·科研方法专题·

计数资料回归分析基础知识

胡良平[1,2*]

(1. 军事科学院研究生院,北京 100850;
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
* 通信作者:胡良平,E-mail:lphu812@sina.com)

【摘要】 本文目的是介绍与“计数资料回归分析”有关的基础知识。首先,介绍资料类型,因为它是合理选择统计分析方法的重要基础;第二,介绍二项分布、泊松分布和负二项分布,因为这三个离散型随机变量概率分布是对计数资料建立回归模型的重要依据;第三,介绍计数资料五个明显的分布特点:①方差小于均值的低离散型计数资料;②方差近似等于均值的一般计数资料;③方差明显大于均值的过离散型计数资料;④离散型随机变量在“0”处取值的概率非常大(简称为零膨胀)且取“非0正整数”时服从泊松分布的计数资料;⑤离散型随机变量在“0”处取值的概率非常大(简称为零膨胀)且取“非0正整数”时服从负二项分布的计数资料。前述的基础知识,是下一步建立合适的计数资料回归模型的必要基础。

【关键词】 资料类型;计数资料;二项分布;泊松分布;负二项分布;低离散;过离散;零膨胀
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2018.05.001

The fundamental knowledge for the regression analysis of the count data

Hu Liangping[1,2*]

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

【**Abstract**】 The purpose of this paper was to introduce the fundamental knowledge related to the regression analysis of the count data. Firstly, the data type was introduced, since it was the basic knowledge for selecting the methods of statistical analysis rationally. Secondly, binomial distribution, Poisson distribution and negative binomial distribution were given, for the probability distributions of the three kinds of discrete random variables were the important evidence for building the regression model of the count data. Finally, the five distinct distributional features of the count data were stated, including: ①the low-dispersion count data with the variance to be less than the arithmetic mean; ②the general count data with the variance to be equal to the arithmetic mean; ③the over-dispersion count data with the variance to be greater than the arithmetic mean; ④the count data in which the discrete random variable taken two kinds of values, one was zero value with the big probability, and the other was non-zero positive integer values with the Poisson distribution, the count data were called the zero-inflated count data; ⑤the count data in which the discrete random variable taken two kinds of values, one was zero value with the big probability, and the other was non-zero positive integer values with the negative binomial distribution, the count data were also called the zero-inflated count data. Based on the knowledge mentioned above, it was very useful for building a suitable regression model of the count data.

【**Keywords**】 Data type; Count data; Binomial distribution; Poisson distribution; Negative binomial distribution; Low-dispersion; Over-dispersion; Zero-inflation

1 资料类型[1]

1.1 资料类型的种类

资料类型的种类数量取决于“划分方法”,“传统划分方法”将资料类型划分为以下三类,即计量资料、计数资料和等级资料;而“现代划分方法”将资料类型划分为以下四类,即计量资料、计数资料、名义资料(包括二分类定性资料和多分类定性资料)和有序资料。其中,常简称“计量资料”与“计数资料”为“定量资料”;简称“名义资料”与“有序资料”为“定性资料”。

值得注意的是:“传统划分方法”中“计数资料”的概念是不正确的,其把诸如“性别”“血型”“职业”这样的定性(二分类或多分类)变量及其取值错误地定义为“计数资料”,其理由是:在所考察的变量(如:性别)各水平组之下,受试者数目是“正整数”,故将其视为“计数资料”,这是“形而上学”思维导致的“错误定义”!

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

1.2 资料类型的识别

正确识别资料类型的方法是看变量在每个个体上的取值或表现。如:“性别”在每个个体上的取值不是“男性”就是“女性”,所以“性别”是一个定性变量或名义变量或二值变量;“血型”是一个定性变量或多分类变量;显然,两种以上分级的“疗效”(如:治愈、显效、好转、无效、死亡)就应被称为“有序资料或等级资料”;而“身高”“体重”“收缩压和舒张压”“体重指数”这些指标的测定结果应被称为“计量资料”,因为它们可以取小数,还有“度量衡单位”;至于某街道各户居民家庭人口数资料、某地区 1 000 例癫痫患者治疗出院后,医生对他们进行 1 年的随访观察,收集每例患者在随访 1 年内癫痫复发次数的资料,就应被称为“计数资料”,因为其数值为“0”或“非 0 的正整数”。

2 常用随机变量的概率分布[1]

2.1 随机变量、离散型与连续型随机变量

2.1.1 随机变量

变量的取值在试验或观察之前是无法准确知道的,例如:假设让某人连续打靶 10 次,设 X 代表打中“靶心”的次数,于是,X 就是一个随机变量。

2.1.2 两种离散型随机变量

若某随机变量只能取实数域某区间内的“0”或“整数”(在多数实际问题中,为“正整数”),则这类随机变量就被称为“离散型随机变量”,例如:某市每户家庭人口数等。若某随机事件可能发生、也可能不发生,为研究方便,令该事件发生为“1”、不发生为“0”,于是,该随机事件也转化成了一个离散型随机变量。

在实际问题中,“离散型随机变量”有以下两种类型:①计数型。变量在每个个体上的取值要么为“0”,要么为“非 0 的整数”,多数场合下,为“正整数”,例如:用某种药物治疗某种难治性疾病的1 000 例患者,治愈例数 X 就是一个计数型随机变量。②定性型。变量在每个个体上的取值只是一个“类别”或“等级”,由研究者对其进行重新“赋值”,例如:将“疗效”中的“有效”赋值为“1”,“无效”赋值为“0”;同理,研究者可以给五种疗效等级分别赋值为“1,2,3,4,5”或“1,4,9,15,24”。但在更多场合下,会将具有五个等级的“疗效”转换为四个“哑变量”,每个“哑变量”都是一个“二值变量”,它们都以某一个疗效等级(例如:死亡)为“基准”。这些问题中的“随机变量”都属于“定性型的随机变量”。

2.1.3 连续型随机变量

若某随机变量可取实数域某区间内的任何值,则该随机变量就被称为“连续型随机变量”,例如:某地区正常成年人的体重、血压、体重指数数值等。

2.2 随机变量概率分布的定义

2.2.1 频率与概率及其相互关系

在样本中,随机事件 A 出现的可能性大小的度量,被称为事件 A 发生的频率。通常,在 n 次独立重复试验中,若事件 A 出现了 k 次,则称式(1)为事件 A 出现的频率。

$$f_n(\mathrm{A}) = \frac{k}{n} \times 100\% \quad (1)$$

在总体中,随机事件 A 出现的可能性大小的度量,被称为事件 A 发生的概率。通常,当 $n \to \infty$ 时,用 n 次独立重复试验的频率 $f_n(\mathrm{A})$ 作为概率 $P(\mathrm{A})$ 的估计值,见式(2)。

$$f_n(\mathrm{A}) \xrightarrow{n \to \infty} P(\mathrm{A}) \quad (2)$$

频率与概率都是用来描述随机事件发生可能性大小的度量,频率是对样本而言,而概率则是总体的属性。

2.2.2 率与率的标准误

率,通常划分为百分率、千分率,仅仅是基数不同;有时还称为样本率、总体率,这完全取决于计算率时所对应的分母。若分母是由样本中的全部个体组成,则应被称为样本频率,简称样本率;若分母是由总体中的全部个体组成,则应被称为总体概率,简称概率。

只有样本率才有标准误,因为总体率是一个固定的常数,不存在抽样误差。那么,何为率的标准误呢?试想,从一个无限总体中反复有放回地抽取样本大小为 n 的个体组成样本,记录导致随机事件 A 发生的个体数量,设为 k_1,则样本率 $P_1 = k_1/n$;若从事先定义的总体中,再随机抽取 n 个个体,又可计算出第二批试验所对应的样本率 $P_2 = k_2/n$;…;这样反复抽样,假定共抽了 m 批($m \geqslant 2$)。不难想象,$P_i, i = 1,2,\cdots,m$,这 m 个样本率不完全相等。度量它们波动大小的变异指标被称为率的标准误,由数理统计知识可知,率的标准误见式(3)。

$$S_P = \sqrt{\frac{P(1-P)}{n}} \quad (3)$$

2.2.3 概率分布的含义

对于非数学工作者来说，“概率分布”一词是比较难以理解的。但当人们把“概率”暂时比作“100 斤大米”，把“分布”暂时理解成“分配”，于是，“概率分布”就变成“分配 100 斤大米”了。问题是：分配给谁？按什么“方案”进行分配？例如：按“相等原则”把“100 斤大米”分配给“10 个人”。若用“X”代表这 10 个人的通用名，用“1～10”代表他们的具体名字、用 W 代表每人分到的大米重量、用 P 代表每人分到的大米占全部大米的比重，则可用如下形式呈现前面的“分配方案”。

X：	1	2	3	4	5	6	7	8	9	10
W：	10	10	10	10	10	10	10	10	10	10
P：	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

在上面的“分配方案”中，最后一行的全部数值之和为“1”，它被称为“总概率”，常简称为“概率”。实际上，就是把概率“1”分配给“10 个人”，故“概率分布”就是“概率分配”之意。

接下来，将上面的具体问题“抽象化”，使其适用于任何所研究的事物或问题。把“X”视为某事物中所关心的结果所代表的“随机变量”，需要列出“X”的所有可能取值（如：上例中，有 10 个人参加分配大米）；把“P”视为按“分配原则”所决定的“分配比例”或“X”取各个特定值对应的“概率”。将一个实际问题中的“X”与“P”及其具体取值均呈现出来，就是这个实际问题所对应的“概率分布”。对于上面的例子，其概率分布如下。

X：	1	2	3	4	5	6	7	8	9	10
P：	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

2.2.4 离散型随机变量概率分布的定义

将某个指定的离散型随机变量的所有可能取值一一列举出来，再将该随机变量取每个特定值的可能性，即概率也一一列举出来，将这两部分信息同时呈现出来，就称其为该离散型随机变量的概率分布。

离散型随机变量只能取有限个或可列个数值，其取值分别为 $x_1, x_2, \cdots, x_i, \cdots$，相应的概率为 $p_i = P(X = x_i)$，$i = 1, 2, \cdots$，它的概率分布常以分布列的形式表示，见式(4)：

$$\begin{pmatrix} X \\ P \end{pmatrix} \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix} \tag{4}$$

任一离散型随机变量的分布列$\{p_i\}$都应满足下面的两个公式：

$$① \ p_i \geq 0, i = 1, 2, \cdots; \tag{5}$$

$$② \sum_{i=1}^{\infty} p_i = 1 \tag{6}$$

在式(4)中，若每个 p_i 都可通过一个公式计算出来，则称该计算公式为该离散型随机变量“X”的“概率函数”。由此可知，要掌握一个实际问题中的离散型随机变量“X”的变化规律，必须提供两方面的信息：①X 的一切可能取值；②X 的概率函数或 X 取每一个特定值所对应的概率。

在“2.3”小节中将介绍两种在统计建模中常用的“离散型随机变量概率分布”。

2.2.5 连续型随机变量概率分布的定义

2.2.5.1 概述

由于连续型随机变量的取值充满一个区间，无法一一列出，因此对于连续型随机变量的概率分布，不能像离散型随机变量那样使用分布列去描述。刻划连续型随机变量概率分布的一个方法是用概率分布函数去描述，但在理论和实际中更方便、常用的是“概率密度函数”。

2.2.5.2 定义

设连续型随机变量 X 有概率分布函数 $F(x)$，则 $F(x)$ 的导数 $f(x) = F'(x)$ 称为 X 的概率密度函数。

“密度函数”这个名词的来由可解释如下：取定一个点 x，则按分布函数的定义，事件$\{x < X < x + h\}$的概率（$h > 0$ 为常数）应为 $F(x + h) - F(x)$。所以比值 $[F(x + h) - F(x)]/h$ 可以解释为在区间 $(x, x + h]$ 内，随机变量在特定区间取值的概率。令 h 趋向于 0，则这个比的极限为 $F'(x) = f(x)$。

连续型随机变量 X 的概率密度函数 $f(x)$ 具有以下三个基本性质：

$$① \ f(x) \geq 0 \tag{7}$$

$$② \int_{-\infty}^{+\infty} f(x)\,d(x) = 1 \tag{8}$$

③ 对任意常数 $a < b$ 有

$$P(a \le X \le b) = F(b) - F(a) = \int_a^b f(x)dx \qquad (9)$$

2.2.5.3 “概率密度函数”的概念

对于非数学工作者，很难理解“概率密度函数”的真实含义。首先，在直角坐标系中，它的图像是一条“曲线”。例如：标准正态分布的概率密度曲线。见图 1。

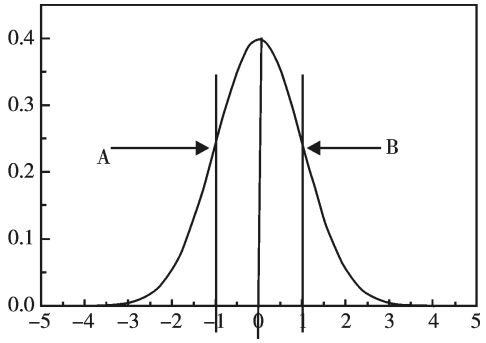

图 1 标准正态分布概率密度函数的图形

从图 1 可观察到如下特点：

①在 $x=0$ 处，曲线有唯一的峰值，为 $\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}=0.398942$；

②图形关于直线“$x=0$”对称；

③图形以 x 轴为其渐近线；

④曲线上有两个对称的拐点 A 和 B，拐点到直线“$x=0$”的距离等于 1；

⑤曲线的位置由期望值 μ 决定，μ 为位置参数；

⑥曲线的“胖瘦”由方差 σ^2 来决定，σ 为形状参数，也叫做随机变量 X 的标准差；

⑦用变量变换的方法可以使一般正态分布转变成标准正态分布，图 1 就是标准正态概率密度函数图（$\mu=0$、$\sigma=1$）。

横坐标轴上的变量就是所研究的“连续型随机变量 X”，曲线在横坐标轴上所覆盖的“范围”就是“连续型随机变量 X”的“取值区间”［对于标准正态分布而言，整个区间为（$-\infty$，$+\infty$）］。整个曲线下的“面积”为“1”，这个“1”就是“连续型随机变量 X”在全部取值区间上取值的“概率”。于是，若在该曲线上任取两点，从这两点向横坐标轴作垂线，其交点分别记为“X_1”与“X_2”，设它们之间的曲线下面积为“P*”，则这个“P*”就是“X 在闭区间 $[X_1, X_2]$ 上取值的概率”。

由此可知，所谓“概率密度函数”，就是通过它可以计算出“连续型随机变量 X”在任何一个指定区间内取值的“概率”。在横坐标轴的不同位置上，取几个相等宽度的“区间”，“连续型随机变量 X”在这些区间上取值的概率是不相等的，形象地理解成：落入各区域上的“雨点”的“密集程度”不同。所以，“概率密度函数”中的“密度”两个字，其真实含义是“大小”，即计算“连续型随机变量 X”在不同区间上取值的“概率大小”的“函数”被数学家和统计学家称为“概率密度函数”。

服从特定概率分布规律的“连续型随机变量概率分布”有很多，其中，最常见的有“正态分布”“t 分布”“χ^2 分布”和“F 分布”，因篇幅所限，此处从略。

2.3 统计建模中三种常用离散型随机变量概率分布举例

2.3.1 二项分布

2.3.1.1 引言

有很多随机现象或试验，每进行一次观察或试验只有两种对立结果中的一种出现。如成功与失败、生存与死亡等。假定在群体中，丝虫病的患病率 $p=0.1$，不患丝虫病的概率 $q=1-p=0.9$。若随机抽查三人，则可能出现下面四种情形之一。

抽查结果	对应的概率计算公式	概率值
全是阳性	$ppp(=p^3)$	0.001
两阳一阴	$ppq+pqp+qpp(=C_3^2p^2q^1)$	0.027
一阳两阴	$pqq+qpq+qqp(=C_3^1p^1q^2)$	0.243
全是阴性	$qqq(=q^3)$	0.729

如果用 X 表示随机抽出的三人中患丝虫病的人数，则 $X=i$ 的概率可概括地表达为：

$$P(X=i)=C_3^ip^iq^{3-i}=C_3^i0.1^i0.9^{3-i},i=0,1,2,3$$

若独立重复上述试验 n 次，称为 n 重 Bernoulli 试验，各种结果的出现有一定的分布规律，称为 Bernoulli 分布，这是因为此分布最初由瑞士数学家和统计学家 J. Bernoulli（1654 年 - 1705 年）发现。又因此分布的概率函数是二项展开式中的一项，故此分布又被称为二项分布。

2.3.1.2 定义

设某随机试验只有两个对立的结果，每次试验的结果要么是事件 A 发生，要么是其对立事件 $\bar{A}$ 发生。并且，$P(A)=p$，$P(\bar{A})=q$，$p+q=1$，$0<p<1$。又设在一次试验中，事件 A 发生的次数为 Y，则称 Y 服从二点分布，其分布列见式（10）。

$$\begin{pmatrix}0 & 1\\ q & p\end{pmatrix} \qquad (10)$$

若在相同条件下，进行 n 次独立重复试验，用 X 表示 n 次试验中事件 A 发生的总次数，即 $X=\sum_{i=1}^{n}Y_i$，各 Y_i 均服从相同的二点分布，则称 X 服从二项分布，并记为 $X\sim b(m;n,p)$ 或 $X\sim B(n;p)$，其概率函数见式(11)：

$$P(X=m)=C_n^m p^m q^{n-m},m=0,1,2,\cdots,n \quad (11)$$

式中 C_n^m 为组合数，即 $C_n^m=\frac{n!}{m!(n-m)!}$，$n!=1\times2\times3\times\cdots\times n$，读作 n 的阶乘。

服从二项分布的离散型随机变量 X 的分布函数见式(12)：

$$P(X\le k)=\sum_{m=0}^{k}C_n^m p^m q^{n-m} \quad (12)$$

2.3.1.3 性质

(1)概率函数 $b(m;n,p)$ 的图形

当 n 和 p 取不同值时，概率分布的形状就有所不同。当随着 n 增大且 p 接近 0.5 时，二项分布逐渐接近正态分布。下面仅给出 $n=20$ 且 p 分别取 0.25、0.50 和 0.75 三种情况下二项分布概率函数图形。见图 2。

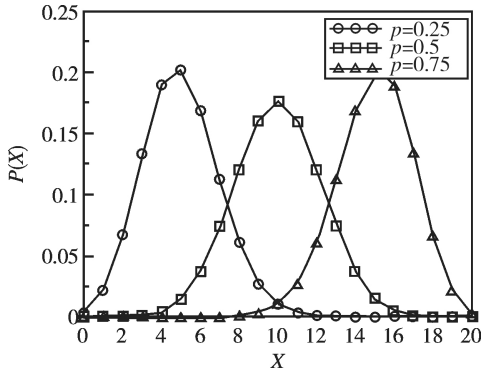

图 2 $n=20,p=0.25$、0.50、0.75 时二项分布的概率函数折线图

(2)期望值与方差

设 $X\sim b(m;n,\pi)$，π 为总体率，则 X 可视为 n 个相互独立的服从二点分布的随机变量 $X_i(i=1,2,\cdots,n)$ 之和，即 $X=X_1+X_2+\cdots+X_n$，二点分布的期望值与方差分别为：

$$E(X_i)=0\times(1-\pi)+1\times\pi=\pi \quad (13)$$

$$Var(X_i)=E[X_i-E(X_i)]^2=E(X_i-\pi)^2$$
$$=(0-\pi)^2(1-\pi)+(1-\pi)^2\pi=\pi(1-\pi) \quad (14)$$

根据期望值与方差的性质可得：

$$E(X)=E(X_1+X_2+\cdots+X_n)$$
$$=E(X_1)+E(X_2)+\cdots+E(X_n)=n\pi \quad (15)$$

$$Var(X)=Var(X_1+X_2+\cdots+X_n)$$
$$=Var(X_1)+Var(X_2)+\cdots+Var(X_n)$$
$$=n\pi(1-\pi) \quad (16)$$

若令样本率 $p=\frac{x}{n}$，则有：

$$E(p)=\pi \quad (17)$$

$$Var(p)=\frac{\pi(1-\pi)}{n} \quad (18)$$

(3)二项分布的可加性

若 $X_1,X_2,\cdots,X_m$ 独立，且 $X_i\sim B(n_i;p)$，$i=1,2,\cdots,m$，$X=X_1+X_2+\cdots+X_m$，则 $X\sim B(n;p)$，其中 $n=n_1+n_2+\cdots+n_m$。

2.3.2 泊松(Poisson)分布

2.3.2.1 引言

在自然界中，有一系列看起来彼此互不相干的随机变量，它们却遵从同一种分布规律。如某交换台在某一段时间内所接到的呼唤次数；某公共汽车站在等长的一段时间内的乘客数；每米布上的瑕疵点数；每件钢铁铸件上的缺陷数；放射性分裂落到某区域内的质点数；显微镜下落在某区域中的血球或微生物的计数；细菌、血细胞、粉尘等在单位面积或容积内的计数；在单位空间中的某些野生动物或昆虫数；在一定人群中某种患病率很低的非传染性疾病的患病数或死亡数等。上述离散型随机变量，一般认为，它们的分布规律是由法国数学家 Simeon Denis Poisson(1781－1840)于 1837 年发现的，故称为 Poisson 分布(然而，有证据表明：在此之前约一个世纪，此分布可能已被 DeMoivre 发现)。这种分布常用于描述单位时间或平面或空间中罕见“质点”总数的随机分布规律，可视为 n 很大，π 很小时二项分布的极限情形。

2.3.2.2 定义

若离散型随机变量 X 的取值为非负整数，且相应的概率函数为：

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda},k=0,1,2\cdots,\lambda>0 \quad (19)$$

则称随机变量 X 服从 Poisson 分布，记作 $X\sim P(k;\lambda)$。

2.3.2.3 性质

(1)概率函数 $P(k;\lambda)$ 的图形

当 λ 取不同值时，概率分布的形状就有所不

同。当随着其均值 λ 不断增大，泊松分布逐渐接近正态分布。下面仅给出 λ 分别取 2.5、5.0 和 10.0 三种情况下泊松分布概率函数图形，见图 3。

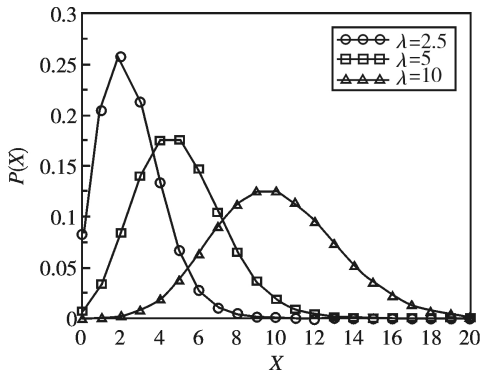

图 3 $\lambda=2.5,5.0,1.0$ 时泊松分布的概率函数折线图

(2)均值与方差

$$E(X)=\lambda \quad (20)$$

$$Var(X)=\lambda \quad (21)$$

(3)极值

$$\begin{cases} P(k;\lambda)>P(k-1;\lambda), k<\lambda \\ P(k;\lambda)<P(k-1;\lambda), k>\lambda \end{cases} \quad (22)$$

当 λ 不是整数时，$P(k;\lambda)$ 在 $\lambda=[\lambda]$（这里中括号表示取整运算）处达到极大值；当 λ 是整数时，$P(k;\lambda)$ 在 $k=\lambda$ 及 $k=\lambda-1$ 处同时达到极大值。

(4)用正态分布近似处理泊松分布资料的条件

泊松分布是非对称的，但当 λ 愈大时非对称性愈不明显；当 $\lambda=10$ 时，分布已相当对称了。一般来说，当 $\lambda\geq 20$ 时，泊松分布的资料可按正态分布处理。

(5)用泊松分布近似处理二项分布资料的公式

当 n 很大，p 很小，$np=\lambda$ 为一常数时，二项分布近似于泊松分布。p 愈小，近似程度愈好。即

$$C_n^m p^m (1-p)^{n-m} \xrightarrow{n\to\infty, p\to 0, np=\lambda} \frac{\lambda^k}{k!}e^{-\lambda} \quad (23)$$

(6)泊松分布的可加性

如果相互独立的 k 个随机变量均服从泊松分布，则它们之和仍服从泊松分布，且其均数为 k 个随机变量的均数之和。

2.3.3 负二项分布

2.3.3.1 引言

在二项分布中，独立重复试验的次数是固定的，如果让指定结果（如阳性）发生的次数固定，则第 r 次发生指定结果时，所需要的试验次数 X 是一随机变量，它的概率分布就是负二项分布。负二项分布中，重复试验相互独立并且发生某一事件的概率保持不变，这与二项分布一致。

负二项分布常用于描述生物的群聚性，如钉螺在土壤中的分布、昆虫的空间分布等。医学上可用于描述传染性疾病的分布和致病生物的分布，在毒理学的显性致死试验或致癌试验中也都有应用。

2.3.3.2 定义

如果离散型随机变量 X 的概率函数为下面的式(24)：

$$P(X=k)=\binom{k-1}{r-1}p^r q^{k-r}, k=r, r+1, \cdots \quad (24)$$

则称随机变量 X 服从负二项分布，记作 $X\sim Nb(r,p)$。式中 p 与 r 分别为指定结果发生的概率与次数。

实际中常用未发生指定结果的次数 Y 代替试验次数 X，这时记 $Y=X-r$，$m=k-r$，式(24)可改写为下面的式(25)：

$$P(Y=m)=\binom{m+r-1}{m}p^r q^m, m=0,1,\cdots \quad (25)$$

2.3.3.3 性质

(1)概率函数 $Nb(n,p)$ 的图形

该分布中有两个参数，分别为 r 和 p，它们取不同值时，分布图形各异。下面给出 $r=5$，$p=0.3$、0.5 和 0.7 条件下负二项分布概率函数图。见图 4。

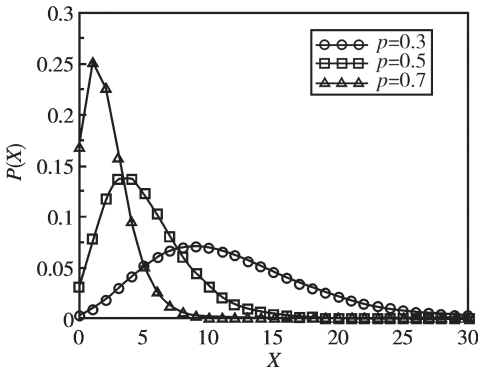

图 4 $r=5$，$p=0.3$、0.5、0.7 时负二项分布的概率函数折线图

(2)期望值与方差

$$E(X)=\frac{r}{p} \quad (26)$$

$$Var(X)=\frac{rq}{p^2} \quad (27)$$

对于随机变量 Y，期望值和方差分别为

$$E(Y)=\frac{rq}{p} \quad (28)$$

$$Var(Y) = \frac{rq}{p^2} \quad (29)$$

(3)负二项分布的退化与扩展

当 $r = 1$ 时,负二项分布退化为几何分布。此外,若 $X_1, \cdots, X_m$ 是相互独立同分布的随机变量,其分布为几何分布,则它们的和 $X_1 + \cdots + X_m$ 服从负二项分布 $Nb(n,p)$。

在式(16)中,如果允许 r 为任意正实数,则称此分布为一般的负二项分布,这个分布已被证实在众多应用中能很好地拟合观测数据。当 r 为正整数时,负二项分布被称为 Pascal 分布。

(4)聚集指数

r 值的大小可以衡量分布的离散程度,即聚集趋向的程度,被称为聚集指数,在很多文献中常用字母 k 来表示它。

3 计数资料的五种分布特征

3.1 方差近似等于均值的计数资料

文献[2]中提供了表 1 前两列数据,试估计每个细胞单位内白细胞数的算术平均值和方差。

表 1 每个细胞单位内的白细胞数及频数

每细胞单位内的白细胞数 X	细胞单位数 f	基于泊松分布估计数
0	64	59.88
1	171	169.06
2	239	238.66
3	220	224.62
4	155	158.55
5	83	89.53
6	46	42.13
7	20	16.99
8	6	6.00
9	3	1.88
10	0	0.53
11	1	0.18
合计	1008	1008.01

基于表 1 中的前两列数据,按下面的公式可以计算出“每个细胞单位内白细胞数的算术平均值和方差”:

$$\bar{X} = \frac{1}{1008}(64 \times 0 + 171 \times 1 + \cdots + 1 \times 11) = \frac{2846}{1008} = 2.8234127 \approx 2.82$$

$$S^2 = \frac{1}{1008}[64 \times (0 - 2.82)^2 + 171 \times (1 - 2.82)^2 + \cdots + 1 \times (11 - 2.82)^2] \approx 2.99$$

因 X 的算术平均值近似等于其方差,由上面的式(20)和式(21)可知:本例中的“每细胞单位内的白细胞数 X”近似服从均值 $\lambda \approx 2.82$ 的泊松分布。于是,将此均值代入式(19)就可计算出 X 分别取表 1 中第 1 列各数值时对应的概率;再乘以其频数(见表 1 中第 2 列数据),就可获得表 1 中第 3 列的数据,该列数据就是按 $\lambda \approx 2.82$ 的泊松分布计算出来的。

3.2 方差明显大于均值的计数资料

文献[3]中有关于“马蹄形蟹及其伴随者”的观测数据,其简略形式见表 2。

表 2 雌蟹伴随者个数的平均值和方差

雌蟹背夹宽度	雌蟹个数	伴随者个数	算术均值	方差
<23.25	14	14	1.00	2.77
23.25~24.25	14	20	1.43	8.88
24.25~25.25	28	67	2.39	6.54
25.25~26.25	39	105	2.69	11.38
26.50~27.25	22	63	2.86	6.88
27.50~28.25	24	93	3.87	8.81
28.50~29.25	18	72	3.94	16.88
>29.25	14	72	5.14	8.29

在表 2 中,人们关心的“变量”是“每个雌蟹周围有几个伴随的雄蟹”。倒数第 2 列为不同“背夹宽度”的雌蟹平均有几个伴随的雄蟹,最后一列为该变量的“方差”。若假定该变量服从泊松分布,那么,该变量的“方差”就明显大于其“均值”了。在统计学上,称这种“计数资料”为“过离散的计数资料”。

3.3 方差明显小于均值的计数资料

在文献[4]中,为了了解和预测人体吸入氧气的效率,收集了 30 名中年男性的健康状况调查资料。共调查了 7 个指标,分别是:吸氧效率(y),年龄(x_1,岁),体重(x_2,kg),跑 1.5 km 所需时间(x_3,min),休息时的心率(x_4,次/分),跑步时的心率(x_5,次/分),最高心率(x_6,次/分)。见表 3。

表 3 30 名中年男性的健康状况调查资料

id	y	x_1(岁)	x_2(kg)	x_3(min)	x_4(次/分)	x_5(次/分)	x_6(次/分)
1	44.609	44	89.47	11.37	62	178	182
2	45.313	40	75.07	10.07	62	185	185
3	54.297	44	85.84	8.65	45	156	168
…	…	…	…	…	…	…	…
30	47.920	48	61.24	11.50	52	170	176
31	47.467	52	82.78	10.50	53	170	172

注:id,观测对象编号;y,吸氧效率;x_1,年龄;x_2,体重;x_3,跑 1.5 km 所需时间;x_4,休息时的心率;x_5,跑步时的心率;x_6,最高心率

试求出表 3 中最后 3 列的“均值”与“方差”。计算结果如下:

Obs	mx4	vx4	mx5	vx5	mx6	vx6
1	53.4516	58.0559	169.645	105.103	173.774	83.9806

以上输出结果中,“mx”与“vx”分别代表“均值”与“方差”。可以看出:x_5 和 x_6 的方差均明显小于其均值,尤其是 x_6,其“均值”约为“方差”的 2.07 倍,或者说,“方差”不到“均值”的一半。

3.4 零膨胀计数资料

在 SAS 帮助信息[5]中提供了一组数据:一个名叫“William Sealy Gosset”(简称“W. S. Gosset”,其曾以笔名“Student”发表了著名的“t 分布”)的化学家采用血细胞计量器计算了同样大小器皿中“酵母细胞”的个数,个数范围为 0~5,对应的器皿数(简称频数)如表 4 所示。

表 4 “W. S. Gosset s”酵母细胞计数

细胞计数	0	1	2	3	4	5
频 数	213	128	37	18	3	1

一般认为:细胞计数近似服从泊松分布,但统计学家 Karl Pearson 研究了表 4 资料后,认为该组资料不服从泊松分布,他利用两个二项分布构成的混合模型来刻划此资料的分布规律。后来,统计学工作者逐渐认识了这种数据分布规律,称它为“零膨胀计数资料”,采用改进的泊松分布模型来描述它,此类模型被称为“零膨胀计数资料泊松分布回归模型”;事实上,还有一种类似的模型,被称为“零膨胀计数资料负二项分布回归模型”。这两个“零膨胀计数资料回归模型”的共同点为取“0”值的次数很多;不同点为取“非 0 正整数”值的那部分计数资料,分别服从泊松分布或负二项分布。

3.4 计数资料回归模型的基本构想

3.4.1 概述

所谓计数资料回归模型就是采用一个回归模型或方程来描述计数的因变量随影响因素或自变量变化而变化的依赖关系。也就是说,因变量一定是“计数变量”,而且,至少要有一个自变量。

3.4.2 适合选用“二项分布回归模型”的场合

当离散型随机变量 Y 的方差明显小于其均值[分别见式(16)与式(15)]时,适合选用“二项分布回归模型”。例如:在表 3 中,若以 x_6 为“因变量”,以 x_1-x_5 为“自变量”,建立“计数资料回归模型”时,宜选用“二项分布回归模型”。

遗憾的是,表 3 的数据结构不符合拟合“二项分布回归模型”的要求。因为拟合此回归模型时,因变量要求是“二值的”或“Y/N”(即以分组形式呈现的“各组阳性数/各组观察或试验总例数”)。

3.4.3 适合选用“泊松分布回归模型”的场合

由于服从泊松分布随机变量的方差等于均值[分别见式(21)与式(20)],此时,常习惯采用泊松分布回归模型。

3.4.4 适合选用“负二项分布回归模型”的场合

当离散型随机变量的方差明显大于其均值时,资料的分布就偏离泊松分布,此时,采用“负二项分布回归模型”取代“泊松分布回归模型”可以较好地提高模型对计数资料的拟合效果。这是由于服从负二项分布的离散型随机变量 Y 的方差[见式(27)]与均值[见式(26)]之比为“$1/q$”,又由于 $q(1-p)$ 是一个小于 1 的数,故方差一般会大于其均值,其程度取决于 q 值的大小。例如,设 q 分别为 0.1、0.4 和 0.8 时,则方差分别是均值的 10.00 倍、2.50 倍

和 1.25 倍。

3.4.5 适合选用“零膨胀泊松分布回归模型”的场合

当离散型随机变量在“0”处取值的次数较多，且在“非 0 正整数”范围内取值所对应的方差近似等于均值时，宜选用“零膨胀泊松分布回归模型”。

3.4.6 适合选用“零膨胀负二项分布回归模型”的场合

当离散型随机变量在“0”处取值的次数较多，且在“非 0 正整数”范围内取值所对应的方差明显大于均值时，宜选用“零膨胀负二项分布回归模型”。

具体地说，如何针对以上提及的具有 5 种分布特点的计数资料进行相应的回归建模，参见本期专题的其他三篇文章。

参考文献

[1] 胡良平. 面向问题的统计学——(1)科研设计与统计基础[M]. 北京：人民卫生出版社，2012：235－270.

[2] 方开泰，许建伦. 统计分布[M]. 北京：科学出版社，1987：88.

[3] Alan Agresti. 类别数据分析导引[M]. 陈家鼎，陈奇志，译. 中国统计出版社，2006：81.

[4] 胡良平. 医学统计学——运用三型理论进行现代回归分析[M]. 北京：人民军医出版社，2010：105－106.

[5] SAS Institute Inc. STAT SAS 9.3 User´s Guide[M]. Cary，NC：SAS Institute Inc，2011：2437－2548.

（收稿日期:2018－10－10）

（本文编辑:唐雪莉）

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 48 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。