应用敏感问题调查模型对大学生作弊情况的统计与分析

郝诚 ，刘蕾 ，唐一博

北京师范大学数学系，北京（100875）

Email: haocheng05@mail.bnu.edu.cn

摘 要: 本文中，我们将在 Warner 模型与 Simmons 模型基础上建立新的敏感问题调查模型。我们模型的优势在于它可以在保持对被访者的隐私具有保护性的同时进一步减小模型的方差。这是 Warner 与 Simmons 模型所不能的。我们还将使用此模型对北京师范大学学生的作弊情况进行统计与分析。

关键词: 敏感问题调查，Warner 模型，Simons 模型

中图分类号: O213

1. 引 言

考试，作为学校教学活动中评估学生掌握知识水平和检测质量的主要方法，在我国高等教育活动中始终扮演着极为重要的角色。但是，考试过程中作弊现象的存在，一直是严重危害教育活动有序开展的难题。国外的众多研究表明，大学生考试作弊相当普遍，其比例在13%一 95%之间；在[1]中，我国教育工作者的调查也发现，30.8%的大学生承认自己在大学期间有作弊行为，10.6%的大学生承认自己有作弊意图。

北京师范大学学生处 2000 年对大学生作弊问题的数据调查报告[2]显示，学生考试作弊情况严重。然而，由于在调查过程中可能会出现的学生出于对自身的保护而对调查不配合或隐瞒真实信息的存在，如何获得尽可能接近真实情况的数据成为对学生作弊现状进行分析的基础。本文试图通过改进 Warner 的方法与无关问题方法，得到一种更完善、可靠的调查方法，且用此方法对北京师范大学学生作弊情况进行调查，并对结果进行分析。

2. 敏感问题调查的已有模型及新模型的建立

2.1 背景与已有模型

在敏感问题调查中，为了保护被调查者的隐私，以及减少被调查者故意欺骗回答对调查结果的干扰，Warner [3] 在 1965 年的论文 *A Survey Technique for Eliminating Evasive Answer Bias* 中提出了“随机回应模型”。其方法主要思想为：

若要确定被调查人群中具有某一（隐私）属性 A 的人数比例 π_A，抽取人群中的一个样本数为 n 的随机样本。对样本中的每位被调查者，他将有 $p\,(p \neq 0.5)$ 的几率被要求回答问题：“你是具有属性 A 吗？”；$1-p$ 的几率被要求回答问题：“你不具有属性 A 吗？”。而调

查者将不会知道被调查者最终是被要求回答的哪个问题，从而被调查者给出的答案（是/否）将不会泄露其是否具有属性 A，继而起到保护被调查者隐私的作用。

这时，被调查者回答“是”的概率λ为

$$\lambda = \pi_A p + (1-\pi_A)(1-p)$$

其中 π_A 为具有属性 A 的人数比例。由此，可以得到 π_A 的表达式

$$\pi_A = \frac{p-1+\lambda}{2p-1}$$

设 n' 为样本中回答“是”的人数，则 λ 可以用样本均值 n'/n 进行估计。事实上，如果被调查者都如实回答，这样的估计是无偏的，见 [3]。于是，π_A 的无偏估计为

$$(\pi_A)_W = \frac{1}{2p-1}\{p-1+\frac{n'}{n}\}$$

方差为

$$Var(\pi_A)_W = \frac{\pi_A(1-\pi_A)}{n} + \frac{p(1-p)}{n(2p-1)^2}$$

在 Warner 的论文发表两年后，SIMMONS [4] 在论文 *The Unrelated Question Randomized Response Model: Theoretical Framework* 中提出了另外一种隐私调查方法方法——无关问题方法。其方法主要思想为：

同样要确定被调查人群中具有某一（隐私）属性 A 的人数比例 π_A，抽取人群中的一个样本数为 n 的随机样本。考虑另外一种非隐私属性 Y，使得当被调查者被问到是否具有此属性时，被调查者将会如实回答。（例如：调查者是否在 8 月出生）。设具有属性 Y 的人数比例为 π_Y，π_Y 并不需要预先知道，如果已知，可以检验最终估计的正确性，见 [3]。

将样本分成两个子样本 1，2(样本数分别为 n_1 和 n_2)。在样本 i(i = 1,2)中，被调查者将以 p_i 的几率回答问题:“你是具有属性 A 吗？”$1-p_i$ 的几率回答问题:“你具有属性 Y 吗？”

设两个样本中被调查者回答“是”的概率为 λ_1 和 λ_2，则

$$p_1\pi_A + (1-p_1)\pi_Y = \lambda_1$$
$$p_2\pi_A + (1-p_2)\pi_Y = \lambda_2$$

联立求解得到

$$\pi_A = \frac{\lambda_1(1-p_2)-\lambda_2(1-p_1)}{p_1-p_2}$$

$$\pi_Y = \frac{\lambda_1 p_2 - \lambda_2 p_1}{p_1-p_2}$$

设样本 i(i=1,2)中回答“是”的人数为 n_i'，用 n_i'/n_i 去估计 λ_i，就可以得到 π_A 与 π_Y 的估计

$$(\pi_A)_U = \frac{\frac{n_1'}{n_1}(1-p_2)-\frac{n_2'}{n_2}(1-p_1)}{p_1-p_2}$$

$$\pi_Y = \frac{\frac{n_1'}{n_1}p_2 - \frac{n_2'}{n_2}p_1}{p_1-p_2}$$

与 Warner 的随机回应方法一样，无关问题方法在被调查者如实回答的情况下是无偏的，且方差为

$$Var(\pi_A)_U = \frac{1}{(p_1-p_2)^2}\{\frac{\lambda_1(1-\lambda_1)(1-p_2)^2}{n_1}+\frac{\lambda_2(1-\lambda_2)(1-p_1)^2}{n_2}\}$$

2.2 新模型的建立

在 [2]、[3] 中，作者讨论了如果选取 p， p_i 及 π_Y 的不同选取方式对方差的影响。为使方差获得最小值， p 及 p_i 需要取 1 或 0，但是这将不能起到保护被调查者隐私的作用。为此，我们通过结合 Warner 的模型与 Simmons 的无关问题模型，提出的新模型来试图解决这个问题。

要确定被调查人群中具有某一（隐私）属性 A 的人数比例 π_A，抽取人群中的一个样本数为 n 的随机样本。同时考虑属性“非 A”及另外一种非隐私属性 Y。属性 Y 的选取要求当被调查者被问到是否具有此属性时，被调查者将会如实回答。设具有属性 Y 的人数比例为 π_Y，且 π_Y 已知。（例如，属性 Y 为“被调查者在 8 月出生”，则当样本容量足够大时，可以认为 π_Y =1/12）

对样本中的每位被调查者，他将有 p 的几率被要求回答问题：“你是具有属性 A 吗？”；q 的几率被要求回答问题：“你不具有属性 A 吗？”；$1-p-q$ 的几率被要求回答问题：“你具有属性 Y 吗？”。

设 λ 为样本中回答“是”的概率，则有

$$\lambda = \pi_A p + \pi_Y(1-p-q)+(1-\pi_A)q$$

其中π_A为具有属性 A 的人数比例，π_Y为具有属性 Y 的人数比例，π_Y已知。由此，得到π_A的表达式

$$\pi_A = \frac{\lambda - q - (1-p-q)\pi_Y}{p-q}$$

设 n'为样本中回答“是”的人数，则根据极大似然的原理，λ的无偏估计为$\hat{\lambda} = \frac{n'}{n}$，所以$\pi_A$的无偏估计为：

$$\hat{\pi}_A = \frac{\frac{n'}{n} - q - (1-p-q)\pi_Y}{p-q}$$

方差为

$$Var(\pi_A) = \frac{\lambda(1-\lambda)}{n(p-q)^2}$$

其中$\lambda = \pi_A p + \pi_Y(1-p-q) + (1-\pi_A)q$。

计算$Var(\pi_A)$的偏导数：

$$\frac{\partial}{\partial p}Var(\pi_A) = \frac{(p-q)\frac{\partial}{\partial p}[\lambda(1-\lambda)] - 2\lambda(1-\lambda)}{n(p-q)^3}$$

$$\frac{\partial}{\partial q}Var(\pi_A) = \frac{(p-q)\frac{\partial}{\partial q}[\lambda(1-\lambda)] + 2\lambda(1-\lambda)}{n(p-q)^3}$$

为使得方差取最小值，需且仅需$\frac{\partial}{\partial p}Var(\pi_A) = \frac{\partial}{\partial q}Var(\pi_A) = 0$。

联立方程，得到方差取到最小值的充分条件为：

$$\pi_Y = 0.5$$

$$p - q = \frac{2\lambda(1-\lambda)}{(\pi_A - \pi_Y)(1-2\lambda)} \qquad (*)$$

其中$\lambda = \pi_A p + \pi_Y(1-p-q) + (1-\pi_A)q$。将$\lambda$带入，则（*）式为$p, q$的二次方程。在数学软件中（如 Matlab）可对（*）式对于不同的π_A求解。

由此，对于预先可估计的 π_A，可以取得适当的 p,q 使得模型保持具有保护隐私作用的同时减小方差。这是 Warner 与 Simmons 的模型所不能的。

3. 模型的检验及数据处理

在 2008 年 2-4 月间，我们用上述建立的新模型对北京师范大学的学生进行了调查统计。调查采用了在线调查的方式（survey.512j.com），这样被调查人员不用担心自己的身份被暴露出来，所选问题答案的真实性可以得到较高的保证。在这次调查中，总共有 4 院系的同学总计 106 人参加了这次调查。调查问卷见附录A。

3.1 两次被调查群体的相同性检验

为对用上述方法调查的结果与 2000 年北京师范大学学生处对学生作弊情况报告中的统计结果（见 [2]）进行对比，我们需要说明指出两次结果的可比性。为此，我们需要验证被调查者群体是相同的（视此期间学生作弊情况没有发生变化）。

这里，我们主要应用 T 检验的方法来给出验证。下面为检验方法，对数据的检验过程将在下一小节给出。

首先，我们要验证正态性假定是否成立。 正态性假定是否成立，对于数据分析结论的可靠性至关重要。当样本容量较小时，考虑 Shapiro-Wilks 的 W 统计量来检验正态性。W 统计量是基于次序统计量线性组合平方的方差最佳估计与通常校正平方和估计之比（0<W<1）。当样本来自正态总体，由样本构成的 W 统计量的值应接近于 1。在单变量过程中不仅计算 W 统计量的值 w，还给出显著性概率 P=P{W<w}的值。若 P 值小于显著性水平值，则否定正态性假设，即在某显著性水平上认为这组数据与正态总体的样本有显著性差异，若 P 值大于显著性水平，则在该显著性水平上没有足够证据认为这组数据与正态总体的样本有显著性差异。P 值接近 1，数据的正态性假定越能肯定。

使用的方法主要是经验分布拟合优度检验发，由样本计算得到的经验分布函数 $F_n(x)$ 可作为总体累计分布 $F(x)$ 的估计，所以考虑 $F_n(x)$ 与原假设指定的分布 $F_0(x)$ 间的差异用以检验原假设。一下几个统计量都是用以度量假定分布 $F_0(x)$ 与经验分布 $F_n(x)$ 这两个函数之间的差异：

Kolmogorov-Smirnov 统计量：

$$D=\sup_x |F_n(x)-F_0(x)|$$

Anderson-Darling 统计量：

$$A^2=n\int_{-\infty}^{+\infty}(F_n(x)-F_0(x))^2[F_0(x)(1-F_0(x))]^{-1}dF(x)$$

Cramer-von Mises 统计量：

$$W^2=n\int_{-\infty}^{+\infty}(F_n(x)-F_0(x))^2 dF_0(x)$$

当原假设成立时，这几个统计量应取较小的值。这几个统计量取很大数值是极端情况，故度量这几个统计量取极端情况的相应的 P 值若小于给定的显著性水平，则有足够的证据否

定正态性假设。

若正态性检验成立，我们便可以对两次的样本的相同性（是否来自同一总体）进行验证。分析的手段是通过比较两组的均值来判断它们是否有显著差别。

两样本 T 检验是比较独立组的一种参数检验。此检验的一般假设是两组的均值相等（零假设）和均值不等（对立假设）。做此假设要求数据满足三个假定：一是观测是独立的，二是每组观测来自正态总体的样本，三是两个独立组方差相等。

设 $\{x_{i1}\}(i=1,2,\cdots n_1)$ 和 $\{x_{j2}\}(j=1,2,\cdots n_2)$ 是来自两个独立组 X_1 和 X_2 的随机样本（假设 $X_1 \sim N(\mu_1,\sigma_1^2)$， $X_2 \sim N(\mu_2,\sigma_2^2)$，且 $\sigma_1^2=\sigma_2^2$）。检验假设 $H_0:\mu_1=\mu_2$。

两个独立组的样本均值分别记为 $\overline{x_1}=\overline{x_2}$，检验其总体均值是否相等的统计量为

$$t=\frac{\overline{x_1}-\overline{x_2}}{\sqrt{s^2(\frac{1}{n_1}+\frac{1}{n_2})}}$$

其中 s^2 是合并方差：

$$s^2=\frac{[(n_1-1)s_1^2+(n_2-1)s_2^2]}{n_1+n_2-2}$$

s_1^2 和 s_2^2 分别是两组的样本方差：

$$s_1^2=\frac{1}{(n_1-1)}\sum_i(x_{i1}-\overline{x_1})^2,\quad s_2^2=\frac{1}{(n_2-1)}\sum_j(x_{j1}-\overline{x_2})^2$$

当两总体均值相等的假设成立时，统计量 t 服从自由度为 n_1+n_2-2 的 t 分布。

上述这个 t 统计量是建立在两总体方差相等（$\sigma_1^2=\sigma_2^2$）基础。如何检验两总体方差相等呢？使用以下形式的 F 统计量 F' 来检验方差相等的假设：

$$F'=\frac{\max(s_1^2,s_2^2)}{\min(s_1^2,s_2^2)}$$

F' 是双边 F 检验统计量。在两总体方差不等假设下，使用的是以下近似 t 统计量：

$$t'=\frac{\overline{x_1}-\overline{x_2}}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

3.2 调查结果与数据分析

在线调查共有两个问题：

- 第一个问题与[2]中调查问卷的第一个问题相同，为：“您对我们学校考试作弊现象程度的看法是什么？” 备被选五个答案：很普遍、普遍、较少、没有、说不清。设置此问题的目的为检验这次调查的样本群与文章[2]中调查的样本群是否为同一总体。
- 第二个问题将在三个备选问题中按预设概率随机挑选出一个供被调查对象回答。三个问题分别为：“您在大学考试中是否有过欺骗行为？”、“您在大学考试中是否从未有过欺骗行为？”及“您周围的同学在大学考试中是否有过欺骗行为？”。三个问题出现的概率分别为 0.2，0.5 及 0.3。

由此，我们将大致的估计在北京师范大学考试作弊的比率。

对于第一个问题的结果与[2]中调查问卷里同一问题的结果，我们依据 3.1 中的方法使用了 SAS 软件来对数据进行分析，程序见附录 B：

1) 验证正态性假定是否成立：调用单变量 （UNIVARIATE）过程，并规定选项 normal 要求进行正态性检验。结果证明，结果是符合正态性检验的。

The UNIVARIATE Procedure
Variable: d

Moments

N	69	Sum Weights	69
Mean	2.75362319	Sum Observations	190
Std Deviation	0.9299873	Variance	0.86487639
Skewness	-0.3859751	Kurtosis	-0.1065458
Uncorrected SS	582	Corrected SS	58.8115942
Coeff Variation	33.7732231	Std Error Mean	0.11195731

Basic Statistical Measures

Location		Variability	
Mean	2.753623	Std Deviation	0.92999
Median	3.000000	Variance	0.86488
Mode	3.000000	Range	4.00000
		Interquartile Range	1.00000

Tests for Location: Mu0=0

Test		-Statistic-	-----p Value------
Student's t	t	24.5953	Pr > |t| <.0001
Sign	M	34	Pr >= |M| <.0001
Signed Rank	S	1173	Pr >= |S| <.0001

2) 接下来利用 SAS 软件进行两组 t 检验。在此之前，首先确定检验的显著性水平 $\alpha = 0.05$；然后检查上述三个假定是否成立：观测是独立的，这点很显然；上述的正态性检验确定了样本是来自与正态总体的；下面将检查独立组方差是否相等。最后进行检验。

在 SAS 中，我们利用 TTEST 过程进行检验。

TTEST 过程首先检验两独立组的方差是否相等，根据输出 p 为 0.6132，这说明方差无显著差异。然后给出方差相等假定满足时的精确两样本 t 检验及方差相等条件

不满足时的近似检验，当方差相等时，p 值为 0.9987，这说明两次样本无显著性差异。

The TTEST Procedure

Statistics

Variable	n	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
per	1	5	-5.871	20.02	45.911	12.493	20.852	59.918	9.3251
per	2	5	0.2991	20.02	39.741	9.5158	15.883	45.64	7.1029
per	Diff (1-2)		-27.03	-4E-15	27.031	12.519	18.534	35.508	11.722

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
per	Pooled	Equal	8	-0.00	1.0000
per	Satterthwaite	Unequal	7.47	-0.00	1.0000

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
per	Folded F	4	4	1.72	0.6109

至此，我们通过检验方法得到两次样本属于同一总体，

第二题的回答中，回答“是”的有 62 人，“否”的 44 人，故

$$\hat{\lambda} = 1 - \frac{44}{106} = 1 - 0.4151 = 0.5849$$

由大学生的作弊现状，我们不妨设对无关问题“您周围的同学在大学考试中是否有过欺骗行为？”回答均为是，即$\pi_Y = 1$。

代入

$$\hat{\pi}_A = \frac{\frac{n'}{n} - q - (1-p-q)\pi_Y}{p-q}$$

得到曾经有过作弊行为的学生人数所占比例为：

$$\pi_A = \frac{\frac{n'}{n} - q - (1-p-q)\pi_Y}{p-q} = \frac{0.5849 - 0.5 - 0.3*1}{0.2 - 0.5} = 0.417$$

4. 数据对比与结论

在北京师范大学学生处 2000 年关于学生作弊情况的数据报告（见 [2]）中指出：

“……问卷调查表明，（北京师范大学）学生作弊是一个较普遍的现象。“想作弊”者占

被调查人数的 22.7%，“作过弊”者占 21%。1994 年在全国 23 省 77 所高校 6617 名大学生的范围内做过一项调查，这项调查显示，有过考试作弊行为者达 40%，其中经常作弊的有 1.3。从这两项调查的结果来看，数字是相当惊人的.....”

对比两次调查的结果可以发现，用敏感问题调查方法得到的结果要明显高于直接调查得到的结果。这也说明敏感问题调查方法确实有助于被调查者减少顾虑，配合调查者提供真实信息。

由此我们可以得出结论，新建立的敏感问题调查模型是具有实效性的。

若要进一步测量用此敏感问题方法调查得到的结果准确性，还需要更多（独立）的以此方法进行调查得到的统计数据，及在更大范围内，更大的样本容量下进行调查统计。

参考文献

[1]．廖平胜．考试学[M]．上海：华中师范大学出版社，1998．

[2]．李雪莲，崔恒建．我校大学生违纪分析、预测及对策[Z]．跨世纪管理干部基金课题，2000．

[3]．S L Warner． Randomized response: a survey technique for eliminating evasive answer bias [J]. Journal of the American Statistical Association, 60 [1965] 63-69．

[4]．G Bernard, Simmons, R Walt, et al．The Unrelated Question Randomized Response Model: Theoretical Framework [J]．Jour. Amer. Stat. Assoc, 64(Jun 1969), 520-39．

Statistical Analysis of Cheating Phenomena in Universities Using Randomized Response Model

Cheng Hao, Lei Liu, Yibo Tang

Department of Mathematical Sciences

Beijing Normal Universities

Beijing, P.R.China 100875

Abstract

In this paper, we give a new randomized response model based on Warner's and Simmons' models. The advantage of the new model is that it can minimize the variance without losing protections of respondent's privacy. We will further use this new model to analyze the cheating phenomenon in Beijing Normal University.

Keywords: *randomized response, Warner's model, Simmons' model*

附录 A

在线调查网页

Survey@Net——北京师范大学学生考试诚信问题调查

同学：您好！

非常感谢您对我们调查的帮助！ 我们是北师大科研基金项目《调查学生敏感问题的统计模型与可靠性分析》的小组成员。这个在线调查是此项目的一部分，旨在调查学生在考试中出现欺骗行为所占的比例。在检验所提供调查方法的准确性与合理性的同时，我们也希望与同学们在这个敏感热点问题上进行真诚沟通。

在此在线匿名调查中，第二题我们采取了 Warner 的调查方法，在备选的三个问题中利用 Javascript 随机产生一道，三个备选问题分别是，“您在大学考试中是否有过欺骗行为？”；“您在大学考试中是否从未有过欺骗行为？”；“您周围的同学在大学考试中是否有过欺骗行为？”由于肯定否定答案在各题中含义各不相同，而我们并不会知道您抽取的是哪道题，故同学们不必担心自己的身份及答案外泄，放心真实填写即可。

具体操作是：选择院系，年级。一题答毕，请点击“获取问题”继续答题。

再一次感谢您的支持！如果愿意留下您的联系方式，您将收到我们的礼品。

请选择院系... 请选择年级..

您对我们学校考试作弊现象程度的看法是什么？

很普遍 普遍 较少 没有 说不清

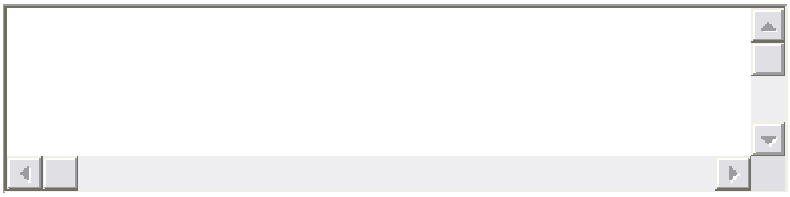
获取问题

XXXXXXXXXXXXXXXXXXXXXXXXX

是

否

你对考试作弊的看法及建议，也可以留下联系方式，以便我们发送礼品(不是必须要填)



提交

注：在网页中“获取问题”按钮为一 JAVA 程序，点击后将在“XXXXXXXXXXXXX”处按预设概率随机出现问题。

附录 B

拟合检验的 SAS 程序：

正态性假设检验：

```
data paper;
   input a b c d @;
    cards;
1   2   0   3
19  3   0   2
19  4   1   2
1   2   1   3
19  4   0   2
18  4   0   1
18  4   0   1
18  4   0   2
18  4   1   2
18  4   0   1
10  1   1   2
10  2   0   3
1   1   1   3
17  2   0   0
19  3   1   2
1   3   0   2
3   4   1   2
3   1   1   2
3   1   1   4
3   1   0   2
19  4   0   2
19  4   0   2
3   3   0   2
3   2   0   2
3   3   1   3
3   4   0   2
3   4   1   2
3   3   1   2
6   1   1   2
22  2   1   4
22  2   1   4
22  2   1   4
21  2   0   4
13  4   1   4
22  2   1   4
2   2   0   4
5   3   1   4
```

```
2   3   0   3
5   2   0   3
5   3   0   3
2   3   0   3
5   4   0   3
2   4   1   3
2   4   1   3
2   3   0   3
5   1   1   3
6   1   1   3
6   1   1   3
6   1   0   4
2   4   0   3
2   4   1   4
2   4   1   4
2   4   1   4
22  1   0   3
2   1   1   4
2   1   1   4
5   1   1   3
6   1   1   4
2   2   0   3
6   3   0   2
19  4   1   3
1   2   1   3
1   2   1   3
6   3   1   3
6   3   1   3
6   3   1   3
6   4   0   1
6   4   0   2
6   4   0   2
;
proc univariate data=paper normal;
   var d;
run;
```

T 检验：

```
data com  ;
input n per  @@;
cards;
1 9.5 1 27.2 1 53.4 1 2.6 1 7.4
2 1.5 2 5.8 2 31.9 2 37.7 2 23.2
;
proc ttest data=com  ;
class n  ;
var  per ;
    run;
```