

# Table of Contents

统计学若干问 ..... 1



# 统计学若干问

- 异常值应该如何处理？
  1. 首先检查是否有数据录入错误；
  2. 不能因为异常值比期望过高或过低而移除；
  3. 使用敏感度分析，比较异常值对结果的影响，如影响较大：
    1. 转换数据（对数、平方根等等）；
    2. 非参数检验。

- 何时需要转换数据？如何转换数据？
  1. 转换数据的目的：
    - 使数据正态分布；
    - 使两个变量呈线性关系；
    - 使方差稳定。
  2. 常见的转换方法：
    - 对数
      - 单个变量，分布右偏，取对数可接近正态分布（常见于部分生化指标）；
      - 两个变量，指数关系，因变量取对数可接近线性关系；
      - 分组数据，数值大的组方差大，取对数可使方差相似。
    - 平方根
      - 和对数转换类似（常见于发生次数少的事件数）；
      - 方差/平均值为常数时可使方差相似。
    - 倒数
      - 用于生存分析，对存活时间取倒数；
      - 方差/平均值的四次方为常数时可使方差相似。
    - 平方
      - 单个变量，分布左偏，取平方可接近正态分布；
      - 两个变量，上凸，取平方可接近线性关系；
      - 分组数据，数值小的组方差大，取平方可使方差相似。
    - Logit转换
      - $\text{logit}(p) = \ln \frac{p}{1-p}$  为
      - 因变量p范围 $[0, 1]$  S型曲线，取Logit可使其线性化。

- 何时需要非参数检验？

参数检验	非参数检验	需要使用非参数检验的条件
成对t检验	Wilcoxon配对检验	差异不是正态分布
两样本t检验	Mann-Whitney U检验□Wilcoxon检验的两样本情形)	方差不等；样本数少时非正态分布也有影响
方差分析ANOVA F检验	Kruskal-Wallis检验	方差不等；严重偏离正态分布

- 如何判断数据是否正态分布？方差相等？
  1. 判断正态分布□Shapiro-Wilk检验□Kolmogorov-Smirnov检验；
  2. 判断方差相等□Levene检验□Bartlett检验；
  3. Bartlett检验要求正态分布□Levene检验则没有要求。

• 如何进行相关性分析？

1. 当散点图提示两个变量为接近线性关系时，可以计算Pearson相关系数 $r$ 的绝对值越接近1，则相关性越强；
2. 如果出现以下情形，则 $r$ 值可能不正确：
  - 两个变量的关系不是线性（如二次函数关系）；
  - 每个个体出现一个以上的观察值；
  - 出现至少一个异常值；
  - 数据分布呈现次群组，且这些次群组至少有一个变量的平均值不同。
3. 如果出现以下情形的一项，应该计算Spearman相关系数而不是Pearson相关系数：
  - 至少一个变量是等级数据（非连续数据）；
  - 两个变量均不是正态分布；
  - 样本数很小；
  - 两个变量不是线性关系。

• 如何确定样本量？

1. 确定如下三个变量：
  - 效力，即 $1-\beta$ 一般至少80%；
  - 显著水平，即 $\alpha$ 一般为0.05或0.01；
  - 有临床意义的平均值的最小差异 $\delta$
2. 计算标准化的差异，见下表，如无标准差结果则可能需要预试验；
3. 初步计算样本量 $N$ 
  - 使用Altman计算图表，将标准化的偏差和效力连线，取得交叉点的数值，即总样本量；
  - 或者使用快速公式，即Lehr公式，用于计算效力80%、显著水平0.05的各组样本量
$$N = \frac{16}{(\text{Standardized Difference})^2}$$
如果效力为90%，则将分子16换乘21。如果标准化的差异小，该方法易高估。
  - 或者使用软件计算,如中提供的power.t.test等。
4. 校正样本量
  - 除了Lehr公式外，以上计算结果 $N$ 为总样本量，每组平均分配；
  - 在病人数少或药物有限等情况下，可减少一组人数而增加另一组人数；
  - 人数不平衡时，效力会下降，因此需要增加总样本数；
  - 设某组（通常为对照组）的人数是另一组（通常为实验组）的 $k$ 倍，则校正后的总样本数为
$$N' = \frac{N(1+k)^2}{4k}$$
故人数少的一组有 $\frac{N'}{1+k}$ 人，人数多的一组有 $\frac{kN'}{1+k}$ 人。

检验方法	标准化的差异	说明
成对t检验	$\frac{2\delta}{\sigma_d}$	$\sigma_d$ 为差异的标准差
非成对t检验	$\frac{\delta}{\sigma}$	$\sigma$ 为两组数据假设的相等的标准差
卡房检验	$\frac{p_1 - p_2}{\sqrt{\overline{p}(1-\overline{p})}}$	$p_1 - p_2$ 为两组的成功比例有临床意义的最小差异

• 样本量有限时如何增加效力？

- 收集更详细的信息，数值数据（如：血压的值）优于分类数据（如：正常血压/高血压）；
- 进行不同形式的分析，如：参数检验比非参数检验更有效力；
- 收集数据时减少随机误差，如：标准化、训练受试者；
- 修正研究设计。

From:  
<https://exile.tk/> - 漂流記

Permanent link:  
<https://exile.tk/doku.php?id=zh:misc:statfaq>

Last update: **2019/06/11 22:25**

