一般常用的missing data的处理方法，如：完整的案例分析、取中值等，可能导致bias

In [statistics](https://en.wikipedia.org/wiki/Statistics), **imputation** is the process of replacing [missing data](https://en.wikipedia.org/wiki/Missing_data) with substituted values

---- 用替代值替代丢失数据的过程

# MissingData的类型

* MCAR：missing completely at random，即丢失数据是完全随机的
* MNAR：missing not at random，丢失数据依赖于某个未被观察到的信息
* MAR ：missing at random ，丢失数据依赖于某个已经被观察到的信息，e.g.当满足某条件的时候，某些属性没有意义。
* 直接替换：从相同的数据集合中，随机抽样，用其对应的属性值替代丢失数据

注意：抽样相同的数据集合时，要考虑避免bias。例如，如果只有 “没有生病男病人“才可能丢失数据，则补充没有生病的男病人的缺失值，被随机抽样的数据应该是所有没有生病的男病人。

* Imputation using the overall sample mean：直接采用所有原始输入记录的平均值

--- 增加了相关性，额外引入了bias。**不建议使用**

* missing-indicator method：对每一个存在缺失值的属性都增加一个指示字段，指示原始属性是否为缺失，并为原始属性增加一个新值代表“缺失”，以后处理中指示字段和原始字段总是成对出现。

--- 这个方法有问题，因为增加的指示字段和原始字段强相关，应用线性回归等方法时可能造成bias，**不建议使用**

* single imputation procedure：imputation是基于丢失数据的记录的其他已知的属性，而不仅仅是整个数据集。即取所有关联属性类似的、但不缺失值的记录，计算后替代缺失值
  + 平均值
  + 计算缺失变量的分布，如：正态分布，然后在抽样 --- 要验证标准差和p-value
  + 线性回归
* Multiple imputation：多个独立的single imputation procedure组合