**Statistics and Application** 统计学与应用, 2016, 5(4), 359-364
Published Online December 2016 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2016.54038

Hans汉斯

# Application of Gibbs Sampling in Data Missing and Execution in R

**Xia Ding**

School of Economics & Management, Shanghai Maritime University, Shanghai
Email: 1632198623@qq.com

Received: Nov. 29th, 2016; accepted: Dec. 12th, 2016; published: Dec. 22nd, 2016



## Abstract

**Data missing is a common problem in statistical research. Based on summarizing the common solutions, this paper proposes to solve the problem by Gibbs sampling, and achieve this process through the R language, so as to provide a new method. The experimental results show that Gibbs sampling is an ideal method to deal with missing data.**

## Keywords

**Gibbs Sampling, Data Missing, R Language**

# 吉布斯抽样在数据缺失中的应用及其R实现

丁　霞

上海海事大学经济管理学院，上海
Email: 1632198623@qq.com

收稿日期：2016年11月29日；录用日期：2016年12月12日；发布日期：2016年12月22日

## 摘　要

**数据缺失是统计研究中经常遇到的问题，文章在总结常见缺失数据的处理方法的基础上，提出了用Gibbs**

文章引用: 丁霞. 吉布斯抽样在数据缺失中的应用及其 R 实现[J]. 统计学与应用, 2016, 5(4): 359-364.
http://dx.doi.org/10.12677/sa.2016.54038

**抽样方法来解决数据缺失问题，并通过R语言来实现这一过程，从而为数据缺失提供一种新的解决思路。实验结果表明，Gibbs抽样是一种效果比较理想的处理缺失数据的方法。**

## 关键词

**Gibbs抽样，数据缺失，R语言**

## 1. 引言

在进行统计学研究过程中，一般都要求数据是完整的、及时的、可靠和准确的。但在实践中，由于客观条件的限制，统计数据的及时、可靠准确问题往往容易解决，但数据完整性问题很难解决，统计数据往往是有缺失的。如果不对统计数据缺失问题进行处理，容易造成分析结果的偏差，甚至错误，从而降低统计结果的准确性和效率。所以进行统计分析最重要的一步就是对不完整的统计数据进行预处理，对缺失数据进行补充，形成一个完整的统计数据集，然后再对补充完整后的统计数据集进行统计分析，这样的分析结果才能够准确。

## 2. 传统缺失数据处理方法

最原始的处理缺失数据的方法是直接删除法，它把存在缺失值的样本删掉，那么数据就变成完整数据了。如果某个变量有许多缺失值，那么也可以直接将该变量删去不要。如果缺失数据是非完全随机缺失，可以对完整的数据进行加权，从而减小偏差。删除法是对缺失数据处理的最简单最直接的方法，但是这会损失样本的信息，如果缺失比例比较大，就不能使用删除法。

除了删除法，常见的处理缺失数据的方法有两种[1]，第一种是直接取最可能的值去代替缺失值，这一方法主要有：

1) 人工填写法。

让最了解数据的人去为缺失数据做补充，效果比较理想。但是当缺失值较多时，该方法会耗费大量人力物力，实操性并不强，所以在实际生活中一般不会选择这个方法。

2) 均值填充法。

如果缺失值是数值，我们采用与该缺失值相同属性的其他值的平均值来填充，如果缺失值是非数值，我们采用与该缺失值相同属性的其他值的众数来填充。

3) 特殊值填充法。

有的时候，如果某类型数据值缺失，则代表了一定的意义，比如个人信用评估中电话号码或者家庭住址的缺失，意味着此人信用存在问题的概率加大。我们可以把缺失值用一个特殊的值如“unknown”来进行填充。

第二种方法是借助于模型来完成对缺失数据的填补。该方法主要有以下几种：

1) 回归填充法

上诉缺失值填充方法并没有考虑与缺失值相关的变量的信息，利用回归填充法，我们可以做一个回归，将缺失变量作为因变量，与它相关的其他变量称为自变量，从而拟合一个线性回归模型，用相关变量的值来估计缺失变量的值。

2) 极大似然估计方法

当缺失数据为随机缺失时，我们可以通过已知信息来求得边际分布，从而对参数进行极大似然估计，

比较常用的是 EM 算法。这种方法适用于样本数量比较大的情况，从而保证估计结果渐近无偏且服从正态分布。

均值填充法和特殊值填充法没有很好利用总体的信息，所以数据填充之后，可能并不能真实反映总体的信息，回归填充法虽然考虑了总体的信息，但是如果缺失值没有相关的变量或者相关的变量值也有缺失，那么该方法也就失效了。EM 算法因为适用于大样本情况，而且计算比较复杂，需准确计算数值积分(E 步)和进行导数运算(M 步)，所以当样本量并不大时，该方法也不太理想。而 Gibbs 抽样可以解决以上问题，而且实施过程简单易行，对真实值的拟合效果也比较理想[2] [3]。

## 3. 吉布斯方法概述

1984 年，D. Geman 和 S. Geman 提出了 Gibbs 算法[4]，发展至今，Gibbs 抽样已经成为了一种应用广泛而且方法简单的 MCMC 抽样方法，Gibbs 抽样的基本思想是通过满条件分布来构造马氏链转移核从而进行抽样。例如，如果要从二维联合分布 $f(x,y)$ 中进行随机抽样，但是 $f(x,y)$ 的表达式并不清楚，或者表达式过于复杂，不方便直接抽样，与此同时，条件分布 $f(x|y)$ 与 $f(y|x)$ 已知而且表达式相对比较简单，我们则可以依照以下迭代步骤进行抽样，从而获得二维联合分布 $f(x,y)$ 的随机样本[5]。

Gibbs 抽样的基本步骤：

1) 在满足条件分布 $f(x|y)$ 的取值要求下选择 $x_0$

2) 从条件分布 $f(y|x_0)$ 中随机选取一个值 $y_0$ (此时产生了第一对随机点列 $(x_0,y_0)$)

3) 从条件分布 $f(x|y_0)$ 中随机选取一个值 $x_1$

4) 从条件分布 $f(y|x_1)$ 中随机选取一个值 $y_1$ (此时产生了第二对随机点列 $(x_1,y_1)$)

…

重复以上步骤 N 次之后，去掉最开始的一部分抽样值，从而避免初始值的影响，即可得到二维联合分布 $f(x,y)$ 的随机抽样样本。

## 4. 模拟研究以及效果评价

### 4.1. 数值实验及其 R 实现

本文利用一组完整数据进行缺失数据的填补实验。首先按照 20%的缺失比例，随机删除部分数据，然后利用 Gibbs 抽样，对缺失数据进行填充，将填补后的数据与完整数据做对比，从而衡量 Gibbs 抽样处理缺失数据的效果。

以上海某高校大学男生(1000 人)的身高体重数据为例，已知成年男子的身高体重服从二维正态分布，记身高为 $X$，体重为 $Y$，则 $(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，其中 $\mu_1,\mu_2$ 分别指身高和体重的均值，$\sigma_1^2,\sigma_2^2$ 分别指身高和体重的方差，$\rho$ 是身高和体重的相关系数。

当随机变量 $(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，对应的条件分布为：

$$X|Y=y\sim N\left(\mu_1+\rho\frac{\sigma_1}{\sigma_2}(y-\mu_2),\sigma_1^2(1-\rho^2)\right)$$

$$Y|X=x\sim N\left(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1),\sigma_2^2(1-\rho^2)\right)$$

随机删除 200 名男生的数据，使数据出现全数据缺失。对余下的数据进行计算，得到满条件分布，选定不完整数据的身高和体重的均值 $(172.57, 62.73)$ 为初始值，便可用 R 语言开始迭代过程，产生马氏链。程序如下[6]：

```
TT<-10000
chain<-matrix(numeric(),nr=2,nc=TT)
chain[,1]<-c(172.57,62.73)
for(i in 2:TT){
chain[1,i]<-rnorm(1,172.57+0.6873*(chain[2,i-1]-62.73),sqrt(23.9*(1-0.4862)))
chain[2,i]<-rnorm(1,62.73+0.7074*(chain[1,i]-172.57),sqrt(24.6*(1-0.4862)))
}
```

为了避免初始值的影响，从而产生更优质的随机数，在迭代 10000 次后，剔除前面的 9800 对随机数，保留余下的 200 对。

## 4.2. 效果评价

表 1 给出了抽样填补后的数据和原始数据的对比，我们可以看到，通过 Gibbs 抽样模拟之后获取的身高和体重的均值分别为 173.31 和 63.70，方差分别为 23.17 和 23.26，$X$ 和 $Y$ 的相关系数为 0.6467。填充后数据的统计值和真实完整数据的统计值相差很小。

再看填充后数据的二维散点图(图 1)，该图显示模拟产生的随机数具有正态分布的特点，而且填充后数据的中心与原数据的中心(红色虚线)基本吻合。

身高和体重分别服从正态分布，作出填充后数据的身高和体重的直方图，并根据完整数据的分布作出概率密度函数曲线。从图 2 和图 3 我们可以看出，直方图的形态与概率密度函数曲线的形态较为相似，这说明 Gibbs 抽样填补缺失数据的模拟效果比较理想。

接下来再对原始数据和填补后的数据进行单因素方差分析，原假设 $H_0$ 为两种数据之间的差异不显著 ($\alpha = 0.05$)，分析结果如表 2 和表 3 所示。

**Table 1.** Comparison between the original data and the populated data
**表 1.** 原数据与填充后数据对比

| | X 均值 | X 方差 | Y 均值 | Y 方差 | X 和 Y 的相关系数 |
|---|---|---|---|---|---|
| 原数据 | 173.11 | 23.58 | 63.19 | 25.33 | 0.7042 |
| Gibbs 抽样数据 | 173.31 | 23.17 | 63.70 | 23.26 | 0.6467 |

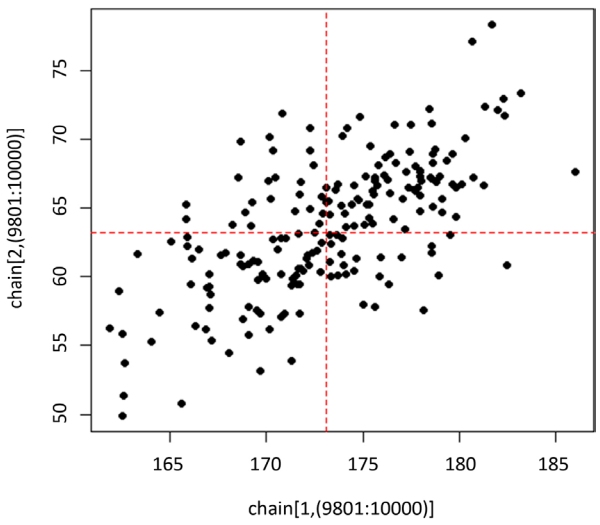


**Figure 1.** The planar scatter plot of populated data
**图 1.** 填充后数据的二维散点图

从表 2 和表 3 中可以看到，两组单因素方差分析的 P 值均大于 0.05，所以不拒绝原假设，两组数据之间无显著差异。这说明通过 Gibbs 抽样进行填充的数据拟合效果较好。

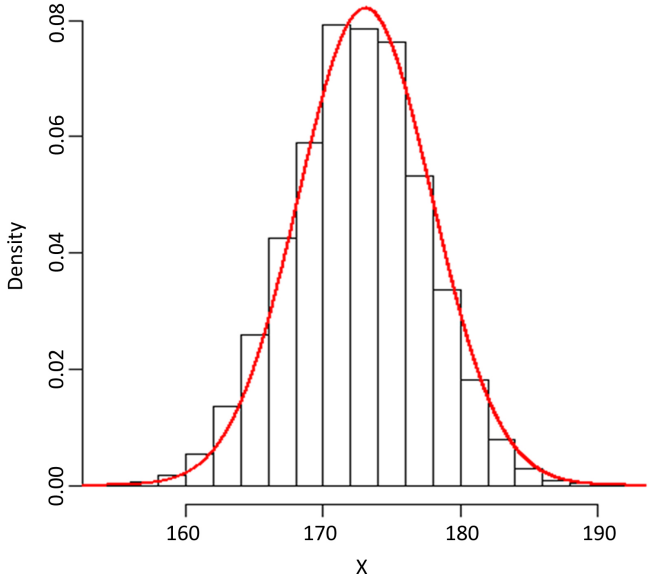


**Figure 2.** The histogram of populated height data
**图 2.** 填充后的身高数据直方图

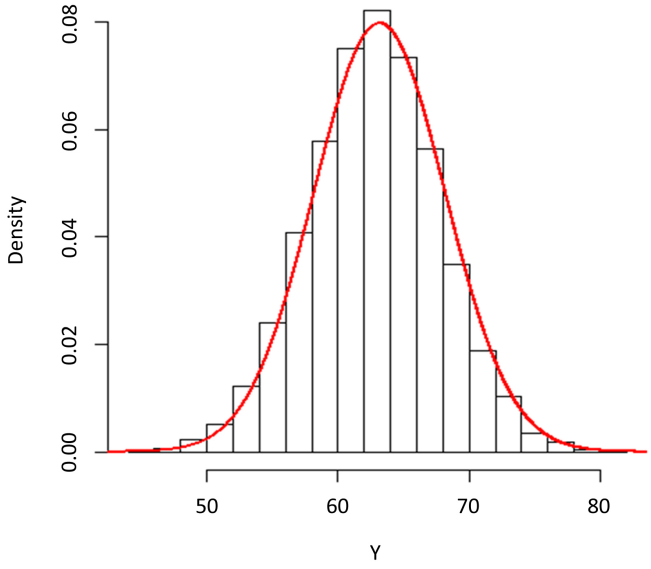


**Figure 3.** The histogram of populated weight data
**图 3.** 填充后的体重数据直方图

**Table 2.** ANOVA (height)
**表 2.** 单因素方差分析(身高)

| 差异源 | SS | df | MS | F | P-value | F crit |
|---|---|---|---|---|---|---|
| 组间 | 51.51623 | 1 | 51.51623 | 2.256341 | 0.133226 | 3.846117 |
| 组内 | 45617.85 | 1998 | 22.83176 | | | |
| 总计 | 45669.37 | 1999 | | | | |

**Table 3.** ANOVA (weight)
**表 3.** 单因素方差分析(体重)

| 差异源 | SS | df | MS | F | P-value | F crit |
|---|---|---|---|---|---|---|
| 组间 | 78.31466 | 1 | 78.31466 | 3.217855 | 0.07299 | 3.846117 |
| 组内 | 48626.39 | 1998 | 24.33753 | | | |
| 总计 | 48704.71 | 1999 | | | | |

## 5. 结论

本文总结了一些常用的处理数据缺失的方法，并提出了将 Gibbs 抽样应用到缺失数据的填补中。在实际操作中，只要通过已知数据得到满条件分布，就可以开始迭代产生随机数，这一迭代过程可以轻松地通过 R 语言实现，而且 Gibbs 抽样的结果与已知数据的统计特征吻合程度比较高，可以比较理想地完成缺失数据的填补过程。综上，Gibbs 抽样具有简单易操作，可充分利用总体的信息，不受限于小样本，且填补结果拟合程度高的优点。

但是需要指出的是，并没有哪种处理方法是普遍适用的，在实际情况中还是要根据具体问题，充分考虑每种方法的优点和缺点以及适用情况之后再选择最合适的方法进行数据填补。还有一点值得注意的是，Gibbs 抽样并非真正的随机抽样，它每一步的抽样都是上一步的函数，只是上一步的抽样是随机的，所以利用 Gibbs 抽样得到的数据实际上是 Markov 链。

## 参考文献 (References)


[1] 曾莉, 辛涛, 张淑梅. 2PL 模型的两种马尔可夫蒙特卡洛缺失数据处理方法比较[J]. 心理学报, 2009, 41(3): 276-282.

[2] 张香云. Gibbs 抽样在不同缺失率下的参数估计[J]. 统计与决策, 2008(4): 23-24.

[3] 陈晓林. 基于 Gibbs 抽样和 EM 算法的生物保守序列 motif 识别[D]: [硕士学位论文]. 苏州: 苏州大学, 2007.

[4] Geman, S. and Geman, D. (1984) Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6**, 721-724. https://doi.org/10.1109/TPAMI.1984.4767596

[5] Hastings, W.K. (1970) Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**, 97-109. https://doi.org/10.1093/biomet/57.1.97

[6] 候雅文, 王斌会. 统计实验及 R 语言模拟[M]. 北京: 北京大学出版社, 2015.