

# 数据分析的统计方法选择小结

## 目 录

数据分析的统计方法选择小结.....	1
目 录 .....	1
●资料 1 .....	2
完全随机分组设计的资料.....	2
配对设计或随机区组设计 .....	3
变量之间的关联性分析.....	4
●资料 2 .....	5
1.连续性资料.....	5
1.1 两组独立样本比较.....	5
1.2 两组配对样本的比较.....	5
1.3 多组完全随机样本比较.....	6
1.4 多组随机区组样本比较.....	6
2. 分类资料.....	6
2.1 四格表资料.....	6
2.2 2×C 表或 R×2 表资料的统计分析.....	7
2.3 R×C 表资料的统计分析.....	7
2.4 配对分类资料的统计分析.....	7
●资料 3 .....	8
一、两个变量之间的关联性分析 .....	8
二、回归分析.....	9
●资料 4 .....	9
一、统计方法抉择的条件.....	9
1. 分析目的.....	10
2. 资料类型.....	10
3. 设计方法.....	11
4. 分布特征及数理统计条件.....	12
二、数据资料的描述.....	12
1. 数值变量资料的描述.....	13
2. 分类变量资料的描述.....	13
三、数据资料的比较.....	14
1. 假设检验的基本步骤.....	14
2. 假设检验结论的两类错误.....	15
3. 假设检验的注意事项.....	15
4. 常用假设检验方法.....	16
四、变量间的相关分析.....	17
1. 数值变量（计量资料）的关系分析.....	18
2. 无序分类变量（计数资料）的相关分析.....	18
3. 有序分类变量（等级资料） 等级相关 .....	18

## ●资料 1

### 完全随机分组设计的资料

#### 一、 两组或多组计量资料的比较

##### 1. 两组资料:

1) 大样本资料或服从正态分布的小样本资料

(1) 若方差齐性, 则作成组  $t$  检验

(2) 若方差不齐, 则作  $t'$  检验或用成组的 Wilcoxon 秩和检验

2) 小样本偏态分布资料, 则用成组的 Wilcoxon 秩和检验

##### 2. 多组资料:

1) 若大样本资料或服从正态分布, 并且方差齐性, 则作完全随机的方差分析。如果方差分析的统计检验为有统计学意义, 则进一步作统计分析: 选择合适的方法(如: LSD 检验, Bonferroni 检验等) 进行两两比较。

2) 如果小样本的偏态分布资料或方差不齐, 则作 Kruskal Wallis 的统计检验。如果 Kruskal Wallis 的统计检验为有统计学意义, 则进一步作统计分析: 选择合适的方法(如: 用成组的 Wilcoxon 秩和检验, 但用 Bonferroni 方法校正  $P$  值等) 进行两两比较。

#### 二、 分类资料的统计分析

##### 1. 单样本资料与总体比较

###### 1) 二分类资料:

(1) 小样本时: 用二项分布进行确切概率法检验;

(2) 大样本时: 用  $U$  检验。

2) 多分类资料: 用 Pearson  $\chi^2$  检验 (又称拟合优度检验)。

##### 2. 四格表资料

1)  $n > 40$  并且所有理论数大于 5, 则用 Pearson  $\chi^2$

2)  $n > 40$  并且所有理论数大于 1 并且至少存在一个理论数  $< 5$ , 则用校正  $\chi^2$  或用 Fisher's 确切概率法检验

3)  $n \leq 40$  或存在理论数  $< 1$ , 则用 Fisher's 检验

##### 3. $2 \times C$ 表资料的统计分析

1) 列变量为效应指标, 并且为有序多分类变量, 行变量为分组变量, 则行评分的 CMH  $\chi^2$  或成组的 Wilcoxon 秩和检验

2) 列变量为效应指标并且为二分类, 列变量为有序多分类变量, 则用趋势  $\chi^2$  检验

3) 行变量和列变量均为无序分类变量

(1)  $n > 40$  并且理论数小于 5 的格子数  $<$  行列表中格子总数的 25%, 则用 Pearson  $\chi^2$

(2)  $n \leq 40$  或理论数小于 5 的格子数  $>$  行列表中格子总数的 25%, 则用 Fisher's 确

确切概率法检验

#### 4. R×C 表资料的统计分析

1) 列变量为效应指标, 并且为有序多分类变量, 行变量为分组变量, 则 CMH  $\chi^2$  或 Kruskal Wallis 的秩和检验

2) 列变量为效应指标, 并且为无序多分类变量, 行变量为有序多分类变量, 作 none zero correlation analysis 的 CMH  $\chi^2$

3) 列变量和行变量均为有序多分类变量, 可以作 Spearman 相关分析

4) 列变量和行变量均为无序多分类变量,

(1)  $n > 40$  并且理论数小于 5 的格子数 < 行列表中格子总数的 25%, 则用 Pearson  $\chi^2$

(2)  $n \leq 40$  或理论数小于 5 的格子数 > 行列表中格子总数的 25%, 则用 Fisher's 确切概率法检验

### 三、Poisson 分布资料

#### 1. 单样本资料与总体比较:

1) 观察值较小时: 用确切概率法进行检验。

2) 观察值较大时: 用正态近似的 U 检验。

#### 2. 两个样本比较: 用正态近似的 U 检验。

## 配对设计或随机区组设计

### 四、两组或多组计量资料的比较

#### 1. 两组资料:

1) 大样本资料或配对差值服从正态分布的小样本资料, 作配对 t 检验

2) 小样本并且差值呈偏态分布资料, 则用 Wilcoxon 的符号配对秩检验

#### 2. 多组资料:

1) 若大样本资料或残差服从正态分布, 并且方差齐性, 则作随机区组的方差分析。如果方差分析的统计检验为有统计学意义, 则进一步作统计分析: 选择合适的方法 (如: LSD 检验, Bonferroni 检验等) 进行两两比较。

2) 如果小样本时, 差值呈偏态分布资料或方差不齐, 则作 Fredman 的统计检验。如果 Fredman 的统计检验为有统计学意义, 则进一步作统计分析: 选择合适的方法 (如: 用 Wilcoxon 的符号配对秩检验, 但用 Bonferroni 方法校正 P 值等) 进行两两比较。

### 五、分类资料的统计分析

#### 1. 四格表资料

1)  $b+c > 40$ , 则用 McNemar 配对  $\chi^2$  检验或配对边际  $\chi^2$  检验

2)  $b+c \leq 40$ , 则用二项分布确切概率法检验

#### 2. C×C 表资料:

- 1) 配对比较：用 McNemar 配对  $\chi^2$  检验或配对边际  $\chi^2$  检验
- 2) 一致性问题 (Agreement)：用 Kap 检验

## 变量之间的关联性分析

### 六、 两个变量之间的关联性分析

1. 两个变量均为连续型变量
  - 1) 小样本并且两个变量服从双正态分布，则用 Pearson 相关系数做统计分析
  - 2) 大样本或两个变量不服从双正态分布，则用 Spearman 相关系数进行统计分析
2. 两个变量均为有序分类变量，可以用 Spearman 相关系数进行统计分析
3. 一个变量为有序分类变量，另一个变量为连续型变量，可以用 Spearman 相关系数进行统计分析

### 七、 回归分析

1. 直线回归：如果回归分析中的残差服从正态分布（大样本时无需正态性），残差与自变量无趋势变化，则直线回归（单个自变量的线性回归，称为简单回归），否则应作适当的变换，使其满足上述条件。
2. 多重线性回归：应变量 (Y) 为连续型变量（即计量资料），自变量 ( $X_1, X_2, \dots, X_p$ ) 可以为连续型变量、有序分类变量或二分类变量。如果回归分析中的残差服从正态分布（大样本时无需正态性），残差与自变量无趋势变化，可以作多重线性回归。
  - 1) 观察性研究：可以用逐步线性回归寻找（拟）主要的影响因素
  - 2) 实验性研究：在保持主要研究因素变量（干预变量）外，可以适当地引入一些其它可能的混杂因素变量，以校正这些混杂因素对结果的混杂作用
3. 二分类的 Logistic 回归：应变量为二分类变量，自变量 ( $X_1, X_2, \dots, X_p$ ) 可以为连续型变量、有序分类变量或二分类变量。
  - 1) 非配对的情况：用非条件 Logistic 回归
    - (1) 观察性研究：可以用逐步线性回归寻找（拟）主要的影响因素
    - (2) 实验性研究：在保持主要研究因素变量（干预变量）外，可以适当地引入一些其它可能的混杂因素变量，以校正这些混杂因素对结果的混杂作用
  - 2) 配对的情况：用条件 Logistic 回归
    - (1) 观察性研究：可以用逐步线性回归寻找（拟）主要的影响因素
    - (2) 实验性研究：在保持主要研究因素变量（干预变量）外，可以适当地引入一些其它可能的混杂因素变量，以校正这些混杂因素对结果的混杂作用
4. 有序多分类有序的 Logistic 回归：应变量为有序多分类变量，自变量 ( $X_1, X_2, \dots, X_p$ ) 可以为连续型变量、有序分类变量或二分类变量。
  - 1) 观察性研究：可以用逐步线性回归寻找（拟）主要的影响因素
  - 2) 实验性研究：在保持主要研究因素变量（干预变量）外，可以适当地引入一些其它可能的混杂因素变量，以校正这些混杂因素对结果的混杂作用

5. 无序多分类有序的 Logistic 回归: 应变量为无序多分类变量, 自变量 ( $X_1, X_2, \dots, X_p$ ) 可以为连续型变量、有序分类变量或二分类变量。

1) 观察性研究: 可以用逐步线性回归寻找 (拟) 主要的影响因素

2) 实验性研究: 在保持主要研究因素变量 (干预变量) 外, 可以适当地引入一些其它可能的混杂因素变量, 以校正这些混杂因素对结果的混杂作用

## 八、生存分析资料:

(要求资料记录结局和结局发生的时间, 如: 死亡和死亡发生的时间)

1. 用 Kaplan-Meier 方法估计生存曲线

2. 大样本时, 可以寿命表方法估计

3. 单因素可以用 Log-rank 比较两条或多条生存曲线

4. 多个因素时, 可以作多重的 Cox 回归

1) 观察性研究: 可以用逐步线性回归寻找 (拟) 主要的影响因素

2) 实验性研究: 在保持主要研究因素变量 (干预变量) 外, 可以适当地引入一些其它可能的混杂因素变量, 以校正这些混杂因素对结果的混杂作用

## ● 资料 2

### 1. 连续性资料

#### 1.1 两组独立样本比较

1.1.1 资料符合正态分布, 且两组方差齐性, 直接采用 t 检验。

1.1.2 资料不符合正态分布

(1) 可进行数据转换, 如对数转换等, 使之服从正态分布, 然后对转换后的数据采用 t 检验;

(2) 采用非参数检验, 如 Wilcoxon 检验。

1.1.3 资料方差不齐

(1) 采用 Satterthwate 的  $t'$  检验;

(2) 采用非参数检验, 如 Wilcoxon 检验。

#### 1.2 两组配对样本的比较

1.2.1 两组差值服从正态分布, 采用配对 t 检验。

1.2.2 两组差值不服从正态分布, 采用 wilcoxon 的符号配对秩和检验。

## 1.3 多组完全随机样本比较

1.3.1 资料符合正态分布，且各组方差齐性，直接采用完全随机的方差分析。如果检验结果为有统计学意义，则进一步作两两比较，两两比较的方法有 LSD 检验，Bonferroni 法，tukey 法，Scheffe 法，SNK 法等。

1.3.2 资料不符合正态分布，或各组方差不齐，则采用非参数检验的 Kruskal-Wallis 法。如果检验结果为有统计学意义，则进一步作两两比较，一般采用 Bonferroni 法校正 P 值，然后用成组的 Wilcoxon 检验。

## 1.4 多组随机区组样本比较

1.4.1 资料符合正态分布，且各组方差齐性，直接采用随机区组的方差分析。如果检验结果为有统计学意义，则进一步作两两比较，两两比较的方法有 LSD 检验，Bonferroni 法，tukey 法，Scheffe 法，SNK 法等。

1.4.2 资料不符合正态分布，或各组方差不齐，则采用非参数检验的 Friedman 检验法。如果检验结果为有统计学意义，则进一步作两两比较，一般采用 Bonferroni 法校正 P 值，然后用符号配对的 Wilcoxon 检验。

需要注意的问题：

(1) 一般来说，如果是大样本，比如各组例数大于 50，可以不作正态性检验，直接采用 t 检验或方差分析。因为统计学上有中心极限定理，假定大样本是服从正态分布的。

(2) 当进行多组比较时，最容易犯的错误是仅比较其中的两组，而不顾其他组，这样作容易增大犯假阳性错误的概率。正确的做法应该是，先作总的各组间的比较，如果总的来说差别有统计学意义，然后才能作其中任意两组的比较，这些两两比较有特定的统计方法，如上面提到的 LSD 检验，Bonferroni 法，tukey 法，Scheffe 法，SNK 法等。绝不能对其中的两组直接采用 t 检验，这样即使得出结果也未必正确。

(3) 关于常用的设计方法：多组资料尽管最终分析都是采用方差分析，但不同设计会有差别。常用的设计如完全随即设计，随机区组设计，析因设计，裂区设计，嵌套设计等。

## 2. 分类资料

### 2.1 四格表资料

2.1.1 例数大于 40，且所有理论数大于 5，则用普通的 Pearson 检验。

2.1.2 例数大于 40，所有理论数大于 1，且至少一个理论数小于 5，则用校正的检验或 Fisher's 确切概率法检验。

2.1.3 例数小于 40，或有理论数小于 2，则用 Fisher's 确切概率法检验。

## 2.2 $2 \times C$ 表或 $R \times 2$ 表资料的统计分析

2.2.1 列变量 & 行变量均为无序分类变量，则

(1) 例数大于 40, 且理论数小于 5 的格子数目 < 总格子数目的 25%, 则用普通的 Pearson 检验。

(2) 例数小于 40, 或理论数小于 5 的格子数目 > 总格子数目的 25%, 则用 Fisher's 确切概率法检验。

2.2.2 列变量为效应指标, 且为有序多分类变量, 行变量为分组变量, 用普通的 Pearson 检验只说明组间构成比不同, 如要说明疗效, 则可用行平均分差检验或成组的 Wilcoxon 秩和检验。

2.2.3 列变量为效应指标, 且为二分类变量, 行变量为有序多分类变量, 则可采用普通的 Pearson 检验比较各组之间有无差别, 如果总的来说有差别, 还可进一步作两两比较, 以说明是否任意两组之间的差别都有统计学意义。

## 2.3 $R \times C$ 表资料的统计分析

2.3.1 列变量 & 行变量均为无序分类变量，则

(1) 例数大于 40, 且理论数小于 5 的格子数目 < 总格子数目的 25%, 则用普通的 Pearson 检验。

(2) 例数小于 40, 或理论数小于 5 的格子数目 > 总格子数目的 25%, 则用 Fisher's 确切概率法检验。

(3) 如果要作相关性分析, 可采用 Pearson 相关系数。

2.3.2 列变量为效应指标, 且为有序多分类变量, 行变量为分组变量, 用普通的 Pearson 检验只说明组间构成比不同, 如要说明疗效或强弱程度的不同, 则可用行平均分差检验或成组的 Wilcoxon 秩和检验或 Ridit 分析。

2.3.3 列变量为效应指标, 且为无序多分类变量, 行变量为有序多分类变量, 则可采用普通的 Pearson 检验比较各组之间有无差别, 如果有差别, 还可进一步作两两比较, 以说明是否任意两组之间的差别都有统计学意义。

2.3.4 列变量 & 行变量均为有序多分类变量

(1) 如要做组间差别分析, 则可用行平均分差检验或成组的 Wilcoxon 秩和检验或 Ridit 分析。如果总的来说有差别, 还可进一步作两两比较, 以说明是否任意两组之间的差别都有统计学意义。

(2) 如果要做两变量之间的相关性, 可采用 Spearson 相关分析。

## 2.4 配对分类资料的统计分析

2.4.1 四格表配对资料

(1)  $b+c>40$ ，则用 McNemar 配对检验。

(2)  $b+c<40$ ，则用校正的配对检验。

#### 2.4.1 C×C 资料

(1) 配对比较：用 McNemar 配对检验。

(2) 一致性检验，用 Kappa 检验。

## ●资料 3

在研究设计时，统计方法的选择需考虑以下6方面的问题：(1)看反应变量是单变量、双变量还是多变量；(2)看单变量资料属于3种资料类型（计量、计数及等级资料）中的哪一种；(3)看影响因素是单因素还是多因素；(4)看单样本、两样本或多样本；(5)看是否是配对或配伍设计；(6)看是否满足检验方法所需的前提条件，必要时可进行变量变换，应用参数方法进行假设检验往往要求数据满足某些前提条件，如两个独立样本比较  $t$  检验或多个独立样本比较的方差分析，均要求方差齐性，因此需要做方差齐性检验。如果要用正态分布法估计参考值范围，首先要检验资料是否服从正态分布。在建立各种多重回归方程时，常需检验变量间的多重共线性和残差分布的正态性。

表 1-1

	连续因变量	分类因变量
连续自变量	回归分析	Logistic 回归
分类自变量	方差分析 (ANOVA)	表格检验 (比如卡方检验)

不同的统计分析方法都有其各自的应用条件和适用范围。实际应用时，必须根据研究目的、资料的性质以及所要分析的具体内容等选择适当的统计分析方法，切忌只关心  $p$  值的大小（是否 $<0.05$ ），而忽略统计分析方法的应用条件和适用范围。

## 一、两个变量之间的关联性分析

### 1. 两个变量均为连续型变量

1) 小样本并且两个变量服从双正态分布，则用 Pearson 相关系数做统计分析

2) 大样本或两个变量不服从双正态分布，则用 Spearman 相关系数进行统计分析

2. 两个变量均为有序分类变量，可以用 Spearman 相关系数进行统计分析

3. 一个变量为有序分类变量，另一个变量为连续型变量，可以用 Spearman 相关系数进行统计分析



## 二、回归分析

1. 直线回归：如果回归分析中的残差服从正态分布（大样本时无需正态性），残差与自变量无趋势变化，则直线回归（单个自变量的线性回归，称为简单回归），否则应作适当的变换，使其满足上述条件。

2. 多重线性回归：应变量（Y）为连续型变量（即计量资料），自变量（X1, X2, …, Xp）可以为连续型变量、有序分类变量或二分类变量。如果回归分析中的残差服从正态分布（大样本时无需正态性），残差与自变量无趋势变化，可以作多重线性回归。

1) 观察性研究：可以用逐步线性回归寻找（拟）主要的影响因素

2) 实验性研究：在保持主要研究因素变量（干预变量）外，可以适当地引入一些其它可能的混杂因素变量，以校正这些混杂因素对结果的混杂作用

3. 二分类的 Logistic 回归：应变量为二分类变量，自变量（X1, X2, …, Xp）可以为连续型变量、有序分类变量或二分类变量。

1) 非配对的情况：用非条件 Logistic 回归

(1) 观察性研究：可以用逐步线性回归寻找（拟）主要的影响因素

(2) 实验性研究：在保持主要研究因素变量（干预变量）外，可以适当地引入一些其它可能的混杂因素变量，以校正这些混杂因素对结果的混杂作用

2) 配对的情况：用条件 Logistic 回归

(1) 观察性研究：可以用逐步线性回归寻找（拟）主要的影响因素

(2) 实验性研究：在保持主要研究因素变量（干预变量）外，可以适当地引入一些其它可能的混杂因素变量，以校正这些混杂因素对结果的混杂作用

4. 有序多分类有序的 Logistic 回归：应变量为有序多分类变量，自变量（X1, X2, …, Xp）可以为连续型变量、有序分类变量或二分类变量。

1) 观察性研究：可以用逐步线性回归寻找（拟）主要的影响因素

2) 实验性研究：在保持主要研究因素变量（干预变量）外，可以适当地引入一些其它可能的混杂因素变量，以校正这些混杂因素对结果的混杂作用

5. 无序多分类有序的 Logistic 回归：应变量为无序多分类变量，自变量（X1, X2, …, Xp）可以为连续型变量、有序分类变量或二分类变量。

1) 观察性研究：可以用逐步线性回归寻找（拟）主要的影响因素

2) 实验性研究：在保持主要研究因素变量（干预变量）外，可以适当地引入一些其它可能的混杂因素变量，以校正这些混杂因素对结果的混杂作用。

## ● 资料 4

### 一. 统计方法抉择的条件

在临床科研工作中，正确地抉择统计分析方法，应充分考虑科研工作者的分析目的、临

床科研设计方法、搜集到的数据资料类型、数据资料的分布特征与所涉及的数理统计条件等。其中任何一个问题没考虑到或考虑有误，都有可能导致统计分析方法的抉择失误。

此外，统计分析方法的抉择应在科研的设计阶段来完成，而不应该在临床试验结束或在数据的收集工作已完成之后。

对临床科研数据进行统计分析和进行统计方法抉择时，应考虑下列因素：

## 1. 分析目的

对于临床医生及临床流行病医生来说，在进行统计分析前，一定要明确利用统计方法达到研究者的什么目的。一般来说，统计方法可分为描述与推断两类方法。一是统计描述(descriptive statistics)，二是统计推断(inferential statistics)。

统计描述，即利用统计指标、统计图或统计表，对数据资料所进行的最基本的统计分析，使其能反映数据资料的基本特征，有利于研究者能准确、全面地了解数据资料所包涵的信息，以便做出科学的推断。统计表，如频数表、四格表、列联表等；统计图，如直方图、饼图，散点图等；统计指标，如均数、标准差、率及构成比等。

统计推断，即利用样本所提供的信息对总体进行推断（估计或比较），其中包括参数估计和假设检验，如可信区间、t 检验、方差分析、 $\chi^2$  检验等，如要分析甲药治疗与乙药治疗两组的疗效是否不相同、不同地区某病的患病率有无差异等。

还有些统计方法，既包含了统计描述也包含了统计推断的内容，如不同变量间的关系分析。相关分析，可用于研究某些因素间的相互联系，以相关系数来衡量各因素间相关的密切程度和方向，如高血脂与冠心病、慢性宫颈炎与宫颈癌等的相关分析；回归分析，可用于研究某个因素与另一因素（变量）的依存关系，即以一个变量去推测另一变量，如利用回归分析建立起来的回归方程，可由儿童的年龄推算其体重。

## 2. 资料类型

资料类型的划分现多采用国际通用的分类方法，将其分为两类：数值变量(numerical variable)资料和分类变量(categorical variable)资料。数值变量是指其值是可以定量或准确测量的变量，其表现为数值大小的不同；而分类变量是指其值是无法定量或不能测量的变量，其表现没有数值的大小而只有互不相容的类别或属性。分类变量又可分为无序分类变量和有序分类变量两小类，无序分类变量表现为没有大小之分的属性或类别，如：性别是两类无序分类变量，血型是四类无序分类变量；有序分类变量表现为各属性或类别间有程度之分，如：临床上某种疾病的“轻、中、重”，治疗结果的“无效、显效、好转、治愈”。由此可见，数值变量资料、无序分类变量资料和有序分类变量资料又可叫做计量资料、计数资料和等级资料。

资料类型的划分与统计方法的抉择有关，在多数情况下不同的资料类型，选择的统计方法不一样。如数值变量资料的比较可选用 t 检验、u 检验等统计方法；而率的比较多用  $\chi^2$  检验。

值得注意的是，有些临床科研工作者，常常人为地将数值变量的结果转化为分类变量的临床指标，然后参与统计分析，如患者的血红蛋白含量，研究者常用正常、轻度贫血、中度贫血和重度贫血来表示，这样虽然照顾了临床工作的习惯，却损失了资料所提供的信息量。换言之，在多数情况下，数值变量资料提供的信息量最为充分，可进行统计分析的手段也较为丰富、经典和可靠，与之相比，分类变量在这些方面都不如数值变量资料。因此，在临床实验中要尽可能选择量化的指标反映实验效应，若确实无法定量时，才选用分类数据，通常不宜将定量数据转变成分类数据。

### 3. 设计方法

在众多的临床科研设计方法中，每一种设计方法都有与之相适应的统计方法。在统计方法的抉择时，必须根据不同的临床科研设计方法来选择相应的统计分析方法。如果统计方法的抉择与设计方法不一致，统计分析得到的任何结论都是错误的。

在常用的科研设计方法中，有成组设计（完全随机设计）的 t 检验、配对 t 检验、成组设计（完全随机设计）的方差分析、配伍设计（随机区组设计）的方差分析等，都是统计方法与科研设计方法有关的佐证。因此，应注意区分成组设计（完全随机设计）与配对和配伍设计（随机区组设计），在成组设计中又要注意区别两组与多组设计。最常见的错误是将配对或配伍设计（随机区组设计）的资料当做成组设计（完全随机设计）来处理，如配对设计的资料使用成组 t 检验、配伍设计（随机区组设计）使用成组资料的方差分析；或将三组及三组以上的成组设计（完全随机设计）资料的比较采用多个 t 检验、三个或多个率的比较采用四格表的卡方检验来进行比较，都是典型的错误。如下表：

表 1 常见与设计方法有关的统计方法抉择错误

设计方法	错误的统计方法	正确统计方法
两个均数的比较（成组设计、完全随机设计）	成组设计的 t 检验	成组设计的秩和检验
多个均数的比较（成组设计、完全随机设计）	多个成组设计的 t 检验	完全随机设计的方差分析及 q 检验、完全随机设计的秩和检验及两两比较
数值变量的配对设计	成组设计的 t 检验	配对 t 检验、配对秩和检验
随机区组设计（配伍设计）	多个成组设计的 t 检验、完全随机设计的方差分析	随机区组设计的方差分析及 q 检验、随机区组设计的秩和检验及两两比较
交叉设计	成组设计的 t 检验、配对 t 检验、配对秩和检验	交叉设计的方差分析、交叉设计的秩和检验

## 4. 分布特征及数理统计条件

数理统计和概率论是统计的理论基础。每种统计方法都要涉及数理统计公式，而这些数理统计公式都是在一定条件下推导和建立的。也就是说，只有当某个或某些条件满足时，某个数理统计公式才成立，反之若不满足条件时，就不能使用某个数理统计公式。

在数理统计公式推导和建立的条件中，涉及最多的是数据的分布特征。数据的分布特征是指数据的数理统计规律，许多数理统计公式都是在特定的分布下推导和建立的。若实际资料服从（符合）某种分布，即可使用该分布所具有的数理统计规律来分析和处理该实际资料，反之则不能。在临床资料的统计分析过程中，涉及得最多的分布有正态分布、偏态分布、二项分布等。

许多统计方法对资料的分布有要求，如：均数和标准差、t 和 u 检验；方差分析都要求资料服从正态分布，而中位数和四分位数间距、秩和检验等，可用于不服从正态分布的资料。所以，临床资料的统计分析过程中，应考虑资料的分布特征，最起码的要求是熟悉正态分布与偏态分布。

例如：在临床科研中，许多资料的描述不考虑资料的分布特征，而多选择均数与标准差。如某妇科肿瘤化疗前的血象值，资料如下表：

某妇科肿瘤化疗前的血象值

表格 2

指标名	例数	均数	标准差	偏度系数	P 值	峰度系数	P 值
血红蛋白(g/L)	98	111.99	18.82	0.180	0.459	0.025	0.958
血小板( $\times 10^9/L$ )	98	173.58	87.11	1.353	0.000	1.843	0.000
白细胞( $\times 10^9/L$ )	98	6.7930	2.767	1.207	0.000	1.202	0.013

从上结果可见，若只看三项指标的均数和标准差，临床医生也许不会怀疑有什么问题。但是经正态性检验，病人的血红蛋白服从正态分布，而血小板和白细胞两项指标的偏度和峰度系数均不服从正态分布（ $P < 0.05$ ）。因此，描述病人的血小板和白细胞平均水平正确的指标是中位数，而其变异程度应使用四分位数间距。

除了数据的分布特征外，有些数理统计公式还有其它一些的条件，如 t 检验和方差分析的方差齐性、卡方检验的理论数(T)大小等。

总之，对于临床科研工作者来说，为正确地进行统计方法的抉择，首先要掌握或熟悉上述影响统计方法抉择因素；其次，还应熟悉和了解常用统计方法的应用条件。

## 二. 数据资料的描述

统计描述的内容包括了统计指标、统计图和表，其目的是使数据资料的基本特征更加清晰地表达。本节只讨论统计指标的正确选用，而统计图表的正确使用请参阅其他书籍。

## 1. 数值变量资料的描述

描述数值变量资料的基本特征有两类指标，一是描述集中趋势的指标，用以反映一组数据的平均水平；二是描述离散程度的指标，用以反映一组数据的变异大小。各指标的名称及适用范围等见表 2。

表 3 描述数值变量资料的常用指标

表格 3

指标名称	用 途	适用的资料
均 数 ( $\bar{X}$ )	描述一组数据的平均水平，集中位置	正态分布或近似正态分布
中 位 数 (M)	与均数相同	偏态分布、分布未知、两端无界
几何均数 (G)	与均数相同	对数正态分布，等比资料
标准差 (S)	描述一组数据的变异大小，离散程度	正态分布或近似正态分布
四分位数间距	(QU-QL) 与标准差相同	偏态分布、分布未知、两端无界
极 差 (R)	与标准差相同	观察例数相近的数值变量
变异系数 (CV)	与标准差相同	比较几组资料间的变异大小

从表中可看出，均数与标准差联合使用描述正态分布或近似正态分布资料的基本特征；中位数与四分位数间距联合使用描述偏态分布或未知分布资料的基本特征。

这些描述指标应用时，最常见的错误是不考虑其应用条件的随意使用，如：用均数和标准差描述偏态分布、分布未知或两端无界的资料，这是目前在临床研究文献中较为普遍和典型的错误。

## 2. 分类变量资料的描述

描述分类变量资料常用的指标有死亡率、患病率、发病率等。

临床上，这类指标的应用较多，出现的错误也较多。这些错误归纳起来大致有两类：一是以比代率，即误将构成比 (proportion) 当做率 (rate) 来描述某病发生的强度和频率，如用某病的病人数除以就诊人数（或人次）得到“某病患病率”或“某病发病率”，就是典型的以比代率的例子。二是把各种不同的率相互混淆，如把患病率与发病率、死亡率与病死率等概念混同。

需要指出的是，单纯利用医院常规资料，最易得到的指标是构成比。而描述疾病发生强度和频率的指标的率反映如患病率、发病率、死亡率等，很难利用医院的常规资料（如医院医院的病例档案）获得。因为，医院常规资料无法得到计算这些率所需的分子和分母的资料。所以，一旦研究者利用的是医院常规资料，则无法衡量疾病对人群的危害程度。常用描述指标如表 3。

表 4 描述分类变量资料的常用指标

表格 4

指标名称	计算公式	意义
率	$\frac{\text{发生某现象的观察单位数}}{\text{可能发生某现象的观察单位总数}} \times K$	描述事件发生的强度和频率
构成比	$\frac{A}{A+B+\cdots} \times 100\%$	事物内部各组成部分所占的比重
相对比	$\frac{A}{B}$	A 指标为 B 指标的若干倍或百分之几

### 三. 数据资料的比较

在众多的科研研究方法中，归纳起来最基本的手段有两种，一是对研究对象的全体进行研究，在实际工作中往往难以实现；二是从总体中抽取一定数量的样本进行抽样研究，但要考虑抽样误差对结果的影响。因此，若用样本信息去推断其所代表的总体间有无差别时，需要使用假设检验(hypothesis testing)或称显著性检验(significance test)。

#### 1. 假设检验的基本步骤

(1) 建立检验假设。

建立假设的过程应有三个内容。即无效假设  $H_0$  (null hypothesis)、备择假设  $H_1$  (alternative hypothesis)和检验水准 (size of test)。无效假设  $H_0$  是研究者想得到结论的对立事件的假设，对于差异性检验而言，研究者想得到的是“有差别”的结论，故首先应假设各总体间无差别；备择假设  $H_1$  是其对立的假设，即是“有差别”的假设；此外，还应确定有统计意义的概率水平，通常取 0.05。建立检验假设的通常格式为：

$H_0$ ：多个样本来自同一总体，各样本间的差别是由于抽样误差所致

$H_1$ ：多个样本来自不同的总体，各样本间的差别是由于不同总体所致  
 $\alpha = 0.05$

(2) 计算统计量。

根据资料的类型、分布特征、科研设计方法等条件，**选择**不同的统计量计算方法，如 t 检验、u 检验等**统计方法**。

(3) 根据统计量的值得到概率(P)值；再按概率(P)值的大小得出结论。其结论只有两种情况，若  $P \leq \alpha$  时，即概率小于我们事先确定好的检验水平概率（如  $P \leq 0.05$ ），我们就拒绝其无差别假设  $H_0$ ，而接受  $H_1$ ，认为差别有统计学意义，各样本来自不同总体，样本间的差别是总体的不同所致；若  $P > \alpha$  时，其概率大于我们事先确定好的检验水平（如  $P > 0.05$ ），我们就不拒绝其无差别的假设  $H_0$ ，还不能认为各总体间有差别，样本来自同一总体，即差别没有统计学意义。

## 2. 假设检验结论的两类错误

在假设检验的两种结论中无论做出何种结论，都有可能犯错误。

当  $P \leq \alpha$  时，做出“拒绝其无差别的假设，可认为各总体间有差别”的结论时就有可能犯错误，这类错误称为第一类错误（I 型错误，type I error），其犯错误的概率用  $\alpha$  表示，若  $\alpha$  取 0.05，此时犯 I 型错误的概率小于或等于 0.05，若假设检验的 P 值比 0.05 越小，犯一类错误的概率就越小。

当  $P > \alpha$  时，做出“不拒绝其无差别的假设，还不能认为各总体间有差别”的结论时，就有可能犯第二类错误（II 型错误，type II error），其犯错误的概率用  $\beta$  表示，在通常情况下犯 II 类错误的概率未知，虽然  $\beta$  是个未知数，但假设检验 P 值越大，犯二类错误的概率就越小。

表 5 假设检验的两类错误

表格 5

真实情况	假设检验结果	
	拒绝 $H_0$	不拒绝 $H_0$
样本来自同一总体	推断不正确（ $\alpha$ ）	推断正确（ $1-\alpha$ ）
样本来自不同总体	推断正确（ $1-\beta$ ）	推断不正确（ $\beta$ ）

## 3. 假设检验的注意事项

（1）假设检验比较的对象是总体，而研究的方法是抽样研究，即通过对样本提供的信息去推断总体间有无差别。不能误认为假设检验是样本间的比较，更不能将此体现在结论中。如果研究方法是普查时，由于不存在抽样误差，也不存在用样本提供的信息去推断总体的问题。因此，在这种情况下也就不能使用假设检验的统计方法。

（2）当  $P \leq \alpha$  时，概率（P）越小，越有理由拒绝无差别的假设，即拒绝假设的可信程度就越大，这时概率（P）越小，其结论的可靠性就越好。当  $P > \alpha$  时，概率（P）越大，越有理由不拒绝无差别的假设，即不拒绝无差别假设的可信程度就越大。这时概率（P）越大，其结论的可靠性就越好。因此，无论概率  $P \leq \alpha$ ，还是  $P > \alpha$  时，都不能说明组间差别的大小。

（3）假设检验的结论不能绝对化。假设检验的结论是根据概率（P）的大小得出的，事实上当  $P \leq \alpha$  时，我们拒绝其无差别的假设，可认为各总体间有差别，但是，只要  $P \neq 0$ ，我们无法完全拒绝无差别的假设，即不能肯定各总体间有差别；同理，当  $P > \alpha$  时，我们不拒绝其无差别的假设，还不能认为各总体间有差别，但是，只要  $P \neq 1$ ，我们无法完全接受无差别的假设，即不能肯定各总体间无差别。因此，在做出统计结论时，要避免使用绝对的或肯定的语句，如当  $P \leq \alpha$  时，使用“拒绝假设，可认为各组间有差别”；而当  $P > \alpha$  时，使用“不拒绝假设，还不能认为各组间有差别”的语言进行描述。

（4）假设检验的方法与设计方案和分布特征有关，如：两组比较的方法有 t 检验、u 检验、两组秩和检验、四格表和校正四格表的  $\chi^2$  检验等，这些方法只能用于两组比



较，而不能用于多组的比较。在实际工作中错误地使用两组比较的方法代替多组比较的情况并不少见，如，三个均数比较用三个  $t$  检验、四个均数比较用六个  $t$  检验等。多组比较可用方差分析、多组秩和检验、行乘列  $\chi^2$  检验等。 $t$ 、 $u$  检验和方差分析用于正态分布的资料，不服从正态分布的资料可用秩和检验。

## 4. 常用假设检验方法

### (1) 计量资料的假设检验

表 6 常用计量资料假设检验方法

表格 6

比 较 的 目 的	应用条件	统计方法
样本与总体的比较	例数 ( $n$ ) 较大, (任意分布)	$u$ 检验
	例数 ( $n$ ) 较小, 样本来自正态	$t$ 检验
两组资料的比较	(完全随机设计) 例数 ( $n$ ) 较大, (任意分布)	$u$ 检验
	例数 ( $n$ ) 较小, 来自正态且方差齐	成组设计的 $t$ 检验
	例数 ( $n$ ) 较小且非正态或方差不齐	成组设计的秩和检验、或成组设计的 $t'$ 检验、或成组设计的中位数检验
配对资料的比较 (配对设计)	例数 ( $n$ ) 较大, (任意分布)	对设计的 $u$ 检验
	例数 ( $n$ ) 较小, 差值来自正态	配对设计的 $t$ 检验
	例数 ( $n$ ) 较小, 差值为非正态	配对设计的秩和检验
多组资料的比较 (完全随机设计)	各组均数来自正态且方差齐	成组设计的方差分析
	各组为非正态或方差不齐	成组设计的秩和检验
配伍资料的比较 (配伍设计)	各组均数来自正态且方差齐	配伍设计的方差分析
	各组为非正态或方差不齐	配伍设计的秩和检验

### (2) 计数资料的假设检验

表 7 常用计数资料假设检验方法



表格 7

比 较 的 目 的	应用条件	统计方法
样本率与总体率的比较	N 较小时	二项分布的直接法
	$np > 5$ 且 $n(1-p) > 5$	二项分布的 u 检验
两个率或构成比的比较（完全随机设计）	$np > 5$ 且 $n(1-p) > 5$	二项分布的 u 检验
	$N \geq 40$ 且 $T \geq 5$	四格表的 $\chi^2$ 检验
	$N \geq 40$ 且 $1 \leq T < 5$	校正四格表的 $\chi^2$ 检验
	$N < 40$ 或 $T < 1$	四格表的确切概率法
配对四格表比较（配对设计）	$B+c \geq 40$	配对 $\chi^2$ 检验
	$B+c < 40$	校正配对 $\chi^2$ 检验
多个率或构成比资料的比较（完全随机设计）	全部格子 $T \geq 5$ 或少于 $1/5$ 的格子 $1 \leq T < 5$	行 $\times$ 列表 $\chi^2$ 检验 (列联表 $\chi^2$ 检验)
	若有 $T < 1$ 或有多于 $1/5$ 的格子 $1 \leq T < 5$	行 $\times$ 列表的确切概率法 (列联表确切概率法)
注：n 为例数；T 为列联表中各格子的理论数；p 为样本率		

## (3) 等级资料的假设检验

表 8 常用等级资料假设检验方法

表格 8

比较目的	统计方法
两组比较（完全随机设计）	两组比较的秩和检验
多组比较（完全随机设计）	多组比较的秩和检验
配对设计	符号秩和检验
配伍设计	配伍设计的秩和检验

## 四．变量间的相关分析

数据资料的比较，是同一指标的不同处理组间的比较。在临床研究工作中，常常涉及疾病危险因素的研究和疾病病因的探索，即分析某个因素与疾病间的关系，如口服女性素避孕药是否是宫内膜癌的危险因素；高血脂症是否是冠心病心肌梗塞的危险因素。如果研究结果证明了它们是某种疾病的危险因素或与某种疾病有相关关系的话，还不能肯定其是因果关系，只有当某个因素导致某个肯定的结果，若该因素消除后，其相应的结果也不复存时候，这时，因果关系才能被肯定。

## 1. 数值变量（计量资料）的关系分析

表 9 常用数值资料的关系分析方法

表格 9

比较目的	应用条件	统计方法
两变量间的依存关系	正态单变量资料*	直线回归（I 型）
——	正态双变量资料**	直线回归（II 型）
两变量间的相互关系	正态双变量资料	直线相关
——	两变量都不服从正态	等级相关

注：\*为两变量中有一个变量服从正态分布的资料；\*\*为两变量都服从正态分布的资料。

## 2. 无序分类变量（计数资料）的相关分析

（1）前瞻性研究

相对危险度（RR）=  $\frac{\text{暴露于危险因素组的总体患病率}}{\text{未暴露于危险因素组的总体患病率}}$

归因危险度（AR）=  $\frac{\text{暴露于危险因素组的患病率} - \text{未暴露于危险因素组的患病率}}{\text{暴露于危险因素组的患病率}}$

（2）回顾性研究

比值比（OR）=  $\frac{ad}{bc}$

2×2 表：列联系数和四格表的  $\chi^2$  检验

行×列表：列联系数和行乘列表的  $\chi^2$  检验

## 3. 有序分类变量（等级资料） 等级相关

（略）