利用SPSS软件进行统计分析前需做哪些数据准备工作？

作者：高乐,龚志忠,张耀文 来源：医咖会

本文原地址：https://www.iikx.com/news/statistics/7821.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

利用SPSS软件进行统计分析前需做哪些数据准备工作?

提到数据分析，首先会想到的可能是t检验、回归分析等各种具体的分析方法，但对于经历过完整数据分析的同学来说，最复杂、最耗费时间的步骤往往是数据的清理，也就是将数据整理成为能够进行上述统计分析的格式。因此，本文将对数据的准备进行简要介绍，重点介绍数据审核，其次会用少量篇幅简单介绍数据在进行分析时的适用性。

在数据审核方面，主要考虑的是**数据的完整性和合理性**，也就是
对缺失数据和离群值进行识别和处理。

对缺失值的处理

在很多情况下，研究中所收集的数据会出现缺失情况，缺失的类型大致可以分为以下三种：

①完全随机缺失(Missing completely at random，MCAR)，数据缺失随机发生，与自身及其他变量均无关，任何变量的每一条记录发生缺失的概率相同。例如由于设备故障、样品运输丢失等导致的数据缺失，可视为MCAR[1]。这是最理想的情况，但在许多领域中这种情况并不合理;

②随机缺失(Missing at random，MAR)，是一种较为合理的情况。缺失值与自身变量无关，但与其他研究变量相关。假设老师的职称越高，提供其工资信息的可能性越低，那么每个职称分组中可认为老师工资信息缺失是随机发生的，可以通过加权的方法进行解决;

③非随机缺失(Missing not at random，MNAR)，即缺失值与自身变量有关。例如一项研究中对受教育程度情况进行了调查，受教育程度较低的个体可能存在该变量的缺失，这就是非随机缺失。

对缺失值最好的处理方法是预防缺失的发生，即
通过合理的研究
设计、预试验的开展、调查员培
训等方法尽量保证数据的完整性
。但当缺失值不可避免时，就需要通过一些统计学方法对其进行处理：

①缺失值删除

(a)删除缺失数据行，适用于MCAR数据的处理，在大样本量且缺失较少的情况下很有效。该方法不会影响结果估计的准确性，但样本量会因此减小，从而影响结果的精确性;

(b)删除缺失变量，适用于存在另一个无缺失的变量能够代替有缺失变量的情况，通常不建议采用这种方法，因为“保留数据总比删除数据好”;

(c)选择性删除，在研究不同组合变量的相关关系时，可选择该组合内所有可用的数据进行估计，但由于模型不同部分的样本量可能不同，会导致研究结果的解释存在一定困难。

②缺失值填补

(a)均值、中位数和众数填补：根据数据分布，选择使用样本均值、中位数或众数对缺失值进行填补，没有考虑时序特征及变量间关系。该方法较为简单，但有明显缺陷，例如降低了数据方差;

(b)多重填补：基于贝叶斯方法，创建多个填补数据集，即根据现有观测数据为每个缺失数据生成若干个可供填补的数值，结合填补后不同的结果，得出平均估计结果并考察缺失数据的不确定性[2];

(c)回归填补：包括线性回归和Logistic回归。首先识别缺失变量的预测变量，其次使用无缺失记录生成预测方程，对缺失值进行预测：

(d)虚拟变量设置：将是否缺失设置为虚拟变量，这是处理分类变量缺失较为简单的一种方法，但估计精度会下降;

(e)线性内插法[3]：若缺失值与未缺失值间存在线性关系，根据缺失值的前一个和后一个观测值对缺失值进行计算;

(f)临床试验中常用方法[4]：末次观察前推法(Last observation carried forward，LOCF)，前次观察值后推法(Next observation carried backward，NOCB)，基线值后推法(Baseline observation carried forward，BOCF)，最差观测值推进法(Worst observation carried forward，WOCF)和将缺失值视为治疗失败法(Missing value treated as failure，MVTF)等。

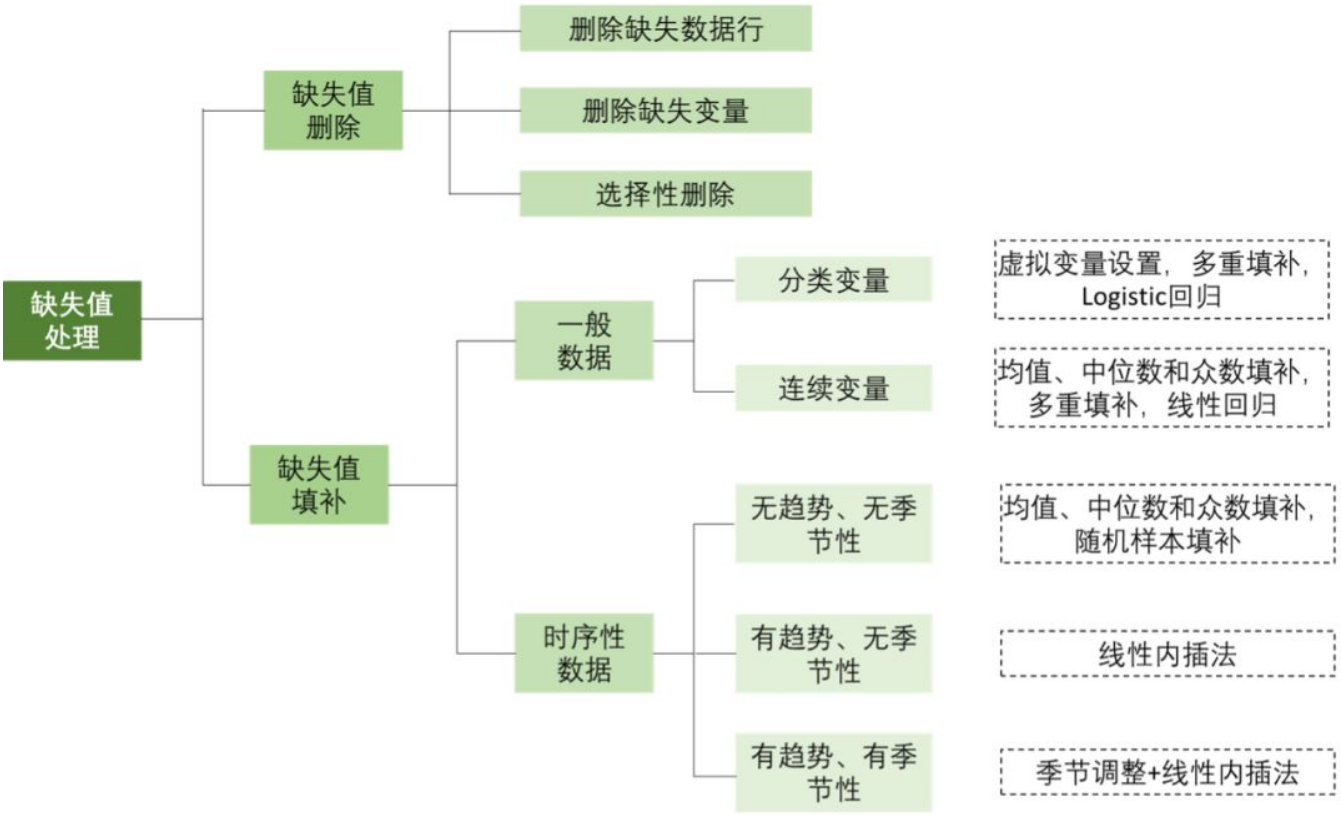


图1 缺失值处理方法的选择

(来源：https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4)

对于完全随机缺失，简单的删除缺失数据行就可得到无偏的估计结果;对于随机缺失，一些复杂的统计方法可能会得到无偏估计结果;而对于非随机缺失，无法得到无偏估计结果，只能通过复杂的统计方法减小估计值的偏倚。

对离散值的处理

在数据检查与整理时，
通常会发现一些完全偏离其他数据的数值，称
之为**离群值**
，产生的原因可能是数
据错误(测量或记录错误)或真正离群值，主要
分为**单变量离群值**和**多变量离群值**
。离群值的识别没有精确的、能够称之为金标准的方法，通常经过一些统计学方法进行识别。

对单变量离群值的识别，主要的方法为**直方图**(
落在直方图两端较远距离数据)和**箱式图**(
距离第25百分位数Q1或第75百分位数Q3的距离是四分位数间距IQR的1.5-3倍为轻度离群值，距离Q1或Q3的距离大于IQR的3倍为极端离群值)[5]。

对多变量离群值的识别，可通过马氏距离进行判断，首先计算一个点到某一数据分布之间的距离

，随后根据卡方分布确定临界值，若某个个体的马氏距离大于该临界值，则可认为是离群值[6]。

对于离群值的处理，有以下几种方法：

①数据检查
：检查是否为客观失误造成的数值异常，如果存在数据收集或录入错误，要及时更正;

②转换变量
：如果能够确定某一离群值是正确且真实的，为避免数据分布严重倾斜，可以对数据进行转换，这样不会改变原有数值间的相对大小，但会使数据分布更为集中;

③删除数据行
：适用于某一个案例出现了多个变量异常的情况，或含有异常值的个体所占比例很小，可以考虑删除整条信息;

④删除变量：若多个案例的某一变量均发现异常，可以根据实际情况考虑删除该变量;

⑤将离群值视为缺失数据处理，可以进行数据填补等操作;

⑥改变数值
：若想对离群值进行保留，可以对其数值进行调整使之更接近均值，也就是说可以设定一个百分位阈值，将超过该百分位范围的数据替换为该百分位数值。

数据适用性

不同的统计分析方法对数据有不同的要求，例如正态分布及方差齐性假设等，因此在分析前，需要对数据进行检查，来判断是否符合相应假设，这些假设见下表所示：

表1 数据假设的使用情境及检验方法

数据假设	使用情境	检验方法
正态性	t检验，方差分析，Person相关性，线性回归，其它参数检验 [7]	SPSS教程：判断数据正态分布的超多方法！
方差同质性	方差分析ANOVA（t 检验 and F 检验）[8]	Bartlett’s检验，Levene’s检验，Brown-Forsythe检验 [9]
方差齐性	相关和回归分析	Breusch-Pagan检验 [10]
方差/协方差矩阵的同质性	多元方差分析MANOVA，判别函数分析，多变量回归 [11]	Box’s M检验
球行检验	重复测量的ANOVA	Mauchly's检验
回归同质性	协方差分析ANCOVA	因变量与协变量散点图，自变量与协变量交互分析（广义线性模型）
线性相关	线性回归分析	散点图

此外，在进行回归分析时，还需要对变量之间的多重共线性进行判断。当模型中的一个变量可以被其他变量或其他变量的线性组合较好的预测时，就会出现该情况。

多重共线性
一般由容忍度和方差膨胀因子(VIF)进行
判断。一般认为
如果容忍度<0.2或VIF>5，则提示变量之间存在多重共线性
。一般解决方法有剔除模型中一个或多个预测变量，将模型中多个预测变量整合为一个复合变量，采用逐步回归的方法进行变量筛选，或进行主成分分析。

通常，收集到的数据都会存在各种各样的问题，为了得到“干净”且适合分析的数据，通常需要经过很多步骤，来达到最终的目的。因此我们要掌握并选择最适合自己数据的处理方法，来得到

更精确的分析结果。

[1]Kang, H. (2013). The prevention and handling of the missing data. Korean journal of anesthesiology, 64(5), 402.

[2] Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ, 338, b2393.

[3] Available athttps://www.lexjansen.com/nesug/nesug01/ps/ps8026.pdf.

[4] Available athttp://onbiostatistics.blogspot.com/2010/08/locf-bocf-wocf-and-mvtf.html.

[5] Available athttp://www.psychwiki.com/wiki/Detecting_Outliers_-_Univariate.

[6]Available athttps://en.wikiversity.org/wiki/Multivariate_outlier.

[7] Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. International journal of endocrinology and metabolism, 10(2), 486.

[8] Available athttps://methods.sagepub.com/reference/encyc-of-research-design/n179.xml.

[9] Available athttp://www.math.montana.edu/jobo/st541/sec2e.pdf.

[10] Available athttps://en.wikipedia.org/wiki/Homoscedasticity.

[11] Available athttp://www.introspective-mode.org/data-assumption-homogeneity-of-variance-covariance/.

更多 统计方法 请访问 https://www.iikx.com/news/statistics/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发