{p_1^2n}+\frac{p_2(1-p_2)-2\pi p_1p_2}{p_1^2n}.$$

又由于 $E\hat{\pi}=\pi$,

$\therefore$ $\hat{\pi}$是 π 的无偏估计,

$$\widehat{Var}(\hat{\pi})=\frac{\phi(1-\phi)}{np_1^2}.$$

2.6 隐含的随机化回答模型

这一模型的优点是不需要使用任何随机化的实验装置,但它又不失随机化的特征性.设 π_A 是总体中具有敏感性 A 的人所占的真实比例.根据是否具有敏感性 A 可把总体分为两类:属于 A 或属于 $\bar{A}$(即不具有敏感性 A).又根据与敏感性 A 无关的问题把总体分为三类:如喜欢红色的(Ⅰ),喜欢兰色的(Ⅱ),喜欢除红,兰以外其他颜色的(Ⅲ).

抽取三个相互独立有放回简单随机样本,容量分别为 n_1,n_2,n_3.每一被抽中的人,根据如下规则真实地回答"1"或"0".

表1 回答规则

Tab.1 Rules of answer

类型	样本1		样本2		样本3	
	A	$\bar{A}$	A	$\bar{A}$	A	$\bar{A}$
Ⅰ	1	0	1	0	0	1
Ⅱ	1	0	0	1	1	0
Ⅲ	0	1	1	0	1	0

若是第一个样本,他应该作如下回答:当他属于 A 且属于类型Ⅰ或属于类型Ⅱ时,回答"1";属于 $\bar{A}$ 且属于类型Ⅰ或属于类型Ⅱ时,回答"0";属于 $\bar{A}$ 且属于类型Ⅲ时,回答"1";属于 A 且属于类型Ⅲ时,回答"0".

若是第二个样本,他应作如下回答:属于类型Ⅰ或类型Ⅲ,若同时又属于 A,则回答"1",若属于 $\bar{A}$,则回答"0";属于类型Ⅱ,若同时具有敏感性 A,则回答"0",不具敏感性 A,回答"1".

对第三个样本,回答规则是:属于类型Ⅱ或Ⅲ,若同时又属于 A,则回答"1",若属于 $\bar{A}$,则回答"0",属于

类型Ⅰ，若同时属于 $\bar{A}$，则回答“1”，若属于 A，则回答“0”.

由上述回答规则可知，虽然被调查者做出了真实的回答，但对调查者来说，他给出的答案“1”或“0”是不确定的.也就是说调查无法根据他的回答来判断他是否具有敏感性.这便起到了保护被调查者隐私的目的，从而可获得他们的合作.

令 λ_i 是第 i 个样本中个体回答“1”的概率.显然有：

$$\lambda_1 = \pi_{A1} + \pi_{A2} + \pi_{A3},$$

$$\lambda_2 = \pi_{A1} + \pi_{A2} + \pi_{A3},$$

$$\lambda_3 = \pi_{A1} + \pi_{A2} + \pi_{A3}.$$

又 $\because\ \lambda_1 + \lambda_2 + \lambda_3 = \pi_A + 1.$

又设 n_{i1} 是第 i 个样本中回答“1”的人数，令 $\hat{\lambda}_i = n_{i1}/n_i, (i = 1,2,3)$，则 π_A 的一个无偏估计是：

$$\hat{\pi}_A = \hat{\lambda}_1 + \hat{\lambda}_2 + \hat{\lambda}_3 - 1.$$

其方差为：

$$Var(\hat{\pi}_A) = Var(\hat{\lambda}_1) + Var(\hat{\lambda}_2) + Var(\hat{\lambda}_3) = \frac{\lambda_1(1-\lambda_1)}{n_1} + \frac{\lambda_2(1-\lambda_2)}{n_2} + \frac{\lambda_3(1-\lambda_3)}{n_3}.$$

它的一个无偏估计为：

$$\hat{Var}(\hat{\pi}_A) = \frac{\hat{\lambda}_1(1-\hat{\lambda}_1)}{n_1} + \frac{\hat{\lambda}_2(1-\hat{\lambda}_2)}{n_2} + \frac{\hat{\lambda}_3(1-\hat{\lambda}_3)}{n_3}.$$

而当 $n_1 : n_2 : n_3 = \sqrt{\lambda_1(1-\lambda_1)} : \sqrt{\lambda_2(1-\lambda_2)} : \sqrt{\lambda_3(1-\lambda_3)}$ 成立时方差最小.

此时方差的最小值为：

$$Var(\hat{\pi}_A) = \left[\sum_{i=1}^{3} \sqrt{\lambda_i(1-\lambda_i)}\right]^2 / n, \quad n = n_1 + n_2 + n_3.$$

由此可以看出，一旦总样本量 n 确定，为了使方差达到最小，抽样时三个样本容量比应满足 $n_1 : n_2 : n_3 = \sqrt{\lambda_1(1-\lambda_1)} : \sqrt{\lambda_2(1-\lambda_2)} : \sqrt{\lambda_3(1-\lambda_3)}$，但又因为 $\lambda_1, \lambda_2, \lambda_3$ 是未知的，故可先做预调查或根据以往资料来确定 $\lambda_1, \lambda_2, \lambda_3$ 的值.

3　模型的比较和选择

在采用随机化回答方法时，假设所有的被调查者都作出了真实的回答，但是在采用各种模型时一般要从无偏性和有效性两个基本方面进行比较.

(1)关于无偏性，从上面的结论可以分析出，已有的沃纳模型、西蒙斯模型、两个样本的无关模型和改进的随机化模型中的 $\hat{\pi}$ 都是 π 的无偏估计.

(2)关于有效性的希望是 $\hat{\pi}$ 尽可能接近真实的数值 π，由切贝谢夫不等式：

$$P(|\hat{\pi} - \pi| \geq \varepsilon) \leq E(\hat{\pi} - \pi)^2/\varepsilon^2 = MSE(\hat{\pi})/\varepsilon^2.$$

可以知道 π 的均方误差(Mean Square Error) $MSE(\hat{\pi})$ 越小，$\hat{\pi}$ 偏离 π 的大于指定值 ε 的概率就越小，$MSE(\hat{\pi})$ 和 $Var(\hat{\pi})$ 的关系如下：

$$MSE(\hat{\pi}) = Var(\hat{\pi}) + [E(\hat{\pi}) - \pi]^2.$$

则当 $\hat{\pi}$ 是 π 的无偏估计时，由上式可得 $MSE(\pi) = Var(\pi)$.因而，方差最小的无偏估计才是“最佳”估计.

4　应用举例

设在总体中有 10% 的人有过一种非法行为，在采用随机化回答方法时，假设所有的被调查者都作出了真实的回答，假设 $n = 500$ 时，按以下各种方法计算出的 $MSE(\hat{\pi})$ 加以比较：

(1) 直接提出一个敏感性模型，如果全体有过这种行为的人当中有(a)15%；(b)20%；(c)75% 的人否认曾经有过这种行为，

(2) 沃纳模型中 $p = 0.8$，

(3) 西蒙斯模型中,已知 $\pi_y = 0.2$,

(4) 两个样本无关问题方法,$p_1 = 0.8, p_2 = 1 - p_1 = 0.2$,

(5) 改进的随机化回答模型中,$p_1 = 0.4, p_2 = 0.3, p_3 = 0.3$,

(6) 隐含的随机化回答模型.

计算有:

(1) 直接提出一个敏感性方法:本题中有过这种行为的人可能否认自己有过这种为,但是没有这种行为的人都讲了真话,可知 $P_A = 1$,

$\therefore E(\hat{\pi}) = \pi P_A$.

因而有:$Var(\hat{\pi}) = \dfrac{\pi P_A(1 - P_A)}{n}$,

$\therefore MSE(\hat{\pi}) = (\pi P_A - \pi)^2 + \dfrac{\pi P_A(1 - P_A)}{n}$.

(a) 此时有过这种行为并说真话的比例为 $P_A = 1 - 15\% = 85\%$,

$\therefore MSE(\hat{\pi}) = (\pi P_A - \pi)^2 + \dfrac{\pi P_A(1 - P_A)}{n} \approx 3.81 \times 10^{-4}$,

(b) $MSE(\hat{\pi}) = (\pi P_A - \pi)^2 + \dfrac{\pi P_A(1 - P_A)}{n} \approx 5.47 \times 10^{-4}$,

(c) $MSE(\hat{\pi}) = (\pi P_A - \pi)^2 + \dfrac{\pi P_A(1 - P_A)}{n} \approx 7.64 \times 10^{-4}$.

(2) 沃纳方法:

$$\phi = p\pi + (1 - p)(1 - \pi) = (2 \times 0.8 - 1) \times 10\% + (1 - 0.8) = 0.26,$$

$$MSE(\hat{\pi}) = \frac{\pi(1 - \pi)}{n} + \frac{p(1 - p)}{(2p - 1)^2 n} = \frac{\phi(1 - \phi)}{(2p - 1)^2 n} = \frac{0.26 \times (1 - 0.26)}{500 \times (2 \times 0.8 - 1)^2} \approx 10.64 \times 10^{-4}.$$

(3) 西蒙斯方法:

$$\varphi = p\pi + (1 - p)(1 - \pi_y) = 0.8 \times 0.1 + 0.2 \times 0.2 = 0.12,$$

$$MSE(\hat{\pi}) = \frac{\varphi(1 - \varphi)}{np^2} = \frac{0.12 \times (1 - 0.12)}{500 \times 0.8^2} \approx 3.3 \times 10^{-4}.$$

(4) 两个样本无关问题方法:$p_1 = 0.8, p_2 = 1 - p_1 = 0.2$,应用最优 n_1, n_2 得:

$$\phi_1 = p_1\pi + (1 - p_1)\pi_y = 0.8 \times 0.1 + (1 - 0.8) \times 0.2 = 0.12,$$

$$\phi_2 = p_2\pi + (1 - p_2)\pi_y = 0.2 \times 0.1 + (1 - 0.2) \times 0.2 = 0.18.$$

$$MSE(\hat{\pi}) = \frac{1}{n(p_1 - p_2)^2}[(1 - p_2)\sqrt{\phi_1(1 - \phi_1)} + (1 - p_1)\sqrt{\phi_2(1 - \phi_2)}]^2 =$$

$$\frac{1}{500 \times 0.8^2}[0.8 \times \sqrt{0.12 \times 0.88} + 0.2 \times \sqrt{0.18 \times 0.82}]^2 \approx 6.3 \times 10^{-4}.$$

(5) 改进的随机化回答方法:

$$\phi = \frac{m}{n} = \pi p_1 + p_2 = 0.1 \times 0.4 + 0.3 = 0.34,$$

$$Var(\hat{\pi}) = \frac{\pi(1 - \pi)}{n} + \frac{p_1(1 - p_1)}{p_1^2 n} + \frac{p_2(1 - p_2) - 2\pi p_1 p_2}{p_1^2 n} =$$

$$\frac{0.1 \times (1 - 0.1)}{500} + \frac{0.4 \times (1 - 0.4)}{0.4^2 \times 500} + \frac{0.3 \times (1 - 0.3) - 2 \times 0.1 \times 0.4 \times 0.3}{0.4^2 \times 500} \approx 5.5 \times 10^{-3}.$$

(6) 隐含的随机化回答模型:

假设:$n_1 = 200, n_2 = 200, n_3 = 100; \lambda_1 = \pi_{A1} + \pi_{A2} + \pi_{A3} = 0.4; \lambda_2 = \pi_{A1} + \pi_{A2} + \pi_{A3} = 0.4; \lambda_3 = \pi_{A1} + \pi_{A2} + \pi_{A3} = 0.064$.

$$Var(\hat{\pi}_A) = \Big[\sum_{i=1}^{3}\sqrt{\lambda_i(1 - \lambda_i)}\Big]^2/n = \frac{1}{500}(\sqrt{0.4 \times 0.6} + \sqrt{0.4 \times 0.6} + \sqrt{0.064 \times 0.0936})^2 \approx$$

(下转第 120 页)

2003.
[3] 朱志宇,张冰,刘维亭. 基于模糊支持向量机的语音识别方法[J].计算机工程,2006,32(2):180-182.
ZHU Zhi-yu, ZHANG Bing, LIU Wei-ting. Speech recognition based on fuzzy support vector machine[J]. Computer Engineering, 2006, 32(2):180-182.
[4] 张翔,肖小玲,徐光祐.基于样本之间紧密度的模糊支持向量机方法[J]. 软件学报,2006,17(5):951-958.
ZHANG Xiang, XIAO Xiao-ling, XU Guang-you. Fuzzy support vector machine based on affinity among samples[J]. Journal of Software, 2006,17(5):951-958.
[5] CHAPELLE O, VAPNIK V, BOUSQUET O, et al. Choosing multiple parameters for support vector machines[J]. Machine Learning, 2002, 46:131-159.
[6] VAPNIK V. The nature of statistical learnin theory[M]. New York: Springer-Verlag, 1999.
[7] HSU C W, LIN C J. A comparison of methods for multiclass support vector machines[J]. IEEE Trans on Neural Networks, 2002, 13(2):415-425.

(编辑:陈燕)

(上接第56页)

$$\frac{1}{500}(0.49+0.49+0.245)^2 \approx 3.0\times 10^{-3}.$$

5 总体评价

由以上的简单举例中不难发现,对应于不同的假设数据自然会得到对不同随机化回答模型的不同的均方误差,因此就会产生对不同随机化回答模型的不同评价.通常,西蒙斯模型较沃纳模型可取,但是对于西蒙斯模型和双样本模型,我们无法一言而概其优劣.一般,当调查对象的数目不是很大时,调查对象的有关资料又轻易查到、且计算量不是很大时,西蒙斯模型以设计简便、误差小而比双样本模型可取.然而,当调查对象过多或调查对象具有不确定性时,从资料调查到计算都不方便或不可能做到时,则可以选用双样本模型.而改进的随机化回答模型虽然简单易于操作,但是有时误差也会相对大一些也有其缺点.隐含的随机化回答模型操作方面又显烦琐要达到方差的最小值存在较大的困难.而对于用直接提出一个敏感性问题的方法,则因为难以估计曾经有过此种行为而又予以否认的比例,所以也难以保证估计的准确性.

作为现代统计的一个极其重要的分支,随机化回答技术在我国的调查实践中的使用将日趋广泛,笔者有理由相信随机化回答模型将日臻完善,敏感性问题的调查水平也会不断提高和进步.

(编辑:董程英)

(上接第115页)

[5] 王建朝,徐常威,何凤荣,等.酰胺-尿素-NaBr熔体中电沉积Tb-Co合金的研究[J].中国稀土学报,2003,21(5):584-588.
WANG Jian-chao, XU Chang-wei, HE Feng-rong, et al. Electrodeposition of Tb-Co Alloy Film in acetamide-urea-NaBr melt[J]. Journal of The Chinese Rare Earth Society, 2003, 21(5):584-588.
[6] BARD A J, FAULKNER L R. Electrochemical methods, foundamentals and applicatoions[M]. New York: John Wliey & Sons, 1980. 222-223,253.

(编辑:董程英)