单样本t检验：检验样本均值与总体均值的差异是否显著

双样本t检验：检验两个样本的均值差异是否显著

双样本t检验考虑条件：

1. 观测之间独立，即观测之间不能互相影响
2. 两组均服从正态分布，即样本分布正态
3. 两组样本的方差是否相同，视其是否相同会采用不同统计量进行检验

- 方差齐性检验(F统计量)来检验两个方差是否相同

\* F统计量由两组样本方差中最大值除以最小值得到

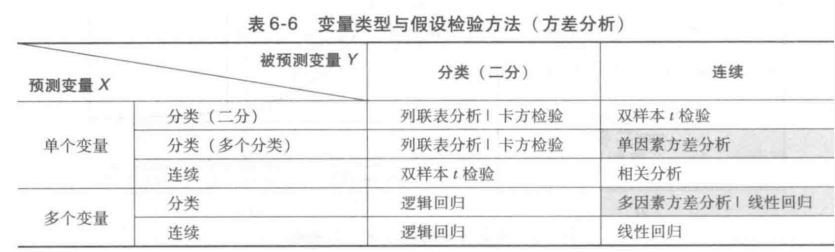


双样本t检验流程

1. 获得两组样本数据，计算其均值
2. 进行方差齐性检验
3. 若方差齐，则进行方差齐双样本t检验；若不齐，则进行方差不齐的双样本t检验

**方差分析**

方差分析用于检验多个样本均值是否有显著差异，故其用于分析多于两个分类的分类变量与连续变量的关系



1. 单因素方差分析

单因素可得到不同因素对观测变量的影响程度，这里因素的不同水平表示因素不同的状态或等级。

原假设：所有组方差相等，备择假设为至少有两组方差不等。

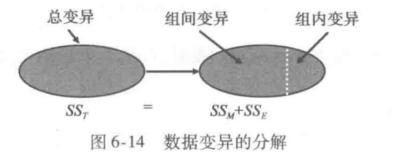
数据总误差分为组内误差和组间误差

\* 组间变异(SS\_m)：同类别下数据的离均差平方和，代表同类别下数据变异程度

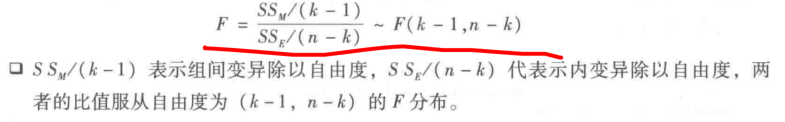
\* 组内变异(SS\_e)：组内均值与总均值的离均差平方和，代表不同类别数据的变异程度

组间变异与组内变异之和为总变异。

- 组内离差平方和受随机误差影响；组间离差平方和受不同水平的影响

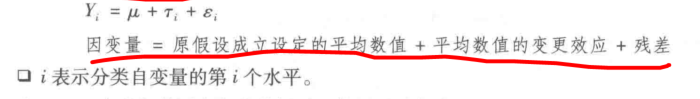


组内差异与组间差异的统计量：



当F值越大，说明组间差异越大，越拒绝原假设，即组间有差异；反之，组间无差异。

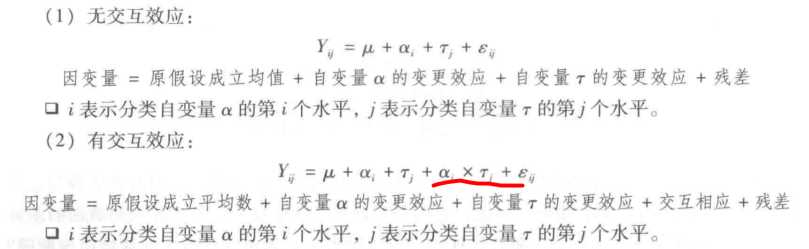
单因素方差分析另一种方法类似于回归：



1. 多因素方差分析

单因素方差分析可检验一个分类变量与一个连续变量间的关系，多因素方差分析可检验多个分类变量与一个连续变量的关系。

多因素方差分析中，除考虑多个分类变量对连续变量影响外，还应考虑分类变量之间的交互效应。



**相关分析（两连续变量关系检验）**

两个连续变量之间的关系可以使用相关分析



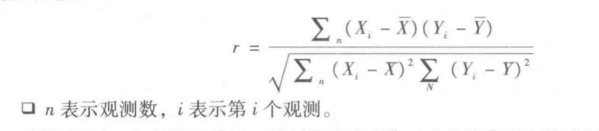
两个连续变量之间关系可划分为线性关系和非线性关系：

\* 线性关系

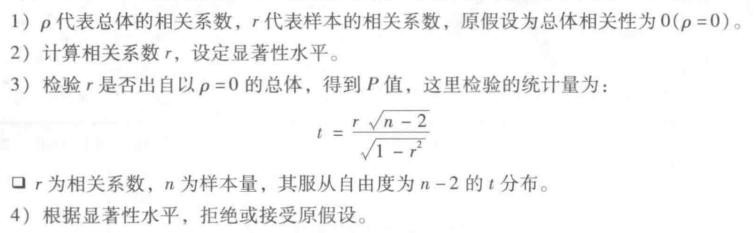
- 可用皮尔逊相关系数对两变量线性相关关系进行探索

- pearson相关系数适合计算两个独立连续的线性相关变量相关程度，前提是变量服从

正态分布



- 计算出pearson相关系数后，可用t检验检查该系数是否有统计学意义



简单相关分析是研究两个变量之间相关关系的方法，按变量性质的不同，采用的相关分析方法也不同。

连续变量可用Pearson相关系数描述变量的相关关系；对于有序变量，则用Spearman相关系数描述。

- Spearman相关系数称为秩相关系数，其使用排序信息而不是变量观测的取值信息进行相关分析，优点是不用假设变量服从正态分布，故使用较广泛

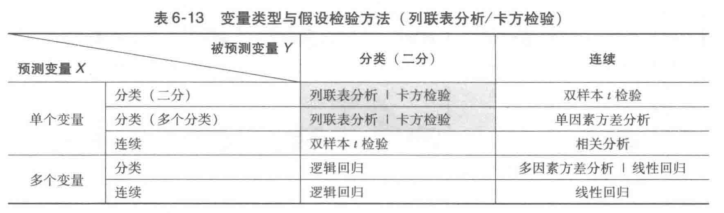
\* 非线性关系

- kendall相关系数主要用于探索两连续变量间的非线性关系



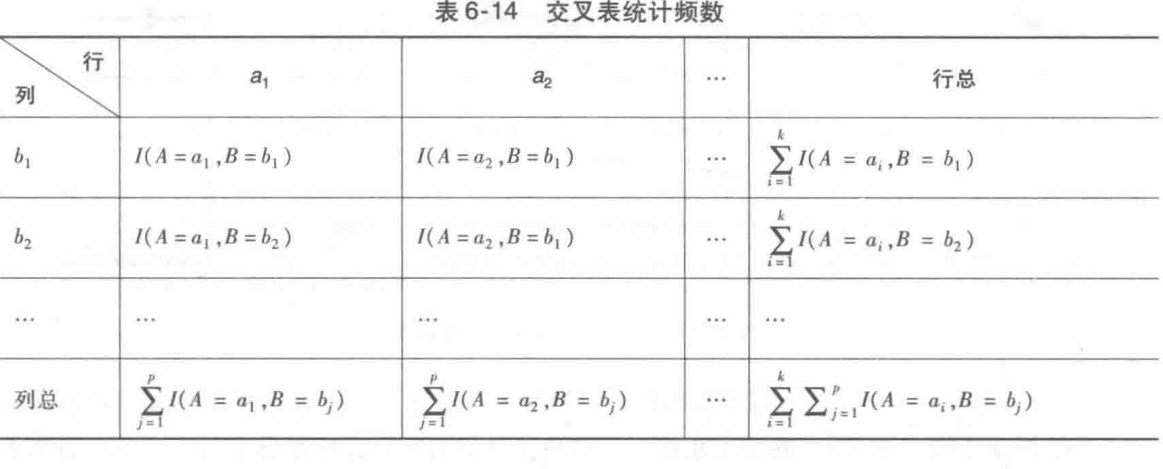
**卡方检验（二分类变量关系检验）**

若一个分类变量的分布随着另一个分类变量水平不同而发生变化时，则两个分类变量就有关系，反之则没有关系。



\* 列联表

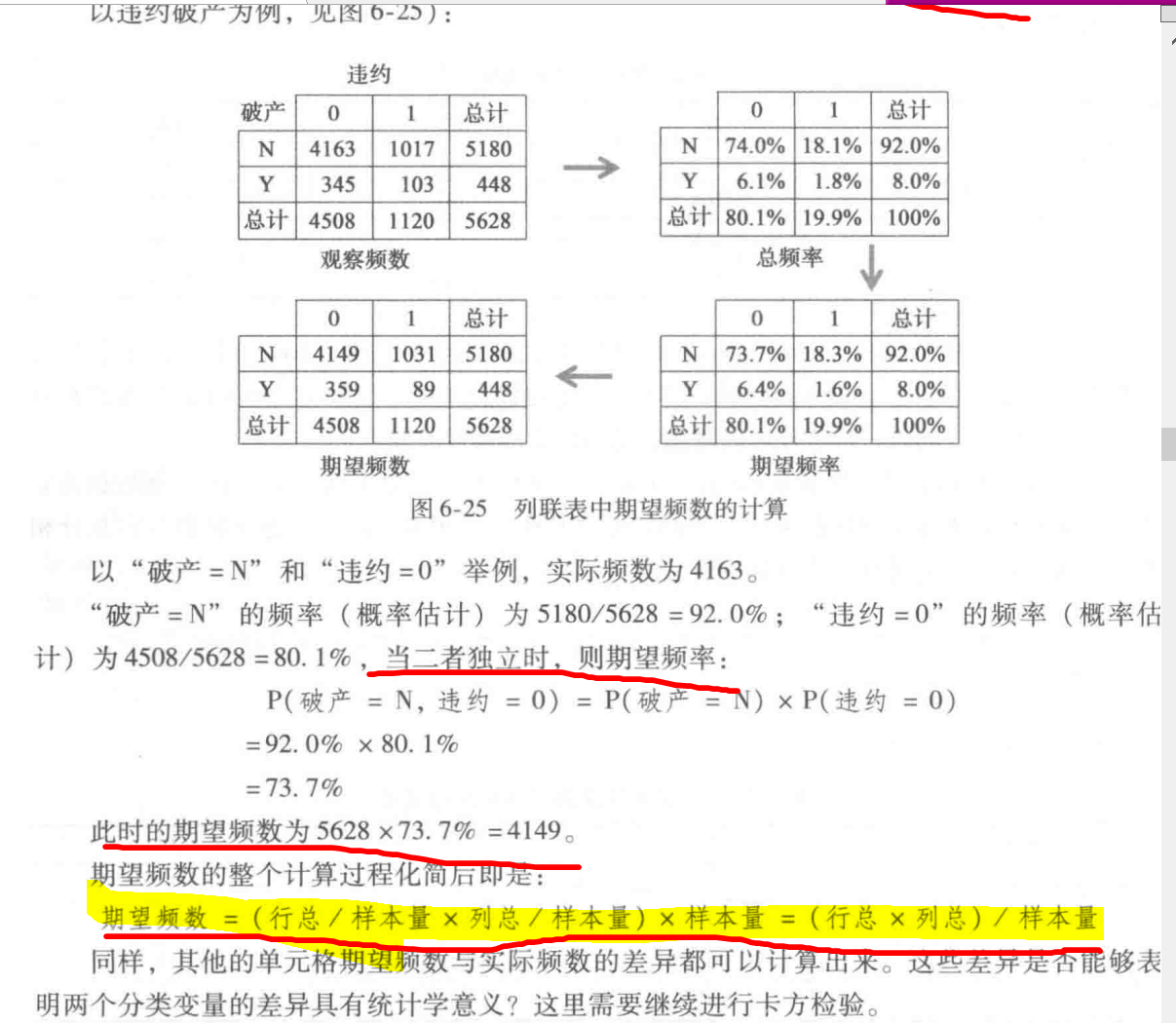
一种常见的分类汇总表，该表将待分析的两类变量中一个变量的每个类别设为列变量，另一个变量的每个类别设为行变量，中间对应不同变量不同类别下的频数



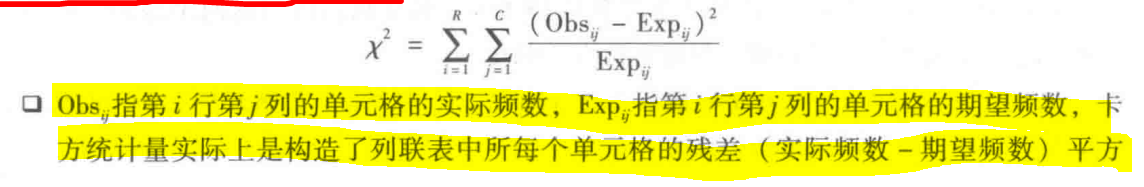
在生成列联表后，虽然能对比出差异，但这种差异是否有统计学意义就需要进行检验，检验方法为卡方检验，其检验统计量可以从列联表的频数计算出来。

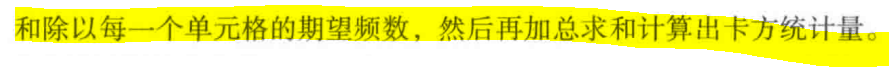
- 卡方检验

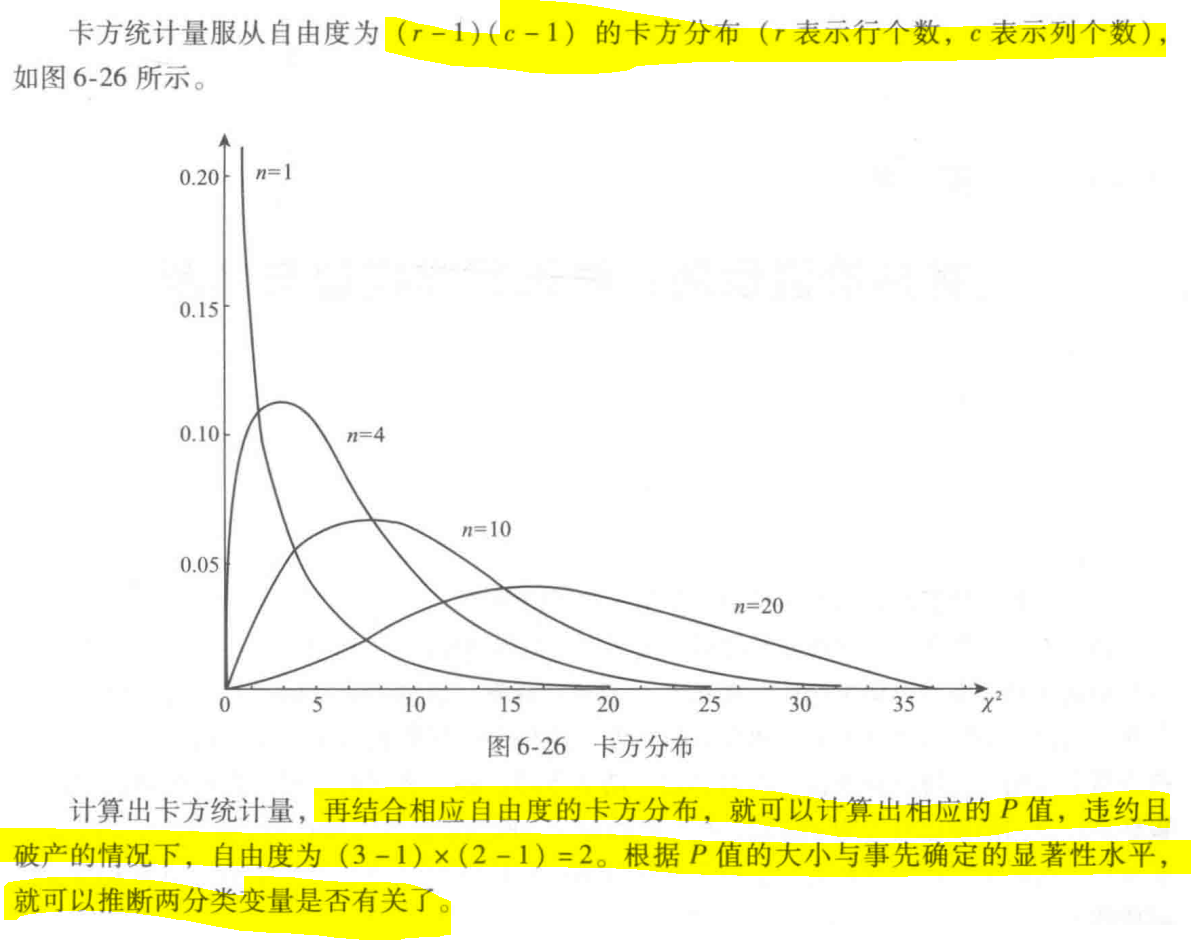
- 思想在于比较期望频数与实际频数的吻合程度



卡方检验原假设是期望频数等于实际频数，即两个分类变量无关，备择假设为期望频数不等于实际频数，即两个变量有关。



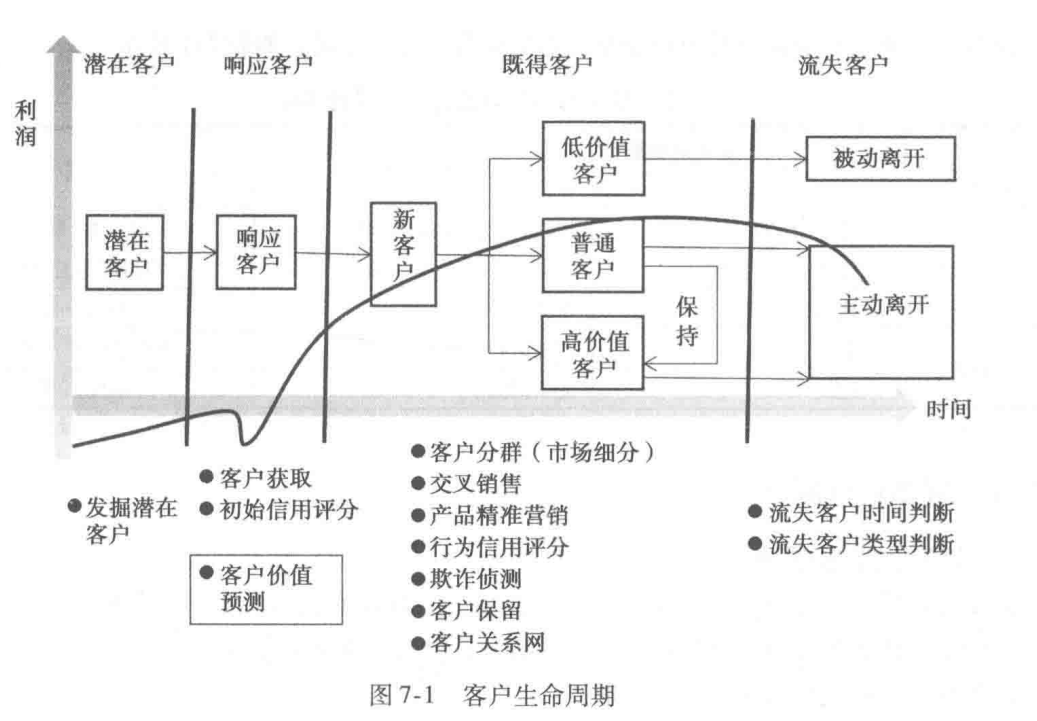




注：卡方检验不能展示两个分类变量相关性的强弱，只能展现两个分类变量是否相关。

**客户生命的四个周期**

1. 潜在客户：挖掘潜在客户位置
2. 响应客户：获得客户，并预测客户价值
3. 既得客户：客户分群，交叉销售，深入挖掘客户需求，精准营销
4. 流失客户：流失客户时间和类型判断



**线性回归**

在两个连续变量的相关性分析中，可用pearson相关系数衡量两连续变量相关行的强弱；若两连续变量的线性相关程度较强，可用线性回归进一步探讨两个变量的关系。

