一、缺失值产生的原因

缺失值的产生的原因多种多样，主要分为机械原因和人为原因。机械原因是由于机械原因导致的数据收集或保存的失败造成的数据缺失，比如数据存储的失败，存储器损坏，机械故障导致某段时间数据未能收集（对于定时数据采集而言）。人为原因是由于人的主观失误、历史局限或有意隐瞒造成的数据缺失，比如，在市场调查中被访人拒绝透露相关问题的答案，或者回答的问题是无效的，数据录入人员失误漏录了数据。

二、缺失值的类型

缺失值从缺失的分布来讲可以分为完全随机缺失，随机缺失和完全非随机缺失。完全随机缺失（missing completely at random,MCAR）指的是数据的缺失是随机的，数据的缺失不依赖于任何不完全变量或完全变量。随机缺失(missing at random,MAR)指的是数据的缺失不是完全随机的 ,即该类数据的缺失依赖于其他完全变量。完全非随机缺失(missing not at random,MNAR)指的是数据的缺失依赖于不完全变量自身。

从缺失值的所属属性上讲，如果所有的缺失值都是同一属性，那么这种缺失成为单值缺失，如果缺失值属于不同的属性，称为任意缺失。另外对于时间序列类的数据，可能存在随着时间的缺失，这种缺失称为单调缺失。

三、缺失值的处理方法

对于缺失值的处理，从总体上来说分为删除存在缺失值的个案和缺失值插补。对于主观数据，人将影响数据的真实性，存在缺失值的样本的其他属性的真实值不能保证，那么依赖于这些属性值的插补也是不可靠的，所以对于主观数据一般不推荐插补的方法。插补主要是针对客观数据，它的可靠性有保证。

1.删除含有缺失值的个案

主要有简单删除法和权重法。简单删除法是对缺失值进行处理的最原始方法。它将存在缺失值的个案删除。如果数据缺失问题可以通过简单的删除小部分样本来达到目标，那么这个方法是最有效的。当缺失值的类型为非完全随机缺失的时候，可以通过对完整的数据加权来减小偏差。把数据不完全的个案标记后，将完整的数据个案赋予不同的权重，个案的权重可以通过 logistic 或 probit 回归求得。如果解释变量中存在对权重估计起决定行因素的变量，那么这种方法可以有效减小偏差。如果解释变量和权重并不相关，它并不能减小偏差。对于存在多个属性缺失的情况，就需要对不同属性的缺失组合赋不同的权重，这将大大增加计算的难度，降低预测的准确性，这时权重法并不理想。

2.可能值插补缺失值

它的思想来源是以最可能的值来插补缺失值比全部删除不完全样本所产生的信息丢失要少。在数据挖掘中，面对的通常是大型的**数据库**，它的属性有几十个甚至几百个，因为一个属性值的缺失而放弃大量的其他属性值，这种删除是对信息的极大浪费，所以产生了以可能值对缺失值进行插补的思想与方法。常用的有如下几种方法。

(1)均值插补。数据的属性分为定距型和非定距型。如果缺失值是定距型的，就以该属性存在值的平均值来插补缺失的值；如果缺失值是非定距型的，就根据统计学中的众数原理，用该属性的众数(即出现频率最高的值)来补齐缺失的值。

(2)利用同类均值插补。同均值插补的方法都属于单值插补，不同的是，它用层次聚类模型预测缺失变量的类型，再以该类型的均值插补。假设 X=(X1,X2…Xp)为信息完全的变量，Y 为存在缺失值的变量，那么首先对 X 或其子集行聚类，然后按缺失个案所属类来插补不同类的均值。如果在以后统计分析中还需以引入的解释变量和 Y 做分析，那么这种插补方法将在模型中引入自相关，给分析造成障碍。

(3)极大似然估计（Max Likelihood ,ML）。在缺失类型为随机缺失的条件下，假设模型对于完整的样本是正确的，那么通过观测数据的边际分布可以对未知参数进行极大似然估计（Little and Rubin）。这种方法也被称为忽略缺失值的极大似然估计，对于极大似然的参数估计实际中常采用的计算方法是期望值最大化(Expectation Maximization，EM）。该方法比删除个案和单值插补更有吸引力，它一个重要前提：适用于大样本。有效样本的数量足够以保证 ML 估计值是渐近无偏的并服从正态分布。但是这种方法可能会陷入局部极值，收敛速度也不是很快，并且计算很复杂。

(4)多重插补（Multiple Imputation，MI）。多值插补的思想来源于贝叶斯估计，认为待插补的值是随机的，它的值来自于已观测到的值。具体实践上通常是估计出待插补的值，然后再加上不同的噪声，形成多组可选插补值。根据某种选择依据，选取最合适的插补值。

多重插补方法分为三个步骤：①为每个空值产生一套可能的插补值，这些值反映了无响应模型的不确定性；每个值都可以被用来插补数据集中的缺失值，产生若干个完整数据集合。②每个插补数据集合都用针对完整数据集的统计方法进行统计分析。③对来自各个插补数据集的结果，根据评分函数进行选择，产生最终的插补值。

假设一组数据，包括三个变量 Y1，Y2，Y3，它们的联合分布为正态分布，将这组数据处理成三组，A 组保持原始数据，B 组仅缺失 Y3，C 组缺失 Y1 和 Y2。在多值插补时，对 A 组将不进行任何处理，对 B 组产生 Y3 的一组估计值（作 Y3 关于 Y1，Y2 的回归），对 C 组作产生 Y1 和 Y2 的一组成对估计值（作 Y1，Y2 关于 Y3 的回归）。

当用多值插补时，对 A 组将不进行处理，对 B、C 组将完整的样本随机抽取形成为 m 组（m 为可选择的 m 组插补值），每组个案数只要能够有效估计参数就可以了。对存在缺失值的属性的分布作出估计，然后基于这 m 组观测值，对于这 m 组样本分别产生关于参数的 m 组估计值，给出相应的预测即，这时采用的估计方法为极大似然法，在计算机中具体的实现算法为期望最大化法（EM）。对 B 组估计出一组 Y3 的值，对 C 将利用 Y1,Y2,Y3 它们的联合分布为正态分布这一前提，估计出一组(Y1，Y2）。

上例中假定了 Y1,Y2,Y3 的联合分布为正态分布。这个假设是人为的，但是已经通过验证（Graham 和 Schafer 于 1999），非正态联合分布的变量，在这个假定下仍然可以估计到很接近真实值的结果。

多重插补和贝叶斯估计的思想是一致的，但是多重插补弥补了贝叶斯估计的几个不足。

(1)贝叶斯估计以极大似然的方法估计，极大似然的方法要求模型的形式必须准确，如果参数形式不正确，将得到错误得结论，即先验分布将影响后验分布的准确性。而多重插补所依据的是大样本渐近完整的数据的理论，在数据挖掘中的数据量都很大，先验分布将极小的影响结果，所以先验分布的对结果的影响不大。

(2)贝叶斯估计仅要求知道未知参数的先验分布，没有利用与参数的关系。而多重插补对参数的联合分布作出了估计，利用了参数间的相互关系。

以上四种插补方法，对于缺失值的类型为随机缺失的插补有很好的效果。两种均值插补方法是最容易实现的，也是以前人们经常使用的，但是它对样本存在极大的干扰，尤其是当插补后的值作为解释变量进行回归时，参数的估计值与真实值的偏差很大。相比较而言，极大似然估计和多重插补是两种比较好的插补方法，与多重插补对比，极大似然缺少不确定成分，所以越来越多的人倾向于使用多值插补方法。

四、小结

插补处理只是将未知值补以我们的主观估计值，不一定完全符合客观事实。以上的分析都是理论分析，对于缺失值由于它本身无法观测，也就不可能知道它的缺失所属类型，也就无从估计一个插补方法的插补效果。另外这些方法通用于各个领域，具有了普遍性，那么针对一个领域的专业的插补效果就不会很理想，正是因为这个原因，很多专业数据挖掘人员通过他们对行业的理解，手动对缺失值进行插补的效果反而可能比这些方法更好。缺失值的插补是在数据挖掘过程中为了不放弃大量的信息，而采用的人为干涉缺失值的情况，无论是那种处理方法都会影响变量间的相互关系，在对不完备信息进行补齐处理的同时，我们或多或少地改变了原始的数据的信息系统，对以后的分析存在潜在的影响，所以对缺失值的处理一定要慎重。