第45卷第3期 中国科学技术大学学报 Vol. 45, No. 3
2015年3月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Mar. 2015

文章编号:0253-2778(2015)03-0205-05

稀有变量关联分析中的泊松近似变阈值方法

方红燕

(中国科学技术大学管理学院统计与金融系,安徽合肥 230026)

摘要:稀有基因变量与疾病性状的关联性研究是全基因组关联分析的补充.然而,传统的关联分析检验方法不适用于稀有变量的检验问题,因此众多的针对稀有变量的新检验方法得以被提出和运用.变阈值方法是其中的一种方法,该方法相对于传统的关联分析检验方法在功效上有较大提高.这里从稀有事件的角度看待稀有变量突变等位基因的发生,以泊松近似分布代替二项分布,提出新的检验统计量,并在变阈值方法的基础上给出一个新的检验方法,即泊松近似变阈值方法.模拟试验结果表明该方法比变阈值方法有一致更高的检验功效,筛选变量的效率也得到了提高.

关键词:全基因组关联分析;稀有变量;泊松近似

中图分类号:O212.1 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2015.03.005

2010 Mathematics Subject Classification: Primary 62C12; Secondary 62F05

引用格式: Fang Hongyan. Poisson approximation variable threshold method in rare variants association study[J]. Journal of University of Science and Technology of China, 2015,45(3):205-209,219.
方红燕. 稀有变量关联分析中的泊松近似变阈值方法[J]. 中国科学技术大学学报,2015,45(3):205-209,219.

Poisson approximation variable threshold method in rare variants association study

FANG Hongyan

(*Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei* 230026, *China*)

Abstract: The association study of rare genetic variants and disease traits as a supplement of genome-wide association study (GWAS) has achieved remarkable development and application. But the traditional association testing methods are not suitable for the rare variants association problem. Many new testing methods designed specifically for the rare variants have been proposed and applied. The variable-threshold (VT) method is one of them, which has achieved more power than the traditional association test methods. Here the occurrence of the mutation of rare variants was treated as a rare event. Poisson distribution was taken instead of binomial distribution and a new test statistic was proposed. Base on the VT method, a new test method was proposed, namely, the Poisson approximation variable threshold method (PAVT). Simulation results show that the new method has more power than VT method uniformly, and variant selection is more efficient.

Key words: genome-wide association study; rare variants; Poisson approximation

收稿日期:2014-09-03;**修回日期**:2015-02-26

基金项目:国家自然科学基金(11271346,11201452)资助.

作者简介:方红燕,女,1985年生,博士生/讲师. 研究方向:生物统计. E-mail: fanhoy@ustc.edu.cn

0 引言

近年来全基因组关联分析研究(genome-wide association study,GWAS)成为复杂疾病遗传学研究中最主要的策略和工具,已经获得了巨大的成功.但是GWAS发现的基因成分只能解释人类疾病变差很小的一部分.GWAS方法的前提假设条件是普通疾病-普通变量(common disease common variants, CDCV)假设,作为其补充,人们开始关注普通疾病-稀有变量(common disease rare variants, CDRV)假设,即研究稀有基因变异或变量与疾病的关联.近年来飞速发展的高通量测序技术成功探测到了大量次等位基因频率更低的稀有变量,为CDRV假设下的研究提供了可能.

稀有变量一般指突变等位基因频率(minor allele frequency, MAF)不大于1%的基因变量.相对于普通变量,稀有变量的突变率更低,所含的信息量更少,且每个变量都只独立影响很小一部分的生病个体,因此,运用传统的关联分析检验方法通常功效偏低,稳定性较差,尤其是样本量较少的情况下的表现更差.因此,传统的关联分析检验方法包括单一位点检验方法和多位点检验方法,功效都很低,不适用于稀有变量的检验问题.近年来,研究人员提出了众多的针对稀有变量的检验方法.例如CAST[1], CMC[2], WSS[3], the variable-threshold method (VT)[4], VW-TOW[5], rb-VT[6], C-alpha[7]和SKAT[8]方法等.文献[4]提出的变阈值关联分析VT方法考虑到致病稀有变量的次等位基因频率有别于非功能性变量的次等位基因频率,因此以MAF作为多重阈值来构成不同的稀有变量子集合,对每个子集合运用标准的合并关联分析检验并运用置换方法得到 p 值,免去了阈值的选择性问题,也增加了检验的功效.

在本文中,针对稀有变量的关联性检验问题我们提出泊松近似的观点.稀有变量的突变率不大于1%,这些基因的变异可以视为低发生率的事件.因此,从这个角度出发,我们可以视稀有变量的变异计数为泊松分布来代替二项分布构造模型.在变阈值方法中应用基于泊松近似的新的检验统计量,得到一种新的检验方法,即泊松近似变阈值方法(Poisson approximation variable threshold method,PAVT).我们将通过数据模拟检验比较PAVT方法和VT方法的检验功效,并通过比较筛选致病变量的效率来比较两种方法的优劣.

1 方法介绍

对于一个样本量为 n 发生率为 p 的稀有事件,通常 p 很小而 n 相对很大.因此可以用泊松分布 $P(np)$ 近似逼近二项分布 $B(n,p)$.在过去的几十年中,极值和稀有事件的理论和应用在各种不同的领域如保险、金融、工程学、环境科学和水文地理学中都引起了巨大的兴趣并得到了发展[9].

假设随机变量 X 和 Y 分别独立地服从泊松分布 $P(np_0)$ 和 $P(np_1)$,要检验两个概率是否相同即检验两个随机变量分布是否有差别,我们提出一个新的检验统计量 T:

$$T=\frac{X-Y}{\sqrt{X+Y}},$$

在原假设之下,检验统计量 T 近似服从标准正态分布.

考虑一个基因上的 M 个二值基因SNP位点与疾病性状 Y 的关联性检验问题.对于稀有变量,所有的基因变量在群体中是连锁平衡的[3,10].令 X_{ij} 和 Y_{ik} 分别表示病例个体 j 和对照个体 k 的变量 i 的突变等位基因计数,$i=1,\cdots,M$, $j=1,\cdots,N_1$, $k=1,\cdots,N_0$.在可加模型之下,每个人的单个变量的计数只可能是0,1或者2,但是对于稀有变量来说取值为2的概率极低.

文献[4]通过推导得出对于某一个基因变量,其对数似然近似与 $1/\sqrt{p(1-p)}$ 成正比,其中 p 表示该变量在对照群体中的等位基因频率.因此近似可以认为MAF越小越有可能是致病基因.VT方法正是基于这个直觉出发,认为存在某一个阈值 t,使得MAF小于 t 的变量比MAF大于 t 的变量更有可能是致病基因.但由于该阈值 t 无法确定,因此以对照群体中的所有样本次等位基因频率作为各个阈值,对每一个阈值 t,计算得分

$$Z(t)=\frac{\sum_{i=1}^{M}\xi_i^t\left(\sum_{j=1}^{N_1}X_{ij}(D_j-\overline{D})+\sum_{k=1}^{N_0}Y_{ik}(D'_k-\overline{D})\right)}{\left[\sum_{i=1}^{M}\xi_i^t\left(\sum_{j=1}^{N_1}X_{ij}^2+\sum_{k=1}^{N_0}Y_{ik}^2\right)\right]^{1/2}}.$$

其中,ξ_i^t 是指示变量,在阈值之下的MAF对应的变量取值为1,否则为0.D_j 和 D'_k 分别为病例和对照中样本个体的疾病性状,实际上应有 $D_j=1$, $j=1,\cdots,N_1$; $D'_k=0$, $k=1,\cdots,N_0$. $\overline{D}$ 是所有样本的疾病性状的平均值.对每一个阈值 t,计算得分 $Z(t)$,

然后求取极大值 Z_{max}作为最终的检验统计量，最后用置换疾病性状 K 次的方法计算统计量的显著性水平，

$$p=(x+1)/(K+1),$$

其中，x 表示不低于 Z_{max}的置换得到的极大检验得分值数目.

我们提出的 PAVT 方法采取检验统计量 T 替换原来的检验统计量，得到新的得分 $S(t)$：

$$S(t)=\frac{X_t-\frac{N_1}{N_0}Y_t}{\sqrt{X_t+\frac{N_1^2}{N_0^2}Y_t}},$$

其中，$X_t=\sum_{i=1}^{M}\sum_{j=1}^{N_1}\xi_i^t X_{ij}$，$Y_t=\sum_{i=1}^{M}\sum_{k=1}^{N_0}\xi_i^t Y_{ik}$，系数$\frac{N_1}{N_0}$的引入是为了平衡病例与对照样本总量的不同对基因变量的计数造成的影响. 当病例对照设计中病例与对照的样本量相同时，$S(t)$ 退化为

$$S(t)=\frac{X_t-Y_t}{\sqrt{X_t+Y_t}}.$$

退化后的检验统计量 $S(t)$形式上与 McNemar 检验统计量相同. 我们沿用 VT 方法的框架，在所有的 $S(t)$中求取极大值 S_{max}，然后用置换的方法获得显著性水平.

以上两种方法中的阈值选取通过估算对照样本中的突变等位基因频率得到，因此共分别计算 M 个得分统计量. 当取得极大得分 Z_{max}或者 S_{max}时，相对应的 MAF 在阈值以下的变量集合 S 被认为是致病变量集合，因此我们在做关联性检验的同时，做到了对致病变量的筛选.

2　模拟研究

下面通过模拟试验研究我们提出的 PAVT 方法与 VT 方法在稀有变量与疾病性状的关联性分析中的检验功效.

我们使用 Wright's 方程[11-12]产生对照群体中每个变量的突变等位基因频率：

$$f(p)=cp^{(\beta_s-1)}(1-p)^{(\beta_N-1)}e^{s(1-p)},$$

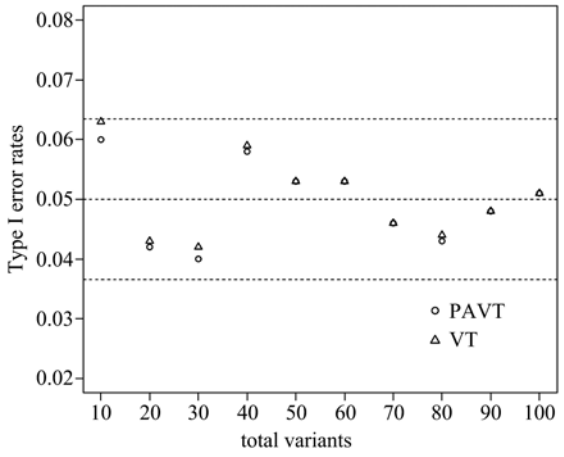
其中，$f(p)$表示以 c,β_s,β_N和 s 为参数的概率密度函数，这里，为了保证得到的是轻度有害的突变，我们选取参数 $\beta_s=0.001$，$\beta_N=\beta_s/3$，$s=12$[3,6,10]. 从 Wright's 方程可以看出，p 越小，$f(p)$取值越大，以 p 为次等位基因频率的可能性越大，因此 Wright's 方程有利于抽取稀有变量. 本文中我们采用可加模型，当 MAF 很小的时候可加模型等价于显性模型. 由于稀有变量间的关联性很小，因此可以认为各个变量间是相互独立的[10,13]. 假设对照群体服从 Hardy-Weinberg 平衡定律，因此可以计算出对照群体中的各基因型概率 q_U. 而病例群体的基因型概率 q_A 可由以下公式计算得到

$$r=\frac{\alpha}{(1-\alpha)q_U}+1,\ q_A=\frac{rq_U}{1+(r-1)q_U}.$$

其中，r 是优势比(odds ratio)，α 表示边际群体归因风险(marginal population attributable risk，PAR)，代表的是变量对疾病的效应大小.

在模拟试验中，所有致病变量的边际 PAR(α)都为同一数值，变量的所有 α 的总和为群体 PAR (gPAR). 我们选定样本量 $N_0=N_1=1\ 000$，置换次数 $K=1\ 000$，各次试验的重复次数1 000.

我们首先研究两种方法的 I 型错误率. 在原假设之下，对照和病例两组不存在差异，相当于设定 $\alpha=0$. 我们选取总变量数分别为 10 到 100 个不等. 对于1 000次的重复抽样，置信水平为 0.05 的 95% 置信区间为(0.036 5，0.063 5). 由图 1 可以看出，两种检验方法的 I 型错误率都处于可控范围内，因此两种检验方法都是可行的.



虚线分别为 0.05 水平以及 95%置信区间的上下界

图 1　PAVT 和 VT 在显著性水平 0.05 下的 I 型错误率比较

Fig. 1　The estimated type I error rates comparison of PAVT and VT methods at significant level 0.05

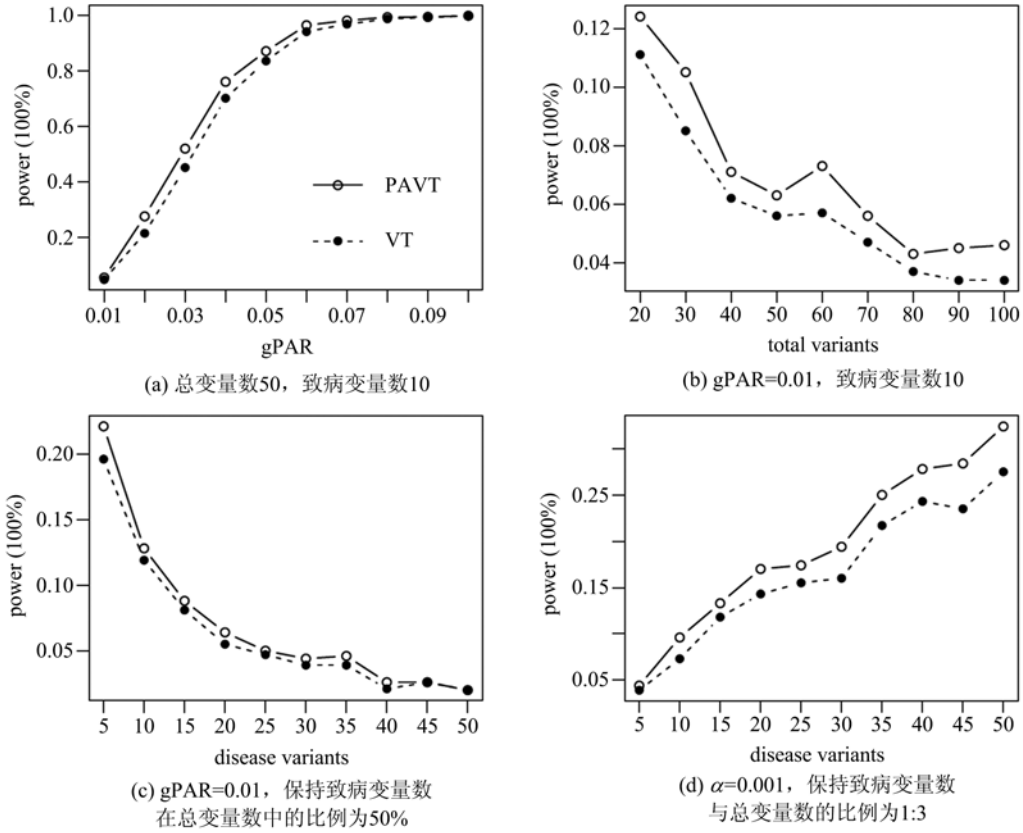
接着我们比较两种检验方法在不同参数选取下检验功效的大小. 根据变量对疾病的作用，可以把变量分为 3 类：对疾病有致病作用的变量称为致病变量，对疾病有抑制作用的变量称为有益变量，对疾病没有作用的变量称为冗余变量. 我们首先研究

相对简单的情形即只存在致病变量和冗余变量的情况.

模拟试验中涉及的参数主要为群体 PAR(gPAR)、边际 PAR(α)、致病变量数(disease variants)和总变量数(total variants).我们分别考察检验方法在以下 4 种参数设置下的功效变化:①固定总变量数为 50,致病变量数为 10,从 0.01 到 0.1 改变 gPAR,在这种情况下 α 也随之而增大;②固定 gPAR 为 0.01,致病变量数为 10 个,改变总变量数;③固定 gPAR 为 0.01,增大致病变量数与总变量数,但保持致病变量数占总变量数 50%,此时相应的 α 在减小;④固定每个致病变量的 α 为 0.001,保持致病变量数与总变量数的比例为 1∶3,改变致病变量数.图 2 展示了两种检验方法在这 4 种参数设置下的功效变化情况.从图 2 可以看出,在变量的 gPAR 越大,α 越大,致病变量的含量越高,冗余变量数越少即冗余信息越少的情况下,两种方法的检验功效都更高.而我们提出的 PAVT 方法的检验功效总是优于 VT 方法的检验功效.

当变量中同时存在致病变量和有益变量的时候,由于两种变量对疾病的作用方向是相反的,因此会导致作用相互抵消,对于稀有变量来说,本来就稀疏的信息更难捕捉到.我们也研究了当 3 种变量同时存在时两种检验方法的功效.在图 2(a)参数设置①的 50 个变量中改变 5 个冗余变量为有益变量,有益变量的 α 为−0.01,MAF 为 0.11.由图 3 可以看出,我们的 PAVT 方法的检验功效依然高于 VT 方法的检验功效.

冗余变量的存在会降低检验的功效,因此从所有变量中筛选出致病变量也是一项重要的工作.PAVT 和 VT 在对关联性做检验的同时也对变量进行了筛选,极大得分值所对应的阈值以下的 MAF 对应的变量被认为是致病的,我们记录了1 000次重复抽样的致病变量数的均值记为 SDV,而其中对应的真实的致病变量数均值记为 RDV.RDV 与 SDV 的比率 ratio 称为筛选效率.我们以参数设置①总变量数 50 个、致病变量数 10 个、gPAR 从 0.01 到 0.1 变化这一试验模拟结果为例进行说明.从表 1 可以看出,PAVT 方法挑选的致病变量数和实际致病变量数都略少于 VT 方法,但是筛选的效率要更高.



(a) 总变量数50，致病变量数10

(b) gPAR=0.01，致病变量数10

(c) gPAR=0.01，保持致病变量数在总变量数中的比例为50%

(d) α=0.001，保持致病变量数与总变量数的比例为1:3

图 2 在不同的参数变化下 PAVT 和 VT 方法在显著性水平 0.01 下的功效对比

Fig. 2 Power comparison of PAVT and VT methods at significance level 0.01 under different parameter settings

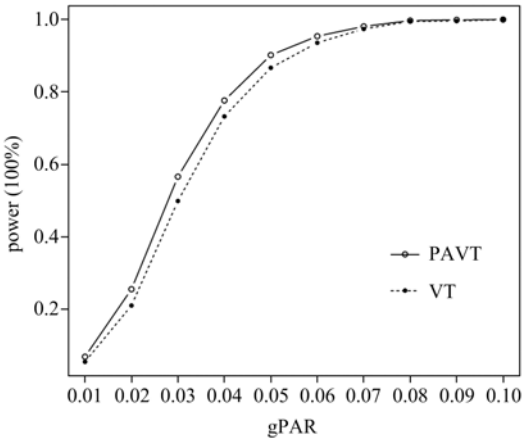


图 3　当 3 种变量同时存在时 PAVT 和 VT 方法在显著性水平 0.01 下的功效对比

Fig. 3　Power comparison of PAVT and VT methods at significant level 0.01 with presence of the three types of variants

表 1　参数设置①时致病变量的筛选情况对比

Tab. 1　Comparison of the disease variants selection under the parameter setting: 50 total variants and 10 disease variants

gPAR	SDV		RDV		ratio	
	PAVT	VT	PAVT	VT	PAVT	VT
0.01	12.119	13.036	2.929	3.07	0.242	0.236
0.02	13.735	15.078	3.576	3.795	0.26	0.252
0.03	14.295	16.043	3.795	4.082	0.265	0.254
0.04	15.349	17.762	4.194	4.606	0.273	0.259
0.05	16.454	18.768	4.474	4.877	0.272	0.260
0.06	17.514	19.939	4.731	5.167	0.270	0.259
0.07	18.349	21.137	4.942	5.443	0.269	0.258
0.08	18.805	21.987	5.041	5.600	0.268	0.255
0.09	20.328	22.794	5.345	5.769	0.263	0.253
0.10	20.396	23.635	5.390	5.972	0.264	0.253

3　结论

针对稀有变量和疾病性状的关联分析的检验问题，我们从稀有事件角度，以泊松分布代替二项分布，提出一个新的检验统计量，并在变阈值 VT 方法的基础上得到了泊松逼近变阈值稀有变量分析 PAVT 方法.

PAVT 方法和 VT 方法都认为变量的次等位基因频率越小越有可能是致病基因，因此都采用对照群体中的次等位基因频率作为阈值来划分所有的基因变量，把 MAF 低于阈值的作为潜在致病变量，得到检验统计量，由于阈值的不确定性，故采用多个阈值得到不同的检验得分值，从中选取最大值，并用置换的方法得到检验的 p 值. 两种方法的不同之处在于 PAVT 方法的出发点不同，提出的检验统计量和得分函数也不同. 文献[6]提出的 rb-VT 方法采用风险比率 $\sum_{j=1}^{N_1} X_{ij} / \sum_{k=1}^{N_0} Y_{ik}$ 作为不同的阈值，并用 VT 方法中的检验统计量得到得分函数，从而选取最大值作为最后的得分. 3 种方法从框架上来看都有相似之处，不过选取的检验统计量和阈值的选取方式有所不同，最终得到的检验效果也不同.

模拟试验结果表明，我们的 PAVT 方法在多重参数设置下都比 VT 方法拥有更高的检验功效. 不管有益变量是否被考虑进模拟中，PAVT 方法都拥有更高的功效. 另外，通过对比两种检验方法对致病变量的筛选结果，PAVT 的筛选效率也略高. 这一结果说明当应用于稀有变量这一特殊的基因变量的检验问题时 PAVT 方法所采取的检验统计量要优于 VT 方法所采用的检验统计量，这与 PAVT 方法的出发点是相吻合的. PAVT 方法基于稀有事件这一出发点，以泊松分布来代替二项分布对基因变量进行计数，更符合稀有变量的数据特点.

在模拟试验中，通过 Wright's 方程抽得的变量的突变等位基因频率不止包含了稀有变量(MAF 不高于 0.01)，也包含了普通变量. 这些变量随机地被分配成为致病变量、有益变量或者冗余变量. 由试验模拟结果可以看出，普通变量存在的情况下，PAVT 方法的检验功效依然更高，因此在实际应用中 PAVT 方法也更加可行.

文献[14]提出了一系列的自适应检验方法：

$$aT = aT(U) = \min_{1 \leqslant m \leqslant k} P_{T(U_{(m)})},$$

$aT(U)$表示的是检验统计量 $T(U)$的自适应形式，不同的检验统计量 T 以及检验统计量中的得分向量 U 的不同排序方式都可以得到不同的检验统计量，例如 PAVT，VT 和 rb-VT 方法就可以视为该自适应方法的一个实际例子. 从这个角度出发，我们将来可以寻求以我们提出的统计量 T 为检验统计量的自适应检验方法，并寻求合适的变量排序，有望得到新的适合稀有变量与疾病性状的关联性检验问题的方法，以得到更高的检验功效.

参考文献(References)

[1] Morgenthaler S, Thilly W G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST) [J]. Mutation Research, 2007, 615:28-56.

(下转第 219 页)