缺失数据统计处理方法的研究进展*

帅 平[1] 李晓松[2△] 周晓华[3] 刘玉萍[1]

临床试验和流行病学调查中经常出现缺失数据[1-2]。一直以来,统计学家们研究的分析方法主要针对完整数据,含缺失值的数据无疑给生物医学者在实际应用分析时带来不少困难[3-4]。Croy[5]等的研究发现,在随机抽取的 25 篇关于质量分析的文献中,仅有 3(12%)篇文章对缺失值进行了处理,采用的方法仅是均值替代、多重回归或根据经验取值替代。Wood[6]等对 2001 年发表在 BMJ、JAMA、Lancet 和 New England Journal of Medicine 期刊上的随机对照试验分析后发现,缺失数据在这些试验中普遍存在,但未得到很好的处理和分析。缺失数据的出现给数据分析和研究推论带来困难,尤其当完全观测数据和不完全观测数据存在系统差异时,常规处理方法得到的结果通常不能代表整体。处理不当时可能导致方差增大,检验效能降低,无法得到科学合理的解释和结论。如何有效处理缺失数据,怎样才能充分利用数据信息,准确地反映研究群体的特征,达到预期研究目的,已成为当前统计研究中的难点和热点问题。本文将就当前国内外缺失数据的处理方法进行一综述。

常见的处理缺失数据的方法

20 世纪 70 年代后期,国外学者对缺失数据问题的研究开始重视并日渐增多。Dempster,Laird & Rubin[7]首先提出了一种有效处理缺失数据的算法-EM 算法,该算法为处理缺失数据带来了新的革命;正是基于这一算法,Rubin[1]在 80 年代末提出了多重填补的方法;Schafer & Olsen[8]在 1998 年提出了对多变量缺失值的多重填补法;Robins、Rotnitzky & Zhao[9]在 1994 年提出了以估计缺失概率为基础的加权法;Qin[10]和 Tang[11]等学者在 2002 年和 2003 年分别提出了两种不同的运用似然函数的半参数方法来处理不可忽略缺失数据机制的问题。我们将这些学者提出的方法大概归为三类,分别是:基于填补的方法,基于参数似然的方法和基于加权调整的方法。

1. 基于填补的方法

填补是处理缺失数据常用的一类技术方法,其优点是:研究者可以对经过填补后的数据集采用完全数据的分析方法,而不需要采用单独的复杂的算法;在一些情况下,填补可以减少由于无应答等造成的估计偏差,尤其是在拥有比较高质量的辅助信息时。但是,填补法也有缺点,填补过程可能很困难且不容易实现,特别是在多维复杂结构下;另外,一些简单的填补可能歪曲数据的分布和变量间的真实关系。根据对每个缺失值的填补个数来分类,可分为单一填补和多重填补。

(1)几种单一的填补方法

①均值填补(mean imputation)

均值填补是用样本中有观测值的均值代替缺失值,可分为非条件均值填补和条件均值填补。非条件均值填补是指对所有的缺失值,用所有观测值的均值进行填补,因此所有填补值都是相同的。条件均值填补是利用辅助信息,对总体进行分层,使各层中的各单元尽可能相似,然后在每层中用该层有响应单位的均值填补该层中的缺失值。分层均值填补比非条件均值填补的填补效果好。但是均值填补通常改变了变量的变异程度,低估填补变量的方差。因此一般情况下均值填补比较适合简单的描述性研究,不适应于较复杂的需要方差估计的分析[12-13]。

②演绎填补(deducive imputation)

演绎填补法是通过可以搜集到的复杂资料,依据逻辑和常规,对缺失数据进行推断,找出填补值。用公式表示就是 $Z_i = f(X_i)$,其中 z_i 为第 i 个缺失数据的填补值,X_i 是辅助变量,$f(*)$ 是根据缺失数据的目标变量 y 与辅助变量 X 之间的逻辑运算关系构造的函数。该方法操作简单,在有高质量的辅助信息下,可以提供准确或近乎准确的填补值,但其效率很大程度上依赖于辅助资料是否充分。

③回归填补(regression imputation)

回归填补是由单元的缺失项对观测项的回归,用预测值代替缺失值。通常由观测变量及缺失变量都有观测的单元进行回归计算。填补中还可以给填补值增加一个随机成分,这种方法称为随机回归填补。它是用回归填补值加上一个随机项,预测出一个缺失值的替代值,该随机项反映所预测的值的不确定性影响。随机回归填补法能够较好的利用数据提供的信息,解决因预测变量高度相关引起的共线性问题[14]。

* 本文获国家自然科学基金项目资助(项目编号:3072819)
1. 四川省医学科学院·四川省人民医院健康管理中心(610072)
2. 四川大学华西公共卫生学院卫生统计教研室
3. 美国华盛顿大学公共卫生学院生物统计系
△通信作者:李晓松,E-mail:lixiaosong1101@126.com

④最近距离填补(nearest neighbor imputation)

最近距离填补法是利用辅助变量,定义一个测量单元间距离的函数,在缺失值临近的回答单元中,选择满足所设定距离条件的辅助变量中的单元所对应的变量的回答单元作为填补值,即在填补类中按匹配变量找到与受者记录最接近的供者记录。用于定义赋值单位的距离函数可以有很多类型,马氏距离就是其中一种。由于距离函数有不同类型,用最近距离函数得到的填补值具有伪随机性,这给考察最近距离填补估计量的性质带来了挑战。

⑤热卡填补(hot deck imputation)

热卡填补中常见的有随机热卡填补法和序贯热卡填补法。随机热卡填补是通过对变量 Y 的回答单元进行有放回的简单随机抽样获得填补值。这里的填补值是随机的,避免了均值填补中方差低估的缺点。序贯热卡填补法首先对数据分层,然后在每层中按照某种顺序对单元排序,对于有数据缺失的单元,用同一层中最后一个被计算机读取的数据进行填补。该方法存在的问题是填补值的选择是由辅助变量决定的,用不同的变量进行排序,得到的序列不同,对某一缺失值来说可能采用的填补值也就不同。因此,应该选择与研究变量性质高度相关的排序变量,使得排列位置相邻的单位在研究性质上也相近[15-16]。

⑥冷卡填补(cold deck imputation)

冷卡填补法是相对于热卡填补而言的,指填补值不是从当前的调查,而是从以往的调查或者其他历史数据中获得的。

上述单一的填补方法通常可能会扭曲目标变量的分布,使填补变量的方差被低估,还可能歪曲变量与变量间的关系,无法得到真实的效应结果[4,15]。另外一个问题是基于填补的数据推断参数,无法解释填补的不确定性。

(2)多重填补(multiple imputation,MI)

多重填补由 Rubin 在 1978 年提出[1],它通过某种方法对每个缺失值都构造 d 个替代值($d\geqslant 2$),以形成 D 个完整的数据集,对每个数据集均采用相同的针对完整数据集的统计方法分析,将得到的结果综合,产生最终的统计推断。与单一的填补方法相比,MI 能反映由缺失数据带来的不确定性,增加了估计的效率。

多重填补中最关键的问题是如何进行有效的填补,从理论上讲缺失值可以从联合后验预测分布中进行抽取。但在实际中尤其是复杂问题中要做到这点并不容易,特别是在多变量数据及涉及非线性关系等情况下。近十年里,逐渐形成了两种最常见的对多元数据进行填补的策略,分别是联合模型法和全条件定义法。

①联合模型法(joint modeling,JM)

JM 在给定数据 Y 和模型参数 θ 下假定参数的多元密度分布为 $P(Y|\theta)$,在给定一个 θ 的适当的先验分布和上述假定下,利用贝叶斯理论从联合后验预测分布 $P(Y_{mis}|Y_{obs})$ 中抽取产生填补值,通常是在可忽略的缺失机制(missing at Random,MAR)下。该方法能产生对参数的有效推断,被认为是适当的填补。JM 通常需要特殊的方法来实现,数据扩张(data augmentation,DA)即是基于此策略的填补方法。

DA 最早由 Tanner & Wong[17] 提出,分为借补步(Imputation,I 步)和后验步(Posterior,P 步)。若在第 t 次迭代时 θ 的一个抽取值为 $\theta^{(t)}$,那么

I 步:抽取 $Y_{mis}^{(t+1)}$,使其具有密度 $p(Y_{mis}|Y_{obs},\theta^{(t)})$

P 步:抽取 $\theta^{(t+1)}$,使其具有密度 $p(\theta|Y_{mis}^{t+1},Y_{obs})$

I 步中的缺失值是从给定已观测数据和当前的参数值后的条件分布进行抽取。P 步中参数的抽取可以看作是从完整数据后验分布的一个抽取。因此,进行数据扩张将产生 Y_{mis} 的后验预测分布的一个抽取值和 θ 的后验分布的一个抽取值。这一迭代过程可以产生给定 Y_{obs} 下 Y_{mis} 和 θ 的联合后验分布中的一个抽取。当 $t\to\infty$ 时,迭代过程收敛到一个给定 Y_{obs} 下(Y_{mis},θ)的联合分布的抽取。

②全条件定义法(fully conditional specification,FCS)

JM 的理论是可靠的,但缺乏对模型设定的灵活性,尤其在数据特征比较特殊时,可能还会导致结论的偏倚。有学者通过模拟研究分析发现 JM 在一些情况下表现不佳,认为“分别进行回归可能比联合模型更有意义”[18-19]。

FCS 由 Van Buuren 等提出[20],它在填补时不考虑被填补变量和已观测变量的联合分布,而是利用单个变量的条件分布建立一系列回归模型逐一进行填补。假设 X 为无缺失变量集,$Y=(Y_1,Y_2,\cdots,Y_j)$ 为 j 个带缺失值的变量,FCS 迭代地从下面形式的条件分布中进行抽取:

$$P(Y_1 \mid X,Y_{-1},\theta_1)$$
$$\vdots$$
$$P(Y_J \mid X,Y_{-J},\theta_J)$$

每一次迭代包括对所有 Y_J 进行抽取的一个循环。具体在第 t 次迭代中,有:

$$\theta_1^{*(t)} \sim P(\theta_1 \mid X,y_1^{obs},y_2^{(t-1)},\cdots,y_J^{(t-1)})$$
$$y_1^{*(t)} \sim P(y_1^{mis} \mid X,y_1^{obs},y_2^{(t-1)},\cdots,y_J^{(t-1)},\theta_1^{*(t)})$$
$$\theta_2^{*(t)} \sim P(\theta_2 \mid X,y_1^{obs},y_2^{(obs)},\cdots,y_J^{(t-1)})$$
$$y_2^{*(t)} \sim P(y_2^{mis} \mid X,y_1^{(t)},y_2^{obs},\cdots,y_J^{(t-1)},\theta_2^{*(t)})$$
$$\vdots$$
$$\theta_J^{*(t)} \sim P(\theta_J \mid X,y_J^{obs},y_1^{(t)},y_2^{(t)},\cdots,y_{J-1}^{(t)})$$
$$y_J^{(t)} \sim P(y_J^{mis} \mid X,y_J^{obs},y_1^{(t)},y_2^{(t)},\cdots,y_J^{(t)},\theta_J^{*(t)})$$

上式在抽取 $\theta_j^{*(t)}$ 时并没有用到 y_j^{mis} 的信息，这与 DA 不同。FCS 通常迭代次数比较少，一般为 5～10 次。n 次迭代全部结束后，取第 n 次的填补值作为最终结果，形成一个完整数据集。要得到 D 个数据集，需要将上面的 n 次迭代独立进行 D 次。

FCS 又称为迭代的单变量填补(iterated univariate imputation)，序列回归(sequential regressions)，链式方程(chained equations)等。其优势在于将一个 K 维问题分解成 K 个一维问题，可以创建更加灵活的除常见多元模型外的其他模型形式，解决在多元密度下难以进行填补的问题，在建立不可忽略缺失机制的模型时也相对较容易[21-22]。FCS 在很多实际应用中表现良好，模拟研究也证明能得到无偏的估计和较好的收敛性[21-25]。

多重填补处理缺失数据的优势被日益重视并得到广泛应用，许多软件开发了相应程序。SAS 中的 PROC MI 和 PROC MIANALYZE[26]、S-Plus6.0、SOLAS[25]，NORM[27]，以及 LISREL[25]等均可以进行多重填补运算。R 软件中含有多个可以处理缺失数据的软件包(package)，如 norm[28]、cat[29]、mix[30]、pan[31]、mi[32]等。此外 R 中的 mice[33]则利用 FCS 的思想，即链式方程的方法进行多重填补。

2. 基于参数似然的方法-极大似然估计法(maximum likelihood estimation，MLE)

极大似然估计法是在总体分布类型已知情况下的一种参数估计方法[4]。在模型假定正确的情况下，若缺失机制为随机缺失，通过已观测数据的边际分布可以对未知参数进行极大似然估计，得到未知参数的准确估计值。

假设变量 $Y=(Y_{obs},Y_{mis})$，M 为缺失数据的指示变量，$M=1$ 表示 y 缺失；$M=0$ 表示 y 观测到。Y_{obs} 和 Y_{mis} 联合分布的概率密度为 $f(Y|\theta)=f(Y_{obs},Y_{mis}|\theta)$。$Y$ 和 M 的联合分布可描述为 Y 的分布密度和给定 Y 和 θ 下 M 的条件分布密度的乘积，即 $f(Y,M|\theta,\varphi)=f(Y|\theta)f(M|Y,\varphi)$，其中 φ 为 M 的参数。

实际观测数据的分布为 $f(Y_{obs},M|\theta,\varphi)=\int f(Y_{obs},Y_{mis}|\theta)f(M|Y_{obs},Y_{mis},\varphi)\,dY_{mis}$。$\theta$ 和 φ 的整个似然为 $L_{full}(\theta,\varphi|Y_{obs},M)\propto f(Y_{obs},M|\theta,\varphi)$。在 MAR 下，基于数据 Y_{obs} 的 θ 的似然为 $L_{mar}(\theta|Y_{obs})\propto f(Y_{obs}|\theta)$。$\theta$ 的推断通常应该基于整个似然 $L_{full}(\theta,\varphi|Y_{obs},M)$。但在 MAR 时，缺失数据的分布不依赖于缺失值 Y_{mis}，有 $f(M|Y_{obs},Y_{mis},\varphi)=f(M|Y_{obs},\varphi)$ 对一切 Y_{mis}。那么可以得到：

$$f(Y_{obs},M|\theta,\varphi)=f(M|Y_{obs},\varphi)\times\int f(Y_{obs},Y_{mis}|\theta)\,dY_{mis}=f(M|Y_{obs},\varphi)f(Y_{obs}|\theta) \quad (1)$$

这样在 MAR 时，由于产生的似然成正比，根据 $L_{full}(\theta,\varphi|Y_{obs},M)$ 对 θ 基于似然的推断与根据 $L_{mar}(\theta|_{obs})$ 对 θ 基于似然的推断是一样的。

当缺失为单调模式或似然函数较简单时，可通过直接公式推导或因子化似然函数的方法求得极大似然值[34]。然而，在一些复杂情况下，尤其是当数据为任意缺失模式，似然函数没有明显形式的解，极大化 $\ell(\theta|Y_{obs})$ 变得非常困难甚至不可能，需要 EM 算法求解极大似然值。

EM 算法由 Dempster 等在 1977 年提出[35]。EM 的基本思想是将 $\ell=(\theta|Y)$ 中出现的缺失数据视为 (θ,Y_{obs}) 的函数，用条件期望替换缺失数据，然后估计参数，假定新的参数是正确的，再估计缺失值，再估计参数，如此迭代直至收敛。具体由 E 步(expectation)和 M 步(maximization)组成。

E 步：在给定已观察的数据和当前参数下，求缺失数据的条件期望，然后用这些条件期望替换缺失数据。令 $\theta^{(t-1)}$ 为第 $(t-1)$ 次迭代时 θ 的估计，第 t 次迭代有：

$$Q(\theta|\theta^{(t-1)})=E[\log f(Y_{obs},Y_{mis}|\theta)|Y_{obs},\theta^{(t-1)}]=\int l(\theta|Y_{obs},Y_{mis})f(Y_{mis}|Y_{obs},\theta=\theta^{(t-1)})\,dY_{mis} \quad (2)$$

M 步：在当缺失数据被替换后像没有缺失数据一样进行极大似然估计。在 M 步对 θ 极大化 $Q(\theta|\theta^{(t-1)})$，找到一个 $\theta^{(t)}$ 满足：

$$Q(\theta^{(t)}|\theta^{(t-1)})\geqslant Q(\theta|\theta^{(t-1)}) \quad \text{对一切 } \theta \quad (3)$$

上述两步反复进行至达到某个停止准则，如两次迭代参数的变化很小时即到达收敛。

在许多应用中，M 步没有一个简单的计算形式，统计学家们又提出了 ECM 算法、ECME 算法、PX-EM 算法等，以提高计算和收敛速度。Ibrahim[36]提出的一种采用加权方法的 EM 算法，可以用于很多参数回归模型，包括广义线性模型，非线性模型，随机效应模型，参数和非参数生存模型等。

直接求似然，因子化似然法和 EM 算法都是计算极大似然估计的方法。从理论上来说，基于似然的方法比直接删除法或单一填补等方法更有吸引力。但是，基于似然的方法仍然有一定的应用条件[3,4,37]。首先，需要有足够大的样本保证得到似然估计值是无偏的。另外，似然函数是基于完整数据某个假定的参数模型，即 $P(Y_{bos},Y_{mis}|\theta)$。在实际应用中，如果模型假定错误，基于似然法的估计可能稳定也可能不稳定。目前有少数软件可以提供 EM 算法求解极大似然值，如 NORM[28]，S-Plus[38]，R 中的 norm，mix 等。

3. 基于加权调整的方法

加权调整是当出现缺失单元时，用某种方式把缺失单元的权数分解到非缺失单元(即观测数据)身上，通过增大样本中有观测数据的权数，以减小由于缺失数据可能对估计量带来的偏差。金勇进[39]介绍了几种简单的加权调整法，如 Politz-Simmons 调整法，加权组调整法，事后分层调整法等，但这些方法存在一定局

限，有时不但不能有效减小估计量的偏差，还可能增大估计量的方差。Robins，Rotnitzky 等人提出一种与极大似然有相似性质的加权估计方程（weighted estimating equations，WEE）处理 MAR 情况下缺失数据的方法[40-41]。该方法是广义估计方程（generalized estimation equations，GEE）的扩展，在理论上被认为估计效率更高，稳健性更好，尤其是在模型假定错误的情况下。

在回归模型中假设 y_i 为结局变量，x_i 为协变量，均值模型可写为 $u_i = u_i(x_i, \beta) = E(y_i \mid x_i, \beta)$。当不存在缺失值时，估计方程为 $u(\beta) = \sum_{i=1}^{n} u_i(\beta) = \sum_{i=1}^{n} d_i v_i^{-1}(y_i - u_i)$。令该方程等于0，可得到极大似然估计，该估计是 β 的无偏估计。

令 r_i 为缺失的指示变量，$r_i = 1$ 表示 x_i 全部观测到，$r_i = 0$ 表示 x_i 有部分值缺失。假定在给定 (y_i, x_i) 下，$r_i = 1$ 的概率为 π_i，有 $\pi_i = \pi_i(\phi) = Pr(r_i = 1 \mid m_i; \phi)$，$m_i$ 是 (y_i, x_i) 的某种函数，ϕ 是 r_i 的参数。当缺失机制为 MAR 时，π_i 仅依赖于 $x_{obs,i}$，即

$\pi_i = Pr(r_i = 1 \mid y_i, x_{obs,i})$。

当存在缺失值时，若仅用观测到的数据估计 $\hat{\beta}_{CC}$，估计是有偏的。假设 π_i 能够有效估计到，将 r_i 替换为 r_i/π_i，权重变为 r_i/π_i，加权估计方程为：

$$u_{WEE}(\beta) = \sum_{i=1}^{n} \frac{r_i}{\pi_i} u_i(\beta) = \sum_{i=1}^{n} \frac{r_i}{\pi_i} d_i v_i^{-1}(y_i - u_i) \tag{4}$$

在 MAR 假设下，因加权的作用，公式4可以得到 β 的无偏估计。

公式4中仅用到了 $r_i = 1$ 的观测值，估计效率较低。为了提高估计效率，Robins[42-43] 等建议加入未观测值的信息。如果 π_i 被正确估计，$E_{r_i \mid y_i, x_i}\left(\frac{r_i}{\pi_i}\right) = 1$ 成立。那么这个更有效的无偏估计方程为：

$$u_{WEE2}(\beta) = \sum_{i=1}^{n} \left[\frac{r_i}{\pi_i} d_i v_i^{-1}(y_i - u_i) + \left(1 - \frac{r_i}{\pi_i}\right) q(y_i, x_{obs,i}; \beta, a) \right] \tag{5}$$

与公式4相比，该方法增加了信息，提高了对 β 的估计效率。根据公式5，在协变量缺失情况下，该最优函数还需要对协变量的分布 $p(x_{mis,i} \mid x_{obs,i}, \alpha)$ 的设定，因此需要另一组类似公式5的估计方程来估计 α。令 $\gamma = (\beta, \alpha, \varphi)$，加权估计方程为：

$$S(\gamma) = \sum_{i=1}^{n} S_i(\gamma) = \sum_{i=1}^{n} \begin{bmatrix} S_{1i}(\beta, \alpha, \varphi) \\ S_{2i}(\beta, \alpha, \varphi) \\ S_{3i}(\varphi) \end{bmatrix}$$

$$= \sum_{i=1}^{n} \begin{bmatrix} \frac{r_i}{\pi_i} u_{1i}(\beta) + \left(1 - \frac{r_i}{\pi_i}\right) E_{x_{mis,i} \mid y_i, x_{obs,i}} u_{1i}(\beta; y_i, x_{obs,i}, x_{mis,i}) \\ \frac{r_i}{\pi_i} u_{2i}(\alpha) + \left(1 - \frac{r_i}{\pi_i}\right) E_{x_{mis,i} \mid y_i, x_{obs,i}} u_{2i}(\alpha; x_{obs,i}, x_{mis,i}) \\ m_i(r_i - \pi_i) \end{bmatrix} \tag{6}$$

其中 $u_{1i}(\beta) = u_{1i}(\beta; y_i, x_{obs,i}, x_{mis,i})$，$u_{2i}(\alpha) = u_{2i}(\alpha; x_{obs,i} x_{mis,i})$，$\varphi$ 是 r_i 的参数。令 $S(\hat{\gamma}_{WEE}) = 0$，则可得到 r 的无偏估计量。

Robins 等认为上述加权估计方程具有双重稳健性（Double Robustness）。公式6中含有三个模型：目标参数模型，缺失机制的模型和在给定已观测值下，缺失变量的条件分布模型。当缺失机制的模型假定错误，但给定观测值下缺失变量的条件均值模型是正确时，用公式6仍然可得到 β 的有效估计；当给定观测值下缺失变量的条件均值模型假定错误时，但缺失机制的模型是正确时，也能得到 $\hat{\beta}$ 的有效估计。也就是说当 π_i 或 $P(x_{mis,i} \mid x_{obs,i})$ 其中任一个被正确假定时，无论另一个是否正确，对 $\hat{\beta}$ 的估计均是渐进无偏的。

Parzen 等讨论了当实际是 logistic 模型的缺失协变量被错误地假设为多元正态模型时，用 WEE 方法仍然能得到良好稳健的估计。此外，WEE 方法还被用到处理不可忽略缺失数据[44]，含缺失应变量的重复测量数据[45]。Beunckens[46] 等提出用多重填补与 WEE 结合的方法处理含缺失值的纵向数据。由于 WEE 方法计算和程序编写非常复杂，目前还没有任何可以直接应用的软件。

不可忽略的缺失数据的处理

上述处理缺失数据的方法和研究大多是基于 MAR 的假定，但实际应用中，不可忽略的缺失（missing not at random，MNAR）也经常存在。例如在临床试验中，当试验对象脱落的原因与缺失的结局测量指标密切相关时，脱落者与非脱落者之间可能非常不同。在 MNAR 情况下，处理比较复杂，需要对模型有非常强的假定。目前主要处理 MNAR 数据常见的模型有选择模型（selection model）[14,47]，组型混合模型（pattern-mixture model）[47-48]，共享参数模型（shared parameters model）[47] 等。但有学者指出这些方法有的假定太强，有的模型很不稳定，假定分布的微小变动就会引起结论的很大改变[49]。有学者提出可以通过在研究设计阶段的方法改善和技巧，将 MNAR 情况转变为 MAR[4]。还有学者建议在处理 MNAR 数据时，由于结果依赖于不同的模型假定，不同假定下结果可能千差万别，还需进行敏感性分析[13]。

总　结

由于缺失数据的普遍存在以及绝大多数统计分析

方法主要针对完整数据，缺失数据的处理给研究者和应用者都带来了巨大的挑战，同时也带动了这一领域研究的发展。鉴于将缺失数据直接删除和单一填补方法的局限与不足，目前比较流行的有三类处理方法，分别是极大似然法，多重填补和加权估计方程。ML 在模型假定正确时可以提供比 MI、WEE 方法更高效率的估计，但要求在大样本情况才能获得有效地估计；MI 考虑了缺失数据的不确定性，可以提供完整的数据集，方便研究者采用不同的统计模型和方法对数据进行分析，但是计算处理比较耗时，并且大多数要求数据为 MAR 的缺失机制；WEE 的优势在于其稳健性，但是估计效率相对于 ML 和 MI 较低，而且计算和程序编写非常复杂，尚未有任何针对这一方法的软件出现。三类方法各自的优缺点提示研究者需要根据数据的特征和实际的情况来选择合适的方法。由于对 MI 的热衷和广泛应用，目前有很多 MI 的商业和免费软件可以使用。针对 EM 算法求解极大似然估计的软件不太多，而针对 WEE 方法的软件则几乎没有。后两种方法在应用软件上的开发和研究是今后发展的方向。

参 考 文 献

1. Rubin D. Multiple imputations for nonresponse in surveys. 1987, New York: John Wiley & Sons, Inc
2. Abraham W, Russell D. Missing data: a review of current methods and applications in epidemiological research. Current Opinion in Psychiatry, 2004, 17(4): 315.
3. Graham J. Missing data analysis: making it work in the real world. Annu Rev Psychol, 2009, 60: 549-576.
4. Schafer J, Graham J. Missing data: our view of the state of the art. Psychol Methods, 2002, 7(2): 147-177.
5. Croy C, Novins D. Imputing missing data. J Am Acad Child Adolesc Psychiatry, 2004, 43(4): 380.
6. Wood A, White I, Thompson S. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clin Trials, 2004, 1(4): 368-376.
7. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 1977, 39(1): 1-38.
8. Schafer J, Olsen M. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. Multivariate Behavioral Research, 1998, 33(4): 545-571.
9. Robins J, Rotnitzky A, Zhao L. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, 1994, 89(89): 864-866.
10. Qin J, Leung D, Shao J. Estimation with survey data under nonignorable nonresponse or informative sampling. Journal of the American Statistical Association, 2002, 97(457): 193-200.
11. Tang G, Little R, Raghunathan T. Analysis of multivariate missing data with nonignorable nonresponse. Biometrika, 2003, 94(4): 747-764.
12. Allison P. Missing data. 2001, Thousand Oaks, Calif.: Sage Publications.
13. Molenberghs G, Kernward M. Missing data in clinical studies, 2007, Chichester: Wiley.
14. Bello A. Imputation techniques in regression analysis: Looking closely at their implementation. Computational statistics & data analysis, 1995, 20(1): 45-570.
15. 金勇进. 缺失数据的插补调整. 数理统计与管理, 2001, 20(6): 47-53.
16. 金勇进. 调查中的数据缺失及处理(I)—缺失数据及其影响. 数理统计与管理, 2001, 20(1): 59-62.
17. Tanner M, Wong W. The Calculation of posterior distributions by data augmentation. Journal of the American Statistical Association, 1987, 82(398): 528-540.
18. Schenker N, Taylor J. Partially parametric techniques for multiple imputations. Computational Statistics and Data Analysis, 1996, 22(4): 425-448.
19. Belin T, Hu M, Young A, et al. Performance of a general location model with an ignorable missing-data assumption in a multivariate mental health services study. Statistics in Medicine, 1999, 18(22): 3123-3135.
20. Buuren V, Boshuizen H, Knook D. Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine, 1999, 18(6): 681-694.
21. Buuren V, Brand J. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation, 2006, 76(12): 1049-1064.
22. Buuren V. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research, 2007, 16(3): 219-242.
23. Brand J, Buuren S, Gelsema E, et al. A toolkit in SAS for the evaluation of multiple imputation methods. 2003, 57(1): 36-45.
24. Raghunathan T, Lepkowski J, Hoewyk J, et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology, 2001, 27: 85-96.
25. Horton N, Lipsitz S. Multiple imputation in practice: comparison of software packages for regression models with missing variables. American Statistician, 2001, 55: 244-254.
26. Multiple Imputation for Missing Data. http://support.sas.com/rnd/app/da/new/dami.html.
27. Schafer J. Analysis of incomplete multivariate data. Monographs on statistics and applied probability. London; New York: Chapman & Hall, 1997.
28. norm: Analysis of multivariate normal datasets with missing values. http://cran.r-project.org/web/packages/norm/index.html.
29. cat: Analysis of categorical-variable datasets with missing values. http://cran.r-project.org/web/packages/cat/index.html.
30. mix: Estimation multiple Imputation for Mixed Categorical and Continuous Data. http://cran.r-project.org/web/packages/mix/index.html.
31. pan: Multiple imputation for multivariate panel or clustered data. http://cran.r-project.org/web/packages/pan/index.html.
32. mi: Missing Data Imputation and Model Checking. http://cran.r-project.org/web/packages/mi/index.html.
33. mice: Multivariate Imputation by Chained Equations. http://cran.r-project.org/web/packages/mice/index.html.
34. Little R, Rubin D. Statistical analysis with missing data. 1987, New York: Wiley.
35. Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. 2000, New York: Springer.

(下转第 142 页)

独立的Ⅱ期试验更为合适[19]。

总结与展望

适应性Ⅱ/Ⅲ期无缝设计可以有效缩短药品开发时间，所面临的统计问题已有部分解决方案，但实施之前应特别注意妥善计划，并对其带来的收益与风险进行权衡。总结近二十年来有关适应性Ⅱ/Ⅲ期无缝设计的探讨与实际应用，可以看到这一新型临床试验设计的作用与地位在不断增强。除文中所述内容外，在期中分析时对最终检验的优效性与非劣效性的取舍，以及融合多臂多阶段设计等相关问题上亦有一些研究成果[20-21]。对这些问题的继续研究与完善将进一步开阔适应性无缝设计的应用前景，从而促进有效新药物或新疗法尽早地服务于人类。

参 考 文 献

1. Anonymous(2004). Innovation/Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products. FDA report from March 2004.
2. Lurdes YTI, Peter FT, Donald AB. Seamlessly expanding a randomized phase II trial to phase III. Biometrics, 2002, 58: 823-831.
3. Nigel S, Susan T. Sequential designs for phase III clinical trials incorporating treatment selection. Statistics in Medicine, 2003, 22: 689-703.
4. Peter B, Meinhard k. Combining different phases in the development of medical treatments within a single trial. Statistics in Medicine, 1999, 18: 1833-1848.
5. WHO 及 ICH 相关文件中文译文. 见: 郑筱萸 编.《药品临床试验管理规范》培训教材. 北京: 中国医药科技出版社, 2000, 161.
6. Frank B, Heinz S, Franz K, Amy R, Willi M. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. Biometrical Journal, 2006, 48(4): 623-634.
7. Heinz S, Frank B, Amy R, Willi M. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. Biometrical Journal, 2006, 48(4): 635-543.
8. Bauer PKhne K. Evaluation of experiments with adaptive interim analyses, 1994, 50.
9. 颜虹，夏结来，于莉莉. 临床试验中适应性设计研究进展. 中华预防医学杂志, 2008, 42(z1): 16-25.
10. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. Biometrics, 1995, 51: 1315-1324.
11. Cui L, HMJ H, Wang S. Modification of sample size in group sequential clinical trials. Biometrics, 1999, 55: 321-324.
12. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. Biometrics, 1999, 55: 1286-1290.
13. Bauser P, Kieser M. Combining different phases in the development of medical treatments within a single trial. Statistics in Medicine, 1999, 18: 1833-1848.
14. Shaffer JP. Multiple hypothesis testing. Annual Review of Psychology, 1995, 46: 561-584.
15. Benjamini Y, Hocheberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B(Methodological), 1995, 57(1): 289-300.
16. Hommel G. Tests of individual hypotheses for experiments with interim analyses and adaptive choice of hypotheses. Paper given at the Biometric Colloquium of the German Region of the International Biometric Society, Munich, 1997.
17. Hommer G. Adaptive modifications of hypotheses after an interim analysis. Biometrical Journal, 2001, 43: 581-589.
18. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. Biometrics, 1998, 54: 1014-1029.
19. Bauer P, Einfalt J. Application of adaptive designs-a review. Biometrical Journal, 2006, 48: 1-14.
20. Ohrn F, Jennison C. Optimal group-sequential designs for simultaneous testing of superiority and non-inferiority. Statistics in Medicine, 2010, 29: 743-759.
21. Parmar M, Barthel F, Sydes M, et al. Speeding up the evaluation of new agents in cancer, 2008, 100: 1204-1214.

(责任编辑: 丁海龙)

(上接第 139 页)

36. Ibrahim J. Incomplete Data in Generalized Linear Models. Journal of the American Statistical Association, 1990, 85(411): 765-769.
37. Ibrahim J, Chen M, Lipsitz S, et al. Missing-data methods for generalized linear models: a comparative review. Journal of the American Statistical Association, 2005, 100(469): 332-346.
38. Insightful, S-PLUS(Version 6)〔Computer software〕, 2001: Seattle, WA.
39. 金勇进，缺失数据的加权调整. 数理统计与管理, 2001, 20(5): 61-64.
40. Lipsitz S, Ibrahim J, Zhao L. A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. Journal of the American Statistical Association, 1999, 94(448): 1147-1160.
41. Zhao L, Lipsitz S, Lew D. Regression analysis with missing covariate data using estimating equations. Biometrics, 1996, 52(4): 1165-1182.
42. Robins J, Rotnitzky A, Zhao L. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, 1994, 89(427): 846-866.
43. Robins J, Rotnitzky A, Zhao L. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association, 1995, 90(429): 106-121.
44. Troxel A, Lipsitz S, Brennan T. Weighted estimating equations with nonignorably missing response data. Biometrics, 1997, 53(3): 857-869.
45. Preisser J, Lohman K, Rathouz P. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. Statistics in medicine, 2002, 21(20): 3035-3054.
46. Beunckens C, Sotto C, Molenberghs G. A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. Computational Statistics and Data Analysis, 2008, 52(3): 1533-1548.
47. Daniels M, Hogan J. Missing data in longitudinal studies strategies for Bayesian modeling and sensitivity analysis. 2008, Boca Raton: Chapman & Hall/CRC.
48. Little R, Wang Y. Pattern-mixture models for multivariate incomplete data with covariates. Biometrics, 1996, 52(1): 98-111.
49. Kenward M. Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. Statistics in medicine, 1998, 17(23): 2723-2732.

(责任编辑: 刘　壮)