·讲座·

如何合理选择统计分析方法处理实验资料（VI）

胡良平

编者按

生物医学期刊是宣传和反映生物医学科研与临床研究成果的重要媒体，是培养年轻科研工作者的摇篮，也是一个国家科研实力的重要象征。期刊中学术论文的质量是期刊存在的重要保证，而学术论文质量高低的重要标志之一是科研设计和统计分析质量的高低。本刊在2007年中，拟邀请军事医学科学院生物医学统计学咨询中心主任、博士生导师胡良平教授，以“如何合理选择统计分析方法处理实验资料”为题，撰写6篇文章，每期发表1篇，较系统地介绍在生物医学论文写作中，如何正确地应用医学统计学知识，从而提高学术论文的质量。需要指出的是，论文中统计学应用正确，并不能说明科研课题做得一定正确。广大作者和读者更应高度重视科研工作之前的科研设计的质量。事实上，由一个错误的科研设计产生出来的实验结果，即使其论文写得再漂亮，统计分析方法用得再正确，对于一个国家科技事业的发展和人才培养都是有害无益的。

上期文章介绍了如何选择统计分析方法处理 R × C 表资料和高维列联表资料，本文将介绍如何进行回归分析和相关分析。若资料中涉及两个或多个定量变量，为了考察一个或多个变量随另外一个变量变化的依赖关系时，在统计学上称为回归分析；若希望考察变量之间的相互关系，则应称为相关分析。当用线性方程来描述变量之间的依赖关系时，常称为线性回归分析；当用非线性方程来描述变量之间的依赖关系时，常称为非线性回归分析或曲线拟合。

1 进行回归与相关分析时应把握的要领

（1）拟分析的数据要具有同质性。

（2）以专业知识为依据。

（3）应绘制反映两变量同时变化趋势的散布图。

（4）正确分析散布图，确定应进行直线回归与相关分析，还是应进行某种曲线回归分析。

（5）应对计算的相关系数进行假设检验。

（6）通常认为决定系数 $r^2 \geqslant 0.5$ 才具有一定的实际意义。

2 进行多重回归分析时应把握的要领

当结果变量（常称为应变量）依赖于原因变量（常称为自变量）变化时，研究应变量随多个自变量变化的规律所对应的统计分析方法，称为多重回归分析。进行多重回归分析时应把握的要领如下。

（1）当应变量为近似服从正态分布的随机变量时，常选用多重线性回归分析。

（2）当应变量分别为二值变量、多值有序变量或多值名义变量时，应分别采用一般的多重 logistic 回归分析、有序变量的多重 logistic 回归分析和扩展的多重 logistic 回归分析。

（3）应当有科学的、全面的评价多重回归方程优劣的评价标准。

（4）在进行多重回归分析中，最常犯的错误有：①将性质截然不同的数据混在一起进行多重回归分析；②分析策略错误，仅把那些经单因素分析具有统计学意义的变量纳入多重回归分析；③采取不筛选变量的方法获得多重回归分析结果，以致于结果中仍包含很多无统计学意义的自变量，并据此得出不科学的结论；④仅选用某一种筛选变量的方法（如前进法或后退法），并不能保证回归分析结果是比较理想的。

3 进行简单相关与回归分析时常犯的错误

例1 原文题目：营养低钙与慢性氟中毒大鼠细胞内钙超载相关性的研究。实验方法是将48只大鼠分为2组，1组为常食，1组为低钙饮食，每组再各分为加氟组与未加氟组，分析各组动物肝、肾、脑细胞$[Ca^{2+}]i$水平与血清$i[Ca^{2+}]$、$t[Ca^{2+}]$水平的关系。检测结果见表1、2，分析认为动物肝、肾、脑细胞$[Ca^{2+}]i$水平与血清$i[Ca^{2+}]$、$t[Ca^{2+}]$水平呈负相关（$r = -0.59$）。结论为慢性氟中毒可致机体组织细胞钙超载，低钙可加重氟中毒时的细胞内钙超载，提示细胞内钙超载可能参与指骨症发病机制并起着重要作用。结论可信吗？

辨析与释疑：在进行相关分析时，粗略地说“肝、肾、脑细胞$[Ca^{2+}]i$水平与血清$i[Ca^{2+}]$、$t[Ca^{2+}]$水平呈负相关（$r = -0.59$）”是不对的，因为并未说明是哪个组及哪个器官$[Ca^{2+}]i$水平与血清$i[Ca^{2+}]$、$t[Ca^{2+}]$水平的相关系数为-0.59，违反资料应具有同质性的原则。正确的做法是，对有意义的2组数据先画出散布图，如果有线性相关的关系，

作者单位：100850 北京，军事医学科学院生物医学统计学咨询中心

表 1 大鼠血清钙、磷和碱性磷酸酶测定结果（$\bar{x} \pm s$）

组别	鼠数	i[Ca^{2+}]（mmol/L）	t[Ca^{2+}]（mmol/L）	ALP（U/L）	P^{3+}（mmol/L）
常食对照组	8	1.04 ± 0.18	2.79 ± 0.36	56.25 ± 22.21	1.76 ± 0.15
常食 + 氟组	9	1.13 ± 0.13	2.31 ± 0.26	111.44 ± 45.40	1.96 ± 0.31
低钙对照组	8	1.19 ± 0.06	2.43 ± 0.13	92.25 ± 29.43	1.65 ± 0.23
低钙 + 氟组	8	0.98 ± 0.24*	2.00 ± 0.50*	216.75 ± 84.25**	2.25 ± 0.37**

注：分别与各自对照组比较，*$P < 0.05$，**$P < 0.01$

表 2 大鼠肝、肾、脑细胞中[Ca^{2+}]i 水平变化（$\bar{x} \pm s$）

组别	鼠数	肝细胞	肾细胞	脑细胞
常食对照组	5	86.37 ± 27.55	95.65 ± 8.75	99.20 ± 12.89
尝食 + 氟组	6	124.11 ± 11.35*	127.20 ± 14.40*	136.75 ± 15.91*
低钙对照组	6	103.17 ± 14.29	103.59 ± 11.40	115.26 ± 30.91
低钙 + 氟组	6	151.57 ± 19.61*	162.62 ± 31.41**	158.79 ± 27.65**

注：分别与各自对照组比较，*$P < 0.05$，**$P < 0.01$

再计算相关系数，并且对相关系数进行检验假设。如果有意义且决定系数 $r^2 \geq 0.5$，提示两者相关并具有较大的实用价值；如果两者从散布图上看无线性关系，或者对 $\rho = 0$ 的假设检验无统计学意义，则不能说明两者有线性关系；如果从散布图上看有曲线关系，则应做相应的曲线回归分析。

例 2 原文题目：原发性周围型肺癌对 $^{99}Tc^{m}$- 甲氧基异丁基异腈摄取与多药耐药蛋白表达的相关性研究。原文中表 3 如下，请问：表中存在什么错误？

表 3 Te/N、Td/N、RI% 与三种多药耐药蛋白表达间的相关分析

病理类型	Te/N		Td/N		RI%	
	r	*P*	*r*	*P*	*r*	*P*
腺癌						
P-gp	0.027	0.945	–0.255	0.508	–0.550	0.125
MRP	0.005	0.990	–0.291	0.447	–0.590	0.094
GST π	–0.393	0.295	–0.558	0.119	–0.417	0.264
鳞癌						
P-gp	0.210	0.535	0.044	0.897	–0.634	0.036
MRP	–0.072	0.833	–0.095	0.781	–0.031	0.928
GST π	0.486	0.130	0.557	0.075	0.101	0.768
合计						
P-gp	0.326	0.161	0.010	0.965	–0.683	0.001
MRP	0.157	0.507	–0.062	0.794	–0.449	0.047
GST π	0.161	0.498	0.027	0.909	–0.275	0.241

辨析与释疑：由此表中最后 3 行可知，原作者将患两种癌患者的数据放在一起进行简单相关分析，这是把性质不同的数据强硬地合并在一起进行分析，其结论是不可信的。

4 进行多重回归分析时常犯的错误

例 3 原文题目：妇科恶性肿瘤患者的生存期预测。原作者先用单因素分析方法筛选自变量，即采用单因素方差分析，从初步调查的 19 项临床生化指标中筛选出 9 项对生存时间有显著影响的指标，再对这 9 项指标进行多重 logistic 回归分析，拟合回归方程的过程中采用了“后退法”筛选自变量，最后得到包含“呼吸困难、发热、年龄、KPS 和血尿素氮”5 个自变量的 5 重 logistic 回归方程，其主要结果列在下面的表 4 中。请问：原作者筛选变量的策略正确吗？基于此 5 重 logistic 回归方程得到的结论可信吗？为什么？正确的做法是什么？

表 4 肿瘤患者多因素分析及回归模型

指标	*B* 值	标准误	Wald 值	*P* 值	分值
呼吸困难					
无	0				0
轻度	–1.659	0.860	3.726	0.054	4
重度	–21.746	?	?	?	54
KPS					
≥ 40 分	0				0
30 分	–1.337	1.024	1.707	0.191	3.5
≤ 20 分	–4.104	1.414	12.947	0.000	10
年龄					
≤ 64 岁	0				0
≥ 65 岁	–1.714	0.864	3.931	0.047	4.5
发热					
无	0				0
低热	–0.403	0.870	0.215	0.643	1
高热	–1.518	0.977	2.414	0.120	4
血尿素氮					
正常	0				0
高于正常	–2.007	0.779	6.643	0.010	5
常数	4.424	1.203	13.525	0.000	

辨析与释疑：在分析过程中，原文作者犯了 3 个错误。①筛选变量的策略错误，仅根据单因素分析中有统计学意义（即 $P < 0.05$）的因素建立多重 logistic 回归方程，并不能确保单因素分析中没有统计学意义（即 $P > 0.05$）的因素与其他因素同时存在于多重 logistic 回归方程中时将永远无统计学意义。也就是说，单独作用小而与某些因素之间存在交互作用的因素将没有机会被选入多重回归方程之中去。②仅采取“后退法”这一种方法筛选自变量建立多重回归方程，很难保证结果就是非常理想的。③表中给出的结果存在过失错误或有造假嫌疑。呼吸困难这个自变量之下的“重度”所在行中 3 个“？”处应该有数据，而原表中空缺，这很可能是过失误差所致；发热这个自变量的后两档分别相对于第一档“无”对应的 *P* 值都大于 0.05，说明“发热”对“是否患肿瘤”这个结果变量的影响无统计学意义，不应该将其保留在最终的多重 logistic 回归方程之中。基于此，可以认为原文 5 重 logistic 回归方程得到的结论的可信度较低。

正确的做法是，多选择几种筛选自变量的方法进行变量筛选，通常应选用不少于 3 种方法筛选自变量来建立多重回归方程，并借助一些评价方法，确定其中 1 个最合适的。另外，在给出多重回归分析结果时不要漏项，尽可能提供准确、完整的信息。

例 4 原文题目：影响利培酮治疗 Tourette 综合征疗效的多因素分析。原作者对 106 例单用利培酮治疗 8 周的 Tourette 综合征患者，采用标准评定工具的 43 个临床指标进行定量或半定量评估，应用单因素和多因素分析方法分析利培酮治疗效果的影响因素。试验数据见表 5 和表 6。单因素分析采用 χ^2 检验或 t 检验，然后采用二项分类 logistic 回归模型对相关因素进行 logistic 回归分析。请问：原作者所做的多重 logistic 回归分析正确吗？为什么？

辨析与释疑：原作者所做的多重 logistic 回归分析不正确。首先，原作者在进行多重 logistic 回归分析之前，先采用单因素分析进行变量筛选，这种筛选变量的策略是不正确

表 5 影响利培酮治疗 Tourette 综合征疗效的单因素分析

因素	有效组（76 例）	无效组（30 例）	χ^2 或 t 值	P 值	OR 值
社会功能障碍（例）					
无或轻度	51	3	28.067	0.000	1.340
中或重度	25	27			
CBCL 基线总分	44.42 ± 14.40	58.20 ± 9.68	4.820	0.000	1.477
家族史（例）					
阳性	21	20	13.818	0.000	24.316
阴性	55	10			
家庭周边环境（例）					
无铅污染	55	12	9.690	0.003	2.433
有铅污染	21	18			
起病年龄（岁）	6.0 ± 2.6	4.1 ± 0.6	–3.996	0.000	1.594
家庭教育类型（例）					
民主科学	22	3	4.285	0.044	1.972
非民主科学	54	27			
既往史（例）					
阳性	28	21	9.514	0.003	11.468
阴性	48	9			
病程（年）	5.0 ± 2.1	6.7 ± 2.1	3.681	0.000	3.491

注：CBCL 为 Achenbach 儿童行为量表

表 6 影响利培酮治疗 Tourette 综合征疗效的多因素 logistic 回归分析

因素	偏回归系数	标准误	Wald 值	P 值	OR 值
常数项	5.505	3.188			
中重度社会功能障碍	3.472	1.262	7.562	0.006	32.258
家族史阳性	2.550	0.941	7.343	0.007	12.804
既往史阳性	1.865	0.859	4.712	0.030	6.459
CBCL 基线总分高	0.089	0.034	6.726	0.009	1.093

的。由于变量与变量之间可能存在交互作用，对各变量逐一进行单因素分析时，即使得出 $P > 0.05$，也不能说明该因素及其与其他因素的交互作用对结果的影响没有统计学意义。因此，进行多重 logistic 回归分析之前，先采用单因素分析进行变量筛选，将假设检验时 $P > 0.05$ 的变量予以剔除是没有科学依据的。进行多重回归分析时，正确的做法（或者叫做策略）通常有 2 种：①采用最优回归子集法寻找相对最佳的多重回归方程。当然，这样做计算量会很大。②将自变量全部引入回归模型，多采用几种变量筛选方法，对自变量进行筛选，如果得到的结果比较一致，则这样的结果是比较可靠的。当然，也可能采用不同的变量筛选方法得到的结果很不一致。如果是这样，该选用哪一种结果呢？在确保变量筛选策略正确的前提下，可以采用一套系统、科学、合理的评价标准对多种结果进行评价和比较，从而从中选择一种较好的结果。采用最优回归子集法时，也应采用该评价标准进行评价。其次，原作者对变量或因素表达不清，且因素与因素的水平混淆，指标与影响因素混淆。这可以从表 5 和表 6 中标有“因素”的列清楚地看出来。“CBCL 基线总分”、“起病年龄”、“家庭教育类型”和“病程”是变量或因素，且含义表达清楚；“社会功能障碍”、“家族史”、“家庭周边环境”和“既往史”也是变量或因素，但其含义却表达不清，应修改为“社会功能障碍程度”、“有无家族史”、“家庭周边环境有无铅污染”和“有无既往病史”才能将变量的含义表达清楚。而“无或轻度”、“阳性”、“无铅污染”、“民主科学”、“既往史阳性”、“CBCL 基线总分高”等则不是变量或因素，它们都只是某个变量的一个水平。此外，“社会功能障碍”、“阳性家族史”、“阳性既往史”和“CBCL 基线总分”并不是 4 个指标，它们有的是自变量，有的则是某个自变量的某个水平。

本期广告目次