·基础理论与方法·

应用Empower Stats软件估计基因环境交互作用研究中的样本含量

林林　陈常中

【导读】 在基因环境交互作用的研究中，传统的样本含量估计公式和设置的参数只考虑一种暴露，未考虑两个以上因素的交互作用，依此样本含量而进行的研究，其检验效能往往很低（文中以例1和例2说明）。在基因-环境交互作用研究的设计阶段，可采用随机模拟法估计样本含量及检验效能。Empower Stats软件可以方便地实现上述分析，同时输出R统计模拟程序，并输出对每个模拟数据所做的回归方程参数。

【关键词】 基因环境交互作用；样本含量；检验效能；随机模拟

Sample size estimation of gene-environmental interaction and the actual implementations on the Empower Stats software

LIN Lin[1], *CHEN Chang-zhong*[2]. 1 *Department of Epidemiology, School of Binzhou Medical College, Yantai* 264003, China; 2 *Dana-Farber Cancer Institute, Harvard Medical School, Boston*, USA

Corresponding author: CHEN Chang-zhong, Email: changzhong_chen@dfci.harvard.edu

【Key words】 Gene-environmental interaction; Sample size; Power; Monte Carlo Method

疾病的基因环境交互作用研究可采用横断面调查、前瞻性研究和病例对照研究设计，而其中一个重要因素就是研究所需的样本含量，以确定在规定的统计显著性水平下达到预期检验效能。传统的样本含量估计公式，设置的参数只考虑一种暴露，未考虑两个以上因素的交互作用，因而检验效能往往很低。为此以下介绍在基因环境交互作用研究中采用随机模拟法估计样本含量及检验效能。

基本原理

随机模拟法是近年来随着计算机和软件技术快速发展的一种计算方法，其基本思路是借助随机数发生器，按预定样本含量与假想的各变量分布产生数据，并按照预定的方法对该随机产生的数据进行统计检验，记录是否拒绝无效假设(H_0)，重复以上过程若干次（如1000次），其中得出拒绝无效假设(H_0)的次数所占的比例即为统计检验效率的估计。如一项横断面调查的样本含量为N，假设一个基因G突变率为P_g，一个环境暴露因素E在人群中暴露比例为P_e，E独立于G，即暴露与基因型无关。所观察的结局变量Y服从正态分布(μ,σ)，E与G对Y的作用可用如下回归表达式表示

$$Y_i=\beta_0+\beta_1\times G_i+\beta_2\times E_i+\beta_3\times G_i\times E_i+\varepsilon_i$$

式中i=1，2，…，N，表示调查个体；$G_i=1$表示i个体有突变，0表示未突变；$E_i=1$表示i个体有暴露，0表示未暴露。

根据β_1、β_2、β_3、P_g、P_e与μ可以计算β_0。

对每个假想的个体i(i=1，2，…，N)，首先用随机数发生器按P_g产生G_i；E与G是独立的，直接按P_e产生E_i；若E与G不独立，即$P_e|G=0$不等于$P_e|G=1$，则根据G_i按相应的P_e产生随机数E_i；再根据$\beta_0+\beta_1\times G_i+\beta_2\times E_i+\beta_3\times G_i\times E_i$计算出$\check{y}_i$，即该个体$Y_i$的期望值，再按($\check{y}_i$，$\sigma$)产生随机数$Y_i$。数据生成后，运用上述模型得出对$\beta_3$检验的$P$值。根据所预定的显著水平(0.05)判断是否成功地检测到G与E的交互作用。

上述模型及原理很容易扩展到病例对照研究、前瞻性研究。环境暴露因素可以是两分类变量和连续性变量，结局变量也可以是连续性变量和两分类变量。以下按研究类型、结局变量类型、环境暴露变量类型分别列出不同情况的回归模型和模拟数据所需参数（表1～3）。

DOI: 10.3760/cma.j.issn.0254-6450.2013.03.019

作者单位：264003 烟台，滨州医学院流行病学教研室（林林）；美国哈佛大学医学院Dana-Farber癌症研究所（陈常中）

通信作者：陈常中，Email：changzhong_chen@dfci.harvard.edu

表1 病例对照研究中回归模型与模拟数据所需要的参数

结局变量(Y)	交互作用回归模型	模拟数据	
		暴露变量(X)特征	需要参数
二分类变量	$\log[p/(1-p)]=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2$，其中$e^{\beta_1}$是在$X_2=0$时，$X_1$的效应(*OR*)，称$X_1$的主效应；$e^{\beta_2}$是在$X_1=0$时，$X_2$的效应(*OR*)，称$X_2$的主效应；$e^{\beta_3}$反映在$X_1$、$X_2$均存在时的附加效应，$X_1$和$X_2$的交互作用	模型中包含2个二分类变量X_1和X_2	①X_1在目标人群中的暴露率；②X_2在目标人群的暴露率；③当$X_1=1$时，X_2的优势与$X_1=0$时X_2的比值，如果X_1和X_2相互独立，则定为1；④样本含量(包括病例数、对照数)
		模型中包含1个二分类变量X_1和1个连续性变量X_2	①X_1在总人群的暴露率；②X_2在总人群中的均数、标准差；③X_1和X_2的交互作用，即$X_1=1$和$X_1=0$时X_2均值的差值，如果X_2和X_1相互独立，设为0；④交互作用项的*OR*值；⑤样本含量(病例数、对照数)

表2 前瞻性研究中回归模型与模拟数据所需要的参数

结局变量(Y)类型	交互作用回归模型	模拟数据	
		暴露变量(X)特征	需要参数
二分类变量	$\log[p/(1-p)]=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2$，其中$e^{\beta_1}$是在$X_2=0$时$X_1$的效应(*OR*)，称$X_1$的主效应；$e^{\beta_2}$是在$X_1=0$时$X_2$的效应(*OR*)，称$X_2$的主效应；$e^{\beta_3}$反映$X_1$和$X_2$均存在时的效应(*OR*)，称$X_1$和$X_2$的交互作用	探索二分类暴露变量X_1和二分类危险因素X_2的交互作用	①Y在非暴露人群中的患病率；②X_2在目标人群中的分布情况；③X_2与X_1有交互作用下的*OR*，若两者相互独立，则定为1；④X_1的主效应；⑤X_2的主效应；⑥交互作用项的*OR*；⑦样本含量，包括非暴露组和暴露组的人数
		探索二分类暴露变量X_1和连续性危险因素X_2的交互作用	①Y在非暴露人群中的患病率；②X_2在目标人群中的均数和标准差；③X_2与X_1有交互作用下，暴露组X_2的均数差与非暴露组的比值，若两者相互独立，设为0；④X_1的主效应(回归系数)；⑤X_2的主效应(回归系数)；⑥交互作用项的回归系数；⑦样本含量，包括非暴露组和暴露组的人数
连续性变量	$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2$，回归系数β表示X对Y的效应，其中$\beta_1$是在$X_2=0$时$X_1$的效应，称$X_1$的主效应；$\beta_2$是当$X_1=0$时$X_2$的效应，称$X_2$的主效应；$\beta_3$反映$X_1$和$X_2$均存在时的效应，称$X_1$和$X_2$的交互作用	探索二分类暴露变量X_1和二分类危险因素X_2的交互作用	①Y在非暴露人群中的均数和标准差；②X_2在目标人群中的分布情况；③X_2与X_1有交互作用下的*OR*，若两者相互独立，则定为1；④X_1的主效应(回归系数)；⑤X_2的主效应(回归系数)；⑥交互作用项的回归系数；⑦样本含量，包括非暴露组和暴露组的人数
		模型中包含连续性结局变量，探索二分类暴露类变量X_1和连续性危险因素X_2的交互作用	①Y在非暴露人群中的均数和标准差；②X_2在目标人群中的均数和标准差；③X_2与X_1有交互作用下，暴露组X_2的均数差与非暴露组的比值，若两者相互独立，设为0；④X_1的主效应(回归系数)；⑤X_2的主效应(回归系数)；⑥交互作用项的回归系数；⑦样本含量，包括非暴露组和暴露组的人数

实例分析

［例1］ 以Vandenbroucke等[1]关于口服避孕药(环境因素E)和Leiden因子V基因突变(遗传因素G)与静脉血栓栓塞的病例对照研究资料为例，描述模型中包含2个二分类变量时，采用随机模拟法估计样本含量及检验效能。其中病例组静脉血栓栓塞患者155例，健康对照169例。如以对照组基因突变率和暴露率代表一般人群基因突变率和暴露率，根据资料计算两者分别为0.035(6/169)和0.38(65/169)，E独立于G，G的主效应*OR*值为6.94，E的主效*OR*值应为3.70，交互作用*OR*值为1.35(34.72/3.70/6.94)。按所用样本含量155∶169，1000次模拟数据计算出的检验G的作用检验效能为0.368，检验E的作用的检验效能为0.988，检验G与E交互作用的检验效能仅为0.005。

［例2］ 以香港男性关于吸烟(环境因素E)和肿瘤家族史(遗传因素G)与肺癌的病例对照研究资料为例[2]，描述模型中包含1个二分类变量和1个连续性变量时，采用随机模拟法估计样本含量及检验效能。其中病例组男性肺癌患者1208例，社区人群对照1069例。如果以对照组肿瘤家族史阳性率和吸烟量代表一般人群的暴露率及平均水平，根据资料两者分别为0.125(134/1069)与17.1±29.0，E独立于G，G的主效应*OR*值为1.51，E的主效*OR*值为1.48，交互作用*OR*值为1.08。按所用样本含量1208∶1069，1000次模拟数据计算出的检验G的作用检验效能为0.189，检验E作用的检验效能

表3　横断面研究中回归模型与模拟数据所需要的参数

结局变量(Y)类型	交互作用回归模型	模拟数据	
		暴露变量(X)特征	需要参数
二分类变量	$\log[p/(1-p)]=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2$；$e^{\beta_1}$是当$X_2=0$时$X_1$的效应，称$X_1$的主效应；$e^{\beta_2}$是当$X_1=0$时$X_2$的效应，称$X_2$的主效应；$e^{\beta_3}$反映$X_1$和$X_2$均存在时的效应，称$X_1$和$X_2$的交互作用	探索二分类暴露变量X_1和二分类危险因素X_2的交互作用	①Y在目标人群中的患病率；②X_1在目标人群中的分布情况；③X_2在目标人群中的分布情况；④X_2与X_1有交互作用下的*OR*，若两者相互独立，则定为1；⑤X_1的主效应(*OR*)；⑥X_2的主效应(*OR*)；⑦交互作用项的*OR*，待计算的效应；⑧样本含量，包括研究对象总数
		探索二分类暴露变量X_1和连续性危险因素X_2的交互作用	①Y在目标人群中的患病率；②X_1在目标人群中的分布情况；③X_2在目标人群中的均数和标准差；④X_1存在下X_2的效应，即$X_1=1$和$X_1=0$时X_2均数差比值，若两者相互独立，则定为0；⑤交互作用项的*OR*，待计算的效应；⑥样本含量，包括研究对象总数
连续性变量	$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2$，以回归系数表示X对Y的效应，即β；$\beta_1$是在$X_2=0$时$X_1$的效应，称$X_1$的主效应；$\beta_2$是在$X_1=0$时$X_2$的效应，称$X_2$的主效应；$\beta_3$反映$X_1$和$X_2$均存在时的额外效应，称$X_1$和$X_2$的交互作用	探索二分类暴露变量X_1和二分类危险因素X_2的交互作用	①Y在目标人群中的均数和标准差；②X_1在目标人群中的分布情况；③X_2在目标人群中的分布情况；④X_1存在时X_2的*OR*，即$X_1=1$与$X_1=0$时X_2的*OR*，若两者相互独立时，定为1；⑤X_1的主效应(回归系数)；⑥X_2的主效应(回归系数)；⑦交互作用项的回归系数；⑧样本含量，包括研究对象总数
		探索二分类暴露变量X_1和连续性危险因素X_2的交互作用	①Y在目标人群中的均数和标准差；②X_1在目标人群中的分布情况；③X_2在目标人群中的均数和标准差；④X_1存在时X_2的效应，即$X_1=1$与$X_1=0$时X_2的均数差比值，若两者相互独立时，定为0；⑤X_1的主效应(回归系数)；⑥X_2的主效应(回归系数)；⑦交互作用项的回归系数；⑧样本含量，包括研究对象总数

为1.000，检验G与E交互作用的检验效能为0.043。

［例3］　假定①G的突变频率为0.2，暴露E的频率为0.4，E独立于G；②G的主效应*OR*值为2.0，E的主效应为1.5；③病例组500例，取不同的对照例数(400、500、600、700、800)；④要检测G与E的交互作用大小分别为1.3、1.5、1.8，对每种组合情况模拟次数均为1000次得出的检验效能，结果见表4和图1。Empower Stats(易侑统计)软件可以实现上述分析，同时输出R统计模拟程序，并输出对每个模拟数据所做的回归方程的参数。计算1000个模拟方程各回归系数的均值，通过与预定的效应比较，可以验证模拟过程是否正确。根据1000次回归系数的均数计算的G(x_1)、E(x_2)的*OR*值分别是2.0、1.5，G与E交互作用的*OR*值分别是1.3、1.5、1.8，与预定的效应完全相符(表5)。

表4　不同样本含量与检验效应的检验效能
(病例数，$n_1=500$)

对照例数(n_2)	G和E的交互作用(*OR*=1.3)			G和E的交互作用(*OR*=1.5)			G和E的交互作用(*OR*=1.8)		
	G主效应	E主效应	交互作用	G主效应	E主效应	交互作用	G主效应	E主效应	交互作用
400	0.884	0.751	0.102	0.885	0.731	0.187	0.897	0.756	0.290
500	0.923	0.789	0.111	0.914	0.778	0.209	0.923	0.800	0.351
600	0.946	0.819	0.116	0.935	0.808	0.232	0.947	0.828	0.396
700	0.956	0.838	0.131	0.940	0.841	0.260	0.960	0.853	0.423
800	0.962	0.849	0.146	0.959	0.854	0.278	0.966	0.876	0.473

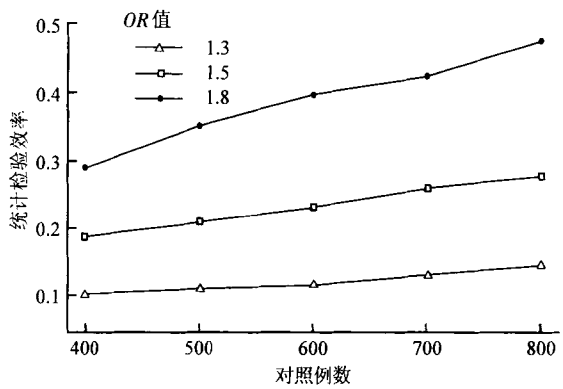

图1　检验效能与样本含量及交互作用大小之间的关系

讨　论

随机模拟作为一种广义的计算方法，利用计算机模拟实际过程，然后加以统计处理并求得实际问题的解。由于模拟的过程可反复进行，系统结构和参数的改变也较为容易，因此应用广泛且适用性强，

表5 1000个模拟方程各回归系数的均值($\bar{x}$)、标准差($\pm s$)和*OR*值(设病例数n_1=500,对照例数n_2=600)

G和E交互作用的预定效应		回归系数(β)		验证效应($OR=e^{\beta}$)
		$\bar{x}$	$\pm s$	
OR=1.3	G主效应	0.6983	0.1987	2.0104
	E主效应	0.4078	0.1418	1.5035
	交互作用	0.2691	0.3326	1.3088
OR=1.5	G主效应	0.6884	0.1985	1.9905
	E主效应	0.4025	0.1416	1.4956
	交互作用	0.4153	0.3423	1.5148
OR=1.8	G主效应	0.7020	0.2024	2.0178
	E主效应	0.4100	0.1375	1.5068
	交互作用	0.5936	0.3489	1.8105

特别在传统理论难以解决的问题上不失为一种有效可行的方法[3]。近年在医药学领域,随机模拟应用研究不断深入[4-6]。使用Empower Stats(易侕统计)与R软件通过随机模拟可以方便计算检验效能(www.empowerstats.com)。该软件设有检验效能模拟功能模块,设计有对两因素交互作用检验效率分析,可以选择不同的研究类型与两种不同的结局变量类型(两分类型与连续型),输入不同的样本含量与要检验的交互作用效应值,输出不同样本含量与不同的交互作用大小的组合情况下的检验效能,并以图形显示检验效能与样本含量及交互作用大小之间的关系。QUANTO是目前较为广泛使用的基因与疾病研究样本量估计的一款软件,在选择了设计类型和提供了相关参数后(检验水准α、统计效能、人群中该病发病率、研究的基因易感型频率、遗传方式、*OR*值等),利用公式计算出研究所需要的样本量[7]。而Empower Stets软件则是通过计算机模拟估算样本量和检验效能。

QUANTO软件专门为基因研究设计,而Empower Stats软件则是通用的,基因暴露可以当成一个两分类变量,相当于显性模型或隐性模型。在流行病学研究中暴露因素的交互作用非常常见,如高血压导致心肌梗死的风险在不同年龄组不同,也就是说高血压效应可以被年龄的效应所改变,年龄和高血压在缺血性心脏病研究中存在交互作用。当年龄的主效应、高血压主效应和年龄-高血压交互作用的预定效应值一定时,按照要求的检验效能,利用Empower Stats软件可以方便的估算出研究所需要的样本量,因此Empower Stats软件较QUANTO软件具有更广泛的适用性。

参 考 文 献

[1] Vandenbroucke JP, Koster T, Briët E, et al. Increased risk of venous thrombosis in oral-contraceptive users who are carriers of factor V Leiden mutation. Lancet, 1994, 344: 1453-1457.

[2] Qiu H, Yu DX, Xie LY, et al. Interaction between continuous variables in logistic regression model. Chin J Epidemiol, 2010, 31(7): 812-814. (in Chinese)
邱红,余德新,谢立亚,等. Logistic回归模型中连续变量交互作用的分析. 中华流行病学杂志,2010,31(7):812-814.

[3] Gao HX. Statistical calculation. Beijing: Peking University Press, 2003: 173-221. (in Chinese)
高惠璇. 统计计算. 北京:北京大学出版社,2003:173-221.

[4] Gao J, Dong W, Gao ES, et al. Comparison between multivariate generalization of the Cox proportional hazards models and Cox proportional hazards model by simulation. Chin J Health Stat, 2007(3): 248-251. (in Chinese)
高峻,董伟,高尔生,等. 多结局生存分析模型与Cox模型的随机模拟比较. 中国卫生统计,2007(3):248-251.

[5] Wang Y, Li W, Cheng XR, et al. Sample size calculation in noninferiority trial by Monte Carlo Method. Chin J Health Stat, 2008(1): 26-28. (in Chinese)
王杨,李卫,成小如,等. 随机模拟法验证非劣效临床试验样本量计算公式. 中国卫生统计,2008(1):26-28.

[6] Chen CS, Xu YY, Yuan TF, et al. Multivariate random coefficients model of repeated measures data in medical research. J Fourth Mil Med Univ, 2004, 25(23): 2182-2185. (in Chinese)
陈长生,徐勇勇,袁天峰,等. 医学多变量重复观测资料的随机系数模型. 第四军医大学学报,2004,25(23):2182-2185.

[7] Gauderman WJ, Morrison JM. QUANTO1.1: A computer program for power and sample size calculations for genetic-epidemiology studies, http://hydra.usc.edu/gxe http://hydra.usc.edu/gxe, 2006.

(收稿日期:2012-09-12)

(本文编辑:张林东)