·科研方法专题·

定性资料的数据结构与分析方法概述

李长平[1,2],胡良平[2,3*]

(1. 天津医科大学公共卫生学院卫生统计学教研室,天津 300070;
2. 世界中医药学会联合会临床科研统计学专业委员会,北京 100029;
3. 军事科学院研究生院,北京 100850
*通信作者:胡良平,E-mail:lphu812@sina.com)

【摘要】 本文的目的是全面介绍定性资料的数据结构与统计分析方法。从4种不同视角来划分定性资料的数据结构:研究类型与设计类型、定性结果变量的表现、表达资料的形式以及观测点个数。共呈现了7类最常用的处理定性资料的统计分析方法,即广义差异性分析、相关分析、关联分析、聚类分析、判别分析、对数线性模型分析和回归分析。

【关键词】 定性资料;数据结构;列联表;数据库;差异性分析;回归分析
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2019.04.001

Overview of the data structure and analysis methods of qualitative data

Li Changping[1,2], *Hu Liangping*[2,3*]

(1. *Department of Health Statistics, School of Public Health, Tianjin Medical University, Tianjin* 300070, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*;
3. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*
**Corresponding author: Hu Liangping, E-mail: lphu812@sina.com*)

【Abstract】 The purpose of this paper was to introduce the data structure and statistical analysis methods of qualitative data. The data structure of qualitative data was divided from four different perspectives, the study type and the design type, the expression of the qualitative results, the form of the expression data and the number of observational points. The seven most commonly used statistical analysis approaches for dealing with qualitative data were as follows: generalized difference analysis, correlation analysis, association analysis, cluster analysis, discriminant analysis, logarithmic linear model analysis and regression analysis.

【Keywords】 Qualitative data; Data structure; Contingency table; Database; Difference analysis; Regression analysis

众所周知,统计资料主要由定量资料与定性资料两种类型组成。尽管原因变量和结果变量也通常涉及这两类资料,但每当人们提起定量资料或定性资料统计分析时,一般是特指结果变量的性质。之前的科研方法专题文章所介绍的各种回归分析方法对应的结果变量的资料类型基本上都属于定量资料(包括计量资料和计数资料)。从本期开始,将介绍结果变量为定性资料的一类回归分析方法,其中,最常用的为多重 Logistic 回归分析。

1 基本概念

1.1 定性资料的种类

1.1.1 概述

定性变量及其取值被称为定性资料,定性变量的取值通常为一些名称或等级,反映了被描述的对象在某些方面的“属性”或“程度”。例如,一次试验的结果有“成功”与“失败”两种可能的表现;又如,某疾病患者治疗结果有5种可能的表现,即“治愈”“显效”“好转”“无效”和“死亡”。

1.1.2 二值资料

只有两个可能取值的定性变量被称为“二值变量”。二值变量及其在全部受试对象上的取值被称为“二值资料”。例如,设SEX代表受试者的性别(通常被视为原因变量),则它就是一个“二值的定性变量”。因为它只有两个可能的取值,即“男”或“女”;再例如,设X代表试验结果(通常被视为结果变量),若X仅有“成功”与“失败”两个可能的结果时,它也是一个“二值的定性变量”。

1.1.3 多值有序资料

多于两个可能取值且取值之间存在程度上差

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

异的定性变量被称为“多值有序变量”。多值有序变量及其在全部受试对象上的取值被称为“多值有序资料”。例如，设 RANK 代表某患者的疾病严重程度（通常被视为原因变量），则它可能有多个不同等级的取值：“极严重”“严重”“中等”“一般”和“弱”5 种可能的结果；再例如，设 Y 代表某疾病患者的治疗结果，若 Y 可取“治愈”“显效”“好转”“无效”和“死亡”5 种可能的结果，则它们都是“多值有序变量”。

1.1.4 多值名义资料

多于两个可能取值且取值之间不存在程度上差异的定性变量被称为“多值名义变量”。多值名义变量及其在全部受试对象上的取值被称为“多值名义资料”。例如，设 ZY 代表受试对象的“职业”（通常被视为原因变量），其取值可能有如下几种：工人、农民、商人、知识分子、军人、自由职业，则它就是一个“多值名义自变量”；再例如，设 XXLY 代表“消息来源”（通常被视为结果变量），其取值可能有如下几种：广播、电视、网络、宣传栏、杂志和报纸，则它就是一个“多值名义结果变量”。

1.2 定性变量的赋值

1.2.1 二值变量的赋值

在进行回归分析时，统计学上要求用具体的“数值”来取代二值变量的两个不同状态。例如，用 SEX 代表性别时，可用“SEX=0”代表男性、“SEX=1”代表女性；反过来赋值也可以。事实上，给它赋任何两个数值（通常取 0 或正整数）都是可以的。

1.2.2 多值有序变量的赋值

在很多文献中，人们常用连续的几个自然数分别代表一个多值有序变量的不同状态。但这样的赋值方法可能存在一定的缺陷。例如，设 RANK 代表某患者的疾病严重程度时，若给 RANK 赋值 1、2、3、4、5，这就意味着疾病的不同等级对患者治疗结果的影响“等效”，这显然不符合实际情况。正确的做法是将其视为“多值名义变量”，采取与其相同的赋值方法，例如进行哑变量变换、优化计分变换或其他变量变换[1-4]。

1.2.3 多值名义变量的赋值

在大多数文献中，人们一般都会对多值名义变量进行哑变量变换。此外，还有多种其他变量变换方法，如优化计分变换、单调变换或变量扩展变换等[1-4]。

1.3 定性资料的数据结构

1.3.1 从研究类型与设计类型角度来划分

大多数科研课题的研究类型可分为以下三种：调查研究（一般包括“横断面调查研究”和“纵向追踪调查研究”）、试验研究和文献研究。在前述的三种研究类型中，再基于对受试对象的分组情况来划分时，常可分为以下几种设计类型：单因素设计[包括单组设计、配对设计、两组（或成组）设计、多组设计]和多因素设计（大多数情况下为多因素析因设计，例如本文中的表 1 资料，可以被视为“三因素析因设计一个二值结果变量的定性资料”）。然而，从前述两种角度来划分定性资料，似乎都不能很准确地刻划出一个科研课题中全部定性资料的“实质”。

1.3.2 从定性结果变量的表现角度来划分

在一个科研课题的全部统计资料中，可以按性质将结果变量划分为两大类：定量结果变量和定性结果变量，本文将着重讨论第二类。在第二类中，无论原因变量是定量的、定性的或混合的，都可统称为“定性资料”。当同时考察的定性结果变量的个数多于一个时，称为“多元定性资料”，这种场合的统计分析需求并不多见；人们最常需分析的是“一元定性资料”，即每次统计分析仅涉及“一个定性结果变量”。此时，根据同时分析的原因变量的个数，可分为“一元定性资料的单因素分析问题”与“一元定性资料的多因素分析问题”。然而，若从定性结果变量的表现来划分定性资料，可分为“二值结果变量的定性资料”“多值有序结果变量的定性资料”和“多值名义结果变量的定性资料”三类。

1.3.3 从表达资料的形式角度来划分

前面只能通过文字描述的方法来呈现定性资料的数据结构，很不直观。事实上，最简单、最直观的方法是利用“列联表”和/或“数据库”两种形式来呈现定性资料。

1.3.3.1 列联表的适用场合与呈现形式

在有一个或多个定性原因变量、一个定性结果变量的定性资料中，当所有定性变量（若包含有定量原因变量，需先将其转化为定性原因变量）的不同水平组合条件下的个体数目≥2 时，可用简洁的形

式呈现全部资料，此形式就被称为“列联表”。见例1、例2、例3。

【例1】二值结果变量的实例：探讨儿茶酚胺含量与冠心病发病与否之间关系的病例对照研究资料。其中，冠心病发病与否是一个“二值结果变量”；儿茶酚胺含量、年龄分层和ECG诊断正常与否是三个“二值原因变量”。见表1。

表1 儿茶酚胺含量等因素对冠心病发病与否影响情况的调查结果

年龄分层	ECG诊断正常与否	儿茶酚胺含量是否高	病例数	对照数
<55岁	是	是	11	17
		否	27	357
	否	是	20	24
		否	25	62
≥55岁	是	是	27	40
		否	25	117
	否	是	30	54
		否	15	37

注：“病例数”代表结果变量的“阳性水平”或取值为“发病”；“对照数”代表结果变量的“阴性水平”或取值为“未发病”

【例2】多值有序结果变量的实例：某课题组欲评价如意金黄散外涂治疗不同程度的化疗药物性静脉炎的临床疗效，为短期内获得足够样本，由甲、乙两个医院选取白血病住院患者化疗静脉给药而导致静脉炎的患者共250例，其中甲医院120例，乙医院130例，两医院患者一般情况比较差异无统计学意义（$P>0.05$）。见表2。“疗效”为“多值有序结果变量”；“不同医院”和“不同程度静脉炎”是两个“定性原因变量”。

表2 如意金黄散外涂治疗不同程度的化疗药物性静脉炎及其疗效

不同医院	静脉炎程度	例数		
	疗效：	治愈	有效	无效
甲医院	Ⅰ级	17	0	0
	Ⅱ级	37	3	0
	Ⅲ级	38	8	2
	Ⅳ级	9	3	3
乙医院	Ⅰ级	13	23	10
	Ⅱ级	6	12	10
	Ⅲ级	22	2	12
	Ⅳ级	8	0	12

注：根据WHO化疗毒性分级标准，0级为无痛；Ⅰ级为无痛，但局部发红；Ⅱ级为轻度疼痛，局部发红；Ⅲ级为中度疼痛，局部轻度肿胀，灼热；Ⅳ级为重度顽固性疼痛，中、重度肿胀

【例3】多值名义结果变量的实例：研究乙型肝炎病毒（Hepatitis B virus，HBV）血清亚型的分布特点，在三个不同地区共280例慢性HBV携带者中，应用聚合酶链反应（PCR）扩增和脱氧核糖核酸（DNA）序列分析，确定HBV血清亚型。见表3。“HBV血清亚型”是一个“多值名义结果变量”；而“调查地区”和“性别”是两个“定性的原因变量”。

表3 三个不同地区乙型肝炎病毒血清亚型分布情况

调查地区	性别	例数		
	HBV血清亚型：	adr	adw2	ayr/w
华南	男性	17	31	5
	女性	16	20	3
西南	男性	13	19	3
	女性	8	16	1
华北	男性	59	6	1
	女性	59	2	1

1.3.3.2 数据库的适用场合与呈现形式

当研究者希望将每个受试对象的全部原因变量及结果变量的取值都清晰地呈现出来，特别是当存在定量原因变量且不将其定性化时，常需采用“数据库形式”呈现资料。见例4、例5、例6和例7。

【例4】二值结果变量的实例：血液病患者的资料共包含11个变量：X_1[SEX（性别，1=男，0=女）]、X_2[AGE（年龄）]、X_3[DIAG（诊断，1=CML，0=非CML）]、X_4[PILOT（预处理，1=TBI，0=非TBI）]、X_5[GVHD（1=GNHD Ⅲ~Ⅵ，0=GNHD 0~Ⅱ）]、X_6[IGM（1=IgM阳性，0=IgM阴性）]、X_7[TRANS（移植类型，1=P型，0=B型）]、X_8[YUFANG（是否采取预防，1=预防，0=未预防）]、Y_1[XZ（血症，1=有血症，0=无血症）]、Y_2[JB（疾病，1=患病，0=未患病）]、Y_3[SW（死亡与否，1=死，0=活）]。见表4。Y_1、Y_2、Y_3是“二值结果变量”；X_1、X_3、X_4、X_5、X_6、X_7、X_8是“定性原因变量”；X_2是“定量原因变量”。

【说明】CML代表慢性粒细胞白血病；TBI代表是否照射；GVHD代表移植物抗宿主病；JB代表巨细胞病毒疾病。

【例5】对535例缺血性中风病患者采集临床信息，包括内风证（含15个条目）、内火证（含25个条目）、痰湿证（含16个条目）、血瘀证（含10个条目）、气虚证（含18个条目）和阴虚证（含13个条目）分量表共计97个条目，希望对《中风病证候要素评价量表》中的条目进行筛选研究。见表5。在表5中，从第2列到最后一列共有97列，被称为“条目”，其实就是“症状”。若将“内风证”“内火证”等视为“证候类型”的不同水平，则每种“证候类型”就有数目不等的“症状”来综合反映或呈现。于是，每种“证候类型”下的“多个症状”就可被视为“多元定性资料（适宜采用潜在类别分析）”或被视为“多个一元定性资料（适宜采用项目反应模型分析）”。

表 4 血液病患者的资料

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Y_1	Y_2	Y_3
1	43	1	1	0	0	P	0	0	0	0
0	16	1	1	0	0	P	0	1	0	0
1	30	1	1	1	1	B	0	1	1	1
…	…	…	…	…	…	…	…	…	…	…
1	30	0	1	0	0	P	1	1	0	0
1	42	1	1	1	0	P	1	1	0	0
1	17	0	1	0	0	P	1	1	0	0

表 5 缺血性中风病患者各条目情况

患者编号	近 48 小时内病情加重或波动	目偏不瞬	目珠游动	瞳神异常	口噤	项强	手足或下颌颤动	抽搐	…	数脉
1	0	0	0	0	0	0	0	0	…	0
2	0	0	0	0	0	1	1	1	…	0
3	0	0	0	0	0	0	0	0	…	0
4	0	0	0	0	0	0	0	0	…	0
5	0	0	0	0	0	0	0	0	…	0
6	0	0	0	0	0	0	0	0	…	0
7	0	0	0	0	0	0	0	0	…	0
8	0	0	0	0	0	0	0	0	…	0
9	0	0	0	0	0	0	0	0	…	0
…	…	…	…	…	…	…	…	…	…	…
535	0	0	0	0	0	0	0	0	…	0

【例 6】多值有序结果变量的实例：一项研究调查本科生是否决定申请读研究生的影响因素，结果为申请意愿（apply）（0=不愿意，1=有意愿，2=非常愿意），考虑的影响因素包括父母亲的受教育程度（pared）（0=父母亲都没有研究生学历，1=父母亲至少有一方有研究生学历）、本科就读院校是公立的还是私立的（public）（0=私立，1=公立）、当前的平均成绩（gpa）。见表 6。申请意愿（apply）是一个“多值有序结果变量”；父母亲的受教育程度（pared）和本科院校是公立的还是私立的（public）是两个“定性原因变量”，而当前的平均成绩（gpa）是一个“定量原因变量”。

表 6 申请意愿的调查数据

apply	pared	public	gpa
2	0	0	3.26
1	1	0	3.21
…	…	…	…
1	0	0	3.26
2	0	0	3.52

【例 7】多值名义结果变量的实例：在某社区随机抽取 100 名育龄妇女，采用问卷调查方法收集其年龄（A）、生育史（B）、收入（C）及避孕方式（Y）等资料，探讨育龄妇女选择避孕方式的影响因素。将调查对象按年龄分组：＜30 岁组（A=1），30～35 岁组（A=2），＞35 岁组（A=3）；按生育史分组：未生育组（B=0）和已生育组（B=1）；按收入水平分组：低收入组（C=1）和高收入组（C=2）；选择的避孕方式有 IUD（Y=1）、口服避孕药（Y=2）和避孕套（Y=3）3 种。见表 7。避孕方法（Y）为“多值名义结果变量”；年龄（A）、生育史（B）和收入（C）是三个“定性原因变量”（注意：“年龄”原本是计量变量，但此处代表的是年龄分组或分层，转换成为“多值有序变量”了）。

表 7 育龄女性选择避孕方式的影响因素

编 号	年龄（A）	生育史（B）	收入（C）	避孕方法（Y）
1	2	1	2	2
2	2	1	1	1
3	1	1	2	2
4	2	1	2	2
5	2	1	1	2
…	…	…	…	…
97	3	1	1	1
98	2	1	1	1
99	3	1	1	1
100	1	0	2	3

1.3.4 从观测点个数角度来划分

上面的表 1 到表 7 中资料都来自“一个观测时间点”且只含一个定性结果变量（除表 5），也就是

人们常说的“横断面研究设计”下收集的一元定性资料；在实践中，研究人员还可能基于纵向追踪的方法收集一元或多元定性资料（或称为重复测量设计一元或多元定性资料）。见例 8、例 9、例 10。

【例 8】以血瘀证候为例，研究者对 993 例缺血性中风病患者在第 1、7、14、28 和 90 天分别观测血瘀证候。其中，年龄、性别（0 为男性，1 为女性）、起病形式（0 表示渐进加重，1 表示即刻达到高峰）和发病距就诊时间（1 表示＜3 小时，2 表示 3～6 小时，3 表示＞6 小时）是临床医生给出的在专业上对证候影响较大的因素。患者在各时间点上血瘀情况见表 8，其他 5 个证候（内风、内火、痰湿、气虚和阴虚）在各时间点上的情况均可整理成与表 8 相同的格式，此处不再一一列出，希望考察对该证候的影响因素，并预测各证候随时间的变化情况。在表 8 中，第 2~5 列为“协变量及其取值”；最后 5 列为在 5 个时间点上对每位患者的重复测量结果（即在每个时间点上，“是否出现血瘀”都是一个“二值结果变量”）。

表 8 缺血性中风病患者在各时间点上出现血瘀与否的观测结果

患者编号	年龄	性别	起病形式	发病距就诊时间	出现血瘀与否				
					时间（天）：	1	7	14	28	90
345	57	1	0	3		1	1	1	1	1
384	74	1	1	2		1	1	1	1	-
3214	68	0	0	3		1	1	1	1	1
8382	85	0	1	3		1	0	0	1	1
9455	67	0	0	2		1	1	1	1	1
9918	69	0	1	1		1	1	1	1	0
11444	71	0	1	3		1	1	1	1	1
12983	76	0	0	2		1	1	1	1	0
…	…	…	…	…		…	…	…	…	…
411443002	67	0	0	3		0	1	1	1	1

注：患者出现血瘀与否是一个二值变量，0 表示未出现此证候，1 表示出现了此证候，缺失以“-”表示

【例 9】在一项关于缺血性中风病住院患者血瘀证候的观察研究中，将患者按性别分为两组（0 为男性，1 为女性）。研究者在第 1、7、14 天三个时间点观测患者血瘀证候情况。血瘀程度由轻到重按 1~4 排序。见表 9。在表 9 中，性别是一个“协变量”；“1、7、14 天”是三个重复测量的时间点；“血瘀证候得分”是一个“多值有序变量”。

表 9 缺血性中风病住院患者在不同时间点血瘀证候情况

编号	性别	血瘀证候得分			编号	性别	血瘀证候得分		
		时间（天）：	1	7	14			1	7	14
1	0		1	2	3	26	1	3	4	3
2	0		4	4	3	27	1	1	2	3
3	0		2	1	3	28	1	3	4	4
…	…		…	…	…	…	…	…	…	…
23	0		3	2	2	48	1	2	2	1
24	0		2	2	2	49	1	2	2	4
25	0		4	2	2	50	1	4	4	4

【例 10】研究者对 993 例缺血性中风病患者在第 1、7、14、28 和 90 天分别观测内风（Y_1）、内火（Y_2）、痰湿（Y_3）、血瘀（Y_4）、气虚（Y_5）和阴虚（Y_6）这 6 个证候。患者各时间点证候情况（性别：0 为男性，1 为女性）为“六元二值定性资料”，具体数据参见文献[5]P_{215} 表 11-1，因篇幅所限，此处从略。

2 定性资料的常用统计分析方法介绍

2.1 广义差异性分析

当结果变量为定性变量且原因变量为一个或两个时，最常选用“广义差异性分析”。如二维列联表资料独立性 χ^2 检验和 Fisher's 精确检验、二维列联表资料线性趋势 χ^2 检验、秩和检验、Kappa 检验（也叫做一致性检验）、对称性检验、三维列联表资料的

CMH 校正χ^2检验以及 CMH 校正秩和检验等。

2.2 相关分析

对于非配对设计的某些二元定性资料，可采用 Spearman 秩相关分析；而对于配对设计扩大形式的某些二元定性资料，可采用 Kendall's Tau-b 秩相关分析。

2.3 关联分析

对于二维列联表资料，可采取 shannon 信息量分析法或定性资料对应分析法进行关联分析。

2.4 聚类分析

对于多元定性资料，可以采取潜在类别分析，即定性资料样品聚类分析（合适的数据结构参见前文例 5，将每种“证候类型”的多个“症状”视为一个“多元定性资料”）。对于重复测量设计多元定性资料，可以采用潜在转移模型分析，即按不同时间点重复测定的多元定性资料对受试对象进行样品聚类分析（合适的数据结构参见前文例 10）。

2.5 判别分析

当定性结果变量将全部受试对象严格地划分成几种不同的类型且分析目的是希望基于现有的资料构建一个统计模型，以较高的精准度对一个未知结局的个体进行分类，此时，就需要进行“判别分析”或“分类分析”。例如，在表 1 中，全部受试者被分为两组：冠心病患者组和正常组。基于这些数据构建一个“统计模型”，若有一些前来就诊的受试者，将根据就诊者的“儿茶酚胺水平”“ECG 诊断结果”和“年龄”取值代入该“统计模型”进行计算，就能较精准地判定每一位就诊者属于“冠心病患者”或“正常人”。所构建的“统计模型”就被称为“判别函数式”。在统计学上，可依据多种不同的计算原理来构造判别函数式，从而，也就有多种不同的判别分析方法。

2.6 对数线性模型分析

在分析 m（≥3）维列联表资料时，若将各网格上的频数取对数变换后视为“定量结果变量 Y 的取值”，并将全部原因变量和定性结果变量都视为“自变量”，研究结果变量 Y 依赖全部自变量及其各级交互作用项的依赖关系，就称为“对数线性模型”[5]。

此分析方法将“多因素方差分析”与“多重线性回归分析”有机地结合在一起，以表 3 为例，此法的基本步骤是：第一步，将表 3 中的两个原因变量和一个结果变量视为一个“3×2×3 的析因设计的框架”（相当于有 18 种水平组合，即 18 个不同的试验条件）；第二步，将此列联表中的 18 个“网格”视为 18 个试验条件，假定在每个试验条件下仅做了一次试验，并进一步假定有一个“定量指标”，其取值为该条件下的频数的对数值（注意：这里所说的“定量指标”是没有真实专业含义的）；第三步，基于表 3 中三个定性变量（假设为 A、B、C）产生出全部各级交互作用项（包括 A×B、A×C、B×C 和 A×B×C），连同 A、B、C，共有 7 个“自变量”；第四步，将原本属于“三维列联表资料的差异性分析”问题转化成为一个具有 7 个自变量的“回归分析”问题。当原因变量较多时，可能的对数线性模型的个数（指包含不同自变量组合）就会很多、非常复杂，并且对分析结果的解释难度很大，故在实践中应慎用此法。

2.7 回归分析

2.7.1 一水平回归分析

在通常情况下，进行回归分析时隐含着一个“假定”，即样本中的全部个体来自同一个总体。具体地说，因变量与自变量之间的依赖关系是相对固定的，一旦被确定下来，在现有各自变量的定义域范围内，基于自变量的取值去预测因变量的取值，其精准度是非常高的。于是，统计学家把此类回归分析称为“一水平一元回归分析”，简称为“回归分析”。

2.7.2 多水平回归分析

事实上，在有些场合下，前述提及的隐含“假定”并不成立，即样本中的全部个体来自若干个彼此不同的总体。具体地说，在不同总体之间，因变量与自变量之间的依赖关系是不完全相同的。例如，研究我国正常成年人体重随身高变化而变化的依赖关系时，基于北方地区获得的样本与基于南方地区获得的样本可能明显不同；再例如，研究我国高中学生毕业成绩与入学成绩之间的依赖关系时，省市之间可能有差别、同一座城市中的不同中学之间可能有差别、甚至同一所中学同一个年级的不同班级之间也有差别。这意味着：回归模

型中的回归系数不再是“常数”，而是一个“变数”，即所谓的“变系数回归模型”或“随机系数回归模型”或“多水平回归模型”[6-9]，其对应的回归分析也就被称为“多水平回归分析”了。

2.7.3 一水平与多水平多重 Logistic 回归模型

若样本来自一个总体，涉及到的原因变量有多个且结果变量为定性变量，就需要拟合“一水平多重 Logistic 回归模型”；反之，就需要拟合“多水平多重 Logistic 回归模型”。若样本来自一个“具有 K 个层次的分层因素”的 K 个总体，其对应的模型就叫做“二水平多重 Logistic 回归模型”[注意：每层中被抽取的多个个体被称为“一水平”上的“观察单位”（简称“单位”，也叫做“水平 1 上的单位”）、“K 个层次”被称为“二水平”上的“单位”，也叫做“水平 2 上的单位”，其“单位”数目就是 K]；若样本来自两个嵌套的“分层因素”的多个总体，其对应的模型就叫做“三水平多重 Logistic 回归模型”[假设：具有 M 个层次的分层因素 A 下面嵌套着具有 K 个层次的分层因素 B，而分层因素 B 的各层下面又嵌套着多个“受试对象”，此时，分层因素 B 的每一层中的“受试对象”被称为“一水平”上的“单位”（简称为“个体水平”上的“单位”或“水平 1”上的“单位”，其“单位”数为 n_i，i=1，2，…，K）；而分层因素 B 中的各层被称为“二水平”上的“单位”（简称为“因素 B 水平”上的“单位”或“水平 2”上的“单位”，其“单位”数为 K）；最高层 A 中的各层被称为“三水平”上的“单位”（简称为“因素 A 水平”上的“单位”或“水平 3 上”的“单位”，其“单位”数为 M）]；以此类推。

2.7.4 Logistic 回归模型之前的“限定词”

一个“Logistic 回归模型”受控于四个要素，其一，设计类型，即安排受试对象的方法，通常分为“非配对设计（即非条件）”与“配对设计（即条件）”；其二，受试对象取自几个“分层因素”，若有（M-1）个分层因素，就被称为“M 水平（因为还有一个‘个体水平’）”；其三，定性结果变量的具体表现，即有“二值结果变量”“多值有序结果变量”和“多值名义结果变量”三种；其四，同时考察的自变量的个数 k（k≥2 被称为“多重”，也包括派生的新自变量，例如某些自变量的平方项、立方项、交叉乘积项等）。于是，一个 Logistic 回归模型之前可能有 4 个“限定词”，例如：配对设计（或条件）二值资料多水平多重 Logistic 回归模型、非配对设计（或非条件）多值有序资料多水平多重 Logistic 回归模型。文献中常见的“Logistic 回归模型”的详细名称为“非配对设计二值资料一水平多重 Logistic 回归模型”，其简化表述为“多重 Logistic 回归模型”。

值得一提的是：项目反应模型可以被视为一种特殊的“一水平 Logistic 回归模型”（合适的数据结构参见前面的例 5，将每种“证候类型”的多个“症状”视为“多个一元定性资料”，它基于“每个症状”拟合出一个 Logistic 回归模型，可以是“单参数”“双参数”或“三参数”的模型）[5]。

3 讨 论

3.1 “多水平”的双重含义

在常规的统计学中，经常会提及“单因素‘多水平’试验设计及其定量资料的方差分析”；而在较复杂的回归分析中，却出现了“‘多水平’多重线性回归模型”或“‘多水平’多重 Logistic 回归模型”。由此可知，“多水平”一词有双重含义。

在“单因素‘多水平’试验设计及其定量资料的方差分析”中，“多水平”特指试验因素有多个具体“表现”或“取值”。例如，在某项试验研究中，拟考察某种药物“剂量”对疗效的影响情况，若取该药物的剂量分别为“10 mg”“20 mg”“30 mg”和“40 mg”形成 4 个试验组，各组均进行 n=10 次独立重复试验，就称为“单因素 4 水平设计”；而在以“省”“市”“县”“学校”和“某年级的班级”为分层因素的调查研究中，考察中学生的毕业成绩依赖入学成绩、性别、年龄等协变量变化而变化的关系时，就被称为“多水平多重线性回归模型”。这里的“多水平”是指对“中学生”进行划分的“分层因素”的“个数”再加上“个体水平”。在前例中，就是“5+1=6”，即需要构建“六水平多重线性回归模型”。值得注意的是：所有的“分层因素”应具有“嵌套”关系，显然，“省”下面嵌套着“市”，而“市”下面嵌套着“县”，最后一层为“班级”，而每个班级里部分被抽取的学生彼此之间形成一个“个体水平”上的最小“单位”。

3.2 定性资料多种统计分析方法的合理选用

在本文的第 2 部分中，介绍了 7 类常用的定性资料统计分析方法，其中，“广义差异性分析”和“回归分析”的使用频率最高。事实上，前者是处

理单因素或两因素一元定性资料的最常用方法；后者是筛选影响因素（即处理多因素一元定性资料）的重要且有效的常用统计分析方法。回归分析方法之所以被高频使用，首先，其他很多统计分析方法几乎都可以被视为是回归分析方法的“特例”；其次，回归分析方法能回答研究人员最为关心的很多问题，如筛选出对结果变量贡献最大的自变量或影响因素、用一个真实的统计模型去刻划一个错综复杂的系统中变量之间的依赖关系、可依据自变量的取值并以一定的置信度去预测因变量的取值。

参考文献

[1] 胡良平. 提高回归模型拟合优度的策略(I)——哑变量变换与其他变量变换[J]. 四川精神卫生, 2019, 32(1): 1-8.

[2] 胡良平. 提高回归模型拟合优度的策略(II)——算术均值变换与其他变量变换[J]. 四川精神卫生, 2019, 32(1): 9-15.

[3] 胡良平. 提高回归模型拟合优度的策略(III)——校正均值变换与其他变量变换[J]. 四川精神卫生, 2019, 32(1): 16-20.

[4] 胡良平. 提高回归模型拟合优度的策略(IV)——优化计分变换与其他变量变换[J]. 四川精神卫生, 2019, 32(1): 21-28.

[5] 胡良平, 王琪. 定性资料统计分析及应用[M]. 北京: 电子工业出版社, 2016: 1-245.

[6] 杨珉, 李晓松. 医学和公共卫生研究常用多水平统计模型[M]. 北京: 北京大学医学出版社, 2007: 69-91.

[7] 王济川, 谢海义, 姜宝法. 多层统计分析模型——方法与应用[M]. 北京: 高等教育出版社, 2008: 128-149.

[8] 王济川, 谢海义, 费舍尔. 多层统计分析模型: SAS与应用[M]. 北京: 高等教育出版社, 2009: 116-136.

[9] 石磊. 多水平模型及其统计诊断[M]. 北京: 科学出版社, 2008: 27-50.

（收稿日期：2019-08-01）

（本文编辑：陈 霞）

科研方法专题策划人——胡良平教授简介

胡良平，男，1955年8月出生，教授，博士生导师，曾任军事医学科学院研究生部医学统计学教研室主任和生物医学统计学咨询中心主任、国际一般系统论研究会中国分会概率统计系统专业理事会常务理事和北京大学口腔医学院客座教授；现任世界中医药学会联合会临床科研统计学专业委员会会长、中国生物医学统计学会副会长，《中华医学杂志》等10余种杂志编委和国家食品药品监督管理局评审专家。主编统计学专著45部，参编统计学专著10部；发表第一作者学术论文260余篇，发表合作论文130余篇，获军队科技成果和省部级科技成果多项；参加并完成三项国家标准的撰写工作；参加三项国家科技重大专项课题研究工作。在从事统计学工作的30年中，为几千名研究生、医学科研人员、临床医生和杂志编辑讲授生物医学统计学，在全国各地作统计学学术报告100余场，举办数十期全国统计学培训班，培养多名统计学专业硕士和博士研究生。近几年来，参加国家级新药和医疗器械项目评审数十项、参加100多项全军重大重点课题的统计学检查工作。归纳并提炼出有利于透过现象看本质的“八性”和“八思维”的统计学思想，独创了逆向统计学教学法和三型理论。擅长于科研课题的研究设计、复杂科研资料的统计分析与SAS实现、各种层次的统计学教学培训和咨询工作。