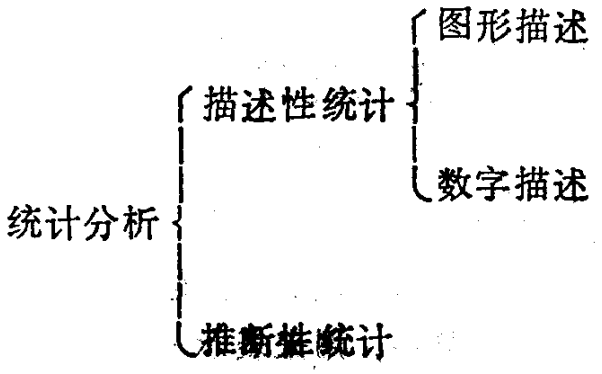
第九章　资料的统计分析

在社会调查中，运用统计的方法进行资料的定量分析和研究，是提高调查质量的必要手段。定量分析的方法是以数理法则的具体测量、计算及分析技术为基础的，它是社会调查测量化发展的产物。应用各种统计分析方法，可以为研究社会诸要素相互作用的复杂关系，提供精确度较高的数据，为更准确地认识社会现象，进行各种比较研究和预测分析提供条件。

统计分析通常可以分成两大类：一类是描述性统计，它主要是对资料进行图形描述和数字描述。在十九世纪和本世纪初，主要是运用描述性统计进行社会现象的分析。另一类是推断性统计，它是在不完全的资料的基础上对总体作出比较精确的决定的科学。推断性统计是在随机抽样的基础上推论有关总体的情况，故亦称为统计推断。

根据统计分析所研究的变量数量的多寡，又可以分成单变量分析，双变量分析以及多变量分析。统计的分类大致如下：



统计分析
- 单变量分析：比例、均值、方差、估计等
- 双变量分析：另加上：相关分析、回归分析等
- 多变量分析；另加上：因素分析、途径分析等

在现代高度发展的社会中，社会生活的复杂多变以及数学和社会统计技术和研究手段的发展，使我们进行社会调查不能再满足于一些简单的计量，因为这种简单的计量分析还不能深入探讨现象之间的数量关系及其变异。现代社会调查要求以精确的数理统计方法，结合计算机的使用，把社会变量之间的关系尽可能地建立数学模型加以分析，以便更深一层地了解动态因素之间的关系及本质，预测未来，提出解决问题的多种比较方案和最佳方案。这些高一层次的统计分析方法需由专门的“社会统计”课程进行讲解，本章只能概要性地介绍最简单、基本的分析方法，而且不作任何数学上的证明，学生只要掌握了公式的使用并了解公式的基本意义，便可以在实际调查中加以运用。

在运用统计分析时还应注意，任何统计分析是否有用完全取决于人们对他们所占有的资料加以说明的能力。这也就是说，定量的分析一定要在定性分析的指导下有的放矢地进行，这样才可能使定量分析循着正确的分析路径展开，才可能达到为提高定性分析的准确性服务的目的。

第一节　描述性统计分析

一、单变量统计分析方法

在统计分析中，有两种数据对我们的调查起重要的意义，第一是某种标志值或变量的集中趋势，用来表示一组数字资料的中

心位置。第二是变量的离散趋势，用来表明数据的差异情况和扩散范围。这两种统计量是相互联系的。仅有集中趋势来反映数据的平均水平还不够，还应结合考虑数据的差异，把两种趋势结合起来考虑，才能正确认识一组数据的全貌。集中趋势的代表性如何要由离散趋势来表明。同时，这两种趋势还可以作为其他统计分析的基础，用以计算其他的统计量，所以这两种统计量是统计分析中最基本的统计量之一。

1．集中趋势测量法

当变量A具有众多不同的数值时（即是一组数据），如果我们可以找到一个数值来代表这众多的数值，该数值就是变量A的集中趋势值。由于这一集中趋势值是可以概括大量数据的代表性数值，反映了这组数据的集中趋势，所以也称之为集中统计量。其中，算术平均数对各数据的差的平方和为最小，我们用这个代表值来估计或予测总体参数时，所发生的误差原则上是最小的，所以可以利用算术平均数进行抽样推论。

反映集中趋势的统计量有：众数、中位数、平均数（又称算术平均数）、几何平均数和调和平均数。在社会调查中最常用的是算术平均数，其次是众数、中位数。

根据不同的计量层次，变量的集中趋势值的计算方法是不同的。换句话，各种集中趋势值都只适用于一定测量层次的变量，不能超出范围使用，这是我们应加以注意的。

1）众数（M_0）

它是用来表示定类以上测量层次变量的集中趋势的统计量。众数是指在一组数据中出现次数最多的指标值，即在分组中哪一类的次数最多，该类就是众数。实际上，众数就是通俗概念上的“多数”。求得众数，就是用多数的特征来代表整体的特征。以众数作预测所犯的错误总数是最小的。

众数的推求：

（1）分组资料。在分组的次数分布表中直接选出次数最多的一组。例如，在表9·1中女性为众数。

表9.1　××厂职工性别分类表

性别	人数
男	235
女	320
合计	550

（2）组距分组资料。采用求近似值的方法。计算公式为：

表9．2　3000户农户年收入次数分布表

年收入（元）	户　数
500—600	240
600—700	480
700—800	1050
800—900	600
900—1000	270
1000—1100	210
1100—1200	120
1200—1300	30
合　计	3000

$$M_0 = L_b + \frac{f_a}{f_a + f_b} \times d$$

上式中，L_b为众数组的下限；f_a为大于众数组的次数；f_b为小于众数组的次数；d为组距。

根据表9·2的资料，已知众数所在组是第三组，根据公式即可求得众数：

$$M_0 = 700 + \frac{600}{600 + 480} \times 100$$

$$= 700 + 0.556 \times 100$$

$$= 700 + 55.6$$

$$= 755.6\text{（元）}$$

另一种计算公式为：

$$M_0 = L + \frac{\triangle_1}{\triangle_1 + \triangle_2} \times d$$

式中，L代表众数组下限；$\triangle_1$代表众数组次数与下一组次数之差；$\triangle_2$代表众数组次数与上一组次数之差；d为组距。

同上资料计算可得：

$$M_0 = 700 + \frac{570}{570 + 450} \times 100 = 700 + 55.9 = 755.9\text{（元）}$$

所以755.9元就是三千户农户的众数。

众数的计算有一定的条件，即当样本数较多而且具有明显的集中趋势时可以用众数来反映数据的集中趋势。如果样本数少，或样本数虽多但分组后无明显的集中趋势，就不宜用众数。

众数适用于测量各种测量层次的变量。其优点在于：概念通俗，简单明了；计算方法简单；不受数据中特殊数值的影响，在数据分布偏态时最能表明数据的实际集中趋势。其缺点是：受次

数和分组的影响，稳定性差；在数据分布出现双峰或矩形时没有实际意义，也不能推算出总体的众数。

2）中位数（Md）

中位数是用来表示定序以上测量层次的变量的集中趋势的统计量。它是指在一组按数值大小顺序排列的数据中，处于中央位置的数值。中位数把所有的数据分成两半。计算中位数，首先要将所有的数据按顺序排列，然后求出中央位置（即中位数所在的位置）。方法为 $\frac{n+1}{2}$ 得数即是中位数的位置。中位数推求方法：

（1）从未分组资料中求中位数，例如，现有数据7个，数值排列顺序为：1，2，3，14，15，16，17，则中央位置是 $\frac{7+1}{2}=4$ 即第4个位置，则数值14为中位数。如果数据个数为偶数，如数值排列为1，2，3，14，15，16，17，18则中央位置为 $\frac{8+1}{2}=4.5$ 即在第4.5个位置的中间，所以中位数是这两个位置数值的平均数，即 $\frac{14+15}{2}=14.5$。

（2）分组资料亦可照上面方法先找出中央位置，再找出与该位置相对应的指标值，即为中位数。

（3）从组距分组资料中求中位数，可以运用公式求得中位数的近似值。

计算公式：

$$M_d = L + \frac{N/2 - f_b}{f_m} \times d$$

式中，L代表中位数所在组的下限；N代表次数总和；f_b代表中位数所在组以下各组的累积次数；f_m代表中位数所在组的次数；

d代表组距。

例：求三千户农户年收入的中位数，资料见表9．3。

表9．3　　三千户农户年收入情况表

年收入（元）	户　数	向上累积cf↑
500—600	240	240
600—700	480	720
700—800	1050	1770
800—900	600	…
900—1000	270	…
1000—1100	210	…
1100—1200	120	…
1200—1300	30	…
合　计	3000	—

已知：中位数位置$=\frac{3000}{2}=1500$即在第三组，第一、二组累积次数为720则：

$$M_d=700+\frac{1500-720}{1050}\times 100=700+74.3=774.3(元)$$

所以三千户农户的年收入中位数为774.3元，说明有一半的农户年收入在774.3元以上，一半在774.3元以下。

中位数的优点是：不受每个数据的影响，因而当一组数据中有数值特别大或特别小的数据时，可以较准确地反映出数据的集中趋势。对于定类变量来讲，众数和中位数均可反映其集中趋势，但由于众数不考虑次序，所以结果不如中位数准确，所以求定序变量的集中趋势，一般采用中位数为好。其缺点是：由于不受两端数据的影响，所以信息损失较多，敏感性较差。同时，它也不能对总体的中位数进行推算。

3）平均数（或算术平均数。$\overline{X}$）

平均数是代表定距、定比类型的变量的集中趋势的统计量，它是对一组数据的数值总和的平均。加权平均数是它的一种特殊形式，用于反映分组后不同次数的全部数据的总和平均。

平均数的推求：

（1）基本计算公式：$\overline{X}=\frac{\Sigma X}{N}$（$\Sigma$为总和符号）。当N数量不多或资料未分组时，可采用此方法。

（2）根据分组资料求加权平均数，计算公式：

$$X=\frac{\Sigma fX}{N}\quad(N=\Sigma f)$$

（3）从组距分组资料中求平均数。公式：

$$X=\frac{\Sigma fX_m}{N}$$

（X_m代表组中值）

例：某校学生××科考试成绩次数分布状况见表9．4，求平均分数。

$$\overline{X}=\frac{55\times 10+65\times 60+75\times 90+85\times 100+95\times 45}{305}$$

$$=78.61\text{（分）}$$

表9．4　××校学生×科考试成绩次数分布表

成绩（分）	组中值（X_m）	人　数
50—59	55	10
60—69	65	60
70—79	75	90
80—89	85	100
90—100	95	45
合　计	—	305

对于定距、定比类型的变量来说，虽然测其集中趋势也可以采用中位数、众数，但由于平均数充分利用了所有的资料，因而一般地讲，其结果较稳定，也最能代表集中趋势。它同时还可以进行代数运算。平均值除了进行比较外，还可以作估计或预测等其他进一步的统计分析。但它的缺点在于容易受极端数值的影响，即当资料中有个别数据的数值异常特殊(如过大或过小)时，则均值就有可能失去代表性。例如在一些资本主义国家内收入贫富悬殊，这样计算平均收入就不能证明什么问题。而若采用中位数就不受个别特殊数值的影响。另外，平均值的大小不仅取决于各个标志值大小，同时也取决于次数的多少。所以在加权平均数中，次数f起到了权衡经重的作用，故亦称之为权数。

计算各种类型的集中统计量必须注意一个最基本的前提，即集中趋势的统计量只有在某一相同特征的总体中才有代表性，即

要在科学分类的基础上才能进行推求，否则不仅没有意义，而且还可能掩盖了事物不同性质的数量表现。

2．离散趋势测量法

要全面地了解某一变量的状况，除了求出它的集中趋势外，还必须了解该变量的离散程度。变量的离散趋势可以表现数据的分布状况，说明各个数据之间的差异程度以及它们与中央趋势的离异状况，同时还可以进行比较。所以离散趋势又称为离中趋势或变动度，它与集中趋势两者相互补充，使我们对数据的认识更加全面。

例如：我们有以下三组数据

甲组工资：66，66，66，67，67，67，68，69 $\overline{X}_1 = 67$元

乙组工资：52，53，61，67，71，72，78，82 $\overline{X}_2 = 67$元

丙组工资：43，44，50，54，67，90，91，97 $\overline{X}_3 = 67$元

以上三组的平均工资都等于67元，但三组中各标志值与集中统计量的离散程度却大不相同。甲组的全部标志值集中在离平均数二个单位的范围内；乙组数据与中央统计量的最大离差为15个单位；丙组甚至达到30个单位。由此可见，仅看变量的集中趋势既不能反映出数据间的差异大小，也不能判断集中趋势的代表性程度。尤其在集中趋势相近的情况下，要真实反映现象的状况，就必须了解变量的离散程度。从上例中我们还可以了解到，一组数据的离散程度越大，说明该组数据的分布越分散，数据与中央趋势的差异就越大，因而其集中趋势的代表性程度就越低，在预测中可能犯的错误也就越大。反之，集中趋势的代表性则越大。

反映离散趋势的统计量有：全距异众比率、四分位差、平均差、标准差、方差、离散系数等。其中标准差是最科学、最完善的离散统计量，其次是方差，再其次是全距与四分位差、异众比

率和离散系数。

同样，各种离散统计量也都适用于一定的测量层次的变量，不可超出范围使用。

1）异众比率（VR）

这是与众数相对应使用的离散统计量，适用于定类以上所有类型的变量。它是指非众数的次数在总体中所占的比率，以此表明众数的代表性大小。VR数值越大，M_0的代表性就越小。异众比率可以粗略地比较两组或两组以上定类变量的离散程度，但当数据分布出现双峰或矩形时则没有意义。其计算公式为：

$$VR=\frac{N-f_{m_0}}{N}$$

式中：f_{m_0}代表众数组次数，N为总体次数。

现有两校学生父亲职业的资料，要求比较两校学生的父亲职业的离散程度。

$$VR甲=\frac{550-288}{550}=0.476 \qquad （众数组次数为288人）$$

$$VR乙=\frac{480-295}{480}=0.385 \qquad （众数组次数为295人）$$

表9．5 甲乙两校学生父亲职业情况

父亲职业	甲校人数	乙校人数
干部	110	95
工人	152	90
农民	288	295
合计	550	480

VR甲大于VR乙，这表明甲校学生父亲的职业差异比乙校大，所以用众数说明乙校的代表性程度较甲校要高。

2）四分位差（Q）

四分位差是与中位数对应使用的离散趋势统计量，它适用于定序以上测量层次的变量。四分位差是指位置之差，它先将顺序排列的数据序列分成四个相等部分，然后再比较各位置间的差异。四分位差的优缺点也类似于中位数，由于该统计量在统计分析中用途较小，这里不作详细介绍。

3）标准差与方差（σ或s，σ^2或s^2）

标准差是与平均数相对应使用的离散趋势统计量，适用于测量定距、定比的测量层次的变量的离散趋势。方差(σ^2或s^2)就是标准差的平方，所以标准差亦为均方差，它是测定各个数据相对于平均数的偏离程度的统计量。由于数列中数值之间的差距不平均，那么通过找出它们之间平均的离差作为标准差距，来衡量该组数据的离散趋势，这一标准差距值便是标准差。

推求标准差的基本公式为：

$$S=\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$

上式中，S为标准差；$\bar{x}$为平均数；x为每个数据的值；n为全部数据总数。

对于分组资料，计算方法同上，但由于每组的数值还取决于该组的次数，所以要加权。对于组距的分组资料，应先求出各组的组中值然后再加以计算，即将x改为x_m。加权公式为：

$$S=\sqrt{\frac{\Sigma f(x-\bar{x})^2}{n}}$$

亦可使用简捷公式：

$$S=\sqrt{\frac{\Sigma fx^2}{n}-\bar{x}^2}$$

σ代表总体的标准差，S代表样本的标准差，在分析资料阶段，一般使用S。当样本数n≤30时，应用公式：

$$S=\sqrt{\frac{\Sigma f(x-\bar{x})^2}{n-1}}$$

n－1为修正值。

例：某单位工人月工资状况见表9．6，求工人工资的标准差。

表9．6 某单位工人工资状况

工资（元）	人 数
50	15
60	28
64	40
70	290
74	160
80	17
合 计	550

$\bar{x}=69.9$（元）

$$S=\sqrt{\frac{\Sigma fx^2}{n}-\bar{x}^2}$$

$$= \sqrt{\frac{15(50)^2+28(60)^2+40(63)^2+290(70)^2+160(74)^2+17(80)^2}{550}-(69.9)^2}$$

$=5.3$（元）

这表明，550名工人的工资标准差为5.3元，即工人的工资相对于平均工资的标准偏差为5.3元。如果我们以69.9元的平均数去估计全体工人中每一个工人的工资数，则发生的平均偏差为5.3元。

标准差有深刻的数理基础，由于全部的数据都参与了计算，所以能够反映全部数据的数值的差异情况，数值稳定，受抽样变动的影响较小，可供进一步统计分析运用。许多复杂的统计分析都离不开标准差，标准差是理想的表明离散趋势的统计量。其缺点也与平均数相似，它受到极端数据的影响较大。

以上简略介绍了六种不同层次的集中趋势和离散趋势的统计量及其推求公式。我们下面列表（表9．7）总结一下不同的计量层次在运用上述方法时的区别，以便运用时加以注意。

表9．7　　集中统计量与离散统计量的应用范围

计量层次	集中统计量			离散统计量		
	众数	中位数	平均数	异众比率	四分位差	标准差、方差
定类	✓	—	—	✓	—	—
定序	✓	✓	—	✓	✓	—
定距、定比	✓	✓	✓	✓	✓	✓

表中：✓表示可用，—表示不可用

二、双变量统计分析方法

1．相关关系

世间的事物是相互联系的，在社会调查中必须重视分析两种以上社会现象或因素之间的关系。社会现象之间的数量关系按其表现形式可以分为两种类型。一类是函数关系，即当一种现象的数量确定以后，另一种现象的数量也就完全确定了。如物体作匀速直线运动时，路程s、速度V和时间t之间的关系为 $s = Vt$。假设速度V是一个常量，那么只要知道了路程或时间中的任何一个变量，也就可以准确地求出第二个变量。第二种关系是相关关系，在这种关系中，现象之间的关系不是完全确定的。虽然两个变量之间有依存关系，一个变量发生变动，另一个变量也会随着变动，但变动不完全确定。因为影响变动还有其他因素，所以两个变量间的相互依存关系并不严格。双变量的统计分析就是为了说明两种现象间的相关关系的。

社会调查的目的之一在于探讨社会现象发生、发展的演变规律。规律本身是一种确定的关系，但在实际生活中，由于受各种随机因素的影响，大量的是不完全确定的相关关系。我们可以在大量观察的基础上，找出这些不易确定的关系在平均值上的联系程度。这种联系程度称之为相关。换言之，如果一个变量发生了变化，另一个变量也有变化，则证明这两者之间有相关关系，即它们之间具有连带性。

相关关系的分析可以在二个变量之间进行，称为双变量相关分析，或二元相关分析，又叫简相关、单相关。同时又可以分析一个变量同时与两个以上变量之间的相关，称为多元相关分析或复相关分析。

相关关系的类型很多，从变量变动的方向来看，有正相关、负相关和零相关。正相关关系表示在变量x、y之间，如果变量x

增加（或减少）时，变量y也相应地增加（或减少），它们之间变动的方向是一致的。负相关表示，变量变动的方向是相反的，即当x增加（或减少）时，y也随之减少（或增加）。零相关则表示两个变量之间不存在相关关系，表现为当x变动时，y不发生变动或变化无规律可循。另外，相关关系还可以分为直线相关和曲线相关。直线相关是指一个变量变动时，另一变量也随之发生大致均匀的变动，在图形上近似地呈现直线形状。曲线相关则是一种不均匀的变动，在图形上呈现为各种不同的曲线。

相关关系的方向和强弱程度要通过计算出一个相应的统计量来进行说明，我们把这些相应的统计量称为相关系数。相关系数的正负号表示相关的方向，绝对值表示关系的程度，其取值范围都在－1.00～＋1.00之间，小数点后至少要保留两位数。

2．消减误差比例（PRE）

由于变量的计量层次的不同，计算相关关系的方法也各不相同而呈现出不同的统计意义。在社会调查的分析阶段，我们需要根据统计量的PRE的大小，来衡量该统计量是否有意义。

PRE即“消减误差比例”，计算公式为：

$$PRE = \frac{\text{已消减的误差}}{\text{全部误差}}$$

我们知道，社会调查研究的主要目标是预测或解释社会现象的变化。在进行预测或解释时，难免有误差（或错误）。假定变量X与Y之间有关系，则我们用X值来解释或预测Y时，理应可以减少若干误差，而且这种关系愈强，所能减少的误差就会越多。反过来说，如果能减少的误差越多，证明两者之间的关系愈强。所以我们可以从消减的误差的大小来反映出X、Y之间相关的强弱程度。

现在假定我们不知道变量X的值，我们在预测Y时所产生的

全部误差为E_1。如果我们知道了X的值，便可以根据X的每个取值来预测Y值。假定这时所产生的误差总数为E_2，那么以X值来预测Y值时所减少的误差就是$E_1 - E_2$。这一数值与原来全部误差的比值，即是PRE。所以，$PRE = \frac{E_1 - E_2}{E_1}$，即$PRE = \frac{\text{已消减的误差}}{\text{全部误差}}$。PRE越大，证明以X值来预测Y值时能够消减的误差所占比例愈大，X、Y的关系越强，因而用X来预测Y也就越准确。

PRE统计值所具有的这些意义，符合社会调查研究的需要。它显示了用一个现象来解释另一个现象时能够消减百分之几的错误。显然，错误减少得越多，证明预测或解释的能力愈强。从以上公式中不难看出，PRE值是介于（0—1）之间的。正如在社会研究中一般不存在完全的相关关系一样（即相关系数正好等于+1或-1），PRE在分析时一般也不可能达到+1。

根据以上分析，我们在选择测量相关关系的方法时，应考虑以下几点：

第一，要注意变量的测量层次，不同的测量层次采用不同的计算方法。

第二，测量相关的统计方法很多，但不全都具备PRE的意义。在社会研究中，我们主要选用具有PRE意义的方法，这样才能显示该相关系数的意义来。

定类、定序、定距这三种测量层次，通过组合，可以获得多种双变量的组合方式，对这些组合方式，都有不同的测量方式。下面简要介绍其中的两种。

3．相关关系测量法

1）相关比率（η^2）

相关比率（η^2）读作eta，是用于测定属于定类——定距（定比）、定序——定距（定比）、非直线的定距——定距的测量类型的变量之间相关程度。

定义公式：（1）$E^2(\eta^2)=\frac{\Sigma(Y-\bar{Y})^2-\Sigma(Y-\bar{Y}_i)^2}{\Sigma(Y-\bar{Y})^2}$

这里显示了$PRE=\frac{E_1-E_2}{E_1}$的统计意义。

运算公式：$E^2(\eta^2)=\frac{\Sigma n_i\bar{Y}_i^2-N\bar{Y}^2}{\Sigma Y^2-N\bar{Y}^2}$

式中，n_i为每类个案数（次数），$\bar{Y}$为全部Y值的平均数，$\bar{Y}_i^2$为每类均值的平方，ΣY^2为全部Y值的平方和，N为全部个案数。

注意在计算Y_i^2时，对于分组的资料，应进行加权。同样，计算$\bar{Y}_i$时也应注意加权问题。

现根据表9．8的调查资料计算39名学生的母亲文化水平（定序）与其所生子女数（定距）之间的关系。因母亲文化程度是定序层次，子女数则是定距层次，所以要用η系数来测量这两者间的相关情况。从表下端可知各类的平均子女数（$\bar{Y}_i$），并可算出$\bar{Y}=4.05$，$\Sigma Y^2=2^2+3^2+4(2)^2+\cdots\cdots+8^2=722$。

表9．8　39名学生母亲文化程度与其所生子女数次数分配表

子女数 / 文化程度	大学	高中	初中	小学	文盲
2	1	4	—	—	—
3	1	1	4	4	—
4	—	1	1	6	4
5	—	1	—	2	3
6	—	—	—	1	2
7	—	—	—	1	1
8	—	—	—	—	1
n_i	2	7	5	14	11
ΣY_i	5	20	16	59	58
$\bar{Y}_i$	2.5	2.9	3.2	4.2	5.3

将以上数字代入公式：

$$E^2=\frac{2(2.5)^2+7(2.9)^2+5(3.2)^2+14(4.2)^2+11(5.3)^2-39(4.05)^2}{722-39(4.05)^2}$$

$$=0.43$$

$$E=\sqrt{0.43}=0.66$$

从E值可以看出母亲的文化程度对生育子女数量有很显著的影响，从E^2可以看出用母亲的文化程度来解释所生子女数，可以消减43%的误差。另外，我们从表下端的$\bar{Y}_i$中可以看到，子女数是随着母亲文化程度的下降而不断增加的。

相关比率还可以用来分析两个定距变量之间的非直线关系，如年龄大小与抽烟数量的关系，工龄与劳动效率的关系等。这些都不是直线关系，而是曲线关系。劳动效率随年龄增加不断直线上升到一定程度后便开始逐年下降，如果用求直线相关的方法则会歪曲事实。

相关比率在社会研究中使用广泛，它还可以用于测量定序与定距的变量关系。由于社会现象大量的是非直线关系，而且定序与定距、定类与定距的这种关系又很多见，所以应该注意掌握好有关的计算方法并了解E^2所表示的意义。

2）积矩相关系数（r）

也称为简单直线（积差）相关系数，它适用于测量有直线关系的两类定距变量间的关系。这也是一种重要的统计方法，但因条件的限制，如要求变量成正态分布、样本数不少于30、变量应为定距（定比）与定距（定比）的关系等，使其用途受到很大的影响。

一般计算公式：

$$r=\frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2}\ \sqrt{\Sigma(Y-\bar{Y})^2}}$$

计算r的公式很多，通常使用以下简捷公式进行运算：

$$r=\frac{N\Sigma XY-(\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2-(\Sigma X)^2}\sqrt{N\Sigma Y^2-(\Sigma Y)^2}}$$（r^2具有PRE意义）

这样，当缺少变量X、Y的平均值$\bar{X}$、$\bar{Y}$的资料时，可以直接从数据中计算出r系数来。

例：表9．9是12人的受教育年限与各人收入的资料。X为受教育年限，Y代表收入。

表9．9　　受教育年限与收入资料

受教育年限（年）（X）	年收入（千英磅／年）Y	X^2	Y^2	XY
10	6	100	36	60
7	4	49	16	28
12	7	144	49	84
12	8	144	64	96
9	10	81	100	90
16	7	256	49	112
12	10	144	100	120
18	15	324	225	270
8	5	64	25	40
12	6	144	36	72
14	11	196	121	154
16	13	256	169	208
Σ146	102	1902	990	1334

将数据代入公式：

$$r=\frac{N\Sigma XY-(\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2-(\Sigma X)^2}\sqrt{N\Sigma Y^2-(\Sigma Y)^2}}$$

$$=\frac{12(1334)-(146)(102)}{\sqrt{12(1902)-(146)^2}\sqrt{12(990)-(102)^2}}=0.75$$

$$r^2=(0.75)^2=0.56$$

计算结果证明，受教育年限与年收入之间具有线性关系，相关系数为0.75，两者关系很密切。从r^2值可以证明，如果用受教

育年限来解释收入问题，可以减少误差56%。

对于定类——定类、定序——定序、定类——定序的关系测量，常用的相关系数有λ（Lambda）系数、G（Gamma）系数等，这里不一一介绍。

进行相关分析，要注意：

第一，相关分析只是为了证明变量之间是否有依存关系，以及这种关系的性质与程度。要注意不应把相关关系当作因果关系来进行理论上的解释，如不能把教育当作是收入的原因，我们只能确定在变量A、B之间有相关关系，一个变动了，另一个也要变动，但这并不意味着它们之间存在因果关系。变量间是否有因果关系非相关分析所能确定的，正因为如此，我们把相关分析作为一种辅助性手段。只有根据理论分析出具体的社会现象各标志值之间关系确实存在的事实和性质后，使用相关分析才有意义。否则，就会发生在数量上计算出相关系数，而在实际生活中两者却完全没有关系的错误来。所以不应离开理论的分析而盲从于某一数字。

第二，相关系数是表示变量间相关程度的量的指标，是一个比率数值，而不是相关量的等单位度量。例如我们不能讲相关系数0.90是0.30三倍，而只能说前者的相关关系程度比后者要密切。同时相关系数－0.20与＋0.20之间所表示的关系程度是相同的，只是方向不同而已，因而不能讲0.20＞－0.20。对于相关程度的解释，一般认为0—±0.30为相关程度低；±0.30－±0.50为相关程度显著，±0.70—±0.90为相关程度高；±0.90—±1.00则为相关程度极高。

第三，考虑到我们目前所具备的研究手段和实际情况，在进行研究设计时，不要设立过多的相关因素，最好找二、三个有代表性的变量进行分析，否则因受人力、时间及计算工具等条件的限制，在统计分析时会带来许多困难。

第二节　推断性统计分析

一、统计推断的作用

社会调查研究所涉及的对象总体往往是具有相当大的数量，甚至有时是具有无限数量的总体，而研究者所具备的人力、物力、时间、经费等各种条件都不允许进行全面的调查，只能采取从总体中选取少数样本进行调查，以此来推论总体的办法。统计推断可以帮助我们从个别的调查结果中去认识规模巨大的总体，它在调查分析中发挥着以下两个重要的功能：（1）估计总体的参数值，即利用一定的数学方法从样本的统计值来概算总体相应的参数的大小；（2）进行统计假设检定。统计假设检定有单变量与双变量两种检定，主要是先对总体的某些特征作出假设，再通过对样本的研究来检验该假设。由于是利用样本的数据进行估计和检验，所以推断过程中的风险和误差是不可避免的。但在统计推断中，可以明确地知道和把握这种风险，以便决定我们对参数估计和假设检验时所具有的把握度。

当我们涉及统计推断时，必须涉及两个概念；参数与样本指标。我们在第三章的第一节已经就总体指标、样本指标问题进行过解释。参数实际上就是总体指标，它是描述有关总体的全部特征的变量的工具，而样本指标仅是对总体的一部分——样本的变量进行描述的工具。总体参数与样本指标是相对应的。如：总体均值$\bar{X}$与样本均值$\bar{x}$，总体方差σ^2与样本方差S^2；总体成数P与样本成数p等等。统计推断实际上就是利用样本指标来推断总体指标的过程。

统计推断所依据的理论基础仍是正态曲线理论。正态曲线是一条数理曲线，它由两个参数决定：（1）均值；（2）方差。曲线下的面积就构成了这些参数发生的概率。当我们反复地从总

体中抽取规模固定的样本时，不论总体分布是否服从于正态分布，其样本指标$\bar{x}$所构成的抽样分布总会近似于正态分布。这些平均数的均值则等于总体参数值$\bar{X}$。所以在标准化的情况下，样本指标的抽样分布构成的面积，就可以成为任何一个样本总体的指标观测值出现的概率。

二、总体参数的估计

利用计算出来的样本的统计值来估计总体参数，有两种方法：点估计和区间估计。点估计又叫单值估计，它是使用一个单一的数值来估计未知的总体参数，对参数进行估计的相应样本的统计量称为估计量。例如，我们对××大学300名学生的月平均消费水平进行调查，计算出样本平均值为70元，然后就用这一平均数数值作为该校全体大学生的月消费水平的估计值，这是一种以点代面的估计方法。点估计虽然简便，但由于没有考虑到抽样误差，既不能说明估计的准确程度，也不能说明估计的可靠程度，所以在社会调查研究中使用不多，一般只用于对总体的参数进行粗略的估算。

区间估计则不同，它不是直接的、简单的估计，而是依据样本的统计值和抽样误差去估计总体参数的可能范围。换句话说，区间估计是在一定的把握度（概率保证度）上对总体参数可能落入的一个数值范围（区间）做出估计。进行区间估计，还可以说明推断的准确程度和把握程度。由于上述优点，区间估计成为统计推断的主要方法。由于存在着抽样误差，区间估计不可能百分之百的准确，它实际上是随机抽样的结果，因而推断的准确程度受样本估计值和抽样误差的影响。所以，区间估计本身就是一个随机变量，对同一总体的不同抽样，可以产生不同的估计区间。这就会产生一个如何判断推断的可靠程度问题，即我们所做估计成功的把握问题。这种把握程度可以用显著性水平的概念来表示

（也可以从置信度角度来理解，即在第四章提到的与t值相对应的概率保证度）。显著性水平是用P值表示。当估计成功的概率为95%时，可写成P0.05，说明显著性水平为5%，即估计错的可能性不超过5%。按一定的把握度求得的估计区间称为置信（可信）区间，表明在一定的把握度上估计总体参数可能存在的范围。所以，区间估计要说明两个内容：（1）指出总体参数可能存在于两个数值的范围内，说明估计的准确程度；（2）说明估计的把握度。

进行总体参数的区间估计，其计算程序为：

第一，选定把握度。在社会调查中通常使用的可信度为90%、95%、99%。

第二，计算标准误差。它是指样本分布的标准差，用来表示估计总体参数的可靠性，各样本统计值都有自己的标准误差，计算结果是不同的。

第三，根据样本估计值和标准误差，推算出总体参数所在的可能范围。

依照上述程序，我们来进行总体参数估计的实例计算。

1．总体平均数的区间估计

这是在一定把握程度上从样本平均数推算总体平均数大小的估计方法。

例：在某工厂随机调查100名工人，计算出他们的月平均收入为80.20元，标准差为11.50元，求当把握度为95%时（即在P0.05的显著水平上）估计总体平均数的置信区间。

1）查附表4，在显著水平P0.05时，Z值为1.96。

2）计算标准误差（$SE_{\bar{x}}$）

公式：$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{S}{\sqrt{n-1}} \approx \frac{S}{\sqrt{n}}$（式中S代表样本标准差）

$$SE_{\bar{x}} = \frac{11.5}{\sqrt{100-1}} = \frac{11.5}{\sqrt{99}} = 1.15$$

3）确定总体平均数的置信区间：

区间公式为$\bar{X} \pm Z \times S\bar{E}_x$

$$\bar{x} - 1.96 \times 1.15 \leqslant \bar{X} \leqslant \bar{x} + 1.96 \times 1.15$$

$$80.20 - 2.25 \leqslant \bar{X} \leqslant 80.20 + 2.25$$

$$77.95 \leqslant \bar{X} \leqslant 82.45$$

在P0.05的显著水平上，总体平均数的置信区间为〔77.95，82.45〕。

结论：当把握度为95%时，总体平均数的置信区间为〔77.95，82.45元〕。或者可以说，总体平均收入落在77.95元～82.45元这一区间的机会是95%。还可以理解为：我们对总体平均收入作77.95元—82.45元的区间估计，其正确的可能性为95%，发生错误的可能性不超过5%。

我们在这里所举的例子是大样本的计算方法，即n＞30，当n≤30时，称为小样本计算，这时与P0.05显著水平相应的不是Z值，而是t值，属t分布。因社会调查多是大样本调查，所以对t分布不作介绍。

2．总体比例和百分比的区间估计

总体比例和百分比的估计是指在一定的把握程度上从样本的比例和百分比来推算总体的比例和百分比。其估计程序同平均数估计相同，但标准误差的计算公式不同：

$$SE_{比例} = \sqrt{\frac{P(1-P)}{n}}$$

$$SE_{\%} = 100\% \times \sqrt{\frac{P(1-P)}{n}}$$

例：对某区青工随机抽取80名调查，了解其中参加业余学习的人数比例（或百分比）。调查已知有65%的青工参加业余学习，现要求在P0.05的显著水平上对总体的百分比进行估计。

1）确定Z值，在P0.05显著水平时，查表可知Z值为1.96。

2）求标准误差

$$SE\% = 100\% \times \sqrt{\frac{0.65 \times 0.35}{80}} = 5.3\%。$$

3）求总体百分比置信区间：

$$0.65 - 1.96 \times 0.053 \leqslant P \leqslant 0.65 + 1.96 \times 0.053$$

$$0.55 \leqslant P \leqslant 0.75 \text{或} 55\% \leqslant P \leqslant 75\%$$

结果表明，在显著性水平P0.05时，该区青年参加业余学习的人数百分比在55%～75%之间。

以上计算是针对二项分布的，即总体只分为参加业余学习与不参加两个部分。如果总体内比例多于两项，在作区间估计时则应化为两项式，逐一分开计算。同时，上述计算也是针对n＞30时的样本而言的，当n≤30时，样本的分布不是正态分布，而是t分布，则要用t值来进行估计。

对总体相关系数的估计，因计算较复杂，本节从略。有兴趣者，可以参考有关书籍。

三、统计假设检验

统计假设是先对总体的某些特征作出假设，再通过对样本的研究来对假设进行检验的方法和过程，它也是统计推断的一个重要内容。

统计假设检验要解决两类问题。第一类是，当我们对某一研究总体的特征已有初步了解后，在此基础上可以提出一种假设，然后用抽样方法对这一假设进行检定。其特点在于在进行资料的检定时，我们并不直接对研究以前所设置的统计假设或研究假设（简称H_1）进行检验，而是首先检定一个与研究假设相对立的假设，它在统计上称为虚无假设（H_0），通过检定H_0来间接地知道H_1正确的可能性。检验的目的是为了排除抽样误差的可能

性，因为任何抽样都有误差，这种误差使调查的结论与原统计假设之间必然存在差异，这种差异究竟是抽样误差的原因还是由于对总体作出的统计假设有错误？如果是前者，就应肯定原先的假设，如果是后者，就应否定原先的假设。检验的目的，就是通过设立H_0来排除抽样误差的可能性。因为抽样误差是建立在H_0基础上的，只要否定了H_0，也就可以否定抽样误差，从而间接地肯定H_1。

假设检定的第二类问题是要对从两个或两个以上的样本总体中计算出来的统计值之间的差异进行显著性的检验，如比较两个平均数、两个相关系数之间的差异，这些差异可能是由于它的各自代表的总体参数不同造成的，也可能是因抽样误差造成的。因此通过设立H_0，对显著水平的检验，如果显著水平高，H_0则被推翻，说明这些统计值的差异确实是代表了总体之间的差异，否则差异则可能是因抽样误差造成的，而非两个总体之间存在差异。

以上分析可见统计假设检验的总思路是，要想肯定H_1，必须否定H_0，通过直接检定H_0，看其显著水平的高低，如果显著水平高，则H_0被否定，H_1得以间接的肯定；显著水平低（即低于检定前所定的标准），则H_0成立，H_1被间接地否定。

假设检验的步骤与具体推理过程较复杂，其内容在有关的统计原理书中均有介绍，这里亦不再专门论述。

必须指出的是，统计假设并不能证明某个假设的正确与否。它仅仅是用概率统计理论来分析样本提供的信息，告诉我们能否推翻原假设，如果推翻不了就只能接受。统计假设检验只是调查研究分析的一种重要手段，但不是唯一的手段，更不是检验一种假设或理论是否成为真理的手段。即使是否定了H_0，也不能说H_1一定对，而只能说“可能”对。因为H_0被否定并不意味着它完全没有出现的可能，只是可能性很小罢了。