讲座

简明生物统计方法

王广仪
（吉林医科大学）

第四讲　相关与回归

相关与回归是分析两个(或多个)变量之间的关系的方法。譬如说某一特定人群的身高与体重，不仅只分析身高或体重的某一方面（或说指标),而是分析身高与体重两个指标之间有无关联,这便是相关与回归所要分析的内容。当然，相关与回归分析也包括根据样本估计总体和显著性测验，即所谓统计推断问题。以下主要介绍相关与回归的统计推断方法。

一、相关分析方法

两个(或多个)变量之间(或者说同一统计单位的两个或多个指标之间)有关联,包含两种情况：(1) 有自变与应变（因变）的从属关系；(2) 分不出自变与应变的从属关系。前者用回归分析,后者用相关分析；当然，必须先确认有相关才能进行回归分析。相关的程度用相关系数（总体相关系数用 ρ，样本相关系数用 r）表示。

直线相关分析方法也有通过变量进行计算的参数统计方法，与通过排列编号的非参数统计方法两种。

例 4-1 随机抽测 15 例健康人 24 小时尿量与甲基尼克酰胺总排出量，结果见表 4-1。试分析二者的相关关系。

表 4-1　尿量与甲基尼克酰胺排出量

编　号	尿　量（毫升/24小时）	甲基尼克酰胺总量（毫克/24小时）
1	1,480	3.82
2	1,085	5.53
3	980	4.15
4	2,310	4.98
5	665	4.16
6	895	2.96
7	1,625	7.15
8	2,235	6.30
9	1,040	2.77
10	1,210	5.00
11	2,626	7.09
12	2,125	5.81
13	1,015	5.09
14	490	3.19
15	1,175	4.12
合　计	20,956	72.12

1. 参数统计的相关分析方法

(1) 求相关系数 (r),公式为:

$$r=\frac{\Sigma(X-\bar{x})(Y-\bar{y})}{\sqrt{\Sigma(X-\bar{x})^2}\sqrt{\Sigma(Y-\bar{y})^2}}=\frac{\Sigma XY-\frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{\Sigma X^2-\frac{(\Sigma X)^2}{n}}\sqrt{\Sigma Y^2-\frac{(\Sigma Y)^2}{n}}} \tag{4.1}$$

对于例 4-1,则有

$$\Sigma XY-\frac{(\Sigma X)(\Sigma Y)}{n}=1{,}480\times 3.82+1{,}085\times 5.53+\cdots+1{,}175\times 4.12-\frac{20{,}956\times 72.12}{15}=9{,}048.7$$

$$\Sigma X^2-\frac{(\Sigma X)^2}{n}=(1{,}480)^2+(1{,}085)^2+\cdots+(1{,}175)^2-\frac{(20{,}956)^2}{15}=5{,}874{,}446.9$$

$$\Sigma Y^2-\frac{(\Sigma Y)^2}{n}=(3.82)^2+(5.53)^2+\cdots+(4.12)^2-\frac{(72.12)^2}{15}=27.1$$

$$r = \frac{9,048.7}{\sqrt{5,874,446.9}\sqrt{27.1}} = 0.72$$

相关系数限定在 -1—0 或 0—1 之间；$r = 0$ 为无相关，$r = 1$ 为完全正相关（两个变量成互为递增或递减关系），$r = -1$ 为完全负相关（关系与前者相反）。本例 $r = 0.72$ 为有较密切的正相关。

（2）相关系数的显著性测验

$r = 0.72$ 是根据样本求出的相关系数，虽然表示存在着较密切的正相关，但因有抽样误差的影响，尚不能贸然断定其所代表的总体亦存在明显的相关关系，必须经过显著性测验加以推断。

相关系数的显著性测验仍用 t 测验法，测验该样本由无相关（$\rho = 0$）的总体中随机抽取的可能性（概率）多大？如果可能性很小（$P < 0.05$），则认为相关显著，否则认为不显著。测验公式为：

$$t_r = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}} \qquad (4.2)$$

$$df = n - 2$$

对于本例，$r = 0.72$，$n = 15$，则有

$$t_r = \frac{0.72 - 0}{\sqrt{\frac{1 - (0.72)^2}{15 - 2}}} = 3.7$$

所得 $t_r = 3.7$，它大于 $t_{r,\ 0.01(df=13)} = 3.012$*，$P < 0.01$，相关非常显著。结论：健康人的尿量与甲基尼克酰胺总排出量之间，存在明显的相关。

为避免计算，求出相关系数（r）之后，亦可直接查表（附表 7）判定测验结果。对于例 4-1，自由度 $df = 13$ 与 $P = 0.01$ 的界限值为 0.641，$r = 0.72$ 大于 $r_{0.01(df=13)} = 0.641$，$P < 0.01$，结论同前。

2. 非参数统计的相关分析方法

（1）先把原变量 X 及 Y 各自排列编号，然后求出对应的顺序号之差（d）及其平方和（Σd^2），据以计算相关系数 r_s（Spearman 等级相关系数符号），见表 4-2。

表 4-2　计算相关系数 r_s 用表

尿量（毫升/24小时）		甲基尼克酰胺总量（毫克/24小时）		d	d^2
x_{i1}	R_{i1}	y_{i1}	R_{i2}	$(R_{i1} - R_{i2})$	
490	1	3.19	3	−2	4
665	2	4.16	7	−5	25
895	3	2.96	2	1	1
980	4	4.15	6	−2	4
1,015	5	5.09	10	−5	25
1,040	6	2.77	1	5	25
1,085	7	5.53	11	−4	16
1,175	8	4.12	5	3	9
1,210	9	5.00	9	0	0
1,480	10	3.82	4	6	36
1,625	11	7.15	15	−4	16
2,125	12	5.81	12	0	0
2,235	13	6.30	13	0	0
2,310	14	4.98	8	6	36
2,620	15	7.09	14	1	1

Σd^2:198

（2）计算相关系数。公式为：

$$r_s = 1 - \frac{6\Sigma d^2}{(n - 1)n(n + 1)} \qquad (4.3)$$

对于本例，则有

$$r_s = 1 - \frac{6 \times 198}{15(15 - 1)(15 + 1)} = 0.65$$

（3）相关系数的显著性测验。可根据 $\Sigma d^2 = 198$ 直接查附表 8 进行判定。$n = 15$，正相关 $P = 0.01$ 的界限值为 211，$\Sigma d^2 = 198$，它小于 211，$P < 0.01$，结论同前。

二、回归分析方法

进行回归分析的目的一般有二：（1）通过自变量以一定的置信度估计应变量；（2）如果存在回归，则可通过引入适当的自变量而缩减应变量的变差（即按照引入的自变量，对应变量进行再分组而缩减其变差）。

两变量（X，Y）间的回归关系，有线性的（直线关系）和非线性的（曲线关系），以下只简介直线回归分析方法，当然，对于那些能通过变量代换（常用者如对数等代换）而使曲线直线化的非线性回归关系亦适用。

* 见参考资料[2]，8 页

直线回归用下列线性方程表示：

$$\hat{Y} = a + bX \tag{4.4}$$

式中，

a——截距，即当 $X = 0$ 时的 $\hat{Y}$ 值，

$$a = \bar{y} - b\bar{x} \tag{4.5}$$

b——回归系数（又称斜率），即自变量 X 每变动一单位，应变量 Y 因之而变动的平均单位数，

$$b = \frac{\Sigma(X-\bar{x})(Y-\bar{y})}{\Sigma(X-\bar{x})^2} = \frac{\Sigma XY - \dfrac{(\Sigma X)(\Sigma Y)}{n}}{\Sigma X^2 - \dfrac{(\Sigma X)^2}{n}} \tag{4.6}$$

此线性方程如系根据样本数据求出的，当然亦存在抽样误差，亦需进行统计推断。首先；回归方程中的 b 是样本的回归系数，和其所代表的总体的回归系数 β 之间存在抽样误差。用 b 去估计 β，亦用区间估计法，公式如下：

β 的 95%（或 99%）置信区间：

$$b \mp t_{0.95(或0.99)} S_b \tag{4.7}$$

式中，

S_b——b 的标准差，

$$S_b = \frac{S_{y\cdot x}}{\sqrt{\Sigma X^2 - \dfrac{(\Sigma X)^2}{n}}},$$

其中

$$S_{y\cdot x} = \sqrt{\frac{\Sigma(Y-\hat{Y})^2}{n-2}} = \sqrt{\frac{\Sigma Y^2 - \dfrac{(\Sigma Y)^2}{n} - b\left[\Sigma XY - \dfrac{(\Sigma X)(\Sigma Y)}{n}\right]}{n-2}};$$

$t_{0.95(或0.99)}$——t 值（同前），自由度 $df = n-2$。

如果需要比较两个样本回归系数时，须进行差异显著性测验，用 t 测验法，公式如下：

$$t = \frac{|b_1 - b_2|}{S_{(b_1-b_2)}} \tag{4.8}$$

式中，

$$S_{(b_1-b_2)} = \sqrt{S^2\left[\frac{1}{\Sigma(X_1-\bar{x}_1)^2} + \frac{1}{\Sigma(X_2-\bar{x}_2)^2}\right]},$$

其中

$$S^2 = \frac{\Sigma(Y_1-\hat{Y}_1)^2 + \Sigma(Y_2-\hat{Y}_2)^2}{(n_1-2)+(n_2-2)};$$

自由度 $df = n_1 + n_2 - 4$。

另外，通过回归方程用 X 去推测 Y 时，只能知道 Y 取值的平均数 $\hat{Y}$，至于每个实际值的 Y，距 $\hat{Y}$ 还会有一定波动的，那么 Y 距 $\hat{Y}$ 可能有多远（亦即用回归方程进行推测的精密度多大）呢？亦可做区间估计，公式如下：

Y 的 95%（或 99%）置信区间：

$$\hat{Y} \mp 1.96\ (或\ 2.58) S_y \tag{4.9}$$

式中，

$$S_y = S_{y\cdot x}\sqrt{1 + \frac{1}{n} + \frac{(X-\bar{x})^2}{\Sigma(X-\bar{x})^2}}$$

例 4-2　用浓硫酸测定小麦种的含水量结果如表 4-3，试进行回归分析。

表 4-3　小麦含水量与硫酸发热测定结果

样　品　号	实测水分(%)	5 分钟后温度升高度数(℃)
1	12.0	7.8
2	14.0	9.6
3	14.3	10.0
4	15.4	10.8
5	16.8	11.7
6	18.8	13.3
7	20.1	14.2
8	20.5	15.2
合　　计	131.9	92.6

为便于计算，先算出表 4-4 结果。

表 4-4　计算回归方程用表

样品号	5 分钟后发热度数(℃) X	实测水分(%) Y	X^2	Y^2	XY
1	7.8	12.0	60.8	144.0	93.6
2	9.6	14.0	92.2	196.0	134.4
⋮	⋮	⋮	⋮	⋮	⋮
8	15.2	20.5	231.0	420.3	311.6
合　计	92.6	131.9	1116.1	2241.6	1580.9

1. 计算回归方程［据公式 (4.4)，(4.5)，(4.6)］

$$b = \frac{1580.9 - \frac{(92.6)(131.9)}{8}}{1116.1 - \frac{(92.6)^2}{8}} = 1.22$$

$$a = 16.5 - 1.22 \times 11.6 = 2.3$$

$$\hat{Y} = 2.3 + 1.22X$$

2. 统计推断

β 的 95% 置信区间：$1.22 \mp 1.96 \times 0.4$*，即 0.44—2.00。

Y 的 95% 置信区间：如某次用浓硫酸测定小麦种含水量时，5 分钟后温度升高 7.8℃，则小麦种的含水量当在：$11.8 \mp 1.96 \times 0.47$，即 11.—12.6 之间[计算过程见公式 (4.9)]。

附表 7　相关系数显著性测验表示例

自由度(df)	5%	1%
1	0.997	1.000
2	0.950	0.990
3	0.878	0.959
4	0.811	0.917
5	0.754	0.874
6	0.707	0.834
7	0.666	0.798
8	0.632	0.765
9	0.602	0.735
10	0.576	0.708
11	0.553	0.684
12	0.532	0.661
13	0.514	0.641
14	0.497	0.623
15	0.482	0.606

附表 8　等级相关显著性测验表示例

n	正相关		负相关	
	5%	1%	5%	1%
5	2	0	40	38
6	6	2	68	64
7	16	6	106	96
8	30	14	154	138
9	48	26	214	192
10	72	44	286	258
11	103	63	377	337
12	143	89	483	429
13	191	121	607	537
14	247	161	748	663
15	313	211	909	807

参考资料

[1] 中国科学院数学研究所统计组编：常用数理统计方法，科学出版社，1973。

[2] 中国科学院数学研究所概率统计室编：常用数理统计表，科学出版社，1974。

[3] Steel, R. G. D. and Torrie, J. H.: Principles and Procedures of Statistics, 1960.

[4] 王广仪：简易医用数理统计方法初稿，吉林省医学科学情报室、吉林医大科研处编：学术活动资料，1964。

* $S_b = 0.4$，计算过程见公式(4.7)

1976 年 第 1 期 更 正

第 11 面毛主席语录应为：**“自力更生为主，争取外援为辅，破除迷信，独立自主地干工业、干农业，干技术革命和文化革命，打倒奴隶思想，埋葬教条主义，认真学习外国的好经验，也一定研究外国的坏经验——引以为戒，这就是我们的路线。”**

1975 年 第 4 期 更 正 表

页	行	误	正
4	21	在与疾病流行	在与疫病流行
4	25	这要比英国人	这比英国人
4	29	是对刘少奇、林彪一伙	是对蒋介石、刘少奇、林彪一伙
4	倒 1	孟轲公然叫嚣：“善战者服上刑”，“辟草，任土地者次之”。	孟轲公然叫嚣：要对“善战者服上刑”，“辟草莱、任土地者次之”。
5	1	“贼工末技”	“贱工末技”
5	13	大都是没有真才实学，	大都是没有真才实学、
6	6	«内经»上就有过这	«内经»上就有这
7	21	非可度量	外可度量
7	22	方可以治	云可以治
2	倒 4	“教人于髑髅堆中，	“教人于髑髅堆中、
57	左倒 11	用 15N 氢氧化钠	用 15N 氢氧化铵
59	左 21	，亦免……	，以……