第一章 绪论

第一节 研究与统计

研究(research)是一个为理论或实际问题寻找答案的系统过程，在此过程中统计分析往往是一个至关重要的环节。为了有助于读者更好地理解统计方面的问题，有必要首先简要讨论一下研究的一些主要方面或参数。

一、方式(approach)

研究方式有综合性(synthetic)与分析性(analytic)之别。综合性的研究把要研究的问题看作一个整体，强调的是各部分之间的相互依赖与联系，研究的目的是对各部分之间的关系有个总括的大致的了解。而分析性的研究则是把各部分分离出来,分别加以研究,强调的是各个构成部分的作用 当然,把对各个部分的研究结果综合起来,也能得到对整体的总括的了解。

二、目的(objective)

研究目的可有探索性(heuristic)的与演绎性(deductive)的两种。探索性的研究往往是归纳性的、描述性的，研究者没有十分明碗具体的研究问题，对要研究的现象只有一个模糊的看法，研究的目的就是对该现象进行观察、记录和描述，以期对其获得更多的了解，为进一步的研究提出具体的问题或假设。因此，这类研究可以说是假设产生性(hypothesis-generating)的。而演绎性的研究往往有一个非常明确具体的问题，而且对研

究结果已经有某种期待，这就是假设；该假设可以基于探索性的研究，也可以基于某种理论。演绎性研究的目的就是来检验这一假设的有效性。因此，这类研究可以说是假设检验性(hypothesis-testing)的。

三、设计(design)

研究设计涉及对研究环境的操纵与控制（manipulation and control）的程度与方式。这是一个连续体，在其一端是对研究环境进行多方面严格控制与操纵的研究设计，而在另一端则是有意对研究环境不加任何控制、操纵和限制的研究设计。控制与操纵的程度直接影响数据的质量、数据的分析、结果的解释以及结果的适用性。研究环境是一个笼统的概念，包括研究的规模或具体程度、变量的控制程度、被试对研究的意识程度，等等。分析性、演绎性研究的限制程度往往高于综合性、探索性的研究。譬如，它所研究的问题较为明确具体，同时为了保证研究结果清楚明确，便于解释，研究者必需采取适当的措施，尽可能控制一切与所研究的中心问题无关的因素。

四、方法(methodology)

研究方法指收集数据的具体步骤。收集数据的方法取决于研究领域、研究方式与目的等因素。显然，要决定如何收集数据，首先必须明确要收集什么样的数据，也就是对数据加以定义，这直接关系着收集数据的具体步骤以及所使用的工具、对变量的控制和对结果的解释。需要注意的是，有些方法有可能把被试的注意力引向所要收集的数据，也就是他们会意识到他们正在被当作研究的对象，或者研究者正在收集什么样的数据。这样一来，他们的行为就有可能受到影响而变得不自然，从而收集到的数据的质量（代表性与可靠性）就会受到影响。

无论什么类型的研究，只要是按照严格的步骤和方法系统地收集数据，进而为某个研究问题提供答案，都属于“实验”研究(但是，为了便于区别，我们不妨把综合性的、探索性的、假设产生性的、较少控制与操纵的研究称作“准实验”研究，而把分析性的、演绎性的、假设检验性的、对研究环境控制与操纵较为严格的研究称为“实验”研究)。

任何严肃的实验研究都必须注意并采取有效措施保证其“内部有效性”(internal validity)与“外部有效性”(external validity)。“内部有效性”是指研究结果的可靠性，即研究结果确实是由研究者所调查的因素（而不是其它无关的因素)所产生的，而“外部有效性”指研究结果能够适用于该

研究环境之外的（类似）环境。前者关系到我们是否有把握接受基于该研究设计所产生的结果；后者关系到我们能否赋予该研究结果以普遍性。显然“内部有效性”是“外部有效性”的前提，假如我们对研究结果的可靠性根本就没有什么把握，不能确定是什么因素导致了这一结果，那么再去把该结果推而广之就没有什么份值了。

第二节　实验设计与统计

统计固然重要，但它毕竟只是整个研究过程的一个环节——对获取的数据进行统计分析，据此对所研究的现象进行解释。要保证研究结果的可靠性，首先必须获取可靠的数据，因此如何获取有效可靠的数据是一项研究需要考虑的中心问题，也是研究成败的关键。为此，研究者在开始一项研究之前必须充分考虑研究的设计问题，对整个研究过程予以周密的计划和考虑，对每个环节都按照科学的要求制定出实施计划，对可能出现的问题要提出预防措施。具体来讲，对于分析性的、演绎性的、假设检验性的、对研究环境控制与操纵较为严格的研究，至少要考虑以下几个方面：

（1）对所要研究的问题涉及的主要方面（变量）及其关系加以界定；

（2）提出明确的假设；

（3）确定研究的具体步骤和方法；

（4）选取有代表性的样本（包括样本的结构、性质与大小）；

（5）操纵和控制适当的因素或变量。

这样才能保证我们所研究的确实是我们打算研究的现象，才能保证研究结果不受或少受无关因素的干扰，进而保证数据的可靠性以及研究结论的适用性。显然，没有严肃周密的设计，就无法获取有意义、有份值的数据，那么进一步的统计分析也就没有多大意义了。从另一个角度来讲，在考虑研究设计时，也要同时考虑统计的问题，即将来准备用什么统计方法分析数据；否则，即使数据本身非常可靠，如果设计不符合统计方法的要求，就会找不到合适的统计方法进行分析，这样的数据是没有多大用处的。总之，任何严肃的研究都要慎重考虑研究的方法和步骤以及统计分析的问题，力求科学、有效而又经济地实现其研究的目的，否则只能意味着时间、人力、物力和财力的浪费。

但是，限于篇幅，本书主要讨论与统计有关的问题，至于研究设计，仅在必要时简略论及。

第三节　统计学的基本内容

依其功能，统计学通常分为两大部分：描述统计与推断统计。

一、描述统计

描述统计的作用是对数据进行整理、归纳和总结，使数据得以压缩，便于研究者把握其一般性的特征或全貌。

当研究者获取一个样本（或较小的有限总体）的一组数据之后，往往要对其进行某种方式的归纳整理：一是因为原始数据往往很庞杂，如不加以整理，就会很难理解和解释，尤其难以一下子抓住能反映所研究的事物性质的一般性特征和全貌，如果把这样的数据报告给读者，读者也难以理解数据的意义以及据此所得出的结论；二是因为对样本的描述统计是进行推断统计的基础，要进行推断统计必须先对样本数据加以描述，计算出有关的统计值。描述统计的主要内容有：利用统计图表以及计算峰值和偏态值来描述数据的分布情况；通过计算一些统计值来反映数据的集中趋势(例如平均数、中数、众数等)和离中趋势(例如全距、标准差等)。

二、推断统计

推断统计的作用是根据描述统计提供的结果，进一步对有关关系加以推断。推断统计的第一个重要方面是根据样本统计值对总体参数进行推断或估计。我们在讨论样本与总体时将会谈到，我们所研究的对象的个体往往非常多，甚至无限多，因而由于时间与经济等方面的限制，通常仅能选取部分个体（即样本）加以研究，然后再用适当的统计手段对所有个体（即总体）的情况进行推断，例如通过样本平均值推断或估计相应的总休平均值。推断统计的另一个重要方面是对事物之间关系的推断，例如比较两种或多种事物之间在某个方面（例如平均值）的差异等。在进行比较时，我们首先要分别分析每个样本的特点，然后再利用特定的统计方法检验它们之间是否存在差异，并碗定这一差异是‘真正的’差异(即由我们所研究的某个因素所造成的）还是由偶然的因素（例如抽样误差）造成的差异，与此同时，我们还可以指出得出这一结论的误差大小（例如 5%，1%等)或把握程度(例如 95%，99% 等），这一比较的过程称为假设检验，是统计学最重要的内容之一。

第四节　数据与变量

统计是与各种数据和变量打交道的，因而对于研究者来说，在进行统计分析之前，必须充分了解所要处理的数据和变量的类型及性质，因为这对于统计方法的正确选择和对统计结果的解释都是至关重要的。

一、自变量与因变量

在语言研究中，我们经常要用实验的方法来调查变量之间的关系。我们往往操纵(如引入、移去、变化)某一变量 X，然后观察并测量其对另外某一变量 Y 所产生的影响(如出现、消失、变化)。所操纵的变量叫做自变量，因为我们感兴趣的是它如何影响别的变量，而不是它受别的什么变量的影响。我们观察和测量的变量叫做因变量，因为它是随着自变量的变化而变化的变量，或者说它的值取决于自变量的值。我们可以把自变量看作起因，把因变量看作自变量的效应或结果；也可以把自变量看作刺激变量或输入或先行条件（先于某一结果的必备条件），把因变量看作反应变量或输出或在所研究的人或现象上产生的结果。

例如，要比较传统教学法与交际教学法在外语教学中的效果，我们可以随机选取两组学生作为被试(subjects)（采用随机抽样旨在保证两组被试的可比性),分别施以其中一种教学法,经过一段时间之后,用适当的方法对其学习成绩加以测量和统计分析，我们就可以知道不同的教学法是否会对外语学习产生明显不同的影响（体现为学习成绩的显着性差异）。这里教学法就是自变量，被试的学习成绩就是因变量。再譬如，我们要研究一下应考经验对考试成绩的影响，这里应考经验和考试成绩就分别为自变量和因变量。

但是,有的时候(例如在研究变量之间的相关时),决定哪个为自变量哪个为因变量却是任意的，因为二者之间没有真正的区别（即没有因果关系）。在这种情况下，一般就不再作出这一区分了。

此外，自变量也可以称作因素，其变化称为水平。在上述有关教学法的例子中，教学法为一自变量或因素，而传统教学法与变际教学法则是该因素的两个水平；假如我们要研究语言难度（自变量或因素）对阅读理解(因变量)的影响,不同的语言难度(难、中、易等)就是该因素的水平。因而，注意不要把一个自变量的几个水平看作几个独立的自变量。

二、计数数据与测量数据

按照获取的方法，实验数据可分为计数数据和测量数据。计数数据是指计算个数的数据，例如长、短句数，高、低智商的人数，男、女生人数，等等。此类数据一般取整数。而测量数据则是指利用某一测量工具而获得的数据，如考试成绩等。

三、连续变量与非连续变量(离散变量)

实验数据按其是否具有连续性可以分为连续变量与非连续变量。

连续变量可以取某一范围内的任何值，其单位可以做非常细微的划分来显示程度上的细微差异，从理论上讲，其精确度是没有任何限制的，例如考试分数、说出某个句子所花的时间等等。而非连续变量或离散变量则只能取某些值，两个单位之间不能再做细微的划分。非连续变量可以是数字型的（一般取整数），例如单词长度：一个单词可以是 1 个字母长、2 个字母长、3 个字母长，等等，但不可能是 1.3 个字母长、2.652 个字母长等，这里单词的长度就是一个数字型的非连续变量（然而平均字母数却是连续变量）。非连续变量也可以是范畴型的，它的值往往是某种特征、接受或不接受某种实验处理等等，例如性别(男、女)、颜色(红、黄、蓝等)、智商(高、低等)、受教育情况(初等、中等、高等)等。

在实验研究中，自变量多是范畴型的非连续变量，而因变量多为连续性变量。

四、称名变量、顺序变量、等距变量及比率变量

变量分类的另一个重要依据是变量的不同测量量表或测量水平，据此可以区分称名变量、顺序变量、等距变量及比率变量。变量的测量水平直接影响典型值的选取、假设检验中统计手段的选取等。

1. 称名量表与称名变量

“称名”即命名，也就是说称名量表实质上并不是在测量，而是在命名，即把个体按照某一特征分成不同的范畴，范畴之间往往只有‘异同’之别，而没有‘多少’之分，即不存在数学上的关系(当然也可以用数字表示不同的范畴，这时可以看作是简单的数学关系），或者说这一量表是定性的而不是定量的，以此方法加以‘测量’的变量称为称名变量。例如，我们要研究智商与学绩的关系，我们可以把被试按照智商的高低分成高智商组和低智商组，这样‘智商’这个自变量就被分成了两个范畴，该变量就是

个称名变量。其它常见的称名变量还有：不同的教学学习方法、对某个问题的反应(例如“是”与“否”、“同意”与“不同意”等)、不同性质的阅读材料(“难”与“易”、“熟悉”与“不熟悉”等)等。

2. 顺序量表与顺序变量

顾名思义，顺序量表的作用是把个体排序或分等，个体之间的关系体现为“大于”与“小于”或“高于”与“低于”等的关系，但是顺序量表仅仅排序，不能指出其间的差别大小，不同序数之间的数值差不一定相同，或者说其单位是不相等的。例如，把某个班的考试分数按照高低排序，就产生一个顺序量表，各分数之间的顺序关系一目了然，但各个名次之间的距离却不一定相同，例如第一名与第二名之间的分数差不一定和第七名与第八名之间的分数差一样。以排序的方法进行测量的变量称为顺序变量。以上述“智商”为例，我们可以不把被试分为高智商和低智商两个组，而是按照各人智商值的高低排序，这时该变量就成为顺序变量了。但是需要指出的是，当我们把被试分成高智商组和低智商组时，由于“高”与“低”也表明了一种顺序关系，所以我们既可以把智商这一变量看作称名变量，也可以看作顺序变量，不过在统计中，当遇到这样仅涉及两个事物排序的情况时，通常把变量当作称名变量而不是顺序变量。

顺序量表所提供的信息要多于称名量表，因为除了范畴的异同之外，它还能告诉我们观察值之间的排列顺序。当然，由于牵涉排序的问题，所涉及的工作量也大一些。

3. 等距量表与等距变量

与顺序量表不同的是，等距量表的单位是相等的，即量表各点之间的间距是相等的，其测量数据具有这种性质的变量称为等距变量。例如，气温 10℃和 20℃之间的差距与 20℃和 30℃之间的差距是相同的，所以气温就是一个等距变量。

在语言研究中，真正的等距变量是不多的，各种等级量表和测试分数可以看作等距变量，因为其单位是相等的，例如 80 分与 85 分之间的距离和 90 与 95 分之间的距离是相等的（我们以后还会谈到，原始分数可以转换为标准分数，以保证其具有等距量表的特点）。

显而易见，等距量表所提供的信息又多于顺序量表，因为它不仅表明了事物之间的顺序，还进一步告诉了我们各点之间的距离。

4. 比率量表与比率变量

比率量表与等距量表相似，所不同的是它具有一个绝对零点（即量表上代表完全不具有某一特征的一点），因而，量表上不同点的比率是可比

的，比率量表便于比较实际数值而不是数值间的距离。测量数据具有这种性质的变量称为比率变量，常见的有距离、时间、身高、体重等。以时间这个比率变量为例，它具有绝对零点，因而，我们可以说，9s（秒）三倍于3s，9s与3s的比率等于6s与2s的比率，但对于像温度这样的等距变量，因其没有绝对零点，我们很难说20℃的热度两倍于10℃的热度，也不能说20℃之于10℃相当于60℃之于30℃。

比率变量一般在自然科学研究中才能遇到，而在行为科学（如语言研究）中，更常见的是称名变量、顺序变量以及等距变量。

五、量表的转换与选择

量表（或测量水平）之间是可以转换的，但通常的做法是从高等向低等转换，即按照等距量表→顺序量表→称名量表的方向转换，而相反方向的转换往往是不可能的。例如，如果要测量一组学生的智商，我们可以为每个学生打分，这样得到的数据就是等距数据，在此基础上，我们可以把得分排序（从高到低或从低到高），进而得到顺序数据，还可进一步按照该顺序把学生分成高低两个智商组，从而获得称名数据。

显而易见，随着测量水平的降低，数据的信息量也跟着减少。所以在决定采用何种测量水平时，一定要权衡利弊。一般来讲，除非出于某种特别的考虑，应尽量采用信息量大的测量水平。其实，在对语言现象的研究中，当数据属于低测量水平时，却常常把它看作高一级的测量水平，例如，当数据其实为顺序数据时，用数字来取代等级，然后对数字进行计算，这样就使顺序数据看起来像是等距数据。在以后的讨论中，我们将会逐渐明白其中的原因，但是简单地说，这与所选取的统计方法有关。我们前面说过，不同类型的变量或数据需选用不同的统计方法，对于等距变量或比率变量，要使用所谓的“参数检验”，而对于称名变量或顺序变量，则要采用“非参数检验”，相比较而言，前者要比后者“威力”更大。

第五节　总体、样本与随机抽样

一、总体

作为研究对象的任何个体的集合或目标群体叫做总体。例如，要研究中国英语本科二年级学生的认知风格，那么，所有中国英语本科二年级学生就是研究的总体；要调查使用者对某教材的意见，那么所有使用该教材的院校或个人就为该调查的总体，等等。

显然，总体有大小之分，即研究对象的规模不同。有些总体是有限的，而有些总体却是无限大的，至少在理论上如此。例如要研究说出某个句子所花的时间，我们就要记录一个人每次说出该句子所用的时间，由于这种测量可以反复地进行下去，所以记录的时间个数（即研究的总体）就将有无限多个。因此，我们又可将总体分为有限总休与无限总体。

二、样本

为了保证研究结果的可靠性及适用性，即保证研究的“内部有效性”与“外部有效性”，最理想的是把总体中的所有个休都加以研究。如果总体非常小，这是可以做到的。但在大多数情况下往往难以做到，或者没有必要这样做。对于无限总休，这显然是做不到的，而对于较大的总体，虽然可以做到，但无疑需要花费大量的时间、精力与财力。所以，在进行各类研究中，一般的做法是在投入与研究结果的可靠性之间求得一种平衡，那就是在保证研究结果相当可靠的前提下，尽可能降低研究成本。至于什么才算“相当可靠”，则要根据研究的目的以及研究者的经验来碗定。如果把握不住这一平衡，就可能陷入所谓的“收益递减率”(the law of diminishing returns)，即超过了一定限度之后，研究的投入与研究结果可靠性的提高就不成比例了，有时二者甚至会成反比，因为数据量过大，在处理过程中就容易产生误差。

鉴于以上情况，惯常的做法是从总体中抽取一部分个体加以研究，在此基础上再对总休做出推断。所抽取的一组个体称为样本。显然，要对总体作出可靠的有效的推断，所抽取的样本必须具有代表性，或者说必须在其结构等主要特征上接近总体，而不能有太大的偏向性。

三、随机抽样

为了保证样本的代表性，就要采用随机抽样的方法来抽取样本。

所谓随机抽样，并不是说抽取样本的过程是杂乱无章的，而是指总体中的每个个体被抽中的概率是相同的，即有同等的机会在样本中得到休现。

进行抽样之前，首先要定义总休，即碗定总体的范围，碗定包括或不包括哪些个体。这直接关系着样本抽取工作的难易度以及研究结果适用范围等问题。总体范围越小则越易于抽取合适的样本，但据此样本得出的结论的适用范围也就越小；反之，随着总体规模的扩大，抽样难度也随之增大，但研究结果的适用范围也同样随之扩大。例如，我们要调查交际

教学法对我国英语教学的适用性，我们可以选取某个特定的院校层次（如高等学校）、专业层次(如英语专业)、教学层次(如基础阶段)、某个教学班次的学生作为研究对象，也可以研究所有院校的所有专业层次的所有教学层次的所有班次的学生。

碗定总体的范围之后,要碗定抽样框架,即把所定义的(有限)总体中的所有个体逐一列出来，并予以顺序编号（如果能保证抽样时很容易地找出所抽取的个体，也可以不予编号）。

随机抽样的具体方法主要有以下几种：

1. 简单随机抽样

简单随机抽样也可称为单级抽样，以别于多级抽样，包括：

(1) 抽签：把总休中所有个体的编号写在一张张纸条上，然后放在一个容器(如盒子、帽子等)里,充分混合后,按照所碗定的样本容量(即样本规模的大小），从中抽出等量的纸条，再把纸条上的编号与抽样框架加以比较，以确定所对应的个休。

(2) 随机数表：如果总体太大，上述方法就会十分费时费力，这时可采用随机数表(见附表 1)来进行。在该表中 0～9 十个数字已经随机地排列起来,并分成若干组,每组 5 个数字(但这只是为了方便使用,组的大小对实际使用没有任何影响）。

现举例说明如何使用随机数表来抽取样本。

首先要明碗总体的规模，以便碗定随机数表中实际以几个数字为一组；然后在随机数表中从任意一个地方开始，从左到右或从上到下顺序寻找，直到找足所需的样本容量为止。凡等于或低于总体规模的数字（重复出现的除外)即为有效数字,记下来(或用某种符号标出)。最后把选取的数字与抽样框架中所列出的名单对照，以碗定所对应的被试。比如，我们要在 450 人中抽取一个 20 人的样本，该总体的规模就是 450,含有 3 位数，因而随机数表中的数字就要分为 3 位数一组。现假设我们从附表 1 中第 2 行的第 3 组数字(即 46 317)开始,从左到右进行。为方便起见,我们把表中的部分数字复录于下（重新编为 3 位数一组)：

463 *178 480 *386 *056 628 *123 *358 470 *391
777 496 490 625 *252 977 *382 *390 *111 *106
868 645 580 822 557 *232 *141 502 *154 *268
*024 *314 *219 *396 960 *196 202 918 805

(“ * ”号表示被选取的有效数字)

最后被选取的被试为（此时数字应理解为被试的编号）：

178 386 056 123 358 391 252 382 390 111
106 232 141 154 268 024 314 219 396 196

把这些数字与抽样框架（总体的所有个体的编号名单）中的编号加以对比，就可确定所对应的被试。

(3) 系统随机抽样（或准随机抽样）：所要抽取的样本的第一个单位按照真正随机的方法选取，余后单位则等距离抽取（抽取间距的大小视样本的大小而定）。例如，我们要从 1 000 个英语句子中抽取 50 个作为样本加以研究，抽取间距为 1 000/50，即 20，我们先从随机数表中选取一个等于或小于 20 的两位数作为样本的第一个单位，假如第一个数为 18，则以后应抽取的句子数依次为 38,58,78,98等，直到 998(包括 998)。由于第二个以及以后的单位并不是独立于第一个单位而真正随机抽取的，所以严格来讲，所抽取的样本并不是真正的随机样本。

2. 分层随机抽样

该方法是首先确定分层参数或变量，即研究所涉及的自变量或控制变量，据此把总体分成若干部分或层（stratum，也可以把这些层看作次总体），然后在各部分中分别进行简单随机抽样，最后把抽取的分样本合并起来，就得到一个总样本。例如，我们要研究学生的性别(自变量)是否会对外语学习（因变量）产生影响，因此所抽取的样本必须既包括男生也包括女生。这时我们可以先把所定义的总体分为男生和女生两部分，然后再在这两部分中分别随机抽取一定量的男生和女生，组成所要研究的样本。

假如从各部分中抽取的分样本在总样本中所占的比例与各部分在总体中所占的比例相同，那么这种抽样就称为比例分层随机抽样。例如，总体中男、女生各占一半，(总)样本中二者也是如此。反之，如果各部分在总体中所占的比例不相同，而在总样本中所占的比例却相同，这种抽样称为非比例分层随机抽样。

显而易见，随机可以防止抽样的偏向性，而分层则能提高抽样的精确性，因为它能使得总体中各部分的比例分布在样本中得到体现。尤其是比例分层随机抽样，更能保证样本的代表性，因为总体的比例结构在样本中得到了充分的系统的体现，而对于后者，在总体中比例较小的部分在样本中就会得到过分的体现，反之亦然。在这种情况下，如果要用非比例样本的结果来估计总体的特征，就要给不同的层以不同的加权，以保证估计的可靠性。

如上所述，分层可以在一个层面上进行，但也可以在多个层面上逐步

进行，例如，如果研究者同时想了解性别和年龄这两个变量对外语学习的影响，就可以先把总体分成男、女生两个部分，再在这两部分中分别分出两个(或多个)年龄层,然后在这四个层中进行(非)比例随机抽样。分层抽样的一个突出的优点是，可以对总体的各个部分进行比较研究。但是如果分层过多，就会使抽样过程复杂化，同时也使每层的个体数（可以看作次总体）大大减少。

3. 多级抽样

顺名思义，多级抽样就是逐级进行抽样，把每一级所抽取的样本看作下一级的总体。例如，我们要在全国所有英语专业院校（系）的一年级学生中进行学习方法问卷调查，准备选取 300 名学生作为样本，有两种抽样方法：一是把所有学生直接作为总体，进行单级简单随机抽样，可想而知，这虽然不是不可能的，但无疑要牵涉太多的时间、人力和物力，因而可行性较差 另一个选择是先把所有这类院校(系)作为总体,显然其规模要小得多，可以很容易地制定抽样框架。假如有 200 所英语专业院校(系),我们可先采用简单随机的方法抽取 20 个，然后再采用简单随机的方法从这 20 个院校(系)中分别抽取 15 名学生，把由此得来的 20 个分样本合并起来，就得到所需的研究样本。

再譬如,我们想从某个作家的作品(如小说)中抽取 10 000 个词加以研究，显然比较经济快捷的抽样方法是先把该作家的所有作品作为总体，用简单随机的方法从中抽取一部分（例如 5 部),然后再在每一部作品中分别随机抽取一定的页数（例如 20 页),最后再在每一页中分别随机抽取 100 个词,这样就可获得所需的10 000词的样本。

四、总体参数与样本统计量

描述总体的某一特征的数值称为参数，而描述样本的某一特征的数值叫做统计量或统计特征值（有时也称为估计值）。例如，我们从某年级 150 个学生中随机抽取 20 个作为样本，测量其词汇量。如果用该样本的平均词汇量来估计该年级 150 个学生的平均词汇量，这个样本的平均词汇量就叫做统计量，全年级的平均词汇量就是参数。统计学中一般用英文字母表示样本统计量，用希腊字母表示总体参数，例如样本的平均数用 $\bar{X}$(读作“ X 杠 ”)表示，总体的平均数用希腊字母 μ 表示。显然,当总体为有限总体，而样本又与总体一样大时,总体参数与样本统计量实质上是同一个统计指标,而当总体为无限总体,样本小于总体时,二者则不同,这时可以把样本统计量作为总体参数的估计值。

思考与练习

1. 进行一项研究一般需要考虑哪些方面的问题？
2. 什么是“内部有效性”与“外部有效性”二者的关系如何？
3. 统计学有哪两个大部分构成？
4. 描述统计与推断统计的作用各是什么？各包括哪些主要内容？
5. 为什么研究中要考虑设计的问题？进行设计时应主要注意哪些问题？
6. 为什么要区分不同的变量和数据？
7. 什么是自变量和因变量？假如要研究句子所含命题数与句子阅读时间的关系，何为自变量，何为因变量？
8. 什么是称名变量、顺序变量和等距变量？就所提供的信息量而言，哪个最佳？为什么？
9. 如何进行测量水平的转换？数据信息量与测量水平的转换有什么关系？
10. 什么是总体和样本？为什么要研究样本？
11. 什么叫“随机抽样”？它有什么作用？
12. 假如一所学校有 15 个系，每个系有 20 个自然班，每个班有 25 人左右，共 7 300 人。要从中选取一个 100 人的样本，那么：
    (1) 最好用什么方法进行抽样？为什么？请说出具体步骤。
    (2) 如果要保证样本中男、女生各占一半，又如何进行抽样？
    (3) 试用随机数表以简单随机抽样的方式从 7 300 名学生中选取该样本。

第一章

数据的初步整理
——统计图表

同任何研究一样，语言研究的目的是为了探讨和说明问题，以便深入地了解事物或现象的本质及其相互关系，而对数据的统计分析是实现这一目的的重要一环。如前所述，原始实验数据往往杂乱无章，如果不加以适当的整理，大量有份值的信息就会被掩盖起来，同时也无法进行进一步的统计分析，这样的数据是说明不了什么问题的。因而，统计分析的第一项重要工作就是对原始数据进行初步整理、归纳和分类，使其最突出、最重要的特征得以显现出来。

对原始数据进行整理的基本方法之一是编制统计图表。统计表把被说明的事物及有关统计数字分门别类地整齐地表示出来，简洁明了，易于比较分析；统计图则使数据的突出特征具体、形象、直观、生动地展示出来，易于理解，且印象深刻。因而，图表的适当应用可以起到去粗取精、化繁为简的作用。

本章将介绍在语言研究中对数据进行整理压缩的常用图表的编制方法和注意事项。

第一节　范畴型数据的整理

在语言研究中经常要把研究对象(人、反应、语言现象等)按某种标准分成相互排斥的类或范畴（或者根据多种标准交叉分类），这类数据叫做范畴型数据(见第一章"称名变量")。对于范畴型数据的整理，主要是进行分类并计算出每一类的观察次数和相对次数（即在总次数中所占的百分比），最后以表和条线图的形式表示出来。例如，我们从学生的英语作业中收集到90个错误，经分析，其中30个是由汉语干扰造成的，25个是由过度概括造成的，20个是由教学方法不当造成的，15个是由其它原因造成的。此数据可以整理如下表（相对次数也可以加括号放在次数之后）：

表 2.1　英语错误分类表

类　别	汉语干扰	过度概括	教学方法	其它	总计
次　数	30	25	20	15	90
相对次数	33.3	27.8	22.2	16.7	100

该数据也可以用条线图进行更直观的表示：

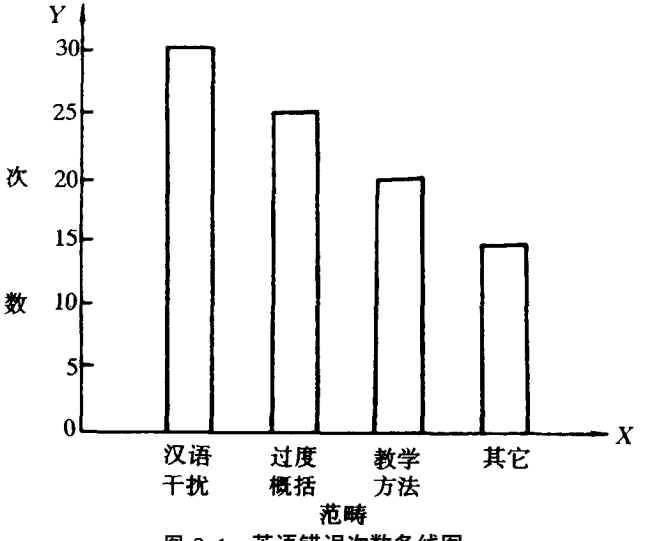

图 2.1　英语错误次数条线图

坐标中横坐标代表范畴或类别，纵坐标代表每个类别的观察次数。

第二节　数值型数据的整理

语言研究中更经常遇到的是数值型数据，譬如考试分数、句子阅读时间、每句单词数等。如果数据量很小(譬如只有几个数值),把它列举出来即可，而不需进行任何整理，但是对于数量较大的数据，则必须利用图表进行初步整理，才能更易看出其中带有规律性的特点，尤其是数据中各数值的分布情况,譬如数据的集中趋势及离中趋势(详细讨论见第三章),即数据的典型数值以及数值之间的差异程度。

一、原始数据

下面一组数据为一篇英语阅读材料中 100 个句子的长度数据（以音

节数表示)：

表 2.2(a)　100 个英语句子的长度值 *

29	40	36	58	20	23	44	18	47	18
22	17	13	12	11	20	21	9	14	6
26	8	17	24	27	9	16	28	12	22
21	56	44	26	35	21	55	17	50	40
48	20	14	12	17	10	25	24	20	18
18	19	20	15	26	23	16	18	12	24
12	40	34	26	13	37	19	22	22	48
17	23	14	16	13	10	19	17	9	16
12	13	9	16	19	19	14	10	11	16
16	15	8	8	15	15	39	44	51	29

* 数据取自笔者的一次阅读研究(Li Shaoshan,1991)

对于这样数值数目较大的数据，如此列举，其用处是很有限的，因为它杂乱无序，难以获得有份值的信息。如果按照数值的大小顺序列举，数据的条理性和清晰度就可以大大提高。

表 2.2(b)　100 个句子的长度（按数值大小排列）

58	29	21	17	13
56	28	20	17	12
55	27	20	16	12
51	26	20	16	12
50	26	20	16	12
48	26	20	16	12
48	26	19	16	12
47	25	19	16	11
44	24	19	16	11
44	24	19	15	10
44	24	19	15	10
40	23	18	15	10
40	23	18	15	9
40	23	18	14	9
39	22	18	14	9
37	22	18	14	9
36	22	17	14	8
35	22	17	13	8
34	21	17	13	8
29	21	17	13	6

表 2.2(b)看起来就清楚多了，稍加分析就可以看出数据的分布情

况,例如,最长句与最短句的长度是多少,二者之间的距离有多大(可粗略表示数据的离散情况），哪些长度的句子出现次数比较多（大体表示数据的集中情况),等等。

二、次数分布表

尽管通过排序，数据的条理性有所提高，但是表 2.2(b)仍然不够简明，不能做到一目了然。

从表中可以看出，数据中数值出现的次数或频率是不同的，有的只出现一次，而大部分是重复出现的，如果把重复出现的数值在表中只列举一次，随后标明其出现的次数，就可以把数据进一步压缩，使其更加条理化。这样的表称为次数分布表。

1. 未分组与分组次数分布表

次数分布是统计学中的一个重要概念，它表示数据的散布情况；而次数分布表则是对数据进行初步整理的重要手段，它能较为直观地表示出数据的分布情况，使人们得以大体上了解数据的平均水平和差异情况等（有关平均水平和差异情况等的更碗切的度量方法，我们将在第三章讨论)。

一般来讲，次数分布表的最左边一列为各个数值，接下来为登记次数，其次为各数值出现的次数。按照统计学中的惯例，在登记次数时一般用 /,//,///,////,~~////~~(分别表示 1,2,3,4,5)作为计数符号，正如同中国人用‘ 正 ’字的笔划表示次数一样。

上述句子长度数据的次数分布表如下：

表 2.2(c)　100 个句子长度的次数分布表

句子音节数	登记次数(句子数)	次数
58	/	1
56	/	1
55	/	1
51	/	1
50	/	1
48	//	2
47	/	1
44	///	3
40	///	3