-

常用变量及统计方法有哪些？

一、变量

了解以下几种变量有助于选择适当的统计分析方法作研究。

按作用划分

- **因变量：**如果一个变量由其他变量来描述，该变量为因变量或反应变量。
- **自变量：**如果一个变量与其他变量一起用于描述因变量， 该变量为自变量或预测变量。

自变量是“原因”，而因变量就是“结果”，自变量和因变量有时是相互转换的。

根据测量尺度划分： 根据变量测量精度不同，变量由低到高划分为

- **定类变量：**又称名义(nominal)变量。它的取值只代表观测对象的不同类别，例如“性别”变量、“职业”变量等都是定类变量。定类变量的取值称为定类数据或名义数据。定类数据的特点是用不多的名称来加以表达， 最常用来综合定类数据的统计量是频数、比率或百分比等。

- **定序变量：**又称为有序(ordinal)变量、顺序变量，它的取值的大小能够表示观测对象的某种顺序关系（等级、方位或大小等），也是基于“质”因素的变量。例如，“学历”变量的取值是： 1—小学及以下、 2—初中、 3—高中、中专、技校、4—大学专科、 5—大学本科、6—研究生以上。由小到大的取值能够代表学历由低到高。最适合用于综合定序数据取值的集中趋势的统计量是中位数。

- **定距变量 ：**又称为间隔(interval)变量，它的取值之间可以比较大小，可以用加减法 计算出差异的大小。例如，“年龄”变量，其取值 60 与 20 相比，表示 60 岁比 20 岁大， 并且可以计算出大 40 岁（60-20）。定距变量的取值称为定距数据或间隔数据。

定距数据是

一些真实的数值，具有公共的、不变的测定单位，可以进行加减乘除运算。

定距数据的基本特点是两个相同间隔的数值的差异相等。例如，年龄的 60 岁与 50 岁之差等于 40 岁与 30 岁之差。对于定距数据，不仅可以规定“等价关系”以及“大于关系” 和“小于关系”，而且也可以规定任意两个相同间隔的比值或差值。如果将每个数值分别乘 以一个正的常数再加上一个常数，即进行正线性变换， 并不影响定距数据原有的基本信息。

因此，常用的统计量如均值、标准差、相关系数等都可直接用于定距数据。

• **定比变量：**又称为比率(ratio)变量，是区别同一类别个案中等级次序及其距离的变量。定比变量除了具有定距变量的特性外， 还具有一个真正的零点， 因而它具有乘与除（×、÷）的数学特质。例如年零和收入这两个变量，固然是定距变量， 同时又是定比变量，因为其零点是绝对的，可以作乘除的运算。如 A 月收入是 60 元，而 B 是 30 元，我们可以算出前者是后者的两倍。智力商数这个变量是定距变量， 但不是定比变量， 因为其 0 分只具有相对的意义，不是绝对的或固定的， 不能说某人的智商是 0 分就是没有智力； 同时， 由于其零点是不固定的，即使 A 是 140 分而 B 是 70 分，我们也不能说前者的智力是后者的两倍，只能说两者相差 70 分。因为 0 值是不固定的，如果将其向上移高 20 分，则 A 的智商变为 120 分而 B 变成 50 分，两者的相差仍是 70 分，但 A 却是 B 的 2.4 倍，而不是原先的两倍了。定比变量是最高测量层次的变量。当前的社会学研究所应用的统计方法还很少要求达到定比变量这一测量层次。

一般定类变量和定序变量用于描述定性数据， 属于定性变量； 而定距变量和定比变量用于描述定量数据，属于定量变量。

同其他分类标准一样， 一个变量在不同分析中可当作不同尺度的变量。例如，“年龄”在某些分析中（如回归分析）当作定距变量，而在另外一些分析中（如方差分析）可通过分组作为定类变量处理。

二、常用统计方法

单变量频次分析：是针对单个变量按照变量取值类型来统计每个取值类型出现的次数。比如，需要分析交通肇事者的年龄分布情况，则可以针对交通肇事类案件犯罪人的年龄或者年龄段的频次分析。 以年龄段作为统计字段， 统计交通肇事类案件中不同犯罪年龄段的犯罪人人数各为多少， 也即十四岁以上不满十六岁、十六岁以上不满十八岁、十八岁以上不满二十五岁、二十五岁以上不满六十岁、六十岁以上各为多少人。

描述（演绎）性统计：是针对数值型变量（定距变量）进行简单的归纳分析，来描 述数据的状况，称之为描述性统计。 描述性统计通过求和、平均值、方差、标准差、众 数、中位数、标准误差、观测数、最大值、最小值、第 K 大值、第 K 小值、峰度、偏度、 区域（全距）、置信度等指标对数据的集中性、分散性、对称性、尖端性进行描述， 归 纳数据的统计特性。比如我们要针对法院在罚金方面的量刑进行分析， 则可对案件中的 罚金字段进行描述性分析， 统计所有案件中法院判处罚金的平均值、最大值、最小值等数据。

卡方分析：通过卡方检验，来分析变量不同取值是否对个数产生影响。比如是否因 为性别的不同会导致犯罪人人数不同， 比如，是否因为审级的不同是否会导致案件数的不同。

交叉分析： 分析两个分类变量之间是否独立的一种统计分析方法。 交叉分析是分析 两个分类变量之间是否独立的一种统计分析方法。 比如，可以分析某类案件未成年人与判决结果类型的关联性。交互分析这种方法固然好， 但是它存在一定的局限性。它最大的一个问题就是： 假定其他关系不存在， 只是一个自变量与一个因变量之间的交叉分析。比如它只分析婚姻是否成败和文化程度的关系， 但是我们都知道婚姻的成败与否不仅仅与文化的程度有关系， 它还和很多其他因素有关系。但是在交互分析中， 它假定排除其他所有因素的影响， 仅仅考虑文化教育程度与婚姻成败的关系， 所以这个分析过程和结论在一定意义上是失真的。因为它将其他的有关系的因素全都屏蔽， 这样就不能描述不同自变量作用的力度。比如婚姻的成败不仅仅与文化程度有关系， 还和地区、家庭等都有关系。当然是可以通过将婚姻的成败作为因变量， 分别与不同的各个自变量进行多次的交叉分析， 但是不能比较哪一个自变量的影响力度更大一些， 哪个自变量的解释力更强一些，这就是交互分析的局限性。总之，多个交互分析的结果不能直接拿来作比较。所以其分析结果的信息量还是很有限的。 不过，我们可以通过回归分析来弥补这种局限性。

回归分析：回归分析显示了多个自变量对一个因变量综合影响情况，并给出被选入 的多个自变量能够多大程度上解释因变量的变化， 每个自变量的回归系数是多少， 每个 自变量对因变量是否存在显著的影响， 影响的方向和大小是多少。比如，我们要分析被害人人身伤害类型等多个因素对刑期的综合影响情况， 即可针对这些字段进行多元线性回归分析。

方差分析：是指当需要检验某一个分类变量是否会因为不同的取值而导致一个数值 型的因变量的取值发生不同的变化时， 通过单因素方差分析来进行分类变量对因变量的影响判断。比如，要分析法院判决的刑期在多大程度上会受到罚金与否的影响， 则可针对“有期徒刑刑期”和“罚金”两个字段进行方差分析。

-

三、 变量与分析方法选用关系表

分析方法	变量类型		
	单一变量分析	**自变量（行变量）**	**因变量（列变量）**
单变量频次分析	1 个定类变量 （定序变量）	……	……
描述分析方法	1 个数值型变量 （定距变量）	……	……
卡方分析方法	1 个定类变量 （定序变量）	……	……
单因素方差分析方法	……	1 个定类变量 （定序变量）	数值型变量
交叉分析方法	……	1 个定类变量 （定序变量）	1 个定类变量 （定序变量）
回归分析方法	……	多个自变量	1 个因变量