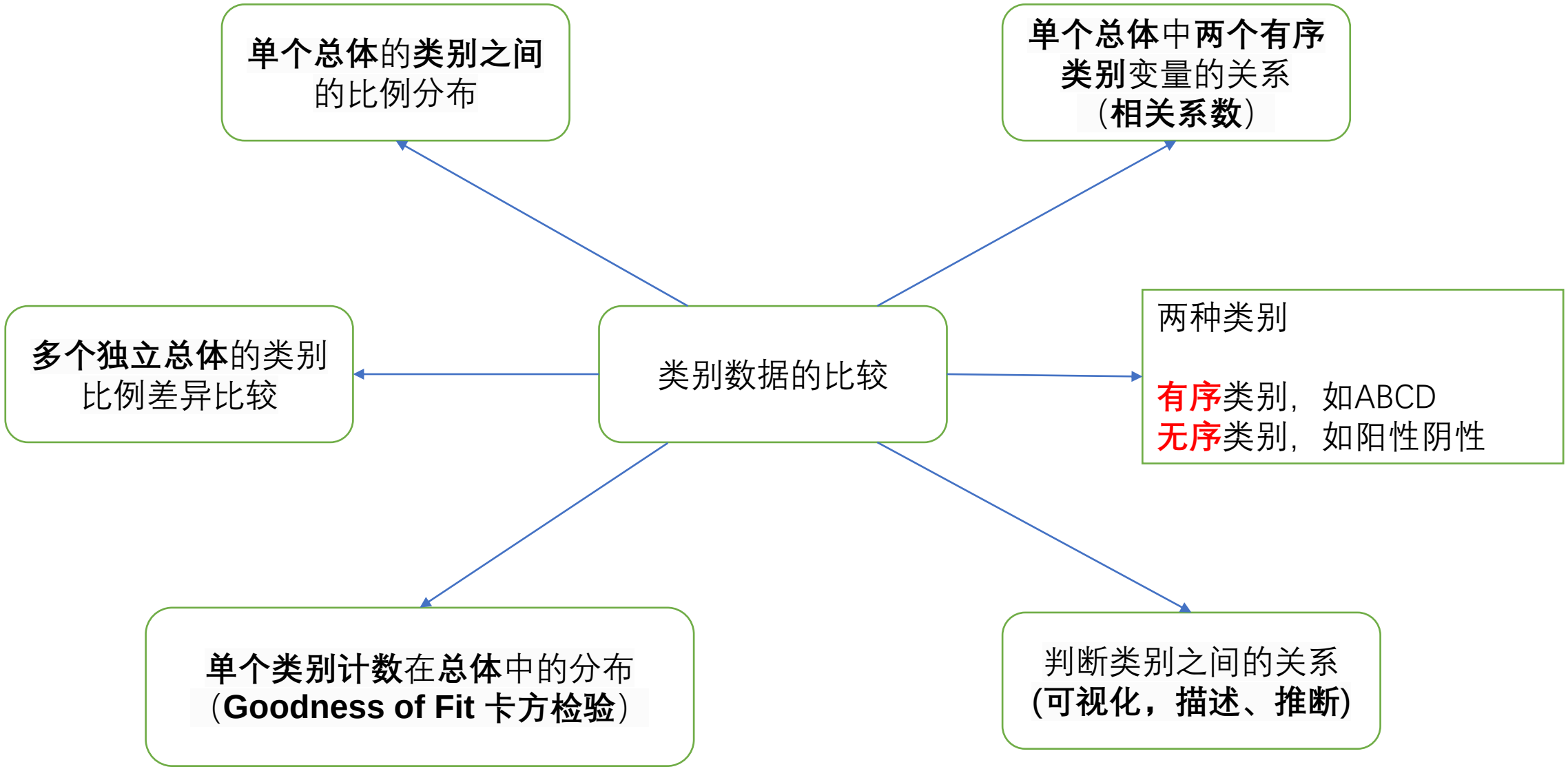


* 默认所有样本具有**随机性**和**足够大**

对样本的附加条件

1. 样本观测数<总体的5%
2. 每个类别的观测量>10

使用CI估计
单个总体的类别之间
的比例分布

$$\mathrm{CI}_{1-\alpha} = p_1 \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p_1(1-p_1)}{n}}$$

$$\mathrm{CI}_{1-\alpha} = p_1 \pm \left(Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p_1(1-p_1)}{n}} + \frac{1}{2n}\right)$$

* 下方为修正后的CI

计算原理

每个类别的观测数$\sum x_i \sim B(n, p_i)$，其中n是样本观测数，p_i 是总体比例，当样本满足附加条件时，$\sum x_i \sim N\left(np_i, np_i(1-p_i)\right)$，得到比例的分布
$p = \frac{\sum x_i}{n} \sim N\left(p_i, \frac{p_i(1-p_i)}{n}\right)$，使用观测到的$p_1 = \frac{\sum x_i}{n}$估计$p_i$，得到枢轴量$p_1 \sim N\left(p_i, \frac{p_1(1-p_1)}{n}\right)$

多个独立总体的类别比例差异比较

两个总体

与单个总体类似的，得到比例之差服从的分布 $(p_1-p_2) \sim N(P_1-P_2, \frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2})$，
其中P指总体，p指样本，用其计算CI

差异的置信区间

$$CI_{1-\alpha} = p_1 - p_2 \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

$$CI_{1-\alpha} = p_1 - p_2 \pm Z_{\frac{\alpha}{2}} \cdot \left(\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}+\frac{1}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right)$$

* 下方为修正后的CI

使用zTest的NHST

枢轴量p_1-p_2，如上所示$H_0: P_1 = P_2$
设u服从Z分布，则有p-value，以下的显著指实际显著性
双侧检验:$p = P(|u| \geq |\frac{p_1-p_2}{\sigma}|)$，当p1，p2**相当**时
左侧检验:$p = P(u \leq \frac{p_1-p_2}{\sigma})$，当p1**显著小于**p2时
右侧检验:$p = P(u \geq \frac{p_1-p_2}{\sigma})$，当p1**显著大于**p2时
不难发现单边检验的**p-value更小，统计显著性更高**

* 仅两个总体可以使用zTest

* 使用z分布求CI和NHST时，样本必须服从**附加条件**(见上一页)

多个总体

RC联表的χ^2检验

属于假设检验，H_0:一个类别占**所有类别的比例在所有总体中相同**
即$p_{11} = p_{12} = p_{13}; p_{21} = p_{22} = p_{23}; \dots\dots$

* 卡方检验**貌似**很少会用在多个类别中，
即$r \geq 3$很少见

	样本1	样本2	样本3
类别A	Obs11	Obs12	Obs13
类别B	Obs21	Obs22	Obs23
类别C	Obs31	Obs32	Obs33

	样本1	样本2	样本3
类别A	Exp11	Exp12	Exp13
类别B	Exp21	Exp22	Exp23
类别C	Exp31	Exp32	Exp33

Obs和Exp分别是**观测数**和**期望数**
，$Exp_{kj} = \sum_i Obs_{ij} \cdot \frac{\sum_i Obs_{ki}}{\sum_i \sum_l Obs_{il}}$
得到$\sum_i \sum_j \frac{(O_i-E_i)^2}{E_i} \sim \chi^2$，df=(r-1)(c-1),
为了保证行列之和不变各自-1，对χ^2进行NHST

卡方检验的效应量

Cramer's V= $\sqrt{\frac{\chi^2}{N \cdot \min(r-1, c-1)}}$ 表征**总体对比例不同的关联**
0.1<V<0.2为弱关联
0.2<V<0.5为中等关联
V>0.5为强关联

卡方检验的条件

* 各个**单元格独立**观测
* 2x2表:
  1. $E_{ij} \geq 10$
  2. $5 \leq E_{ij} < 10$，使用Yate's correction（仍是卡方检验）
  3. $E_{ij} < 5$，使用Fisher's Exact Test（不是卡方检验）
* 大于2x3表:
  1. E<5的单元格小于20%
  2. 可以合并几列再做

检验原理

Obs和Exp分别是**观测数**和**期望数**

Exp_i是利用假设的分布和观测数构造的期望分布

当总体满足假设期望**(H0)**时，$\sum_i \frac{(O_i-E_i)^2}{E_i} \sim \chi^2$

df=c-q-1，其中**c是样本数**，**q**是构造Exp时**用到的Obs**的函数的**个数**(如Exp~N(mean(Obs),SD(Obs),q=2)，**-1**是为了保证**总数不变**

进行NHST

	样本1	样本2	样本3	样本4
某个类别	Obs1	Obs2	Obs3	Obs4

	样本1	样本2	样本3	样本4
某个类别	Exp1	Exp2	Exp3	Exp4

单个类别的**计数**在**总体**中的分布

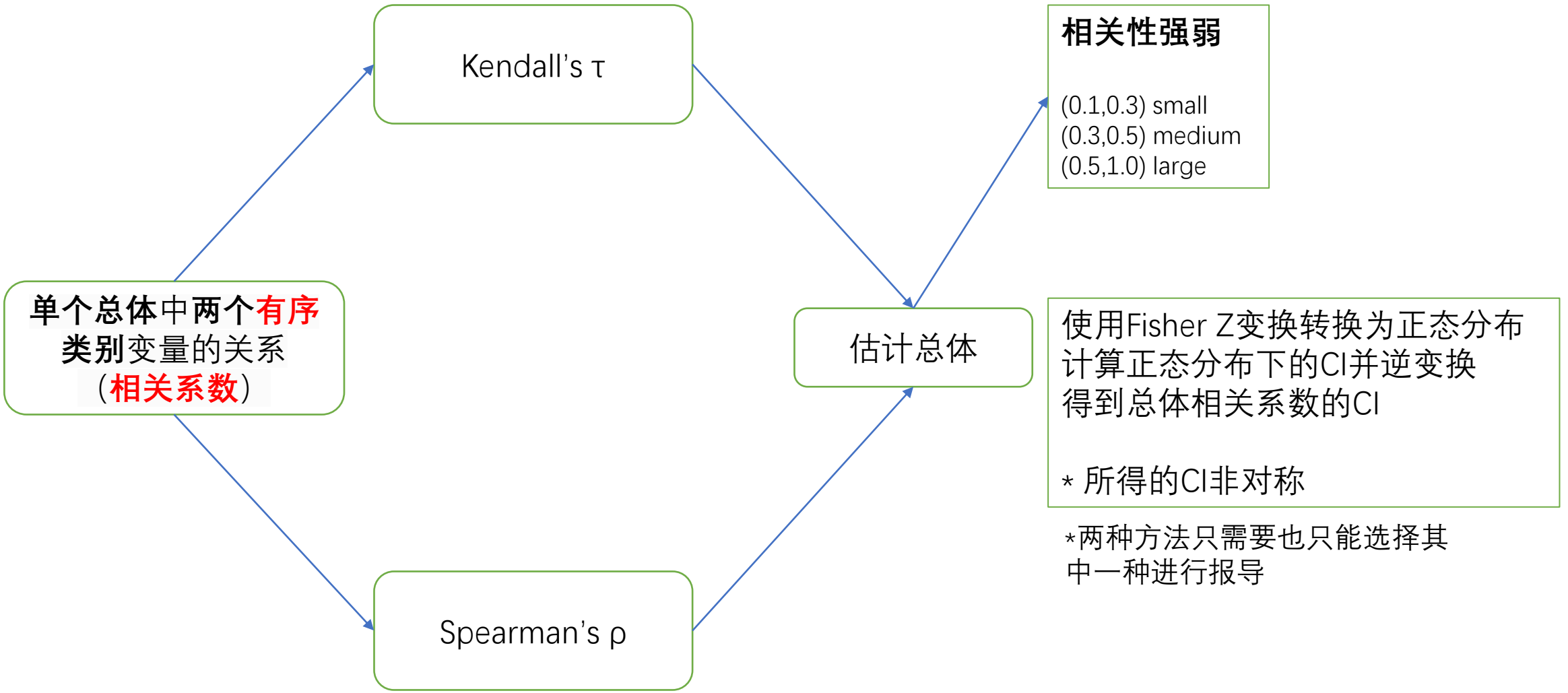
(Goodness of Fit 卡方检验)

每个样本应当是独立的

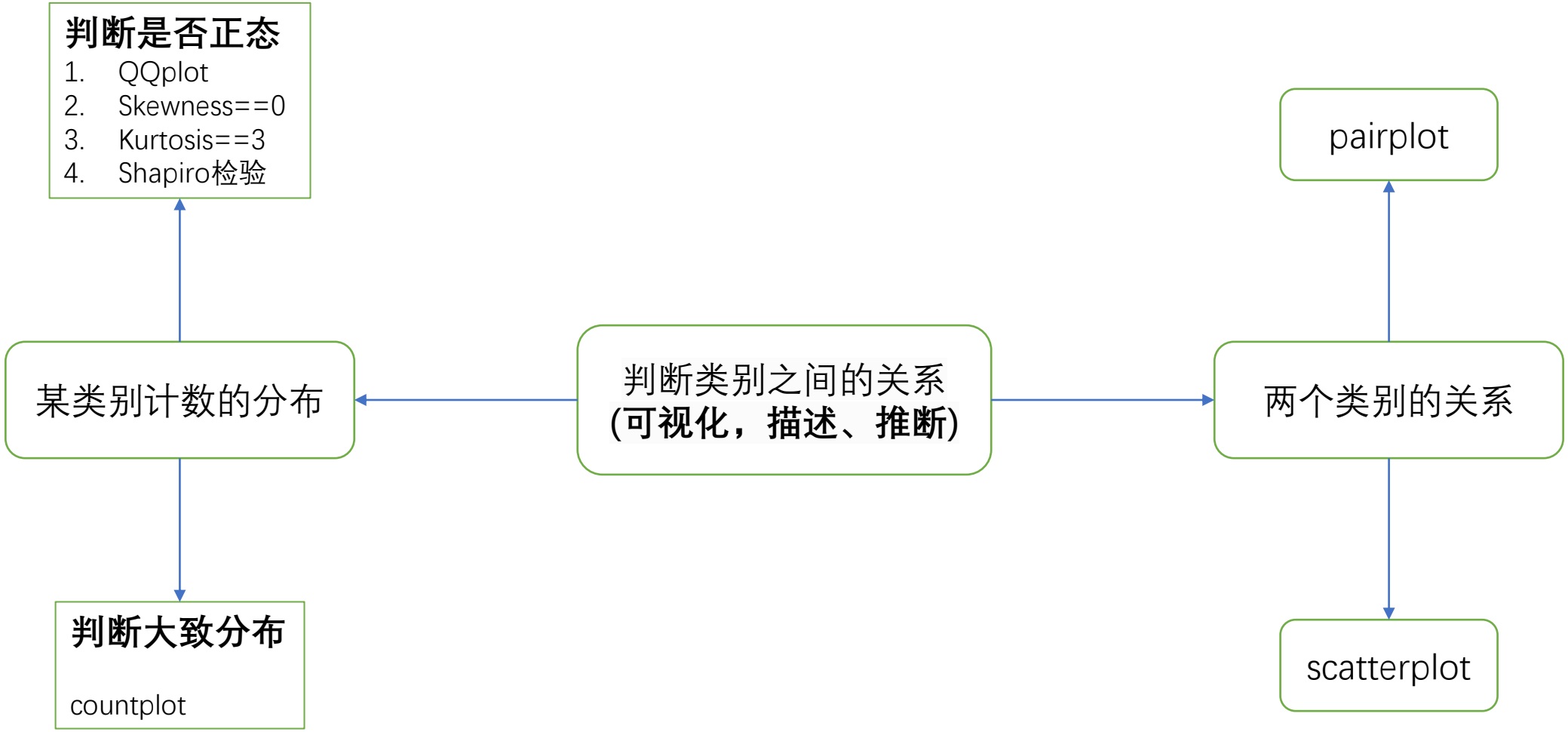
统计人群中身高的分布，样本为不同身高区间的人数

统计不同城市某种疾病的阳性率，样本为不同城市患病人数（在每个城市抽样数一定的条件下）

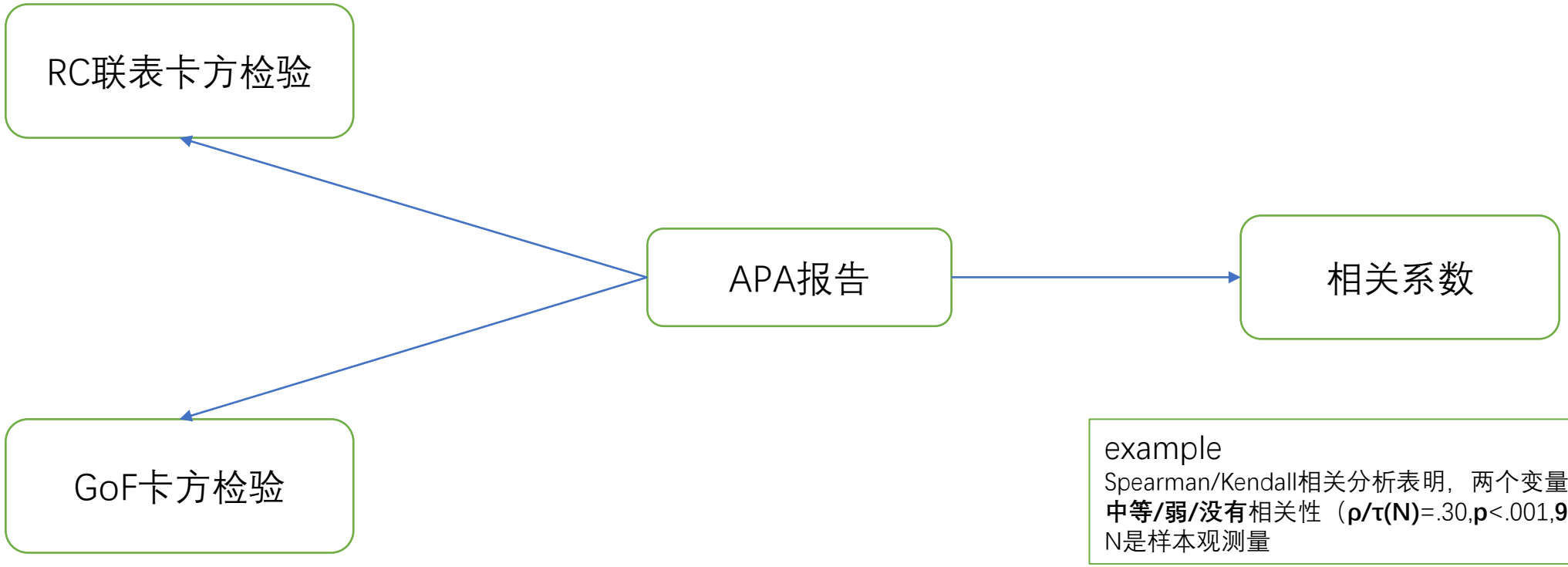
* 使用GoF对正态分布的检验是最准确的


Kendall's τ
单个总体中两个有序类别变量的关系（相关系数）
Spearman's ρ
估计总体
相关性强弱
(0.1,0.3) small
(0.3,0.5) medium
(0.5,1.0) large
使用Fisher Z变换转换为正态分布
计算正态分布下的CI并逆变换
得到总体相关系数的CI
* 所得的CI非对称
*两种方法只需要也只能选择其中一种进行报导

* 如何计算略去


判断是否正态
1. QQplot
2. Skewness==0
3. Kurtosis==3
4. Shapiro检验
某类别计数的分布
判断大致分布
countplot
判断类别之间的关系
(可视化，描述、推断)
两个类别的关系
pairplot
scatterplot

example
卡方检验表面剂量与治疗效果具有**显著**的**弱关联性**（**chi2(df)**=14.23,**p**<.001,**Cramer's V**=0.18）



example
Spearman/Kendall相关分析表明，两个变量之间存在**很强/中等/弱/没有**相关性（**ρ/τ(N)**=.30,**p**<.001,**95%CI**=[0.15,0.35]）
N是样本观测量

example
通过检验，感染人数的分布**显著服从/不服从**XX**分布**，**chi2(df,N)=**xx,**p**=.xxx
其中N是样本观测量