中国农业大学学报 2006,11(2):1-6
Journal of China Agricultural University

序贯多重决策过程及其在基因组研究中的应用

张群远 Michael A. Province
(美国 华盛顿大学医学院 基因组科学中心 统计基因组学部,圣路易斯 63108)

摘 要 降低和控制对大量分子标记进行测验的统计错误,已成为目前基因组研究中的普遍难题。针对此问题,本文根据序贯分析和多重假设测验的理论,介绍了可用于变化样本容量下对大量标记同时进行测验分组的序贯多重决策过程的原理和方法,探讨了该方法在实际应用中的计算精度和计算量优化等关键问题及相应的解决方案,最后以 1 套包括 5 841 个 SNP 和 87 个细胞系的药物遗传学实验数据进行实例分析,并与传统的测验结果比较,表明了序贯多重决策过程在实际应用中的优点和可行性。

关键词 基因组;多重假设测验;统计错误;序贯分析;序贯多重决策过程
中图分类号 Q 348 **文章编号** 1007-4333(2006)02-0001-06 **文献标识码** A

Sequential multiple decision procedures and applications in genome studies

Zhang Qunyuan, Michael A. Province
(Division of Statistical Genomics, Center for Genome Sciences,
Washington University School of Medicine, St. Louis, MO 63108, USA)

Abstract How to reduce and control statistical errors of testing larger numbers of markers has become a common and difficult problem in genome researches. To deal with this issue, under the theories of sequential analysis and multiple hypothesis test, we introduce in this paper a novel method, Sequential Multiple Decision Procedure (SMDP), which can be used for testing and grouping large numbers of makers with sequentially varying samples size. Some crucial practical aspects (computational accuracy and amount, etc.) of this method were investigated, and corresponding strategies were proposed. A set of real data from a pharmaco-genetics experiment containing 87 cell lines and 5841 SNPs was analyzed, in comparison with traditional methods, as an example to demonstrate the feasibility and advantages of SMDP.

Key words genome; multiple hypothesis test; statistical error; sequential analysis; sequential multiple decision procedure (SMDP)

近年来,采用大量的分子标记对全基因组进行“扫描”,以寻找影响和控制人类、动植物重要性状表达的遗传位点或基因,已经成为遗传研究的一个重要策略。其中单核苷酸多态性(SNP)标记和 RNA 表达水平的基因芯片检测是 2 种最重要的技术。这些技术通常对数以万计的标记或基因同时进行检测,最后需要对大量的标记或基因进行统计显著性测验。这时,全基因组水平上的第一类统计错误率会变得很大,导致把很多无效标记或基因判断为有效,即假阳性(false positive)增加。比如,在显著水平 $\alpha = 0.01$ 上对 1 套 HGU133PLUS 基因芯片的约 57 000个基因在 2 组样本中的表达差异进行测验,即使 2 组样本没有任何真实差异,但仅仅由于随机误差的缘故,也大约会有近$57\,000 \times 0.01 = 570$ 个基因被认为在 2 组样本间差异显著,这给有目的地选择有效的基因进行后续试验带来了很大的困难。这

收稿日期:2006-03-15
基金项目:美国 NIH 资助项目(U01 GM63340-05)
作者简介:张群远,博士后,主要从事统计遗传学和统计基因组学研究,E-mail:qunyuan@wustl.edu

时,如何控制和降低假阳性是基因组研究中的重要问题。

降低全基因组水平上的第一类错误率(即假阳性)的最常用办法是采用较小的 α 显著水平[1],或者对无效假设下所得的概率(即 P 值)进行矫正[2-4],但这样做会反过来增加第二类统计错误,导致一些真正的基因检测不到,即假阴性(false negative)增加[5]。那么,如何在假阳性和假阴性之间平衡,成了一件"矛盾"事,是目前基因组研究中普遍存在的难题。从统计理论上来看,这一矛盾是由于比较错误率(test-wise error)和实验错误率(experiment-wise error)的不一致引起的(比较错误率即单个标记或基因的统计推断错误率,实验错误率指全基因组意义上统计推断错误率)。由于我们只是用双假设测验(包含一个无效假设和一个备择假设)的方法来处理多重假设的问题,试图通过单个标记或基因的独立测验来得到全基因水平上的结论,所以,这种不一致难以避免[6]。因此,有必要直接从多重假设的角度来寻找全基因组标记或基因检测的更有效的统计方法。

基因组研究中还有一个重要问题是,由于资源和经费上的考虑,样本容量一般比较小(通常都在几十以内),所以抽样误差较大,降低了统计功效。常规的统计方法只利用少数的几个统计数(比如平均数、方差、t 统计量等),没有充分利用样本内信息。如何挖掘和利用样本内的各种分布信息来提高统计分析的功效,如何以最少的样本容量来获得尽可能准确的结论,也是非常值得研究的问题。

序贯分析中的多重决策理论正好可以用于解决以上 2 个问题。序贯分析(sequential analysis)是近代统计学的一大分支,最早由 Wald(1947)创建[7]。后来,根据序贯分析和多重假设测验的理论,Bechhofer 和 Kiefer 等发展了一系列序贯识别和排名的模型和方法,其中的序贯多重决策过程(sequential multiple decision procedure,SMDP)可用于解决从 k 个处理中选择 t 个最佳处理的问题[8],但长期以来没有得当足够的重视和充分应用。近年来,Province 把 SMDP 发展应用到人类基因组的研究中[9]。随后,Zhang 和 Province 对该方法做了进一步完善[6],改进了其统计精度并提出计算优化的方法。本研究在上述研究基础上,对 SMDP 的理论、方法及其在基因组上的应用做概述性介绍、总结和讨论,以引起对此方法的更多关注、发展和应用。

1 SMDP 的基本原理和一般过程

设某试验对 k 个处理进行重复观测,得到第 i 个处理的第 j 次重复观测值,以 X_{ij} 表示。理论上,k 个处理对应着 k 个总体,如果这些总体都属于同一类型的 Koopman-Darmois 总体[8],根据序贯分析的多重假设测验理论,在试验进行到任何第 m 次重复($m \geq 1$)时,可以得到第 i 个处理对应于 m 的序贯统计量 Y_{im} 为:

$$Y_{im} = \sum_{j=1}^{m} P(X_{ij}) \quad (i = 1, 2, \cdots, k) \qquad (1)$$

$P(X_{ij})$ 为观测值 X_{ij} 的某个特定函数。设在 k 个处理中选择任意 t 个处理的所有可能组合数目为 U,对于第 u 种组合方式,在第 m 次重复时的序贯统计量 $Y_{um}^{(t)}$ 为:

$$Y_{um}^{(t)} = \sum_{i=1}^{t} Y_{im} \quad (u = 1, 2, \cdots, U) \qquad (2)$$

对 U 个 $Y_{um}^{(t)}$ 值从小到大排序列,得到一序列统计量 $Y_{[1]m}^{(t)} \leq Y_{[2]m}^{(t)} \leq \cdots \leq Y_{[U]m}^{(t)}$。对于 $Y_{um}^{(t)}$ 值最大(即 $Y_{um}^{(t)} = Y_{[U]m}^{(t)}$)的 t 个处理的组合,可得到另一个序贯统计量:

$$W_{[U]m}^{t} = \frac{\exp(\delta^{*} \times Y_{[U]m}^{(t)})}{\sum_{v=1}^{U} \exp(\delta^{*} \times Y_{[v]m}^{(t)})} \qquad (3)$$

若 $Y_{[U]m}^{(t)}$ 对应的 t 个处理与其余 $k - t$ 个处理的最小真实差异为 δ,$W_{[U]m}^{t}$ 即为 $\delta \geq \delta^{*}$ 时正确选择到这 t 个最佳处理的概率。根据 $W_{[U]m}^{t}$ 的计算,Bechhofer 和 Kiefer 构建了从 k 个处理中选择 t 个有差异的最佳处理的 SMDP 过程[8]如下:

1)事先根据实际需要确定一个可接受的正确选择的置信概率 P^{*} 和期望最小真实差异 δ^{*};

2)选择一个初始的最小样本容量 m_0(即重复数 = m_0),通常可以从 $m_0 = 1$ 开始;

3)以一定的步长增加样本容量,通常每次增加 1 个重复,构成序贯样本;

4)根据公式(3),每次样本增加后,都计算新样本容量(即重复数 = m)下的 $W_{[U]m}^{t}$ 值;

5)不断循环 3)和 4)步,直到 $W_{[U]m}^{t} \geq P^{*}$,便停止分析。此时,$W_{[U]m}^{t}$ 对应的 t 个处理即为与其余 $k - t$ 个处理差异显著的最佳处理。

2 SMDP 在标记-性状连锁分析中的应用

与大多数假设测验方法一样(如 t 测验和 F 测

验等),SMDP 并不是一种独立的方法,它需要结合特定的模型和 Koopman-Darmois 分布来使用。由于模型分布转化以及计算量等问题,SMDP 多年没有受到充分的发展和应用。Province (2000) 首先把 SMDP 用于人类遗传连锁分析中同胞对设计,把 SMDP 和回归模型结合起来[9]。具体做法是,把每个遗传分子标记看成 1 个处理,把每对同胞看做 1 次重复(同胞对的数目即为样本容量)。根据 Haseman 和 Elston 的同胞对分析原理[10],可以得到标记 i 在同胞对 j 之间的后裔同源概率(IBD)和同胞对 j 之间表型性状差异的平方(D^2)。常规的 Haseman-Elston 分析是利用全部样本,以标记 i 的多个同胞对的 D^2 对 IBD 做回归,来推断该标记是否与性状连锁。SMDP 则是从 1 个小样本开始(比如随机抽取的 10 个同胞对),然后每次增加 1 个新的随机抽取的同胞对,进行序贯分析。

Province 推导了标记 i 在序贯样本 h 下的序贯回归误差平方和($Y_{i,h}$)的计算公式,并解决了 Koopman-Darmois 分布的转化问题,使得 $Y_{i,h}$ 可用于 SMDP 分析。$Y_{i,h}$ 即为公式(1)和(2)中的 Y_{im}(即 $h = m$),其值越小,回归越显著,标记的遗传效应就越大。SMDP 的目的就是利用尽可能少的同胞对来找出 $Y_{i,h}$ 显著最小的标记组。值得注意的是,由于公式(3)中 $W^t_{[U]m}$ 对应的是 Y_{im} 值显著最大的 t 个处理,所以最后找出的实际是 $Y_{i,h}$ 显著最大的标记组(即遗传效应不显著的标记),剩余的标记才是我们需要寻找的标记。

Province 通过模拟研究表明,与其他一些主要的基于显著水平 α 和 P 值矫正的传统测验方法相比,SMDP 方法的 2 类统计错误率都较低,并且只需要较小的样本容量。由于回归是普遍应用的方法,基于回归的 SMDP 事实上也可以应用于同胞设计以外的很多分析,尤其在全基因组与性状关联的标记或基因的检测中。由于式(3)需要计算 k 个标记中 t 个标记的所有可能组合,标记多时存在计算量过大的问题,Province 在其研究中采用了一个最简化的公式去计算式(3)中 $W^t_{[U]m}$,导致 $W^t_{[U]m}$ 的估计偏低,而且标记数目越多,这种偏离越严重,并需要更多的样本作为"补偿"。所以,当我们把这一方法尝试应用于基因组标记检测或基因芯片实际数据时,往往只能识别出效应明显特别强的少数标记。为此,Zhang 和 Province 对该方法做了进一步改进,提出使用更具有普遍性的简化序贯多重决策过程(Simplified Sequential Multiple Decision Procedures, SSMDP),可以对 $W^t_{[U]m}$ 的计算精度和计算量进行优化控制[6]。以下即对 SSMDP 及其相关问题进行介绍和讨论。

3　简化序贯多重决策过程(SSMDP)

基因组分析中,当分子标记和基因的数目(k 值)很大时,即使 t 很小(比如 $k = 50\,000$, $t = 5$),若按公式(3)来计算 $W^t_{[U]m}$,由于需要计算 k 个标记中 t 个标记的所有可能组合,计算量将会非常大,这会妨碍 SMDP 在实际中的有效应用,甚至在计算资源上无法实现。为减少计算量又不过多损失计算精度,Zhang 和 Province 建议采用如下更一般化的公式:

$$W^{[U-S]}_{[U]m} = \frac{\exp(\delta^* \times Y^{(t)}_{[U]m})}{(S-1)\exp(\delta^* \times Y^{(t)}_{[S]m}) + \sum_{v=S}^{U}\exp(\delta^* \times Y^{(t)}_{[v]m})}$$

$$Y^{(t)}_{[1]m} \le Y^{(t)}_{[2]m} \cdots \le Y^{(t)}_{[S-1]m} \le Y^{(t)}_{[S]m} \le Y^{(t)}_{[S+1]m} \cdots$$

$$Y^{(t)}_{[U-1]m} \le Y^{(t)}_{[U]m} \qquad (4)$$

公式(4)只用 $Y^{(t)}_{um}$ 值最大的前 $U - S + 1$ 个(而不是全部 U 个)组合来计算 $W^t_{[U]m}$。若 $S = 1$,公式(4)即等同于公式(3);若 $S = U - 1$,则相当于 Province 最初采用采用的算法[9]。这里,$U - S + 1$ 叫做 $Y^{(t)}_{um}$ 值最大的顶部组合数目(Top Combination Number, TCN),即 $TCN = U - S + 1$。显然,TCN 越大,即公式(4)和(3)越接近,$W^t_{[U]m}$ 的计算精度就越高,但同时也就需要越多的计算量。以上过程称为简化序贯多重决策过程(SSMDP),其优点是可以通过调节 TCN 的大小来平衡 $W^t_{[U]m}$ 的计算量和计算精度,以满足实际分析的需要。

4　SSMDP 在实际中的应用

4.1　有效 TCN 和计算优化

若要把 SSMDP 应用于基因组的实际研究,关键问题是如何选择一个合适的 TCN 值,以实现计算量和计算精度的优化平衡。为解决这一问题,我们利用药物遗传学的一套数据来研究 TCN 和 $W^t_{[U]m}$ 之间的关系。该数据包含 29 个三元家系共 87 个个体的细胞系在人类第 9 染色体上5 841个 SNP(来自人类基因组 HapMap 项目中的 CEPH 数

据[11-12])和这些细胞系对化疗药物的反应数据(来自作者目前参加的正在进行的美国 PGRN 药物遗传学研究项目)。把每个细胞系的 SNP 作为自变量(编码成 0,1,2),药物反应值作为因变量,按随机顺序获得序贯样本。在序贯样本容量为 50(即 $m = 50$)下,获得序贯回归误差平方和 Y_{im},然后利用公式(2)和(4)计算不同 TCN($TCN = 2 \sim 10\,000$)和不同 t($t = 2 \sim 12$)对应的 $W^t_{[U]m}$值。结果(图 1)表明,TCN 变大时,$W^t_{[U]m}$逐渐上升,但 TCN 增加到一定值以后,$W^t_{[U]m}$不再明显增加。我们把 $W^t_{[U]m}$开始稳定时的 TCN 称为有效 TCN(Efficient TCN, ETCN)。ETCN 的发现很重要,这意味着 SSMDP 完全可以使用一个远远小于 U 的 TCN(即 ETCN)来获得 $W^t_{[U]m}$的准确估计,从而有效克服计算量和计准确性之间的矛盾,使 SSMDP 的应用成为可能。

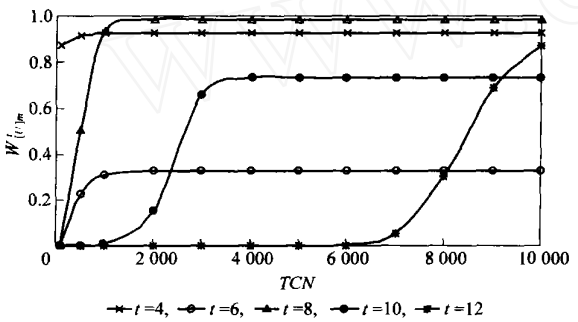

图 1 $W^t_{[U]m}$和 TCN 的变化关系

Fig. 1 Relations of $W^t_{[U]m}$ and TCN

4.2 显著标记(或基因)数目的确定

SSMDP 应用中还有一个重要问题是,不同 t 对应的 $ETCN$ 不同,但由于 t(即真正差异显著的标记或基因的数目)在实际上并不知道,所以也就很难用一个统一的 ETCN 来对数据进行分析。解决这一问题,可采用如下过程:首先,用一个较小的 TCN(比如 $TCN = 100 \sim 500$)来对一个较小的 t(比如 $t = 1 \sim 3$)进行 SSMDP 测验。如果发现有差异显著的标记被检出,则增加 TCN 和 t 值再进行测验,重复此过程,直到即使不断增加 TCN 也没有新的标记被检测出。最后一次检测出的显著标记的数目即为总的显著标记的数目。这样做的优点是在确定显著标记的数目的同时,可以避免使用过大的 TCN 来进行分析,从而节约计算量,这对于大规模基因组数据的分析尤其有用。

4.3 期望最小真实差 δ^* 的确定

与常规的假设测验不同,SMDP 分析并不指定显著水平 α,而是指定正确选择的概率 P^* 和期望最小真实差异为 δ^*。通常使用 $P^* = 0.95$。δ 是度量 2 个有差异的标记组之间的最小距离的一个复杂指标,对于以上回归模型,可根据均值已知但方差未知的正态总体对应的 Koopman-Darmois 分布函数特征来计算[8-9]:

$$\delta = \left| \frac{1}{2\sigma^2_{M-1}} - \frac{1}{2\sigma^2_1} \right|$$

式中:M 为总的标记数目;σ^2_{M-1}表示 $M - t$ 个无效应标记中最小的序贯回归误差方差;σ^2 表示 t 个效应显著的标记中最大的序贯回归误差方差。实际分析中,可以用 σ^2_{M-t}和 σ^2_t 估计值来计算 δ,并以之作为 δ^*,但这样做效果往往并不理想,因为小样本下 σ^2_{M-t}和 σ^2_t 的估计往往不够精确。

如何选择一个合适的 δ^*,是 SMDP 应用的一个难点。理论上,SMDP 是假定真实的 δ 已知,但实际上并非如此。所指定的 δ^* 若比真实的小,则对 SMDP 分辨力要求过高,这一方面需要更大的样本容量,另一方面会识别不出一些真正有效应的标记或基因(即增大 β 错误);δ^* 若比真实的大,则会检测出一些没有真实效应的标记或基因(即增大 α 错误)。因此,SMDP 中 δ^* 的确定需要有客观的依据,并根据实际数据来估计和调节。一个可行的做法是选择不同样本顺序下序贯分析结果比较稳定的 δ^*;或者固定一个比较大的 δ^*,选择那些不同样本顺序下都稳定显著的标记。这样做虽然需要进行多次分析,会增加一定的计算量,但可以有效地降低统计错误。通过 δ^* 来调节统计错误率,是 SMDP 有别于传统统计测验的一个重要特点。

5 SSMDP 的应用实例

最后,为了对 SSMDP 和传统分析方法做一实际的比较,利用 4.1 中的数据,按 4.2 的策略进行 SSMDP 分析($P^* = 0.95$, $\delta^* = 10$),并用 SSMDP 停止后的剩余样本进行常规的回归验证分析;同时也对全部样本做了传统的回归分析,对回归测验的 P 值做了 Bonferroni 和 FDR 矫正[4](表 1)。在5 841 个 SNP 中,传统的回归分析一共发现 72 个 SNP 在 0.01 水平上显著。我们知道,由于第一类统计错误的存在,即使没有任何 SNP 真正具有遗传效应,而是仅仅由于随机误差,也会引起平均5 841 ×0.01≈

58 个左右的标记显著。所以,72 个显著标记中应该有约 58 个左右的假阳性标记,但具体哪些标记是假阳性,回归模型本身无法识别。若对 P 值进行目前常用的 Bonferroni 和 FDR 矫正,则所有的标记无一显著,即使决定系数为 0.136,其标记也被推断为无效,这显然是由于假阴性所致。由此可见传统测验方法在基因组研究中面临着矛盾。

相比之下,SSMDP 以 $TCN = 10^4$ 发现了 11 个显著的最强效应标记,不断增加 TCN 到 10^5 也没有新的标记检出,所以可认为这 11 个标记为具有真正效应的全部标记。这一总的结论的正确性由概率 $P^* = 0.95$ 保证。理论上,应该有 72 - 58 = 14 个左右的真实显著的标记存在,SSMDP 检测出 11 个,其结果与 14 个的期望比较吻合。需要指出的是,这 11 个标记并不一定是传统回归分析中决定系数最大的 11 个标记,决定系数最大的标记并不一定是真实有效的标记。另外,SMDP 对不同标记检出的序贯样本容量不同,平均为 56,为总样本容量(87)的 64 %左右。用剩余的样本对这 11 个标记进行回归验证,其中有 6 个在 0.05 水平上显著。

从以上比较分析可以看出,在面对大量标记的情况下,SMDP 的统计效能更高,所需样本较小,能给出更可靠合理的结果,这表现在它对假阳性的良好控制上。

表 1　固定样本容量的回归分析,SSMDP 及其验证的结果

Table 1　Results of fixed sample regression, SSMDP and validation

SNP 标记名称	固定样本容量回归 ($n = 87$)				SSMDP 所需样本容量	剩余样本的回归验证 P
	R^2	P	Bon	FDR		
rs1039280	0.136	0.000 5	1.0	0.76	47	0.046
rs925223	0.136	0.000 5	1.0	0.76	47	0.046
rs1995788	0.114	0.001 6	1.0	0.76	48	0.029
rs3121619	0.100	0.003 1	1.0	0.76	48	0.024
rs1337753	0.135	0.000 6	1.0	0.76	52	0.089
rs2786798	0.135	0.000 6	1.0	0.76	52	0.089
rs2786797	0.123	0.001 0	1.0	0.76	52	0.089
rs2117464	0.099	0.003 3	1.0	0.76	58	0.054
rs2149342	0.099	0.003 3	1.0	0.76	58	0.054
rs3935846	0.126	0.000 9	1.0	0.76	75	0.030
rs3935053	0.126	0.000 9	1.0	0.76	75	0.030

注: R^2 和 P 分别为传统回归分析的决定系数和显著性测验的 P 值。Bon. 和 FDR 分别为用 Bonferroni 和 FDR 方法对 P 值的矫正所得值。

6　讨　论

SMDP 是一种非常适合于基因组研究中同时对大量标记进行测验的方法,其理论上比较完备,具有较高的统计功效,实际应用上也非常可行。总结起来,该方法具有样本的序贯性和假设的多重性这 2 个重要特点。

样本的序贯性是指 SMDP 并不采用固定的样本进行测验,而是从 1 个小样本开始,不断增加样本容量,边增加样本边进行测验,直到发现所需要的显著差异为止。这样做可以非常有效地利用样本,用最小的样本容量来发现我们期望的差异。这一特点可用于序贯试验设计,节约资源。对于样本数已经固定的数据,则可以利用 SMDP 筛选结束后的"多余"数据来对结论进行验证,或者按不同样本顺序进行多次分析,从而提高统计推断的可靠性。正是由于这一特点,使 SMDP 能够充分利用样本变化过程中的丰富信息,理论上比传统的固定样本容量的分析具有更高的统计功效,在试验设计上也更加灵活。

假设的多重性是指 SMDP 测验时所针对的假设并非传统的双假设——1 个无效假设和 1 个备择假设——而是设多个假设。事实上,从 k 个处理中选择 t 个最佳显著处理的测验中一共有 U 个假设,U 为在 k 个处理中选择 t 个处理的所有可能组合数目,因为任何 t 个处理和其余 $k - t$ 个处理都可能存在真实差异。SMDP 就是在这 U 种可能的假设中寻找最大可能的 1 种,并对其进行测验。由于这一特点,SMDP 用于基因组分析时,并不对每个标记(或两两标记间)独立进行多次测验,而是通过一次性的测验把全部的标记分成有差异的 2 组。这样就避免了单个标记测验和全基因组水平上测验的统计错误的不一致性,使得 SMDP 可以同时有效控制 2 类统计错误,不再需要做 P 值矫正。

与传统的测验方法相比,SMDP 的主要缺点是计算复杂,需要较大的计算量。除前面讨论的 SSMDP 等一些优化方法外,我们还可以寻找各种减少计算量的途径。例如,如果我们对有真实效应标记的最大和最小可能数目有所认识,我们就可以根据 ETCN 曲线把 SSMDP 所用的 TCN 值限定在特定范围内。对于样本容量较大的数据,序贯过程中每次可以按较大的步长(大于 1)来增加样本容量,比如,与步长 = 1 相比,步长 = 5 可节约近 80 %的计算量;对于标记数目过于庞大的数据,可以先用常规的

方法初步滤掉一些极不可能显著的标记,然后才做 SMDP 分析,也是一个可行的途径。总之,由于计算机技术的发展,以及算法上的灵活可控,计算量问题应该不会成为 SMDP 应用的障碍。

目前 SMDP 应用上的主要限制是直接适用的统计模型及分布类型不多。前面提到,SMDP 目前只能用于 Koopman-Darmois 分布家族。Koopman-Darmois 分布家族是具有特定概率密度函数形式的单参数分布族,而常用的正态分布等都具有 1 个以上的参数,需进行复杂的分布转化才能应用于 SMDP。所以,进行分布转化将是 SMDP 应用的重要研究内容。同时,也需要研究非 Koopman-Darmois 分布下的 SMDP 过程。如果这 2 方面的问题能得到不断完善和发展,SMDP 将可成为基因组以及其他类似领域的研究中的非常有前景的统计测验工具。

参 考 文 献

[1] Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results[J]. Natural Genetics, 1995, 11: 214-217

[2] Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance[J]. Biometrika, 1988, 75: 800-803

[3] Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test [J]. Biometrika, 1988, 75: 383-386

[4] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing[J]. Journal of the Royal Statistical Society Series B, 1995, 57: 89-300

[5] Rao D C. CAT scans, PET scans and genomic scans[J]. Genetic Epidemiology, 1998, 15: 1-18

[6] Zhang Q Y, Province M A. Simplified sequential multiple decision procedures for genome scans[M] //Proceedings of American Statistical Association (Biometrics section), 2005: 463-468

[7] Wald A. Sequential analysis[M]. New York: Dover Publications Inc, 1947

[8] Bechhofer R E, Kiefer J, Sobel M. Sequential identification and ranking procedures[M]. Chicago: The University of Chicago Press, 1968

[9] Province M A. A single, sequential, genome-wide test to identify simultaneously all promising areas in a linkage scan[J]. Genetic Epidemiology, 2000, 19: 301-332

[10] Haseman J K, Elston R C. The investigation of linkage between a quantitative trait and a marker locus[J]. Behavioral Genetics, 1972, (2): 3-9

[11] The International HapMap Consortium. The International HapMap Project[J]. Nature, 2003, 426: 789-796

[12] Thorisson G A, Smith A V, Krishnan L, Stein L D. The International HapMap Project Web site[J]. Genome Research, 2005, 15: 1591-1593