网络调查的可信度问题研究

文/朱 胜 周小平 晏正春

摘 要:随着互联网技术的不断发展,网络调查成为数据收集的重要手段。本文从网络调查的可信度出发,分析影响可信度的三个主要问题,进而提出解决问题的思路。

关键词:网络调查;可信度;研究

一、引言

如同任何新生事物所遭遇的那样,网络调查在其发展之初便引起了极大的争议。乐观的人们被它明显的优点所吸引,为它美好的前景所激动,他们甚至断言,网络调查将很快取代传统的调查数据收集方法。谨慎的人们却用审慎的目光注视着它先天的缺憾——抽样框误差,而困扰调查实践的无回答率问题也并未因网络调查的出现而得以有效解决,他们重视历史的智慧,而对网络调查是否能扮演好自己的角色深表怀疑。

然而,网络调查以难以预料的方式快速发展,对调查活动产生了很大影响,其一,网络调查使调查数据收集活动趋于大众化。一定规模的数据收集不再限于专业的研究者,进行大规模数据收集的能力也不再限于那些作为社会权力中心的大型组织,网络调查成了互联网社会中每个成员都可以使用的数据收集工具。其二,网络调查使调查数据收集活动趋于多样化。网络调查的易于使用使其成为实现各种调查数据收集目的的随意性的工具,这在互联网对传统社会所产生的那些结构性影响中显得不足为奇。而互联网所孕育的无尽可能性和技术潜力则使得网络调查的实现方式异样纷呈,并与传统调查方式错列交杂。其三,网络调查使调查数据收集活动趋于劣质化。少数精心设计的高质量的网络调查淹没在其他数量巨大的基于网络的数据收集活动中而好坏难分,众多的网络数据收集活动使人们不胜其烦,从而降低了人们参加调查的意愿,这最终将威胁到网络调查方式本身,正如已经在电话调查发展过程中所发生的那样。换言之,普通意义的网络调查越是易于进行,符合专业标准的网络调查则越难付诸实施。

基金项目:国家社科基金“互联网时代统计数据的搜集及分析方法”项目(05BTJ002)。

它表明网络调查目前处于丛林时代。我们研究网络调查可信度,就是为了网络调查的未来发展建立起信心。

二、网络调查可信度的定义

如上所述,网络调查并非一种单纯的事物,而是一种内容众多的杂合体。它既指众多的基于网络的数据收集活动,也指众多的基于网络的数据收集方式。为研究方便,我们将网络调查当作一种数据收集工具,进行适当的分类并择其紧要者。

可信度是一个由来已久的概念,甚至可以追溯到人类关于事物真实性的最初哲学思考。类似的真实性问题导致了对计量工具性质的研究,人们用信度与效度描述计量工具所具有的或然的与相对的性质。信度被一般地定义为计量工具的计量结果之间所具有的一致性,效度则被一般地定义为计量工具的计量结果与计量目标的真实值的接近程度。常见的用于估计信度的方法有:重测可信度,内部一致性可信度,折半可信度等。而效度的估计则通常分为三种情形进行:内容效度,标准效度和结构效度。

而在统计学的调查理论中,类似的真实性问题则要复杂得多。按照抽样理论,这种真实性问题源自对一有限总体的某个未知参数的概率抽样估计。记为$\hat{\theta}$总体参数 θ 的样本估计量,则估计量的精确度可用估计量的均方误差表示:

$$MSE(\hat{\theta})=E[(\hat{\theta}-\theta)]=D(\hat{\theta})+B^2$$

其中,$D(\hat{\theta})$称为估计量的方差,有些类似于计量研究中信度;B 称为估计量的偏倚,有些类似于计量研究中的效度。B 有两种来源,一是源自估计程序——比较方便而又合适的估计量却是有偏的(比如比率估计量),一是源自调查程序。前者可以从数学上找到偏差的上限,而通过一定手段加以控制。后者通常可分为抽样框偏差、无回答偏差与计量偏差,对它们要找到一个可靠的小的上限,通常是不可能的,而一般的做法是对具体的问题进行实证的研究。

因此,网络调查的可信度由上述误差模型确定。研究网络调查的可信度,一方面,要区分基于概率抽样的网络调查与基于非概率抽样的网络调查,后者因缺乏科学的抽样程序而无法估计 $D(\hat{\theta})$,其可信度无从评价。另一方面,对基于概率抽样的网络调查而言,影响其可信度的主要因素有抽样框偏差、无回答偏差和计量偏差——统称为非抽样误差,它们是

评价其可信度的重要影响因素。

三、抽样框误差及其解决

抽样框误差源自目标总体与抽样总体的不一致，目前它成为基于概率抽样的网络调查的最大威胁。对网络调查而言，导致这种不一致的有三种情形：一是丢失了目标总体单位，即抽样框没有覆盖全部的目标总体单位，有些目标总体单位没有在抽样框中出现，因此也就没有被选入样本从而带来的误差，此类问题又称为样本的覆盖率问题。二是包含了一些非目标总体，即抽样框中包含了一些不属于研究对象的非目标总体单位。三是复合联接，当一个目标总体联接着一个以上的抽样单位时，便会产生复合联接误差。

（一）样本的覆盖率问题

覆盖率可定义为目标总体中可通过网络达及的部分所占的比率。尽管网络调查的支持者将互联网规模的快速增长作为对网络调查持乐观看法的一个依据，但关于互联网未来的扩散仍有疑虑。有许多关于 Web 普及率的估计，相互之间却有差异。即便援引较权威的研究，比如美国人口普查局的 CPS 调查，1998 年也只有 26.2%的美国家庭拥有互联网访问，2003 年虽增至 54.7%，然与美国近 95%的电话普及率相比，仍有较大差距。网络调查是否在未来某个时候对人口总体具有代表性取决于 Web 普及率增长的速度。

调查估计量在网络调查所覆盖者与未覆盖者之间的差异，可简单地用“丢失元素”类型的误差加以说明。

目标总体的 N 个元素由抽样框中的 N_A 个元素（网络调查所覆盖者）与丢失的 N_0 个元素（未覆盖者）组成：$N=N_A+N_0$。待估计的某总体总量 Y 即由两部分组成：$Y=Y_0+Y_A$。如果抽样框中没有包含非目标元素，y_A 是 Y_A 的无偏估计，则以 y_A 估计的均方误差为：$MSE(y_A)=var(y_A)+(Y-Y_A)^2$。测定估计中丢失元素影响的几个公式是：①净偏差，$Y_A-Y=-Y_0$；②相对偏差，$-Y_0/Y$；③标准偏差，$-Y_0^2/MSE(y_A)$。

令 $r=\bar{Y}_0/\bar{Y}_A=(Y_0/N)/(Y_A/N_A)$，$w_0=N_0/N$，则相对偏差可由如下公式给出：

$$\frac{-W_0 r}{rW_0+(1-W_0)}$$

若目标总体为整个人口总体，W_0 即为上述 Web 覆盖率，r 则可视为调查估计量在网络调查所覆盖者与未覆盖者之间的差异，可由之估计它们所引起的调查偏差。比如，当 $W_0=0.5$，$r=0.1$ 时，总量估计的相对偏差则为 -0.0909。

然而，r 之估计却并非易事。对于那些一般的人口特征变量，尽管互联网总的普及率在上升，但对于许多群体而言，数字鸿沟已经加宽，因为普及率在那些作为信息拥有者的群体中比在那些作为信息缺失者的群体中上升更快。

对于那些与调查有关的变量，有关的研究证据却极为缺乏。“互联网总体”在许多方面都不同于一般的人口总体，它与覆盖率一起，构成了目前对网络调查可靠性的最大威胁，其误差与其它调查方式中的误差一样，较难控制。但网络调查可以将所有的参与者作为目标总体，根据统计结果进行一般性的估计性的估计，得到所谓的“大众倾向”，这样就减弱了抽样框的样本代表性问题。这类调查通常适合于对网民关心的话题进行的各种调查，如社会风尚、日常生活观点等片面的民意测验和消费产品的满意程度调查等。随着网络的逐渐普及“样本的覆盖率问题”将逐步得到解决。

（二）抽样框过宽问题

对于第二类问题，网络调查方式更容易形成严格意义上的样本总体抽样框。如软件生产商对登记在案的上网用户进行电脑类产品使用情况的调查，可以采取以 E-mail 地址清册作为样本框，按随机原则抽取样本单位 E-mail 地址。也可以考虑随机 IP 法，即以随机抽取的一批 IP 地址作为样本，采用 IP 自动拨叫技术，向这些 IP 发出呼叫，传输邀请拥护参与调查的信息。这样，在很大程度上可以减少对非目标总体的包含，从而大大减少抽样框误差。

（三）复合联接问题

复合联接误差产生的原因主要在于大多数网民都拥有不止一个的邮箱，当应用随机电子邮件方式进行抽样时，这种误差便产生了。减少这一问题的最好措施是在数据采集过程中，通过问卷设计，设法获得每个样本元素的复合连接状况，如复合次数等，然后采用专门的多重估计量技术进行估计。

四、无问答误差及其控制

无回答误差是指由于种种原因没能从所有样本单位及问卷中的所有问题获得有用的数据，由此而形成的误差。从不同的研究角度，可对无回答误差做出不同的分类，在网络调查形式下，从被调查者的角度看，造成无问答问题的动因有两种，有意识不回答和无意识不回答。通常有意识不回答产生的背景是，被调查者对被调查的内容反感，或采取不回答的态度，例如，如果调查内容涉及敏感性问题，被调查者不愿提供，这时就会出现无回答问题。无意识无回答与调查内容无关，通常是由于被调查者很长期时间不上网或者工作繁忙或其他各种原因不能接受调查而产生单位无回答，或者答卷时由于粗心漏掉某个问题而造成的项目无回答。

与抽样框误差类似，无回答所引起的误差既与无回答率有关，也与调查估计量在回答者与无回答者之间的差异有关。应特别注意的是，对于那些抽样框未加确定的调查，无回答问题难以定义。这也就意味着，只有在抽样框和入样概率已经确定的情形——也就是所谓基于概率的抽样，无回答误差的测定或估计才可以进行。

在此情形下，目标总体（与抽样总体一致）的个元素各有其真实值 $Y_i(i=1,\cdots,N)$，对其进行完全的调查以估计其总量 $\sum_{i=1}^{N} Y_i$。如果第 i 个单位没有提供关于 Y_i 的信息，用一个可能有误差的估计值 Z_i 作为替补。根据无回答的随机论观点，第 i 个单位的标志值是一个随机变量，表示为：$\hat{Y}_i=R_iY_i+(1-R_i)Z_i$，替代的个别随机误差为 $\varepsilon_{\alpha i}=Z_i-Y_i$。以 $\hat{Y}_{pst}=\sum_{i=1}^{N}\hat{Y}_i$ 估计 $\sum_{i=1}^{N} Y_i$，其方差为：

$$var(\hat{Y}_{pst})=\sum_{i=1}^{N} p_i(1-p_i)B_{\alpha i}^2+\sum_{i=1}^{N}(1-p_i)\sigma_{\alpha i}^2$$

其均方误差的偏差部分为：$\sum_{i=1}^{N}(1-p_i)B_{\alpha i}$

而按确定论的观点（即 N 个元素可分为两部分：$N=N_0+N_1$，N_0 个为无回答者），上述偏差简化为：$\sum_{i=1}^{N_0} B_\alpha$。也可以写为：

$N(1-\lambda_1)(\bar{Z}_0-\bar{Y}_0)$

其中，$\lambda_1=N_1/N$，即回答率。若用回答单位的均值作为每个无回答单位的替代值，则偏差又可写为：

$N(1-\lambda_1)(\bar{Y}_1-\bar{Y}_0)$

其中，$1-\lambda_1=N_0/N$，即无回答率。而 $\bar{Y}_1-\bar{Y}_0$ 则为调查估计量在回答者与无回答者之间的绝对差异。这是评估无回答偏差的一个最为简明的模型，它清楚地表明了无回答所引起误差的两个组成部分：无回答率与调查估计量在回答者与无回答者之间的差异。如果在上述模型中假定 $|\bar{Z}_0-\bar{Z}_0\leq|\bar{Y}_1-\bar{Y}_0|$，即为无回答单位所寻找的替代值应该比回答单位的均值更接近于无回答单位的真实值，则可给出相对偏差的一个上限：

$$\frac{(1-\lambda_1)(1-\theta)}{\lambda_1+(1-\lambda_1)\theta}$$

其中，$\theta=\bar{Y}_0/\bar{Y}_1$，为调查估计量在回答者与无回答者之间的相对差异。

如果说随着 W_0（Web 覆盖率）的自然增长以及通过细致的实证研究对 r（调查估计量在网络调查所覆盖者与未覆盖者之间的差异）进行有效的控制，可以指望网络调查中的抽样框误差将得以降低的话，关于网络调查中的无回答误差却没有证据可以支持类似的期望。

降低无回答率的方法称之为预防方法。通常的做法是，先是试图识别影响无回答率的因素（比如，导致无回答的原因），然后对之加以控制。前期的大量研究是具体的实验研究（比如，物质激励是否对回答意愿有显著影响），后来的研究则试图为解释无回答行为提供理论框架。基于社会交换理论的理性行为方法认为有三类基本因素决定是否回答：付出，回报和信任。心理学的方法则强调受访者做决定时的经验规则——他们一般不会为做此类决定而花太多的时间和精力，并识别出几种启发性的因素：互惠准则，助人倾向，服从权威和稀缺感知。

无回答问题已成为网络调查实践中一个备受关注的问题，它对网络调查可信度的影响目前仅次 Web 覆盖率问题。无回答的存在减少了有效样本量，造成估计量方差的增大，尤其当无问答是有意识的不回答时，调查对象的数量特征往往与问答层的数量特征差异较大，造成估计偏差。因此必须对无回答误差给予足够的重视。

在网络调查的形式下，可以很好的减少无问答误差。首先，针对有意识的不回答，网络调查条件中被调查者是在一种相对轻松和从容的气氛，按受调查者，不用与调查者面对面地接触，较好的保全了被调查者的隐私，因此最大限度的保证了调查结果的客观性，也减少了有意识不回答的概率。而在网络调查的交互性下，应答者有机会问及调查问卷的目的、问题的含义及其他与调查有关的问题，这也可以减少由于对问题的不清楚而产生的有意识不问答。其次，同与被调查者联系不上形成了无意识无回答误差相对比，发送电子邮件问卷时，无法传送的邮件马上就被退回了，调查者便可以迅速地查明刚才发送的邮件被接收的日期，并相应做出调整，以便发送给更多的应答者。这也极大的减少了无意识无回答误差。同时自动扫描调查问卷可以分辨出漏答和答错的问题，并在提交之前提醒应答者，这同样有助于减少无回答误差。

五、计量误差的产生和解决

抽样调查中的计量误差问题与心理测试研究中的计量误差问题近乎一致，在研究方法上前者侧重于观察研究，后者侧重于实验研究。

计量误差是指调查中所获得的数据与所欲调查项目的真实值之间不一致而产生的误差。在传统的调查方式中它产生于统计的登记、汇总、计算等过程。在网络调查形式下，由于采用计算机代替手工处理，不需要重新打印或把结果制成表格，调查结果可以直接以电子表格的形式从问卷转化成数据库。因此，人工输入数据时错误即使不能全部消灭也能减少很多。从这个意义上讲，在网络调查形式下，计量误差已经不再具有传统的计量误差的意义。人们在网络调查形式下研究的计量误差通常被分为由调查者引起的计量误差和由被调查者引起的误差。由调查者引起的误差是指由于设计的问卷让被调查者感到模棱两可而难以问答所产生的误差。调查问卷的调查项目含糊不清，繁琐复杂，回答问题所需时间长，问卷中使用了带有倾向性、诱导性的词汇，都会导致被调查者的错误回答，从而产生计量误差。因此，在问卷没计时，应力图避免出现这一类问题。由被调查者引起的误差是指由于网络的虚拟性使人们在回答问题时产生顾虑而回答错误所产生的误差。这一误差与无回答误差有很大的重叠性，因此，我们就把它归纳在于无回答误差里边。

参考文献：

[1]孙伶莉.网络调查中的非抽样误差[J].统计与决策，2003(8).

[2]穆广杰.网络统计调查与传统统计调查比较[J].郑州航空工业管理学院学报，24 卷第 1 期.

作者单位：成都信息工程学院

（责任编辑：叶祥凤）