缺失数据的统计处理

金勇进 邵军/著

中国统计出版社
China Statistics Press

缺失数据的统计处理

金勇进 邵军/著

中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

缺失数据的统计处理/金勇进,邵军著.
—北京:中国统计出版社,2009.1
ISBN 978-7-5037-5627-6

Ⅰ.缺…
Ⅱ.①金… ②邵…
Ⅲ.统计数据—数据处理
Ⅳ.O212

中国版本图书馆 CIP 数据核字(2008)第 213271 号

缺失数据的统计处理

作　　者/金勇进　邵　军
责任编辑/胡文华　余成瑶
装帧设计/艺编广告
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街 57 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号
网　　址/www.stats.gov.cn/tjshujia
电　　话/邮购(010)63376907　书店(010)68783172
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/787×1092mm　1/18
字　　数/190 千字
印　　张/12.5
印　　数/1—2000 册
版　　别/2009 年 1 月第 1 版
版　　次/2009 年 1 月第 1 次印刷
书　　号/ISBN 978-7-5037-5627-6 /O·69
定　　价/20.00 元

前 言

如何对缺失数据进行处理一直是统计学家们感兴趣的话题。缺失数据的影响显而易见，它不仅可能造成估计量的偏差，还会导致估计量方差的扭曲，使传统的统计方法的效率降低。自20世纪70年代以后，国际统计学界对缺失数据问题的讨论日渐热烈，至今方兴未艾，其重要原因在于两个方面，一个是人们对缺失数据的认识在不断深化，不断涌现出的现代统计方法为研究数据规律提供了理论支持；另一个是计算机技术的飞速发展，将原本复杂的统计计算变得不再那么复杂，从而使得该领域的研究向操作层面大大迈进，推进了缺失数据问题研究的应用价值。目前已经有《Statistical Analysis with Missing Data》这样的著作出版，这也表明，与20世纪的研究相比，目前的研究已经进入到一个新的阶段。

缺失数据是一个普遍现象，在我国同样存在。解决统计调查中的缺失数据，是提高数据质量的一个重要方面。目前从国内情况看，我们对缺失数据虽有研究，但还不够广泛和深入，也还没有专门的著作对处理缺失数据的统计方法进行系统的介绍和讨论，本书的写作目的是想填补这一空白。

处理缺失数据的具体方法很多，为便于叙述和讨论，本书将这些方法概括为三类，即加权方法、插补方法和参数似然方法。加权方法的实质是将赋予缺失数据的权数分解到非缺失单元身上，如何进行分解有多种方法，在不同缺失数据机制下，各种方法有不同的特点，本书的第二

章主要讨论这些问题。插补方法是处理缺失数据的另一类方法，其核心问题是为缺失数据寻找一个“替代值”。一般而言，插补不会提高估计的精度，因为我们无法判断插补值与原始值的近似程度，但插补仍然是必要的。通过插补，我们可以利用数据采集者的先验信息得到完整的数据集，为以后采用其他统计方法弥补“缺陷”，插补还可以减少估计偏差，本书第三章中对多种不同的单一插补方法进行了讨论。加权方法和插补方法容易理解、操作简单，但在有些情况下并不是效率最高的方法，似然函数方法处理缺失数据通常能产生更优良的估计量。本书第四章对处理缺失数据的几种基本的参数似然方法进行了分析。在对上述处理缺失数据不同类型的方法讨论后，第五章进入对纵向缺失数据问题的研究。目前，我们使用纵向数据的场合越来越多，纵向数据是另外一种类型的数据，对有缺失的纵向数据进行处理和分析有自己的特点和更为有效的方法，对这类问题，感兴趣的读者可以参阅本书第五章。本书的第六章讨论方差估计问题。我们将方差估计的方法分为三类，分别是直接推导法、多重插补法和重抽样法。本章对不同方法的内容和特点进行比较充分的介绍和分析。

本书的一个特点是内容尽量通俗和简化。有关缺失数据问题研究的文献很多，内容比较繁杂，阅读这些文献需要花费很多时间。本书将各类处理缺失数据的统计方法进行了概括和简化，以通俗的表达方式展现给读者，希望读者能够在很短时间内对处理缺失数据的各类方法有一个梗概的了解。我们认为在对缺失数据整体研究提高厚度和深度的初始阶段，综述性介绍已有研究成果的著作是十分必要的，希望通过本书的出版，有更多的学者参与到该领域问题的研究。

本书的另一个特点是联系实际。与纯粹讨论方法的研究文献不同，本书将方法研究与实际问题的解决结合在一起，第二章、第三章、第五章、第六章的最后都有“案例分

析”一节。这些案例，有些取自于作者所从事的课题项目，有些则是通过模拟试验，论证不同方法的特点，从而使本书内容具有更强的可操作性。我们希望本书的出版对从事数据分析的实际工作者也有所帮助。

我们要对在本书完成过程中给予帮助的人士表示感谢。中国人民大学统计学院的博士研究生谢佳斌、吕萍参与了将作者部分英文手稿翻译为中文的工作；硕士研究生吴潇、马诚提供了部分模拟计算的初稿；谢佳斌协助作者对全书进行了技术性的审核和修订；同时感谢中国统计出版社严建辉社长、杨映霜编审为出版本书给予的大力支持和帮助。

进行这样一本书的写作是富有挑战性的工作。尽管我们十分努力，但由于水平有限，书中肯定存在不少缺欠和疏漏，恳请广大读者提出宝贵意见。

金勇进　邵军

2008.12

目录

目录

目录

目录

第一章

导　论

§1　概　　述

§1.1　问题的提出

统计是一门有关数据的科学，统计方法的魅力是在对数据分析的过程中体现出来的。然而，如果出现数据缺失，对缺失的数据又没有采用任何方法进行补救，统计方法的分析效率将受到影响。

统计数据主要来自于两个方面：调查的数据和实验的数据。统计调查中的数据缺失是影响统计数据质量的一个重要方面，在概率抽样中，缺失数据将导致统计推论中出现估计量偏差和估计方差增大，在其他调查方式中，缺失数据也会对统计数据的质量产生影响，使统计数据的说服力降低。况且，各类调查特别是抽样调查应用的领域越来越广阔，各种干扰调查的因素也逐步凸现，调查中出现缺失数据已经成为不可避免的现象，我们正面临着缺失数据的挑战。实验中的缺失数据也会带来相同的后果。这种现象是普遍性的，正因为如此，对缺失数据问题的研究，一直是国际统计学界热点讨论的课题之一。国内对缺失数据问题的研究近几年虽有所发展，但与国外相比，仍有很大的差距，主要表现在，理论探讨方面缺乏原创性，基本上是介绍国外已有方法，即便如此，介绍的也不够全面和丰富；在应用方面，则几乎是空白。

本书将对处理缺失数据的一些方法进行比较系统的介绍和讨论，以期引起人们对缺失数据问题的关注。本书的定位在于方法的应用，我们将注意力侧重在处理缺失数据不同方法在应用中的特点，并辅之一些案例，希望能够对处理缺失数据的统计实践有所帮助。基于这样的写作目的，书中只有简单公式的推

导，而略去比较复杂的理论证明，在十分必要的地方，一些证明用附录的形式表现，希望使本书具有更强的可读性。当然，对理论推导感兴趣的读者可以参阅有关的参考文献，我们同时希望本书能为在该领域的深入研究抛砖引玉。

§1.2 缺失数据产生的原因

在不同领域，缺失数据产生的原因不同。例如进行农作物试验，目标变量是农作物产量，控制变量有水分、肥料、温度等。试验中可能会出现意外情况，如种子没有发芽，或发芽后被鸟叼啄，造成某些产量数据缺失。

调查中造成缺失数据的原因则更加多样。由于调查中缺失数据的现象更为普遍，数据缺失对统计分析带来的影响更为直观，所以人们对此也更为关注。调查中的数据缺失主要产生于两个方面，一个是调查中的不可使用信息，一个是调查中的无回答。

调查中不可使用信息主要指明显的错误信息。如数据录入中出现错误，多录或少录数据位数；调查过程中的记录错误，出现明显的错项、错填；也包括记录是正确的，但调查结果明显不符合逻辑，也许是被调查者有意或无意的错报，等等。这些错误在数据的逻辑审核中被发现，分析人员将这些错误的数据剔除，造成数据缺失。

调查中的无回答是造成缺失数据的另一个原因。无回答有两种：单位无回答和项目无回答。单位无回答是指调查中没有从样本单元获得任何调查问卷中所需要的信息。出现单位无回答的主要原因有：调查人员没有与被调查者取得接触，如被调查者不在家，或者地址不详，调查人员没有找到被调查者；被调查者拒绝接受调查；被调查者由于各种原因无法接受调查，如工作忙、生病等。项目无回答是指调查虽然进行，但被调查者只提供了调查问卷中的一部分信息，而没有提供调查问卷中的另一些信息，如被调查者反感调查中的某些问题而拒绝回答，或者调查人员粗心漏掉询问某些问题等。区分无回答的两种类型是有意义的，因为处理缺失数据的一些方法更适合于单位无回答，而另一些方法更适合于项目无回答，当然也有一些方法对两类无回答都可以使用。

在一些情况下，调查中的不可使用信息可以转化为调查中的无回答。例如某样本单元的调查数据质量极差，无法使用，将这份调查问卷剔除，就变成了单位无回答；或者可以明确判断某个问题被调查者的回答是错误的，将这个错误答案剔除，就变成了项目无回答。所以调查中的无回答是造成缺失数据的主要原因这种表述具有一般性。本书的目的虽然是研究一般性的缺失数据问题，但是更关注调查领域的缺失数据，也就更关心无回答造成的缺失数据。在后面内

容的讨论中，如果没有特殊的指明，无回答和缺失数据表明同样的意思，为了表述方便，“无回答”、“缺失数据”以及“缺失值”会交互使用。但是，缺失数据的概念更为宽泛，由于无回答造成的缺失数据也只是缺失数据中的一部分，在缺失数据的处理方法有特定的应用领域时，我们也会注意提及，使读者了解这些方法在不同背景下的应用。

§1.3 缺失数据的模式

缺失数据模式描述了在整个数据集中哪些数据被观测到了，哪些数据缺失了。它有助于我们认识数据集中不同变量之间的相互关系，从而为寻找更好的解决方法提供有价值的线索。

将不同的缺失数据的情况加以归纳，可以概括出几种数据缺失的模式，如下图所示。

图 1.1 数据缺失模式

注：①○表示观察值，◎表示缺失值；②为方便起见，这里取变量个数=5。

1)单变量缺失模式

单变量缺失模式如图 1.1 中(a)所示。缺失数据仅仅限于单个变量，如农业试验，y_j 表示粮食产量，存在缺失数据，而 $y_1, y_2, \cdots, y_{j-1}$ 分别代表温度、施肥量、施肥种类及降雨量等影响粮食产量的因素，这些变量都是可以完全观测

到的。调查中的项目无回答也会表现为这种模式，例如在某个敏感性问题上，一些样本单元拒绝提供信息。

2)多变量缺失模式

多变量缺失模式如图 1.1 中(b)所示。从第 j 个变量 y_j 开始，之后变量缺失的项目都相同。图(b)中 $j=3$，y_j 之前的变量可以完全观测到，之后变量的数据缺失。例如，调查中有若干个敏感性问题，一些样本单元在这些敏感性问题上均拒绝提供信息。

3)单调缺失模式(monotone missingness pattern)

单调缺失模式如图 1.1 中(c)所示。对数据资料阵 y 进行适当的行列变换后，可以得到这样一个矩阵，它呈现出一种层级缺失的模式，即：当矩阵中的元素 y_j 缺失时，则对任意的 $p\geqslant j$，元素 y_p 也是缺失的。这种模式最典型的是固定样本点的调查，随着时间推移，一些样本单元逐渐丢失，之后的数据便不复存在。医学上对病人跟踪调查的缺失数据许多也属于这种模式。

4)一般缺失模式

一般缺失模式如图 1.1 中(d)所示。数据缺失具有偶然性，没有规律可循，这是最一般和最典型的数据缺失模式，如抽样调查中经常出现的项目无回答。

§2 缺失数据机制

缺失数据机制描述了缺失数据与数据集中变量值之间的关系，这种描述十分重要，缺失数据机制试图从本质上说明数据是如何缺失的，本书后面讨论的各种处理缺失数据的方法都建立在缺失数据机制的某种假定上。

为了进行说明，需要利用一些符号。假定：

y：感兴趣的目标变量，

x：与目标变量有联系的辅助变量(协变量)，

a：目标变量是否缺失的指示变量，例如 $a=1$ 表示有回答，$a=0$ 表示无回答，

Y：在没有缺失值条件下的目标变量数据集，

X：在没有缺失值条件下的辅助变量(协变量)数据集，

A：缺失数据指示变量数据集。

不同学者对缺失数据机制有不同的划分。本书将缺失数据机制划分为六种类型，它们分别是：完全随机缺失、随机缺失、取决于协变量的缺失、非随机缺失、取决于随机影响的缺失和取决于前期数据的缺失。

§2.1 完全随机缺失

完全随机缺失(Missing Completely at Random)简称 MCAR,是指目标变量集 Y,协变量集 X 和指示变量集 A 相互独立。在 MCAR 机制下,有

$$L(y|a=0)=L(y|a=1)=L(y) \tag{1.1}$$

式中 L 表示分布。式(1.1)的意思是说目标变量的回答数据集与无回答数据集有相同的分布,该分布就是目标变量分布本身。例如,在某项血液研究中,实验室工作人员不慎丢失了装有某一被调查者血样的试管。没有理由表明丢失试管(无回答)这一事件与被调查者的白细胞数量(y)或其他因素(x)有任何关系,也就是说无回答的发生完全随机,此时可以把回答单元看作是从样本单位中简单随机抽取的子样本。

在完全随机缺失机制下,对含有缺失值的数据集采用通常的统计分析方法是可行的,估计量无偏,但不同方法的估计效率存在差别。例如,如果估计均值,不考虑缺失数据,仅仅用有回答的数据进行估计,则点估计为

$$\bar{y}_1=\frac{\sum_{i=1}^{n}a_iy_i}{n_1}$$

n_1 表示样本中的回答单元个数。显然 $\bar{y}_1$ 是目标总体均值 μ_y 的无偏估计,该估计量的方差为

$$V(\bar{y}_1)\approx\frac{\sigma_y^2}{pn}① \tag{1.2}$$

式中 $p=P(a=1)$,即样本单元的回答概率,证明见附录。

而如果采用基于模型的估计方法,即假定 $E(y|x)=\beta x$,$V(y|x)=\sigma^2x$,系数 β 的线性无偏估计为

---

① 本式的证明如下:

$$\begin{aligned}V(\bar{y}_1)&=E[V(\bar{y}_1\mid A)]+V[E(\bar{y}_1\mid A)]\\&=E\left[\frac{\sum a_i^2V(y_i\mid a_i)}{n_1^2}\right]+V\left[\frac{\sum a_iE(y_i\mid a_i)}{n_1}\right]\\&=E\left[\frac{\sum a_i\sigma_y^2}{n_1^2}\right]+V\left[\frac{\sum a_i\mu_y}{n_1}\right]=E\left[\frac{\sigma_y^2}{n_1}\right]\\&\approx\frac{\sigma_y^2}{pn}\end{aligned}$$

$$\hat{\beta}=\frac{\sum_{i=1}^{n}a_iy_i}{\sum_{i=1}^{n}a_ix_i}$$

则点估计为

$$\bar{y}_I=\hat{\beta}\bar{x}_n$$

显然，$\bar{y}_I$ 也是目标总体均值 μ_y 的无偏估计，该估计量的方差为

$$V(\bar{y}_I)\approx\frac{\mu_x\sigma^2+p\beta^2\sigma_x^2}{pn}\ ①\tag{1.3}$$

证明见附录。

由于

$$\begin{aligned}\sigma_y^2&=V(y)=E[V(y|x)]+V[E(y|x)]=\sigma^2E(x)+V(\beta x)\\&=\mu_x\sigma^2+\beta^2\sigma_x^2\end{aligned}$$

所以

$$V(\bar{y}_1)-V(\bar{y}_I)\approx\frac{(1-p)\beta^2\sigma_x^2}{pn}$$

显然约等式的结果大于零，所以基于模型的估计方法比简单的估计方法有更高的效率，特别是当 p 较小，或者 β^2 较大，或者 σ_x^2 较大的时候更是如此。这说明，在确定数据缺失机制后，不同统计分析方法的效率是有差异的。

现实中完全随机缺失的情况有所存在，但并不普遍。

---

① 本式的证明如下：

$$\begin{aligned}V(\bar{y}_I)&=E[V(\hat{\beta}\bar{x}_n|A,X)]+V[E(\hat{\beta}\bar{x}_n|A,X)]=E[\bar{x}_n^2V(\hat{\beta}|A,X)]+V(\beta\bar{x}_n)\\&=E\left[\bar{x}_n^2\frac{\sum_i a_iV(y_i\mid x_i)}{(\sum_i a_ix_i)^2}\right]+\beta^2V(\bar{x}_n)=E\left[\bar{x}_n^2\sigma^2\frac{\sum_i a_ix_i}{(\sum_i a_ix_i)^2}\right]+\frac{\beta^2\sigma_x^2}{n}\\&=\sigma^2E\left[\frac{\bar{x}_n^2}{n_1\bar{x}_1}\right]+\frac{\beta^2\sigma_x^2}{n}\approx\sigma^2\frac{E\bar{x}_n^2}{E(n_1)E\bar{x}_1}+\frac{\beta^2\sigma_x^2}{n}\\&=\sigma^2\frac{\mu_x^2+V(\bar{x}_n)}{pn\mu_x}+\frac{\beta^2\sigma_x^2}{n}=\frac{\mu_x\sigma^2}{pn}+\frac{\sigma^2\sigma_x^2}{pn^2\mu_x}+\frac{\beta^2\sigma_x^2}{n}\\&\approx\frac{\mu_x\sigma^2+p\beta^2\sigma_x^2}{pn}\end{aligned}$$

§2.2 随机缺失

随机缺失(Missing at Random)简称 MAR,是指目标变量 y 是否缺失只是与已经观测到的 y 值有关,而与缺失的 y 值无关。假定:

Y_o:Y 的观测部分(Observed),

Y_m:Y 的缺失部分(Missing),

在 MAR 机制下,有

$$L(A|Y,X)=L(A|Y_o,X) \tag{1.4}$$

式(1.4)的意思是说缺失数据(即是否回答的指示变量数据集 A)只是与 (Y_o,X) 有关,而与 Y_m 无关。例如,对人群进行健康检查,如果某些检测指标超过允许范围,该被检查者就要进入医院进行治疗,是否送入医院取决于已经观测到的数据,而与没有观测的数据无关。

如果没有协变量 X,则随机缺失的典型情况就是单调缺失模式。在单调缺失模式下,如果 $a_i=0$,则 $a_{i+1}=\cdots=a_n=0$,即

$$a_i=\begin{cases}1 & \text{如果 } y_{i-1}>c\\ 0 & \text{否则 } y_{i-1}\leqslant c\end{cases}$$

这时式(1.4)可以写为

$$L(a|y)=L(a|y\text{ 的观测部分}) \tag{1.5}$$

例如,学校定期对毕业生进行跟踪调查,随着时间的推移,有些毕业生的地址发生了变化而无法联系到,联系到的毕业生逐期减少。能否联系到毕业生与前期观察有关(通讯地址),这就属于单调缺失模式,是 MAR 的一种类型。

在 MAR 机制下,似然估计有特殊的意义。假定没有协变量,目标变量也没有缺失数据,可以利用 Y 的分布 $L(Y)$ 对目标参数进行似然估计。

如果目标变量出现数据缺失,但是有观测数据 Y_o 和 A,我们可以通过把似然函数 $L(Y,A)$ 中 Y_m 积分掉的方法对目标参数进行估计,即

$$\begin{aligned}\int L(Y,A)dY_m &= \int L(Y)L(A\mid Y)dY_m = \int L(Y)L(A\mid Y_o)dY_m\\ &= L(A\mid Y_o)\int L(Y)dY_m\end{aligned}$$

如果未知参数在 $L(Y)$ 和 $L(A|Y_o)$ 中是不同的,那么通过使用