第 1 章
多元数据和多元统计分析

学习目标

1. 理解多元数据及多元统计分析与一元统计分析的区别。
2. 掌握数据的计量尺度与数据类型。
3. 了解多元统计分析的应用分类。

案例导入

贫困问题是国际社会长期关注的问题。党的十八大以来，我国 14 个集中连片特困地区成为扶贫主战场，实施精准扶贫方略。经过全国各族人民的共同努力和八年的奋进拼搏，我国脱贫攻坚战取得了全面胜利，现行标准下 9899 万农村贫困人口全部脱贫，832 个贫困县全部摘帽，12.8 万个贫困村全部出列，区域性整体贫困得到解决，完成了消除绝对贫困的艰巨任务，创造了又一个彪炳史册的人间奇迹！

这里所谓的现行标准涉及 6 个维度，即人均纯收入达标、“两不愁”“三保障”。人均纯收入的计算综合考虑了物价水平和其他因素，以 2300 元（2010 年不变价）为基准进行调节。“两不愁”是指贫困人口不愁吃、不愁穿，“三保障”指义务教育、基本医疗、安全住房有保障。这说明我国精准扶贫阶段对贫困人口或贫困户不是仅依据收入单一维度进行界定，而是从多个维度进行考量，在贫困治理上也不是仅消除收入贫困，而是综合考虑了贫困的多维特征。

扩展阅读1-1

习主席在全国脱贫攻坚总结表彰大会上的讲话

1.1 多元数据认知

1.1.1 多元数据的概念

对任何一个现实问题要转化为一个统计问题，首要的工作是要对其特征进行刻画，一

般我们采用随机变量，多个特征采用多个随机变量，如$X_1, X_2, \cdots, X_p$。随机变量一般是抽象的，当随机变量描述的是经济变量时则会有具体的意义，如宏观经济指标 GDP、社会商品零售总额、固定资产投资额、消费、个人可支配收入等，这些指标有其概念、单位、核算方法等。如果我们仅考虑问题的单一特征（一个变量），则是一元统计问题，若要同时考虑多个特征，且要体现多个经济变量（指标）之间的相关性，例如，个人消费与其可支配收入正相关等，则我们不但要分析每一个变量，还要分析它们之间的相依程度，这就需要对一元统计分析方法进行拓展，即同时对诸多变量（如$X_1, X_2, \cdots, X_p$）进行分析，这就是多元统计分析分析问题的构思。

为了对诸变量进行统计分析（描述性的或推断性的），我们需要对其进行重复观察，即通过大量的重复观察结果（数据）捕捉诸变量及其之间的规律。对于具有p个变量的多元统计问题，我们可以采用矩阵工具对其观察数据进行展示，如如下的矩阵X。

$$X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

其中，x_{ij}是第i个个体的第j个变量的观测值；n是观测的次数（或称为观测的个体数，样本容量）；p是变量的个数。如果有几个不同的个体归属于s个不同的群体，则可设s是取值为 1，2，…的分类变量以区分这些群体。

1.1.2 数据计量的尺度与数据类型

数据是对对象进行计量的结果，不同的计量尺度会产生不同的结果（数据）。计量尺度有四种，即定类尺度、定序尺度、定距尺度和定比尺度。

1. 定类尺度

定类尺度亦称为名义尺度（nominal scale），它是测量的最低水平，最常用于定性而非定量的变量。例如，跑鞋的牌子、水果的种类、音乐的种类、月份、宗教信仰、眼睛的颜色等。当使用定类尺度进行计量时，变量被划分为几个类别（categories），通过确定对象所属的类别来“测量”对象。因此，用定类尺度进行测量实际上相当于对对象进行分类，并给出它们所属类别的名称，这也是将其称为名义尺度的缘由。

定类尺度计量层次最低，具有如下特征：

- 对事物进行平行的分类。
- 各类别可以指定数字代码表示。
- 使用时必须符合类别穷尽和互斥的要求。
- 数据表现为“类别”。
- 具有“=”或“≠”的数学特性。

如果我们对一个变量使用定类尺度进行计量，则称这个变量为定类变量，计量（测量）结果称为定类数据。

2. 定序尺度

定序尺度（ordinal scale）具有相对较低的计量层次，但测量水平高于定类尺度，它具有相对低层次的数量特性。例如，社会阶层、对健康的自我感知 (从Ⅰ到Ⅴ编码)、教育水平 (没有接受过学校教育、小学、中学或高等教育) 等。定序尺度具有如下特征：

- 对事物分类的同时给出各类别的顺序。
- 比定类尺度更精确。
- 未测量出类别之间的准确差值。
- 数据表现为“类别”，但有序。
- 具有“>”或“<”的数学特性。

如果我们对一个变量使用定序尺度进行计量，则称这个变量为定序变量，计量（测量）结果称为定序数据。

3. 定距尺度

定距尺度（interval scale）比定序尺度有更高的测量水平，它具有数量的特性且相邻单位等间隔，但没有绝对零点，即零点的位置可任意选择。因此，定距尺度具有定序尺度的性质，且相邻单位之间的间隔相等。术语“相邻单位等间隔”意指相邻单位上变量被测量的值是一样的。

因为间隔尺度具有相邻单位之间变量计量（测量）值相等的性质，所以相同间隔之间的差异也表示变量的测量值具有相同的差异。例如，使用摄氏温度计或华氏温度计测量温度。在某些情况下，类似抑郁、焦虑或智力的测量，当实际难以计量时（实际上也确实难以对其进行准确的测度），则可使用间隔尺度对这些变量进行计量。

如果我们对一个变量使用间隔尺度进行计量，则称这个变量为定距变量，计量（测量）结果称为定距数据，这些数据为数值型数据。

扩展阅读1-2

流调中心
抑郁量表

4. 定比尺度

定比尺度（ratio scale）是最高水平的计量尺度，对这种尺度测量的数据可以分析其相对大小及它们之间的差异，其零点的位置是固定的。例如，年龄、从任何固定事件起算的时间、事件发生的频率、体重、长度等。

如果我们对一个变量使用定比尺度进行计量，则称这个变量为定比变量，计量（测量）结果称为定比数据，这些数据为数值型数据。

在统计学中，我们称定类数据和定序数据为品质型数据（类别数据或定性数据），定距数据和定比数据为数值型数据。不同类型的数据需要不同的统计分析方法，一般适合分析低水平尺度数据的方法也可用于分析高水平尺度数据，反之不一定成立。

例 1.1　对 6 个变量进行 10 次观测（10 个个体）的结果，如表 1-1 所示。表 1-1 可以看作是一个 10×6 阶的数据矩阵，相当于对 6 个变量观测了 10 次。其中，“性别”变量、“忧郁”变量为定类变量，“健康状况”变量为定序变量，“IQ”变量为定距变量，“年

龄”变量、“体重”变量为定比变量。

表 1-1 中的定性信息可采用数值代码表示。例如，我们可以定义定类变量“性别”的取值为：男性 =1，女性 =2；定序变量“健康状况”取值用 1~5 表示，取值为 5 表示很好，取值为 1 表示很差等。但是，这里需要注意的是这些相同的数字代码（如 1）表达完全不同的信息，其与测量的尺度有关。

表 1-1 的另一个特征是它包含缺失值（missing values）即未知（not known，NK）。缺失值的产生有各种各样的原因，对变量的观察值为什么出现缺失进行分析，对研究来说很重要。缺失值会导致本书中介绍的许多分析方法出现问题，缺失值越多问题相对越严重。尽管有很多方法可以处理缺失数据的问题 (有效的和无效的)，但这些方法的讨论超出了本书的范围。然而，一种普遍适用的方法是根据未缺失数据的信息估计缺失值，这种插补方法既有简单的使用非缺失数据的平均值代替缺失值，又有复杂的借助于数据随机性的多重插补（填补）方法（multiple imputation）。

扩展阅读1-3

缺失数据处理相关文献

表 1-1　含有 6 个变量 10 次观测的数据集

个体编号	性别	年龄 / 岁	IQ	忧郁症	健康状况	体重 / 千克
1	男	21	120	是	很好	68
2	男	43	NK	否	很好	72.5
3	男	22	135	否	一般	61.2
4	男	86	150	否	很好	63.5
5	男	60	92	是	较好	49.9
6	女	16	130	是	较好	49.9
7	女	NK	150	是	很好	54.4
8	女	43	NK	是	一般	54.4
9	女	22	84	否	一般	47.6
10	女	80	70	否	较好	45.4

1.2　多元统计分析

1.2.1　多元统计分析认知

多元统计分析是分析多维数据的理论与方法，随着现实问题的需要与数据收集、储存技术的发展，多元统计分析方法也与时俱进，在不断地拓展与发展变化。但是，如果想对多元统计分析给出一个准确的定义一般非常困难，我们很难建立一个既被广泛接受又能对其方法技术进行合适逻辑归类的分类框架。鉴于此，本书从研究现实问题实际需要的视角，

通过归类科学研究的目标以体现多元统计分析的方法与应用。

科学研究的目标或实际需要，特别是经济、管理、社会、教育、心理、医学等领域，一般包括以下几个方面：

- 数据减化或结构简化（data reduction or structural simplification）。在不牺牲有价值信息的情况下，使用尽可能简单的方式对感兴趣的现象开展研究，以期使解释更加容易。
- 分类和聚类（sorting and grouping）。根据测量数据的特征，将“相似的”对象或变量进行归类，或构建规则以将新对象归于事先定义好的类中。
- 研究变量之间的相依关系（investigation of the dependence among variables）。研究者一般会对变量之间的关系感兴趣，经常需要回答是否所有的变量相互独立；还是一个或多个变量依赖于其他的一些变量，如果是这样，原因是什么。
- 预测（prediction）。基于某些变量的观测数据，确立变量之间的关系，以对感兴趣的一个或多个变量的值进行预测。
- 假设的构建与检验（hypothesis construction and testing）。利用多元总体的参数构建统计假设，并对其进行检验，以对问题的假设或竞争性论点进行实证分析。

1.2.2　多元统计分析分类应用简例

为了体现实际问题的分析需要和科学研究的目标，下面通过问题举例呈现多元统计分析方法的应用，学员可在此基础上举一反三，思考研究问题与多元统计分析方法的对应关系与选择。

1. **数据简化或结构约化简例**

- 使用几个与癌症患者放疗反应有关的变量数据，构建一个测度方法以测量患者接受放疗的疗效。
- 基于许多国家运动员的竞赛成绩数据，构建一个指数以测量男女运动员的技术水平。
- 利用高级扫描仪收集的多谱图像数据，在二维平面上呈现海岸线的图像。

2. **分类和聚类简例**

- 基于若干人体生理变量的测量值，开发出一种甄别方法，以区别嗜酒者和非嗜酒者。
- 税务部门使用从纳税申报表中收集的数据，将纳税人分为需要审计和不需要审计两个类别。
- 基于反映不同类型国家发展水平的若干变量数据，判断某一国家的发展方式应该采取粗放型、集约型、粗放集约型、集约粗放型四种发展方式中的哪一种。

3. **变量之间相依关系简例**

- 基于几个变量的数据识别影响聘用外部顾问的企业成功的因素。
- 对一些与公司环境和公司组织有关的变量进行测量，并基此解释为什么有些公司的产品具有创新性，而有些公司的产品不具有创新性。

- 基于公司高管的风险倾向与其社会经济特征之间的关系，评估高管的风险行为与其绩效之间的关系。

4. **预测简例**

- 利用学生的测试分数与体现其高中、大学表现的若干个变量之间的联系，预测学生大学期间的表现。
- 基于若干个会计和财务变量识别财产保险者潜在的破产状况。

5. **假设的构建与检验简例**

- 基于若干与污染有关的变量数据，以确定大城市的污染水平在一周内大概相同，还是在工作日和周末之间存在明显的差异。
- 基于一些与职业结构差异有关的变量数据，验证两种相互竞争的社会学观点的正确性。
- 基于一些变量的数据，判断新兴工业化国家不同类型的企业是否表现出不同的创新模式。

练　习　题

1. 数据的计量尺度包括哪几种？如何进行区分？
2. 多元统计分析应用主要包括哪些方面？

第1章 即测即练