名词解释

统计学是收集、处理、分析、解释数据并从数据中得出结论的科学。统计分析数据所用的方法大体上可以分为**描述性统计**与**推断统计**两大类。

描述统计是研究数据收集、处理和描述的统计学方法。其内容包括如何取得研究所需要的数据、如何用图表形式对数据进行处理和展示，如何通过对数据的汇总、概括与分析，得出所关心的数据特征。

推断统计是研究如何利用样本数据来推断总体特征的统计方法。其内容包括**参数估计**和**假设检验**两大类。参数估计是利用样本信息推断所关心的总体特征，假设检验则是利用样本信息判断对总体的某个假设是否成立。

观测数据是通过调查或观测而收集到的数据，这类数据是在没有对事物人为控制的条件下得到的，有关社会经济现象的统计数据几乎都是观测数据。

实验数据是在实验中控制实验对象而收集到的数据。自然科学领域的大多数数据都是实验数据。

截面数据是在相同或近似相同的时间点上收集的数据，这类数据通常是在不同的空间上获得的，用于描述现象在某一时刻的变化情况。

时间序列数据是在不同时间收集到的数据，这类数据是按时间顺序收集到的，用于描述现象随时间变化的情况。

总体是包含所研究的全部个体（数据）的集合。总体通常由所研究的一些个体组成，组成总体的每个元素称为**个体**。

样本是从总体中抽取的一部分元素的集合，构成样本的元素的数目称为**样本量**。抽样的目的是根据样本提供的信息推断总体的特征，根据样本统计量去估计总体参数。

参数是用来描述总体特征的概括性数字度量。它是研究者想要了解的总体的某种特征值，研究者所关心的参数通常有总体平均数、总体标准差、总体比例等。

统计量是用来描述样本特征的概括性数字度量。它是根据样本数据计算出来的一个量，由于抽样是随机的，因此统计量是样本的函数。研究者所关心的统计量主要有样本平均数、样本标准差、样本比例等。

变量是说明现象某种特征的概念，其特点是从一次观察到下一次观察结果会呈现出差别或变化。（变量的具体取值称为**变量值**。）

分类变量是说明事物类别的一个名称，其取值是分类数据。

顺序变量是说明事物有序类别的一个名称，其取值是顺序数据。

数值型变量是说明事物数字特征的一个名称，其取值是数值型数据。

数值型变量根据其取值不同可分为：

离散型变量是只能取可数值的变量，只能取有限个值，且取值都以整位数断开，可以一一列举。

连续型变量是可以在一个或多个区间中取任何值的变量，其取值是连续不断的，不能一一列举。

定序尺度：定序尺度又称顺序尺度，它是对事物之间等级差或顺序差别的一种测度。该尺度不仅可以将事物分成不同的类别，而且还可以确定这些类别的优劣或顺序。或者说，它不仅可以测度类别差，还可以测度次序差。定序尺度的计量结果虽然也表现为类别，但这些类别之间是可以比较顺序的。

定距尺度：定距尺度又称间隔尺度，它不仅能将事物区分为不同类型并进行排序，而且可以准确地指出类别之间的差距是多少。定距尺度是对事物类别或次序之间间距的测度，该尺度通常使用自然或物理单位作为计量尺度，如收入用元、考试成绩用分、温度用度、重量用克、长度用米等等。因此，定距尺度的计量结果表现为数值。由于这种尺度的每一间隔都是相等的，只要给出一个度量单位，就可以准确地指出两个计数之间的差值。

概率抽样也称随机抽样，是指遵循随机原则进行的抽样，总体中每个单位都有一定的机会被选入样本。

简单随机抽样就是从包括总体 N 个单位的抽样框中随机地、一个一个的抽取 n 个单位作为样本，每个单位的入样概率是相等的。

分层抽样是将抽样单位按某种特征或某种规则划分为不同的层，然后从不同的层中独立、随机地抽取样本。将各层样本结合起来，对总体的目标量进行估计。

整群抽样是将总体中若干个单位合并为组，这样的组称为群，抽样时直接抽取群，然后对中选群中的所有单位全部实施调查。

系统抽样将总体中的所有单位（抽样单位）按一定顺序排列，在规定的范围内随机地抽取一个单位作为初始单位，然后按事先规定好的规则确定其他样本单位。

等距抽样也称系统抽样，是将总体中的所有单位（抽样单位）按一定顺序排列，在规定的范围内随机地抽取一个单位作为初始单位，然后按固定的间隔来确定其他样本单位的抽样方法。

多阶段抽样是采用类似整群抽样的方法，首先抽取群，但并不是调查群内的所有单位，而是再进一步抽样，从选中的群中抽取出若干个单位进行调查。因为取得这些接受调查的单位需要两个步骤，所以将这种抽样方式称为二阶段抽样。在这里，群是初级抽样单位，第二阶段抽取的是最终抽样单位。将这种方法推广，使抽样的段数增多，就称为多阶段抽样。

非概率抽样是相对于概率抽样而言的，指抽取样本时不是依据随机原则，二是根据研究目的对数据的要求，采用某种方式从总体中抽取部分单位对其实施调查。

方便抽样是调查过程中由调查员依据方便的原则，自行确定入抽样本的单位。方便抽样的最大特点是容易实施，调查的成本低，但这种抽样方式也有明显的弱点。在科学研究中，使用方便样本可以产生一些想法以及对研究内容的初步认识，或建立假设。

判断抽样是指研究人员根据经验、判断和对研究对象的了解，有目的地选择一些单位作为样本，实施时根据不同的目的有重点抽样、典型抽样、代表抽样等方式。

重点抽样是从调查对象的全部单位中选择少数重点单位，对其实施调查。

典型抽样是从总体中选择若干个典型的单位进行深入调研，目的是通过典型单位来描述或揭示所研究问题的本质和规律。

代表抽样是通过分析，选择具有代表性的单位作为样本，在某种程度上，也具有典型抽样的含义。

自愿样本指被调查者自愿参加，成为样本中的一份子，向调查人员提供有关信息。自愿样本与抽样的随机性无关，样本的组成往往集中于某类特定的人群，尤其集中于对该调查活动感兴趣的人群，因此，这种样本是有偏的。我们不能根据样本的信息对总体的状况进行估计，但自愿样本仍可以给研究人员提供许多有价值的信息，他可以反映某类群体的一般看法。

滚雪球抽样往往用于对稀少群体的调查。在滚雪球抽样中，首先选择一组调查单位，对其实施调查之后，再请他们提供另外一些属于研究总体的调查对象，调查人员根据所提供的线索，进行此后的调查。这个过程持续下去，就会形成滚雪球效应。

配额抽样类似于概率抽样中的分层抽样，首先将总体中的所有单位按一定的标志（变量）分为若干类，然后在每个类中采用方便抽样或判断抽样的方式选取样本单位。

检验统计量：检验统计量是用于假设检验计算的统计量。在零假设情况下，这项统计量服从一个给定的概率分布，而这在另一种假设下则不然。从而若检验统计量的值落在上述分布的临界值之外，则可认为前述零假设未必正确。

相关系数：相关系数是衡量两个随机变量之闻线性关系强度的指标，取值范围为[-1,1]

KMO 测度：KMO 测度用于探查变量间的偏相关性，比较各变量间的简单相关和偏相关的大小，取值范围在 0~1 之间。

移动平均法：移动平均法是通过对时间序列逐期递移，求得平均数作为预测值的一种预测方法，其方法有简单移动平均法和加权移动平均法两种。

自填式是指在没有调查员协助的情况下由被调查者自己填写，完成调查问卷。

面访式是指现场调查员与被调查者面对面，调查员提问。被调查者回答这种调查方式。

电话式是指调查人员通过打电话的方式向被调查者实施调查。

观察式是指调查人员通过直接观测的方法获取信息。

数据的误差是指通过调查搜集到的数据与研究对象真实结果之间的差异。数据的误差有两类：**抽样误差**与**非抽样误差**。

抽样误差是由抽样的随机性引起的样本与总体真值之间的误差。抽样误差并不是针对某个具体样本的检测结果与总体真实结果的差异而言的，抽样误差描述的是所有样本可能的结果与总体真值之间的平均性差异。

非抽样误差是相对于抽样误差而言，是除抽样误差之外的，由于其他原因引起的样本观察结果与总体真值之间的差异。

抽样框误差：统计推断的错误是由于抽样框不完善造成的，我们把这种误差称为**抽样框误差**。

回答误差是指被调查者在接受调查时给出的回答与真实情况不符。回答误差包括：理解误差；记忆误差；有意识误差。

无回答误差是指被调查者拒绝接受调查，调查人员得到的是一份空白的答卷。无回答误差可以是随机性的，也可以是系统性的。

数据的预处理是在对数据分类或分组之前所做的必要处理，内容包括数据的审核、筛选、排序等。

数据审核是检查数据中是否有错误。对于通过调查取得的原始数据，主要从完整性和准确性两个方面去审核。对于二手数据，应着重审核数据的适用性和时效性。

完整性审核：主要是检查应调查的单位或个体是否有遗漏，所有的调查项目是否填写齐全等。

准确性审核：主要是检查数据是否有错误，是否存在异常值等。

适用性审核：主要是弄清楚数据的来源、数据的口径以及有关的背景材料，以便确定这些数据是否符合分析研究的需要，不能盲目生搬硬套。

时效性审核：是指要保证数据的及时性，如果取得的数据过于滞后，就可能失去研究的意义。

数据筛选是根据需要找出符合特定条件的某类数据。比如，找出销售额在 1000 万元以上的企业；找出考试成绩在 90 分以上的学生。

数据排序是按一定顺序将数据排列，以便研究者通过浏览数据发现一些明显的特征或趋势，找到解决问题的线索。排序还有助于对数据检查纠错，以及为重新归类或分组提供方便。

频数：落在某一特定类别或组中的数据个数。

频数分布：把各类别及落在其中的相应频数全部列出，并用表格的形式表现出来。

抽样分布：从已知的总体中以一定的样本容量进行随机抽样，由样本的统计量所对应的概率分布称为抽样分布。

列联表：由两个或两个以上的变量交叉分类的频数分布表。

交叉表：二维的列联表（两个变量交叉分类）。

比例：也称构成比，它是一个样本或总体中各个部分的数据与全部数据之比，通常用于

反映样本或总体的构成或结构。

百分比：将比例乘以 100 得到的数值，用%表示。

比率是样本或总体中不同类别数据之间的比值，由于比率不是部分与整体之间的对比关系，因此比值可能大于 1。

条形图是用宽度相同的条形的高度或长短来表示数据多少的图形。条形图可以横置也可以纵置，纵置时也称为柱形图。

帕累托图是按各类别数据出现的频数多少排序后绘制的条形图。

饼图是用圆形及圆内扇形的角度来表示数值大小的图形，他主要用于表示一个样本或总体中各组成部分的数据占全部数据的比例，对于研究结构性问题十分有用。

环形图是将饼图叠在一起，挖去中间部分得到的图形。

累计频数是将各有序类别或组的频数逐级累加起来得到的频数，有向上累积和向下累计两种方式。

累计频率也叫累积百分比，是将各有序类别或组的百分比逐级累加起来，他有向上累积和向下累计两种方式。

数据分组：根据统计研究的需要，将原始数据按照某种标准分成不同的组别。分组后的数据称为**分组数据**。

单变量值分组：把每一个变量值作为一组；只适用于变量值较少的离散型变量。

组距分组：将全部变量值一次划分为若干区间，将一个区间的变量值作为一组；适用于变量值较多或连续型变量。在组距分组中，一个组的最小值称为**下限**，最大值称为**上限**；**组距为**一组上限与下限之差。组距相等称为**等距分组**，组距不等称为**不等距分组**。

组中值是每一组中下限值与上限值中间的值，即组中值＝（下限值＋上限值）/2,反映各组数

据的一般水平。（使用组中值代表一组数据时的必要假定条件：各组数据在本组内呈均匀分布或在组中值两侧呈对称分布。）

集中趋势是指一组数据向某一中心值靠拢的程度，它反映了一组数据中心点的位置所在。

众数是一组数据中出现次数最多的变量值，用M_0表示。众数主要用于测度分类数据的集中趋

势，也适用于顺序数据和数值型数据；在数据量较大的情况下，众数才有意义。众数是一个位置代表值，它不受数据中极端值的影响，是具有明显集中趋势点的数值，是一组数据分布的最高峰点所对应的数值；一组数据中众数可能不存在，也可能有两个或多个众数。

中位数：一组数据排序后处于中间位置上的变量值，用M_e表示；中位数主要用于测度顺序数据的集中趋势，也适用于数值型数据，但不适用于分类数据；它是一个位置代表值，不受数据中极端值的影响。中位数位置的确定公式为：

$$\text{中位数位置}=\frac{n+1}{2}$$

式中，n 为数据个数。

四分位数：也称四分位点，是一组数据排序后处于 25%和 75%位置上的值。四分位数是通过 3 个点将全部数据等分为 4 部分，其中每部分包含 25%的数据。很显然，中间的四分位数就是中位数。在 25%的位置上的数值称为**下四分位数**，在 75%位置上的数值称为**上四分位数**。

平均数：也称为均值，是一组数据相加后除以数据的个数得到的结果，是集中趋势的最主要测度值，平均数主要适用于数值型数据，不适用于分类数据和顺序数据。

数据的离散程度是数据分布的另一个特征，它反映的是各变量值远离其中心值的程度。数据的离散程度越大，集中趋势的测度值对该组数据的代表性越差；离散程度越小，其代表性越好。

异众比率：指非众数组的频数占总频数的比例，用V_r表示，其计算公式为：

$$V_r=\frac{\sum f_i-f_m}{\sum f_i}=1-\frac{f_m}{\sum f_i}$$

式中，$\sum f_i$为变量值的总频数，f_m为众数组的频数。异众比率主要用于衡量众数对一组数据的代表程度，主要适用于测度分类数据的离散程度，也适用于顺序数据和数值型数据。异众比率越大，说明非众数组的频数占总频数的比重越大，众数的代表性越差；异众比率越小，说明非众数组的频数占总频数的比重越小，众数的代表性越好。

四分位差：也称为内距或四分间距，是上四分位数与下四分位数之差，用Q_d表示。其计算公式为：

$$Q_d=Q_U-Q_L$$

四分位差反映了中间 50%数据的离散程度，其数据越小，说明中间的数据越集中；其数据越大，说明中间的数据越分散；四分位差不受极值影响。主要适用于测度顺序数据的离散程度，也适用于数值型数据，不适用于分类数据。

极差：也称全距，一组数据的最大值与最小值之差，用 R 表示。其计算公式为：

$$R=\max(x_i)-\min(x_i)$$

式中，$\max(x_i)$和 $\min(x_i)$分别表示一组数据的最大值与最小值。极差是描述数据离散程度的最简单的测度，计算简单，易于理解，但它容易受极端值的影响。由于极差只是利用了一组数据两端的信息，不能反映出中间数据的分散状况，不能准确的描述数据的分散程度。

平均差：也称平均绝对离差，是各变量值与其平均数离差绝对值的平均数，用M_d表示。平均差以平均数为中心，反映了每个数据与平均数的平均差异程度，能全面准确地反映一组数据的离散状况；平均差越大，说明数据的离散程度越大；反之，则说明数据的离散程度越小。

方差：各变量值与其平均数离差平方的平均数，方差的平方根称为**标准差**。样本方差用 s^2 表示。方差开方后即得到**标准差**。与方差不同的是，标准差是具有量纲的，它与变量值的计量单位相同，其实际意义要比方差清楚。在对实际问题进行分析时更多的使用标准差。

标准分数是变量值与其平均数的离差除以标准差后的值。

离散系数：也称为变异系数，是一组数据的标准差与其相应的平均数之比。其计算公式为

$$v_s = \frac{s}{\bar{x}}$$

离散系数是测度数据离散程度的相对统计量，主要用于比较不同样本数据的离散程度。离散系数越大，说明数据的离散程度也大；离散系数越小，说明数据的离散程度也小。

偏态：对数据分布对称性的测度，测度偏态的统计量是**偏态系数**，记作 SK。如果一组数据的分布是对称的，则偏态系数等于 0；如果偏态系数明显不等于 0，表明分布是非对称的。若偏态系数大于 1 或小于-1，称为高度偏态分布；若偏态系数在 0.5~1 或-1~-0.5 之间，称为中等偏态分布；偏态系数越接近 0，偏斜程度就越低。分布对称时，SK=0；当 SK 为正值时，表示正离差值较大，可判断为正偏或右偏；当 SK 为负值时，表示负离差值较大，可判断为负偏或左偏；SK 的数值越大，表示偏斜的程度越大。

峰态：对数据分布平峰或尖峰程度的测度，测度峰态的统计量是**峰态系数**，记作 K。其计算公式如下：如果一组数据服从标准正态分布，则峰态系数等于 0；如果峰态系数明显不等于 0，表明分布比正态分布更平或更尖，通常称为平峰分布或尖峰分布。由于正态分布的峰态系数为 0，当 K>0 时为尖峰分布，数据的分布更集中；当 K<0 时为扁平分布，数据的分布越分散。

事件 A 的概率：描述的是事件 A 在试验中出现的可能性大小的一种度量，可能性数值记为 P(A)。

条件概率：当某一事件 B 已经发生时，求事件 A 发生的概率，称这种概率为事件 B 发生条件下事件 A 发生的条件概率，记为 P(A|B)，一般来说，P(A|B)≠P(A)。

概率函数：在同一组条件下，如果每次试验可能出现这样或那样的结果，并且把所有的结果都能列举出来，即把 X 的所有可能值x_1，x_2,⋯,x_n都能列举出来，而 X 的可能值x_1，x_2,⋯,x_n，具有确定概率 P(x_1),P(x_2),⋯,$P(x_n)$，其中 P(x_i)=P(X=x_i)，称为**概率函数**，则 X 称为 P(X)的随机变量，P(X) 称为随机变量 X 的概率函数。

概率密度函数：由于连续型随机变量可以取某一区间或整个实数上的任意一个值，所以我们不能像对离散型随机变量那样，列出每一个值及其相应的概率，而必须用其他的方法，通常用数学函数的形式和分布函数的形式来描述。当用函数 f(x)来表示连续型随机变量时，我们将 f(x)称为**概率密度函数**。

统计量：设 X_1，X_2，⋯，X_n 是从总体 X 中抽取的容量为 n 的一个样本，如果由此样本

构造一个函数 $T(X_1, X_2, \cdots, X_n)$，不依赖于任何未知参数，则称函数 $T(X_1, X_2, \cdots, X_n)$ 是一个统计量。通常，又称 $T(X_1, X_2, \cdots, X_n)$为**样本统计量**。

次序统计量：设 $X_1, X_2, \cdots, X_n$ 是从总体 X 中抽取的容量为 n 的一个样本，$X_{(i)}$称为第 i 个次序统计量，它是样本$(X_1, X_2, \cdots, X_n)$满足如下条件的函数：每当样本得到一组观测值$x_1, x_2, \cdots, x_n$时，其由小到大的排序$x_{(1)} \le x_{(2)} \le \cdots \le x_{(i)} \le \cdots \le x_{(n)}$中，第 i 个值$x_{(i)}$就作为次序统计量$X_{(i)}$的观测值，而$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$称为次序统计量。其中，$X_{(1)}$和$X_{(n)}$分别为最小和最大次序统计量。

充分统计量：在统计学中，假如一个统计量能把含有样本中有关总体的信息一点都不损失地提取出来，那对保证后边的统计推断质量具有重要意义。统计量加工过程中一点信息都不损失的统计量称为充分统计量。

χ^2分布：设随机变量$X_1, X_2, \cdots, X_n$相互独立，且$X_i(i = 1,2,\cdots,n)$服从标准正态分布 N(0,1)则他们的平方和$\sum_{i=1}^{n} X_i^2$服从自由度为 n 的χ^2分布。

t 分布：设随机变量 X~N（0,1），Y~$\chi^2(n)$，且 X 与 Y 独立，则 $t = \frac{X}{\sqrt{Y/n}}$，其分布称为 t 分布，记为 t(n)，其中，n 为自由度。

F 分布：设随机变量 Y 与 Z 相互独立，且 Y 和 Z 分别服从自由度为 m 和 n 的χ^2分布，随机变量 X 有如下表达式：

$$X = \frac{Y/\mathrm{m}}{Z/n} = \frac{nY}{mZ}$$

则称 X 服从第一自由度为 m，第二自由度为 n 的 F 分布，记为 F(m,n)，简记为 X~F(m,n)。

中心极限定理：设从均值为μ、方差为σ^2（有限）的任意一个总体中抽取样本量为 n 的样本，当充分大时，样本均值的抽样分布近似服从均值为μ、方差为σ^2/n 的正态分布。中心极限定理要求 n 必须充分大，在实际生活中，总体分布未知的情况下，我们要求 n≥30 为大样本，n<30 为小样本是经验说法。

参数估计：用样本统计量去估计总体的参数。比如用样本均值$\bar{x}$估计总体均值μ，用样本比例 p 估计总体比例π，用样本方差s^2估计总体方差σ^2。如果将总体参数笼统地用一个符号θ来表示，而用于估计总体参数的统计量用$\hat{\theta}$表示，参数估计就是如何用$\hat{\theta}$来估计θ。

估计量：在参数估计中，用来估计总体参数的统计量称为估计量，用符号$\hat{\theta}$来表示。

估计值：根据一个具体的样本计算出来的估计量的数值称为估计值。

点估计：用样本统计量$\hat{\theta}$的某个取值直接作为总体参数θ的估计值。比如用样本均值$\bar{x}$直接作为总体均值μ的估计值。

区间估计：在点估计的基础上，给出总体参数估计的一个区间范围，该区间由样本统计

量加减估计误差而得到。与点估计不同，进行区间估计时，根据样本统计量的抽样分布能够对样本统计量与总体参数的接近程度给出一个概率度量。

置信区间：由样本统计量所构造的总体参数的估计区间，其中区间的最小值为**置信下限**，最大值为**置信上限**。

置信水平：一般地，如果将构造置信区间的步骤重复很多次，置信区间中包含总体参数真值的次数所占的比例称为置信水平，也称**置信度**或**置信系数**，表示为 1-α（α是事先确定的一个概率值，也称**风险值**，是总体参数未在区间内的概率）。

无偏性：是指估计量抽样分布的数学期望等于被估计的总体参数。设总体参数为θ，所选择的估计量为$\hat{\theta}$，如果 $E(\hat{\theta}) = \theta$，则称$\hat{\theta}$为θ的无偏估计量。

有效性：指对同一总体参数的两个无偏估计量，有更小标准差的估计量更有效，即在无偏估计的条件下，估计量的方差越小，估计越有效。

一致性：随着样本量的增大，估计量的值越来越接近被估计的总体参数，即一个大样本的给出的估计量要比一个小样本给出的估计量更接近总体参数。

独立样本：如果两个样本是从两个总体中独立抽取的，即一个样本中的元素与另一个样本中的元素相互独立，称为独立样本。

匹配样本：即一个样本中的数据与另一个样本中的数据相对应。匹配样本可以消除由于样本指定的不公平造成的两种方法组装时间上的差异。

假设检验：事先对总体参数或分布形式做出某种假设，然后利用样本信息来判断原假设是否成立；假设检验分为参数假设检验和非参数假设检验。

原假设：待检验的假设，又称“零假设”，用H_0表示（研究者想收集证据予以反对的假设）。

备择假设：如果原假设不成立，就要拒绝原假设，而需要在另一个假设中做出选择，这个假设称为备择假设，表示为H_1（研究者想收集证据予以支持的假设总是有不等号）。

弃真错误：第Ⅰ错误是原假设 H_0 为真却被我们拒绝了，犯这种错误的概率用 α 来表示，所以也称 α 错误或者弃真错误。

取伪错误：第Ⅱ类错误是原假设为伪我们却没有拒绝，犯这种错误的概率用 β 表示，所以也称 β 错误或取伪错误。

P 值：当原假设为真时所得到的样本观测结果或更极端的结果出现的概率。如果 P 值很小，说明这种情况发生的概率很小，而如果出现了，根据小概率原理，我们就有理由拒绝原假设，P 值越小，我们拒绝原假设的理由就越充分。

拟合优度检验：是用χ^2统计量进行统计显著性检验的重要内容之一。它是依据总体分布状况，计算出分类变量中各类别的期望频数，与分布的观察频数进行对比，判断期望频数与观察频数是否有显著性差异，从而达到对分类变量进行分析的目的。

列联表：是由两个以上的变量进行交叉分类的频数分布表。

独立性检验：就是分析列联表中行变量和列变量是否相互独立。

方差分析：通过检验各总体的均值是否相等来判断分类型自变量对数值型因变量是否有显著影响。

组内误差：由于抽样的随机性所造成的随机误差，即来自水平内部的数据误差，反映一个样本内部数据的离散程度，只含有随机误差。

组间误差：来自不同水平之间的数据误差，这种误差可能是由抽样本身形成的随机误差，也可能是由行业本身的系统性因素造成的系统误差，因此，组间误差是随机误差和系统误差的总和，反映不同样本之间数据的离散程度。

随机误差：因素的同一水平（总体）下，样本各观察值之间的差异，由样本本身形成的。

系统误差：因素的不同水平（不同总体)之间观察值的差异，由于行业本身的系统性因素所造成的。

总平方和（SST）：反映全部数据误差大小的平方和，反映全部观测值的离散状况。

组内平方和（SSE）：反映组内误差大小的平方和，也称误差平方和或残差平方和，反映每个样本内各观测值的离散状况。

组间平方和（SSA）：反映组间误差大小的平方和，也称因素平方和，反映样本均值之间的差异程度。

相关关系：变量之间存在不确定的数量关系，称为**相关关系**。

最小平方法:对于第 i 个 x 值，估计的回归方程可表示为：

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

最小平方法也称最小二乘法，通过使因变量的观测值y_i与估计值$\widehat{y_i}$之间的离差平方和达到最小来估计β_0和β_1的方法。

拟合优度：回归直线与各观测点的接近程度。

判定系数：回归平方和占总平方和的比例称为判定系数，记为 R^2。

估计标准误差：度量各实际观测点在直线周围的散布状况的一个统计量，它是均方残差的平方根，用 s_e 来表示。

残差：因变量的观测值y_i与根据估计的回归方程求出的预测值$\widehat{y_i}$之差，用 e 表示，反映了用估计的回归方程去预测y_i而引起的误差。第 i 个观察值的残差可以写为：

$$e_i = y_i - \widehat{y_i}$$

标准化残差：残差除以它的标准差后得到的数值，也称 Pearson 残差或半学生化残差，用 z_e 来表示。第 i 个观察值的标准化残差可以表示为：

$$z_{e_i} = \frac{e_i}{s_e} = \frac{y_i - \hat{y}_i}{s_e}$$

式中，s_e 是残差的标准差的估计。

多重判定系数：多元回归中的回归平方和占总平方和的比例，它是度量多元回归方程拟合程度的一个统计量，反映了在因变量 y 的变差中被估计的回归方程所解释的比例。

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

调整的多重判定系数：用样本量 n 和自变量的个数 k 进行调整的多重判定系数。记为 R_a^2

$$R_a^2 = 1 - \left(1 - R^2\right)\left(\frac{n - 1}{n - k - 1}\right)$$

多重共线性：指当回归模型中两个或两个以上的自变量彼此相关时，则称回归模型中存在**多重共线性**。

容忍度：某个自变量的容忍度等于 1 减去该自变量为因变量而其他 k-1 个自变量为预测变量时所得到的线性回归模型的判定系数，即 $1 - R_i^2$。容忍度越小，多重共线性越严重。通常认为容忍度小于 0.1 时，存在严重的多重共线性。

方差扩大因子：方差扩大因子等于容忍度的倒数，即 $VIF = \frac{1}{1-R_i^2}$。显然，VIF 越大多重共线性就越严重。一般认为 VIF 大于 10 则认为存在严重的多重共线性。

平稳序列：基本上不存在趋势的序列，各观察值基本上在某个固定的水平上波动，虽然在不同的时间段波动程度不同，但并不存在某种规律，其波动可以看成是随机的。

非平稳序列：包含趋势、季节性或周期性的序列，可能只含有一种成分，也可能是几种成分的组合，可分为有趋势的序列、有趋势和季节性的序列、几种成分混合而成的复合型序列。

增长率：也称为增长速度，它是时间序列中报告期观察值与基期观察值之比减 1 后的结果，用%表示。

平均增长率：也称平均增长速度，它是时间序列中逐期环比值的几何平均数减 1 后的结果。

移动平均法是通过对时间序列逐期递移求得平均数作为预测值的一种预测方法。

指数平滑法是通过对过去的观察值加权平均进行预测的一种方法，该方法使得第 t+1 期的预测值等于 t 期的实际观察值与 t 期的预测值的加权平均值。

线性趋势是指现象随着时间的推移而呈现出稳定增长或下降的线性变化规律。

指数：测定多项内容数量综合变动的相对数。这个概念包含两个要点：

①指数的实质是测定多项内容，例如零售价格指数反映的是零售市场几百万种商品价格变化的整体状况。

②指数的表现形式为动态相对数。

季节指数：季节指数刻画了序列在一个年度内各月或各季度的典型季节特征。在乘法模

型中，季节指数是以其平均数等于 100%为条件而构成的，它反映了某一月份或季度的数值占全年平均数值的大小。如果现象的发展没有季节变动，则各期的季节指数应等于 100%；如果某一月份或季度有明显的季节变化，则各期的季节指数应大于或小于 100%。因此，季节变动的程度是根据各季节指数与其平均数（100%）的偏差程度来测定的。

必然事件：在同一组条件下，每次试验一定出现的事件。

不可能事件：在同一组条件下，每次试验一定不出现的事件。

处理：不同的因子水平。

因子：检验的对象，所研究的分类型变量的另一个名称。

交互作用：一个因素和另一个因素联合产生的对因变量的附加效应。

期望值：随机变量 X 的平均取值。

随机现象：在一定条件下，并不总出现相同结果的现象。

随机试验：可重复的随机现象称为随机试验，简称试验。

基本结果ω：随机现象的最简单的结果，它是统计中抽样的基本单元，故又称样本点。

基本空间Ω：随机现象所有基本结果的全体称为这个随机现象的基本空间，又称样本空间。

随机事件：随机现象的某些基本结果的集合，简称事件，常用大写字母 A，B，C 来表示。

贝努里试验：只有两个结果的试验称为贝努里试验。

n 重贝努里试验：由 n 个（次）相同的、独立的贝努里试验组成的随机试验称为 n 重贝努里试验。

条件概率：事件 B（P（B）＞0）已发生下，事件 A 的条件概率

$$P(A/B) = \frac{P(AB)}{P(B)}$$

随机变量 X：取值依赖于（随机现象的）基本结果ω的变量。或者说，随机变量 X 是定义在样本空间Ω={ω}上的实值函数，即

$$X = X(\omega) \quad , \ \omega \in \Omega$$

离散随机变量：仅取有限个或可列个孤立点的随机变量。

连续随机变量：取值充满一个区间（a，b）的随机变量。

分布函数：对随机变量 X 及任意实数 x，事件“X≤x”的概率 F（x）=P（X≤x）是 x 的函数，称它为 X 的累计概率分布函数，简称分布函数。

切比晓夫不等式：对方差存在的随机变量 X，有

$$P\left(\left|X - E(X)\right| \geq \varepsilon\right) \leq \frac{Var(X)}{\varepsilon^2}$$

其中，ε 为任一正实数，这表明，大偏差发生概率被其方差所控制。

贝努里大数定律：设 X_n ～b（n，p），则频率 X_n/n 与概率 p 之间的偏差发生的概率将随着 n 的无限增大而趋于零，即

$$\lim_{n\to\infty} P\left(\left|\frac{X_n}{n} - p\right| \geq \varepsilon\right) = 0$$

X 的变异系数：它是以数学期望为单位去度量随机变量 X 取值波动程度的特征数。

$$C_v = \frac{\sigma(X)}{E(X)}$$

卷积公式：寻求独立随机变量和的分布的运算称为卷积运算，用*表示，相应的公式称为卷积公式。

相关系数：设（X,Y）为二维随机变量，它的两个方差 σ_x^2 和 σ_Y^2 都存在，则称 $Cov(X,Y)/\sigma_X\sigma_Y$ 为 X 与 Y 的线性相关系数，简称 X 与 Y 的相关系数，记为

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

它与协方差的符号相同，若 Corr（X,Y）＞0，称 X 与 Y 之间为正相关；若 Corr（X,Y）＜0，称 X 与 Y 之间为负相关；若 Corr（X,Y）=0，称 X 与 Y 不相关。

简单随机样本：由简单随机抽样所得到的样本称为简单随机样本。

统计量：不含任何未知参数的样本的函数称为统计量。

抽样分布：统计量的分布称为抽样分布。

先验分布：对先验信息进行加工获得的分布称为先验分布。

检验统计量：要判断原假设是否为真，需要构造一个统计量，并用该统计量的分布来进行判断，则该统计量称为检验统计量。

原假设：在统计中常把要检验的假设称为原假设，记为 H_0。

备择假设：它是在 H_0 被拒绝时所接受的假设，记为 H_1。

p 值：在一个假设检验问题中，拒绝原假设的最小显著性水平称为 p 值。

正态性检验：用于判断总体分布是否为正态分布的检验称为正态性检验。

方差分析：检验具有相同方差的若干个正态总体的均值是否相等的一种假设检验方法。

方差齐性检验：设有 r 个正态总体 $N(\mu_i, \sigma_i^2)$，i =1,2，…，r，从第 i 个总体中抽取了容量为 m_i 的样本 y_{i1}，y_{i2}，…，y_{imi}，样本均值为 $\overline{y_i}$，样本的无偏方差为 s_i^2。要检验的假设是：

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_r^2$$

这类检验称为方差齐性检验。