Health Statistics

2023-10-01

---

1 统计描述

首先了解数据的频数分布，选用合适的集中以及离散趋势描述数据。

- 观察单位。被观察或测量对象的最基本单位。
- 变量类型。数值变量/分类变量——有序，无序。
- 同质（homogeneity）与变异（variation）。研究对象具有的相同的状况或属性等共性称同质。对于同质的各观察单位，其某变量之间的差异，称为变异。
- 总体（population）。总体是根据研究目的确定的同质的观察单位全体，确切地说，是同质的所有观察单位某种变量值的集合。
- 样本（sample）。样本是指总体中的一部分观察单位的某项变量值的集合，这一部分必须是对总体具有代表性的。
- 误差（error）。包括系统误差（systermatic error）以及随机测量误差（error of randon measurement）和抽样误差（sampling error）。

1.1 数值变量

频数表的编制：

1. 求全距。
2. 确定组数, $range/n$ 取整 $[x, x+n)$。

3. 列表划计。

- 对称或者正态分布数据选用算术均值描述其均值，方差/标准差/变异系数描述其离散程度。

  标准差，变异系数是同单位的。变异系数用于不同尺度，均值相差较大的数据直接相互比较。方差/标准差/变异系数的计算都依赖于均值的计算。

  1. $CV = \frac{S}{X} \cdot 100\%$

- 对数正态分布数据, 例如抗体的几何滴度，细菌计数等，选用几何均值描述其分布，全距/四分位数描述其离散趋势。

1. $G = (\prod X)^{\frac{1}{n}} \to 10^{\frac{\sum lgX}{n}}$

- 任意其它分布选用中位数，全距/四分位数

1. $M = L + \frac{i}{f_X}(n \cdot X\% - \sum f_L)$

1.2 分类变量

分类变量数据主要依赖于各种相对数描述，包括比例（proportion），速率（rate），相对比（ratio）. 其中 rate 主要涉及到时间的概念，需要注意孕妇死亡率等是相对比指标。

数据的标准化法：

1. 直接化法，按总人口统一人口数，等价于对每组分层数据的率求均值-即每组中每个分层的权重都一致。
2. 间接法,依据标准化率计算不同组的理论值,实际值与理论值之比即得到标准化死亡比(standard mortality ration, SMR），进而求得当地的标准化死亡率 $p' = p \cdot SMR$.

动态数列：定基/环比，变化/增长（-1），平均发展速度/平均变化速度 (-1)。

2 实验设计与调查研究

2.1 实验设计

医学实验的特点：最终对象是人

1. 人具有生物学和社会学属性。
2. 人的个体变异性较大，实验单位的一致性较差，观察结果的离散程度较大。
3. 一般不允许在人体上直接实验，需先进行动物实验。

医学实验设计要素：

1. 处理因素。
2. 受试对象。受试对象要求对处理因素敏感，以及反应稳定。
3. 实验效应。

实验设计的基本原则：

1. 对照。
2. 随机化。包括抽取随机，分配随机，相同机会接受不同的实验顺序三层含义。
3. 重复。
4. 知情同意。

常用的设计方案包括完全随机化，区组化设计，析因设计以及被试内设计。被试内设计通过将个体的差异转化为重复测量的差异，而减少了误差。

2.2 调查研究

调查研究最大特点在于其只能被动观察，以及需要更大样本进而有更大的误差。

调查设计的原则需要完整，可行，经济，时效。

可行性分析可以通过逻辑分析，或者经验判断，或者试调查。

抽样的基本程序包括界定研究总体和调查总体，设计抽样方法，编制抽样框架，抽取样本，评估样本。

调查技术包括问卷，电话，访谈，观察法以及敏感问题调查技术-随机化应答技术。

非抽样误差包括抽样框误差，无回答误差，以及计量误差。

非抽样误差的估计有以下方法：

1. 调查质量控制措施的完善程度和落实情况。
2. 调查的应答率。
3. 比较不同来源的资料。
4. 进行抽样复查。

3 Parametric Statistics

3.1 Sampling Distribution for χ^2, F, t

首先给出各分布构造的定义：

- 若 $\{X_i\}_{i=1}^n$ 独立同分布于 $N(0,1)$，那么 $\sum X_i^2 \sim \chi^2(n)$, 其 $E(\chi^2) = n, Var(\chi^2) = 2n$.
- 若有 $\chi_1^2(m)$, $\chi_2^2(n)$, 那么 $\frac{\frac{\chi_1^2}{m}}{\frac{\chi_2^2}{n}} \sim F(m,n)$.
- 若有 $X \sim \mathcal{N}(0,1)$，以及 $\chi^2(n)$, 那么 $\frac{X}{\sqrt{\frac{\chi^2(n)}{n}}} \sim t(n)$

Theorem 3.1. *设 $\{x_i\}_{i=1}^n$ 是来自正态总体 $\mathcal{N}(\mu, \sigma^2)$ 的样本，其样本均值和方差分别为：*

1. $\bar{x} = \frac{1}{n}\sum x_i$
2. $s^2 = \frac{1}{n-1}\sum (x - \bar{x})^2$

则：

1. $\bar{x} \sim \mathcal{N}(\mu, \sigma^2/n)$
2. $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$
3. $\bar{x}, s^2$ 相互独立

    1. 这里描述的是正态总体样本的均值服从于正态分布，而其样本的方差 s 服从于卡方分布。
    2. 注意 s 是样本的方差，而不是属于样本均值正态分布的方差，样本均值的正态分布的方差为 $\frac{\sigma^2}{n}$

Theorem 3.2. *设 $\{x_i\}_{i=1}^m$ 是来自 $\mathcal{N}(\mu_1, \sigma_1)$ 的样本，$\{y_i\}_{i=1}^n$ 是来自 $\mathcal{N}(\mu_2, \sigma_2)$ 的样本，那么：*

1. $\bar{x} = \frac{1}{m}\sum x, \bar{y} = \frac{1}{n}\sum y$
2. $s_x^2 = \frac{1}{m-1}\sum (x-\bar{x})^2, s_y^2 = \frac{1}{n-1}\sum (y-\bar{y})^2$

则：

1. $\frac{s_x^2/\sigma_1^2}{s_y^2/\sigma_2^2} \sim F(m-1, n-1)$

证明：

1. $\frac{(m-1)s_x^2}{\sigma^1} \sim \chi^2(m-1), \frac{(n-1)s_y^2}{\sigma_2^2} \sim \chi^2(n-1)$
2. $\frac{(1.1)/(m-1)}{(1.2)/(n-1)} \sim F(m-1, n-1)$

Theorem 3.3. 设 $\{x_i\}_{i=1}^n$ 是来自正态总体 $\mathcal{N}(\mu, \sigma^2)$ 的样本，其样本均值和方差分别为：

1. $\bar{x} = \frac{1}{n}\sum x_i$
2. $s^2 = \frac{1}{n-1}\sum (x-\bar{x})^2$

则：

1. $\frac{\bar{x}-\mu}{\sigma\cdot\sqrt{\frac{1}{n}}} \sim t(n-1)$

证明：

1. $\frac{\bar{x}-\mu}{\sigma\cdot\sqrt{\frac{1}{n}}} \sim \mathcal{N}(0,1)$
2. $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$
3. $\frac{(1)}{\sqrt{(2)/(n-1)}} \to \frac{\bar{x}-\mu}{s\cdot\sqrt{\frac{1}{n}}} \sim t(n-1)$

Theorem 3.4. 设 $\{x_i\}_{i=1}^m$ 是来自 $\mathcal{N}(\mu_1, \sigma_1^2)$ 的样本，$\{y_i\}_{i=1}^n$ 是来自 $\mathcal{N}(\mu_2, \sigma_2^2)$ 的样本，那么：

1. $\bar{x} = \frac{1}{m}\sum x, \bar{y} = \frac{1}{n}\sum y$
2. $s_x^2 = \frac{1}{m-1}\sum (x-\bar{x})^2, s_y^2 = \frac{1}{n-1}\sum (y-\bar{y})^2$

设 $\sigma_1^2 = \sigma_2^2 = \sigma^2$，则：

1. $\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{s_p\cdot\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t(m+n-2)$, *Where* $s_p^2 = \frac{(m-1)s_x^2+(n-1)s_y^2}{m+n-2}$

证明：

由两样本独立且正态分布：

1. $(\bar{x}-\bar{y}) \sim \mathcal{N}(\mu_1-\mu_1, (\frac{1}{m}+\frac{1}{n})\sigma^2)$
2. $\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sigma\cdot\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim \mathcal{N}(0,1)$

由卡方变量可加性：

1. $\frac{(m-1)s_x^2}{\sigma_1^2} \sim \chi^2(m-1), \frac{(n-1)s_y^2}{\sigma_2^2} \sim \chi^2(n-1)$
2. $\frac{(m-1)s_x^2}{\sigma^2} + \frac{(n-1)s_y^2}{\sigma^2} \sim \chi^2(m+n-2) \rightarrow \frac{s_p^2\cdot(m+n-2)}{\sigma^2} \sim \chi^2(m+n-2)$

3.2 t/z Test

3.2.1 Assumptions and Its Application Case

From wikipedia:

> For exactness, the t-test and Z-test require normality of the sample means, and the t-test additionally requires that the sample variance follows a scaled χ^2 distribution, and that the sample mean and sample variance be statistically independent. Normality of the individual data values is not required if these conditions are met. By the central limit theorem, sample means of moderately large samples are often well-approximated by a normal distribution even if the data are not normally distributed. For non-normal data, the distribution of the sample variance may deviate substantially from a χ^2 distribution.
>
> However, if the sample size is large, Slutsky's theorem implies that the distribution of the sample variance has little effect on the distribution of the test statistic. That is as sample size

- $\sqrt{n}(\bar{X}-\mu) \xrightarrow{d} N(0,\sigma^2)$ as per the Central limit theorem.
- $s^2 \xrightarrow{p} \sigma^2$ as per the Law of large numbers.
- $\therefore \frac{\sqrt{n}(\bar{X}-\mu)}{s} \xrightarrow{d} N(0,1)$

我们知道 t 分布的构造定义如下：

$$\frac{X}{\sqrt{\chi_{(n-1)}/(n-1)}}, \text{Where} X \sim \mathcal{N}(0,1)$$

从 t 的构造中可知，我们需要一个 $X \sim \mathcal{N}(0,1)$ 的正态变量；而在大样本情况下，依据中心极限定理，样本均值总是符合正态分布。进一步地，依据大数定理，$s^2 \approx \sigma^2$，所以直接用 s 替代 σ 可以进行 z 检验。实际上，在大样本下如下公式总是成立的：

1. $\bar{x} \sim \mathcal{N}(\mu, \frac{1}{n}\sigma^2) \rightarrow$ Central Limit Theory

   Specially, 设二项分布 $X \sim B(n,p)$:

   则：

   1. $\bar{x} \sim \mathcal{N}(np, np(1-p)), \text{When} x \rightarrow \infty$
   2. $\hat{p} \sim \mathcal{N}(p, \frac{1}{n}p(1-p))$ 对于二项分布而言，其总体的期望以及方差实际上描述的就是多次独立的伯努利实验的期望与方差。

而大样本下 t 分布近似于 z 分布。所以 t 分布可以应用于小样本下的正态总体，以及大样本下的任意总体的均值的检验。

在应用 t 检验时，我们往往只有一个或两个样本的均值与方差，计算的关键就在于如何利用样本方差计算得到均值分布的方差。

3.2.2 One Smple t-Test

1. $\bar{x} \sim \mathcal{N}(\mu, \frac{1}{n}\sigma^2) \rightarrow \frac{\bar{x}-\mu}{\sigma \cdot \sqrt{\frac{1}{n}}} \sim \mathcal{N}(0,1)$
2. $\frac{\bar{x}-\mu}{s \cdot \sqrt{\frac{1}{n}}} \sim t(n-1)$

   应用条件：

1. 小样本正态分布，大样本。

3.2.3 Paired t-Test

$t(\upsilon) = \frac{|\bar{d}-0|}{s_{\bar{d}}} = \frac{\bar{d}}{s_{\bar{d}}}$, where $\upsilon =$ 对子数 $-1, s_{\bar{d}} = s_d \cdot \sqrt{\frac{1}{n}}$

证明：

1. $\frac{\bar{d}-\mu_d}{\sigma_d \cdot \sqrt{\frac{1}{n}}} \sim \mathcal{N}(0,1)$

应用条件：

1. 小样本下，d 服从于正态分布。
2. 大样本。

3.2.4 Independent Two Sample t-Test

两个总体的比较需要对方差齐性进行检验：

1. $F = \frac{S_1^2}{S_2^2}, \upsilon_1 = n_1 - 1, \upsilon_2 = n_2 - 1$, 其中 S_1^2 是比较大的那个。

When $\sigma_1 = \sigma_2 = \sigma$:

1. $\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{s_p \cdot \sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t(m+n-2)$, Where $s_p^2 = \frac{(m-1)s_x^2+(n-1)s_y^2}{m+n-2}$

When $\sigma_1 \neq \sigma_2$ or Unknown for it $\rightarrow$ Welch's t-Test:

1. $t'(\upsilon) = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}}$

通过 Satterhwaite 法，对自由度进行校正

1. $\upsilon = \frac{(\frac{s_1^2}{m}+\frac{s_2^2}{n})^2}{\frac{s_1^4}{m^2}/(m-1)+\frac{s_2^4}{n^2}/(n-1)}$

3.3 $B(n,p)/Po(\lambda)$ Test

设事件 A 发生的概率为 π, 那么在 n 次伯努利试验中，该事件发生次数 k:

1. $P(k) = \binom{n}{k}\pi^k \cdot (1-\pi)^{n-k}$, Where $\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$

其中 $\binom{n}{k}$ 正好是牛顿二项展开式 $[(1-\pi)+\pi]^n$ 第 $k+1$ 项。

设 $X \sim B(n,\pi)$, 则：

1. $X \sim \mathcal{N}(n\pi, n\pi(1-\pi))$, When $n\pi > 5$ and $n(1-\pi) > 5$
2. $Z = \frac{X-n\pi}{\sqrt{n\pi(1-\pi)}} = \frac{p-\pi}{\sqrt{\pi(1-\pi)/n}} \sim \mathcal{N}(0,1)$
3. $p \sim \mathcal{N}(\pi, \frac{\pi(1-\pi)}{n})$

当 π 特别小时，设 $\lambda = n\pi$, 则当 $n \rightarrow \infty$ ：

1. $P(X) = \frac{e^{-\lambda}\lambda^X}{X!}$

    1. n 足够大，以至于每 $\frac{1}{n}$ 中只有发生与不发生两种情况。
    2. 每一份 $\frac{1}{n}$ 中其概率都为 $\frac{\pi}{n}$
    3. 每一份中之间是相互独立的。

    则 $X \sim Po(\lambda)$

1. $X \sim \mathcal{N}(\lambda, \lambda) \to \frac{X-\lambda}{\sqrt{\lambda}} \sim \mathcal{N}(0,1), \text{When } \lambda > 20$
2. $Z = \frac{X_1/n_1 - X_2/n_2}{\sqrt{X_1/n_1^2 + X_2/n_2^2}}$

泊松分布最大的特征就是其均值与方差都等于 λ, 我们常用这一点来判断一个分布是否属于泊松分布。此外，泊松分布具有可加性。考虑 $X_1 \sim Po(\lambda_1), X_2 \sim Po(\lambda_2)$, 且互相独立：

1. $X_1 + X_2 \sim Po(\lambda_1 + \lambda_2)$

3.4 Aanlysis of Variance

3.4.1 One Way ANNOVA

该假设的原理如下：

> If the group means are drawn from populations with the same mean values, the variance between the group means should be lower than the variance of the samples, following the central limit theorem - wikipedia

其需要满足的假设如下：

1. 正态总体
2. 方差齐性
3. 样本独立性

> Tiku (1971) found that "the non-normal theory power of F is found to differ from the normal theory power by a correction term which decreases sharply with increasing sample size." The problem of non-normality, especially in large samples, is far less serious than popular articles would suggest - wikipedia

1. $Y_{ij} - \bar{Y} = (Y_{ij} - Y_i) + (Y_i - \bar{Y})$
2. $\sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}\right)^2 = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Y_{ij} - Y_i\right)^2 + \sum_{i=1}^{a} n_i \left(Y_i - \bar{Y}\right)^2$
3. $F = \dfrac{\sum_{i=1}^{a} n_i (\bar{Y_i} - \bar{Y})^2 \Big/ (a-1)}{\sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y_i})^2 \Big/ (N-a)} \sim F_{(a-1,N-a)}$

从公式的推导之中可以看到，其核心在于对变异 (SS, sum of squares) 的分解 $SS_{\text{total}} = SS_{\text{within group}} + SS_{\text{between group}}$。方差分析需要进行 Levene's 方差齐性检验，该检验实际上是对每个值减去该组均值的差异做单因素方差分析：

1. $Z_{ij} = |Y_{ij} - \bar{Y_i}|$
2. $L = \dfrac{\sum_{i=1}^{a} n_i (\bar{Z_i} - \bar{Z})^2 \Big/ (a-1)}{\sum_{i=1}^{a} \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z_i})^2 \Big/ (N-a)} \sim F_{(a-1,N-a)})$

3.4.2 Special Two Way ANNOVA - Random Block Design

当观测变量有多个时，进行方差分析，不仅要考虑每个变量的观测值的影响，还需要考虑变量之间的交互作用对观测值的影响。而随机区块设计通过分组以后再进行随机化，使得我们可以不考虑两者的交互作用，只考虑两个变量分别对结果的影响；

1. $Y_{ij} - \bar{Y} = (\bar{Y_i} - \bar{Y}) + (\bar{Y_j} - \bar{Y}) + (Y_{ij} - \bar{Y_i} - \bar{Y_j} + \bar{Y})$

从上面的公式可以看出，随机区组设计将变异分为两个变量来源（区组，实验因素），以及个体本身的误差，其中误差项的自由度为 $(n-1)(\upsilon - 1)$.

对于混杂因素的控制还有将线性回归与方差分析结合起来的协方差分析，其对个体值的分解如下：

1. $Y_{ij} = \bar{Y_i} + b(\bar{X}_{ij} - \bar{X}) + e_{ij}$

其中 X 是混杂因素。

3.5 Mutiple Hypothesis Test

方差分析对各处理组均数是否相等总的检验，在 H_0 被拒绝以后，需要确定究竟是哪些处理组之间存在差异，此时需要进行均数之间的多重比较，这就涉及到累计 I 型错误率。

当 a 个处理组均数需要两两比较时候，共需要比较 $c = a!/[2!(a-2)!]$。设每次检验的检验水准为 α , 累积 I 型错误概率为 $'\alpha$, 则

1. $'\alpha = 1 - (1-\alpha)^c$

3.5.1 q -Test/student-Newman-Keuls

qTest 用于任意两组之间的相互比较，SNK 法的检验效能介于 Bonferroni 和 Tukey 法之间的；当比较均值的组数较多时，Tukey 法更有效，组数较少时，ferroni 法更有效。

其计算过程如下：

1. 将各组的平均值按由小到大的顺序排列。
2. 计算两个平均之间的差值以及组间跨度 r

则 q 统计量：

1. $q = \frac{\bar{Y_i} - \bar{Y_h}}{\sqrt{\frac{MS_{within\ group}}{2}(\frac{1}{n_i} + \frac{1}{n_h})}}$

其中 $\bar{Y_i}, \bar{Y_h}$ 及 n_i, n_h 分别是两个比较组的均数以及样本例数，$MS_{\text{within group}}$ 为进行方差分析得到的组内均方。

3.5.2 Dunnett-t Test

t_D 统计量处理组与对照组的比较，该统计量的计算如下：

1. $t_D = \frac{\bar{Y_i} - \bar{Y_c}}{\sqrt{MS_{within\ group} \times (\frac{1}{n\ _i} + \frac{1}{n_c})}}$

其中 $\bar{Y_i}, \bar{Y_c}$ 及 n_i, n_c 分别是实验组与对照组的均数以及样本例数，$MS_{\text{within group}}$ 为进行方差分析得到的组内均方。

4 Non-parametric Statistics

4.1 χ^2 Test

卡方检验用于观测变量为无序分类变量时，用于检验零假设下观测频数与理论频数 (将所有组合并为一个组计算出的每个结局的频率分布，然后乘以原组的频数) 之间偏离如此大范围的概率（如果观测变量为有序的分类变量，则该使用秩相关的检验)。该方法应用的条件如下：

1. 不宜有 $\frac{1}{5}$ 的格子数的理论频数小于 5, 或有一个格子的理论频数小于 1，否则将导致分析的偏性。可采取扩大样本含量，或者合并或者删除不符合条件的数据。后两者会损失信息，样本随机性，可能会影响推断结论。
2. 多个样本率（即多分组，观测变量为无序的二分类变量）的比较，显著的差异只能推断出这几组之间有总体的差异，但是不能推出其中两者之间是否存在差异。此时可将不同组的数据两两组合重新进行卡方检验进行推断，此时的检验水准的计算公式为 $\frac{\alpha}{k(k-1)/2+1}$

考虑一个 $R \cdot C$ 的列联表

1. $\chi^2 = \sum \frac{(O-E)^2}{E}$, Where$E_{rc} = \frac{n_r n_c}{n} \rightarrow \chi^2 = n(\sum \frac{O^2}{n_r n_c} - 1)$
    1. 如果为单变量，按样本分组；如果观测变量是无序二分类的，则是率的比较；如果观测变量为无序多分类，则为频数分布的比较。
    2. 如果为两个无序的观测变量，则为两个观测变量关联性检验，此时还需要进一步计算关联系数 C (contigency coefficient) $= \sqrt{\frac{\chi^2}{n+\chi^2}}$

对于称为四格表的资料, 即 $2 * 2$ 的列联表，其计算的简化形式为：

1. $\chi^2 = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$

特别地，当理论频数存在 $1 < E < 5$ 时有如下的 Yate correction for continuity：

1. $\chi^2 = \sum \frac{(|O-E|-0.5)^2}{E} = \frac{(|ad-bc|-\frac{n}{2})^2 n}{(a+b)(c+d)(a+c)(b+d)}$

    实际上 Pearson 卡方值是正态总体中一种连续性的变量，四格表资料的卡方值为不连续的值，卡方分布仅仅是对表格资料的统计量分布的近似分布。当四格表中有小于 5 的期望值时，其卡方值偏大，减去 0.5 进行 Yate 的连续性校正。

若 $n \leq 40$ or $E \leq 1$ 则采用 Fisher 确切概率法。

1. $P = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}} = \frac{(a+b)!(a+c)!(b+c)!(b+d)!}{a!b!c!d!n!}$

   实际上 fisher 确切概率法给出了 $R*C$ 列联表的确切概率，可以用于任意情况的检验；但是这里给出的是某一种情况的概率，为了求得比当前情况都更为极端的差值，在四格表资料中我们比较的两个样本率的比较，因此对于所有率的差值大于原假设的极端情况，我们需要计算所有的概率累加，得到发生如此极端值的概率。

当四格表资料为配对设计时，该检验称为 Mc-Nemar Test, 检验两者阳性率是否一致，该值完全由两组中阳性的频数决定，则：

1. $\chi^2 = \frac{(b-c)^2}{b+c}, \upsilon = 1$

若 $b + c \leq 40$, 则：

1. $\chi^2 = \frac{(|b-c|-1)^2}{b+c}, \upsilon = 1$

因为 Pearson χ^2 能反映实际频数和理论频数的的吻合程度，所以 χ^2 检验可以用作频数分布的拟合优度检验（goodness of fit test），用于判断样本是否符合正态分布，二项分布，Possion 分布等。

4.2 Rank Based Test

用于总体分布未知，且观测变量为数值变量或者有序分类变量情况，需要关注的是不同的秩和检验的秩和 T 是如何计算的以及遇到相等的数据该如何处理（只有数值变量会遇到相同的数据，对于观测变量为有序的等级数据，因为用的是其每一类结果的平均秩次，不存在该问题）。

4.2.1 Wilcoxon Signed Rank Test for Paired Sample

依差值的绝对值从小到大编秩。编秩时遇到差值为 0 的舍去不计，同时样本例数 $n-1$; 遇到绝对值差值相等差数，符号相同则顺次编秩，符号相反则取平均秩次，再给秩次冠以原差值的正负号。分别计算出正负秩次 T_+, T_-, 任取其中一个作为统计量秩和 T

1. $T \sim \mathcal{N}(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}), \text{When } n > 25$
2. $Z = \frac{|T-n(n+1)/4|-0.5}{\sqrt{n(n+1)(2n+1)/24}}$ 其中 0.5 为连续性校正常数。

当相同差值数较多时（不包括差值为 0 的值），校正式

1. $Z = \frac{|T-n(n+1)/4|-0.5}{\sqrt{n(n+1)(2n+1)/24-\frac{\sum(t_j^3-t_j)}{48}}}$ 其中 t_j 是第 j 个相同差值的个数。

4.2.2 Wilcoxon Rank Sum Test/Mann Whitney Test for Independent Two Samples

1. 若观测变量为数值变量：将两组原始数据分别从小到大排队，再将两组数据由小到大统一编秩，若有同组相同数据则顺序编秩，若有不同组别相同数据则取平均秩次。记两组中样本例数较小的为 n_1, 其秩和为统计量 T.
2. 若观测变量为有序的多分类变量：将每个观测单位按观测变量等级排序，则观测变量各个等级的平均秩次为该组观测单位秩次和的均值，则可求得每个分组的秩和，取观测单位数较小的组的秩和作为统计量 T.

则统计量 T:

1. $T \sim \mathcal{N}(\frac{n_1(N+1)}{2}, \frac{n_1n_2(N+1)}{12}), \text{Where} N = n_1 + n_2$
2. $Z = \frac{|T-\frac{n_1(N+1)}{2}|-0.5}{\sqrt{\frac{n_1n_2(N+1)}{12}}}$

当相同秩较多时，有如下校正：

1. $Z_c = Z\sqrt{C}, \text{Where } C = 1 - \sum(t_j^3 - t_j)/(N^3 - N)$

4.2.3 ANNOVA for Rank

单因素的方差分析对应 Kruskal-Wallis Test, 随机区组设计的方差分析对应 Freidman Test.

Kruskal-Wallis Test

构造 H 统计量：假设有 a 个组，第 i 组的样本量 n_i, N 为各组样本量之和，将各组数据合并，编秩次，秩次相同的取平均值。R_{ij} 为第 i 个组的第 j 个个体的秩次，$\bar{R}_i$ 为第 i 个组的平均秩次，$\bar{R}$ 为总平均秩次。

1. $H = \frac{\sum_i^a n_i(\bar{R}_i-\bar{R})^2}{\frac{1}{N-1}\sum_{i=1}^a\sum_{j=1}^{n_i}(R_{ij}-\bar{R})^2}$

从上式可以看出 H 统计量实际上是组间变异与总的变异的比值。

没有相同秩次时，秩次服从均匀分布, 上式可以简化为：

1. $H = \frac{12}{N(N+1)}(\sum\frac{R_i^2}{n_i}) - 3(N+1).$

相同秩次过多时，上述以均匀分布为基础推导的公式需要进行校正：

1. $H_c = H/C$

其中 $C = 1 - \sum(t_j^3 - t_j)/(N^3 - N)$。

n_i 与 a 较小时直接计算或者查表。

n 较大时, H 近似服从于 χ^2_{a-1}.

实际上也可以对数据编秩，然后用数据的秩次代替原数据进行方差分析得到 F 统计量，H 统计量与 F 统计量有如下关系：

$F = \frac{H/(a-1)}{(N-1-H)/(N-a)}$

Friedman Test

在区组（行）内进行编秩，有相同的则取平均秩次。i 代表不同的区组，j 代表不同地处理水平。

1. $M = \frac{\sum_{j=1}^{a} n(\bar{R}_j - \bar{R})^2}{\sum_{j=1}^{a}\sum_{i=1}^{n}(R_{ij} - \bar{R})^2/n(a-1)}$

从上式可以看到 M 统计量是处理水平之间的变异与总的变异的比值。

如果没有相同秩次时，上式可以简化为：

1. $M = \frac{12}{na(a+1)}\sum_{j=1}^{a} R_j^2 - 3n(a+1)$

当相同秩过多时可以进行校正，校正系数为

1. $C = 1 - \sum_{j=1}^{a}\sum_{p=1}^{l_i}(t_{jp}^3 - t_{jp})/[na(a^2-1)]$
2. $M_c = M/C$

当 n 以及 a 较小时直接查表或精确计算。n 较大时 M 近似服从于 χ^2_{a-1}.

同样地，我们也可以直接用编秩后数据代替原始数据的值进行随机区组设计的方差分析，此时的 F 统计量与 M 统计量的关系如下：

1. $F = \frac{M/(a-1)}{(na-n-M)/(n-1)(a-1)}$

5 Correlation ann Linear Regression

5.1 Pearson/Spearman Correlation

考虑两个连续的正态分布变量 X, Y, 那么其样本的 Pearson 积差相关系数 (product-moment correlation coefficient):

1. $r_p = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2 \sum(Y-\bar{Y})^2}}$
2. $t_r = \frac{r-0}{S_r}, \upsilon = n-2$ Where $S_r = \sqrt{\frac{1-r^2}{n-2}}$

对于相关有如下注意的问题:

1. 分层数据合并假象，合并分层不改变其相关性时，才可以合并。
2. 两个变量应该都是随机的，而不是控制一个变量，观察另一个变量的结果。

对于非正态分布变量或者总体分布未知变量考虑使用 Spearman 等级相关 (rank correlation)。将 n 对观察值 X_i, Y_i 由小到大编秩，则有:

1. $r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$
2. $t_r = \frac{r-0}{S_r}, \upsilon = n-2$ Where $S_r = \sqrt{\frac{1-r^2}{n-2}}$ When $n \geq 20$

5.2 Simple Linear Regression

简单线性回归，即只有一个自变量的线性回归，其回归方程形式如下:

1. $\hat{Y} = a + bX$
2. $b = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}$, $a = \bar{Y} - b\bar{X}$

对参数 b 有如下 F 检验:

1. $F = \frac{MS_{\text{regression}}}{MS_{\text{residuals}}}, \upsilon_{\text{regression}} = 1, \upsilon_{\text{residuls}} = n-2$

对参数 b 同时有如下 t 检验:

1. $t_b = \frac{b-0}{S_b}, \upsilon = n-2$ Where $S_b = \frac{\sqrt{\frac{SS_{\text{residules}}}{n-2}}}{\sqrt{\sum(X-\bar{X})^2}}$
2. $t_b \pm t_{\alpha,\upsilon}(S_b)$

对于拟合程度的判断可由决定系数确定：

1. $R^2 = \frac{SS_{\text{regressiion}}}{SS_{\text{total}}}$

对于拟合的回归方程，确定数值 X_0, 那么有：

1. $\mu_{Y|X_0} = \hat{Y_0} \pm t_{\alpha,\upsilon} S \sqrt{\frac{1}{n} + \frac{(X-X_0)^2}{\sum (X-\bar{X})^2}}$
2. $Y_0 = \hat{Y_0} \pm t_{\alpha,\upsilon} S \sqrt{1 + \frac{1}{n} + \frac{(X_0-\bar{X})^2}{\sum (X-\bar{X})^2}}$

利用上面的第二个式子可以进行统计控制，即如果想要将 Y 控制在一定范围，则可以通过控制 X 将个体预测值控制在一定范围之中。

6 关于假设检验中若干问题

6.1 假设检验原理

求得样本统计量的分布，则可求得 H_0 下样本统计量向假设的预期偏离如此大范围的机率，则依据小概率定律，选择拒绝或者接受原假设。

但是实际上要求得统计量的分布并不容易，对于参数统计，我们常常需要假设其分布，然后推理出相关统计量的分布；非参数统计则需要对其编秩，然后推断出秩和的分布。

6.2 假设检验与置信区间

对于 $t/ztest$ 而言，统计量 t/z 的大小表示为向均值偏离多少个标准差, 而置信区间的形式则为均值加上多少个标准差。

1. $t/z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$
2. $\mu_{95\%} = \bar{x} \pm \frac{t}{z_{95\%}} \cdot \sigma_{\bar{x}}$

置信区间的含义在于用 100 个样本均值计算得到置信区间，其范围包括总体均值的个数为 95，但是对个一个样本均值计算出来的区间，其是否包括该总体均值只有是或者不是，而不是有 95% 的概率包括，因为该区间以及总体均值都是确定的值。

6.3 I and II type error and power

H_0 没有差异为阴性结果，H_1 为具有差异，阳性结果。I 型错误是拒绝 H_0 所犯的错误为 α, II 型错误为不拒绝 H_0 所犯的错误为 β, 其含义为在 H_1 得到该统计量的概率，$1-\beta$ 即为检验功效，即在 H_1 为真的情况下，在指定的 α 检出其存在差异的概率，即敏感度 (sensitivity).

6.4 统计量分布计算的 Bootstrap 以及置换检验

但是在实际的假设检验过程，倘若需要求得样本统计量的分布，往往需要知晓总体的概率分布。由此在假设检验过程之中，除了零假设以外，实际上在样本统计量的分布的计算之中，同时隐含了总体的分布以及方差齐性等假设，而这些假设也有对应的检验方法。因此，假设检验真正的概率应该所有假设概率的乘积，而不仅仅是在零假设之下所计算出的概率。以上所考虑的是一定要得到样本统计量的确切分布，进而得到精确的概率，例如正态总体假设下，方差齐性的 $ttest$, 但是很多情况下我们不必得到样本统计量的精确分布，得到一个大概估计即可，例如大样本下的正态检验，$R \times C$ 列表值用卡方分布近似，非参数检验-秩检验等。在计算机时代，我们可以直接利用 Bootstrp 模拟得到假设之下样本统计量的分布。

6.5 假设检验与线性回归参数检验的等效性