诺禾致源微生物扩增子统计文档

1 假设检验的一般步骤 1
2 参数检验和非参数检验 2
3 显著检验与多重假设检验校正.....2
4 T 检验.....3
5 秩和检验 4
6 方差分析 5
7 参考文献 7

1 假设检验的一般步骤

在扩增子分析中经常要对某个假设作出判断，例如判断两个分组的某个指标是否具有差异，差异的程度如何，是否具有统计学意义，这些都需要进行假设检验。在扩增子报告中的统计检验部分，包括组间差异分析，组间差异检验和差异物种分析部分都需要用到假设检验的内容。下面简单介绍一下假设检验的一般步骤。首先看下相关的一些基本概念。

包含所研究的全部个体（数据）的集合，称为总体（population）。

从总体中抽取的一部分元素的集合，称为样本（sample）。

用来描述总体特征的概括性数字度量，称为参数（parameter）。

用来描述样本的概括性数字度量，称为统计量（statistic）。

我们所关心的参数通常有总体平均数，总体标准差和总体比例等。与总体相对应我们通常关心的统计量有样本平均数、样本标准差、样本比例等。由于总体的无限性和总体参数较难确定，通常可以从总体中随机抽取样本，通过计算样本的统计量来估计总体的参数。例如我们要判断污染环境中和非污染环境中的微生物具有差异，我们通常是从两种环境中抽取样本，根据样品的一些指标或统计量来判断两种环境具有差异，这就是用样本来估计总体的例子。

假设检验就是先对总体参数提出某种假设，然后利用样本信息判断假设是否成立的过程。假设检验一般分为 4 个步骤：

（1）提出假设

在假设检验中，首先需要提出两种假设，即原假设和备选假设。**通常将研究者想收集证据予以反对的假设称为原假设，或称零假设，用 H_0 表示。通常将研究者想收集证据予以支持的假设称为备选假设，或称研究假设，用 H_1 表示。**原假设所表达的含义总是表述为组间没有差异，变量之间没有关系。与原假设对立，备选假设通常表述为组间具有差异，药物疗效显著提高等。假设检验的一般思路就是先假定 H_0 成立，然后从逻辑从对立面证明真理。

（2）构造检验统计量

根据样本观测结果计算得到的，并据以对原假设和备选假设作出决策的某个样本统计量，称为检验统计量。在进行假设检验时，根据检验的目的不同检验统计量有不同的计算方法。检验统计量还取决于所抽取的样本数和总体的分布情况。

（3）根据显著水平，确定临界值和拒绝域

当原假设正确时拒绝原假设，所犯的错误称为第 I 类错误。犯第 I 类错误的概率，称为显著性水平记为α。根据显著性水平确定的拒绝域的边界值，称为临界值。当总体分布已知时，例如总体服从正态分布，我们可以根据给定的显著性水平（通常为 0.01 或 0.05）查表获得临界值。当总体分布未知时，可以先用 Permutation test 构造经验分布，再根据显著性水平获得临界值。

（4）做出检验决策

将第（2）步计算出的检验统计量与（3）步获得的临界值进行比较，作出拒绝或不拒绝原假设的决策。

传统的统计量检验的方法是在检验之前确定显著性水平α，也就意味着事先确定了临界值和拒绝域。这样，不论检验统计量的值是大还是小，只要它的值落入拒绝域就拒绝原假设，否则就不拒绝原假设。这种给定显著性水平的方法，无法给出观测数据与原假设之间不一致

北京诺禾致源生物信息科技有限公司 网址：www.novogene.com 邮箱：support@novogene.com 电话：010-8283 7801
地址：北京市海淀区金码大厦 B 座 17 层（邮编：100083）| 天津市武清区创业总部基地 B07 栋（邮编：310700）

程度的精确度量。要测量出样本观测数据与原假设中假设值的偏离程度，则需要计算 pvalue 值。pvalue 值，也称为观测到的显著性水平，它表示为如果原假设 H_0 正确时得到实际观测样本结果的概率。pvalue 值越小，说明实际观测到的数据与 H_0 之间的不一致的程度就越大，检验的结果就越显著[1]。

2 参数检验和非参数检验

扩增子分析中经常遇到某种检验方法为参数检验或非参数检验，例如 T 检验、Tukey 检验、方差分析都属于参数检验，而 Wilcoxon 秩和检验，Anosim 分析、MRPP 分析、Adonis 分析、Amova 分析都属于非参数检验，那么什么叫参数检验和非参数检验，它们之间的区别是什么呢。要理解前面的问题，首先需要明白统计推断的概念。

统计推断是研究如何利用样本数据来推断总体特征的统计学方法，包括参数估计和假设检验两大类。

总体的参数一般是未知的，通常可以用样本统计量来对总体的参数进行估计，例如可以用样本均值对总体均值进行点估计，利用样本均值的分布对总体均值进行区间估计，这些都称为参数估计。

对未知参数的假设进行检验称为参数统计，所用的检验叫做参数检验（Parameter test）。

不依赖总体分布的具体形式，也不对参数进行估计或检验的统计方法，叫做非参数统计，其检验方法就是非参数检验（Non-parametric test）。

参数检验一般要利用总体的信息（总体的分布、总体的一些参数特征如方差等），以总体分布和样本信息对总体参数作出推断。

参数检验和非参数检验的区别：

1 参数检验是针对参数做的假设，非参数检验是针对总体分布情况做的假设，这个是区分参数检验和非参数检验的一个重要特征。例如两样本比较的 t 检验是判断两样本分别代表的总体的均值是否具有差异，属于参数检验。而两样本比较的秩和检验（wilcoxcon 检验及 Mann-Whitney 检验）是判断两样本分别代表的总体的位置有无差别（即两总体的变量值有无倾向性的未知偏离），自然属于非参数检验[2]。

2 二者的根本区别在于参数检验要利用到总体的信息（总体分布、总体的一些参数特征如方差），以总体分布和样本信息对总体参数作出推断；非参数检验不需要利用总体的信息（总体分布、总体的一些参数特征如方差），以样本信息对总体分布作出推断。

3，参数检验只能用于等距数据和比例数据，非参数检验主要用于记数数据。也可用于等距和比例数据，但精确性就会降低。

那么什么时候用参数检验，什么时候用非参数检验呢？非参数检验一般不直接用样本观察值作分析，统计量的计算基于原始数据在整个样本中的秩次，丢弃了观察值的具体数值，因此凡适合参数检验的资料，应首选参数检验。但是不清楚是否合适参数检验的资料，则应采用非参数检验。

3 显著检验与多重假设检验校正

在任何一个严谨的科学测量中，判断两个数值是否具有差异，必须要考虑这个差异

北京诺禾致源生物信息科技有限公司 网址：www.novogene.com 邮箱：support@novogene.com 电话：010-8283 7801
地址：北京市海淀区金码大厦 B 座 17 层（邮编：100083）| 天津市武清区创业总部基地 B07 栋（邮编：310700）

的两方面来源：一是真实存在的差异，二是可能来源于检测或随机误差。而一般的显著检验的目的，就是计算出观测到的差异来源自随机误差的概率。例如，在进行显著检验时先提出假设，原假设为两个分组中 A 物种的丰度没有差异，备选假设为两个分组中 A 物种的丰度具有差异，随后通过假设检验的一系列步骤得到 pvalue 小于 0.05，然后作出判断拒绝原假设选择备选假设（即两个分组中 A 物种的丰度存在差异），这一次判断犯错的概率小于 0.05（这里的 pvalue 就是假阳性率，False positive rate）。

上述只是做了一次假设检验判断，但是在很多科学实验中，例如扩增子的组间差异物种分析的 T-test 分析、MetaStat 分析都必须要做多次的判断。再例如，在 RNAseq 分析以及微阵列分析中我们要判断两组样本对应的 1000 个基因的表达量是否存在组间差异 :分别判断 A 基因是否具有差异？B 基因是否具有差异？C 基因是否具有差异？…..，如此下去，我们要进行 1000 次比较。如果我们以 p value 1% (假阳性的概率是 1%)来作为阈值，并假设每次判断都是彼此独立的，那么即使这 1000 个基因实际上都没有差异，我们也可能会得出有 10 个差异基因的结论。也就是说小概率事件经过多次反复尝试后，变成了一个多次出现的事件。在进行多次检验后（也就是所说的多重检验，multiple test），那么基于单次比较的检验标准将变得过于宽松，使得阳性结果中的错误率（FDR 值 False Discovery Rate）已经大到令人不可忍受的地步。那么怎么办？最好的办法就提高判断的标准（p value），单次判断的犯错概率就会下降，那么总体犯错的概率也将下降。**在多重检验中提高判断标准的方法，我们就称之为“多重检验校正”。**

从 1979 年以来，统计学家提出了多种多重检验校正的方法。最简单严厉的方法要属于 Bonferroni 校正[3]，该方法通过将原阈值 p value 1%，除以多重检验的次数作为新的阈值。虽然这种方法能显著降低总体犯错的概率，但是由于单次校验阈值标准较高，使得真正具有差异基因筛选较少，实际是假阴性率提高了。目前最常用的方法是 FDR 校正。Benjamini 和 Hochberg 在 1995 年[4]第一次提出了 FDR 的概念以及相应的多重检验校正方法。FDR 就是一种控制阳性结果中的假阳性率的思路。假设你挑选了 R 个差异表达的基因，其中有 S 个是真正有差异表达的，另外有 V 个其实是没有差异表达的，是假阳性的。实践中希望错误比例 Q=V/R 平均而言不能超过某个预先设定的值（比如 0.05），在统计学上，这也就等价于控制 FDR 不能超过 5%。以后的多种多重检验校正方法都是 FDR 的控制方法的延伸，John Storey（2003）[5] 提出了一个被校正后的 p value 的概念（比 P value 更严格），称之为 Q value，也广泛应用于目前的多重检验校正。在一般情况下，大家可以简单一些理解，FDR、Q value、Adjusted p-value 指的是一个东西。

4 T 检验

T 检验是假设检验的一种，又叫 student t 检验（Student’s t test），主要用于样本含量较小（例如 n<30），总体标准差 σ 未知的正态分布资料。本部分对应于扩增子报告的组间差异物种分析的 T-test 部分。可分为三种类型:单样本 t 检验，配对样本 t 检验，两独立样本 t 检验。

单样本 t 检验主要应用于推断样本所代表的未知总体均数μ与已知总体均数μ_0有无差别。已知总体均数μ_0一般为理论值、标准值或经大量观察所得到的稳定值。该类型最常见的表述为重复测量一个指标，比较测量的均值与已知的规定值（总体均值）是否具有显著的差异。

配对样本 t 检验，常用于医学研究中。配对设计主要有 4 种情况：同一受试对象处理前后的数据，同一受试对象两个部位的数据，同一样品用两种方法检验的结果，配对的两个受试对象分别接受两种处理后的数据。

北京诺禾致源生物信息科技有限公司 网址：www.novogene.com 邮箱：support@novogene.com 电话：010-8283 7801
地址：北京市海淀区金码大厦 B 座 17 层（邮编：100083）| 天津市武清区创业总部基地 B07 栋（邮编：310700）

两独立样本 t 检验就是根据样本数据对两个样本来自的两独立总体的均值是否有显著差异进行推断。该类型对样本是否配对没有要求，经常用于比较两组计量资料的均数间有无显著差别，也是扩增子分析中应用最多的类型。进行两独立样本 t 检验需要满足两个前提条件：1.两样本应该是相互独立。2.样本来自的两个总体应该服从正态分布。

统计独立性是指随机现象的结果不呈现显著性联系，一般都可以满足。样本来自的总体是否服从正态分布需要进行正态性检验。最常用的正态检验方法为 W 检验，对应于 R 语言中的 shapiro.test()函数。

由于 T 检验中两样本方差代表的总体方差是否相等，对应不同的计算方法。在满足样本独立性和样本总体服从正态分布之后，还需要对两样本的方差齐性进行检测。对于所有方差齐性检验，原假设都为“各样本的总体方差全部相同“，备择假设则为“至少有两个样本的方差不同”。方差齐性一般采用 F 检验的方式，用较大的样本方差S_1^2比上较小的样本方差S_2^2。v1 和 v2 分别为样本的自由度，然后通过查 F 分布表即可获得确定的概率值,判断方差齐性。方差齐性检验对应于 R 语言中的 bartlett.test(),var.test()和 car 包中的 leveneTest()函数。

$$F = \frac{S_1^2}{S_2^2}$$

$$v_1 = n_1 - 1$$

$$v_2 = n_2 - 2$$

经过独立性检验，正态性检验，方差齐性检验后可以对两组独立样本进行 t 检验。两组例数可以相等，也可以稍有出入。检验的方法同样是先假定两组相应的总体均数相等，看两组均数实际相差与此假设是否靠近，近则把相差看成抽样误差表现，远到一定界限则认为由抽样误差造成这样大的相差的可能性实在太小，拒绝假设而接受 H_1，作出两总体不相等的结论。T 检验对应于 R 语言中的 t.test(x,y, var.equal = FALSE)函数，x 和 y 分别是两组样品的数值，var.equal 参数可以选择方差是否相等。

5 秩和检验

秩和检验（rank sum test）是最常用的非参数检验方法，也称为秩转换（rank transformation），该方法在非参数检验中占有重要地位。本部分文档对应于扩增子报告的 Alpha、Beta 多样性指数组间差异分析的 wilcoxon 秩和检验部分。在进行秩和检验时首先将原始数据从小到大，或等级从弱到强转换成秩后，再基于秩次的统计量，进行检验，做出统计推断。秩和检验的特点是用数据的秩代替原始数据进行假设检验，它对总体分布的形状差别不敏感，只对总体分布的位置差别敏感。

秩和检验根据样品是否配对和比较样本的多少可分为四种类型：配对样本比较的 Wilcoxon 符号秩检验，两独立样本比较的 Wilcoxon 秩和检验，完全随机设计多个样本比较 Kruskal-Wallis H 检验，随机区组设计多个样本比较的 Friedman M 检验。

Wilcoxon 符号秩检验（Wilcoxon signed rank test）一般用于两种情况，一是配对样本差值的中位数与 0 的比较，二是单个样本中位数和已知的一个总体中位数比较。两配对样品的符号秩检验对应于 R 语言中的 wilcox.test(x,y,paired=TRUE)函数[6]，x 和 y 分别是两配对样品的数值。单个样品与已知总体中位数的 Wilcoxon 符号秩检验对应于 R 语言中的 wilcox.test(x,mu=a)，x 和 mu 分别为单个样品的数值，mu 为总体的中位数。

两组独立样本比较的 Wilcoxon 秩和检验等效于 Mann-Whitney 检验，目的是推断计量资料或等级资料的两个独立样本代表的两个总体分布是否有差别，常用于比较两组计量资料的中位数有无显著差别。理论上 H_0 为两总体分布相同，即两样本来自同一总体；H_1 为两

北京诺禾致源生物信息科技有限公司 网址：www.novogene.com 邮箱：support@novogene.com 电话：010-8283 7801
地址：北京市海淀区金码大厦 B 座 17 层（邮编：100083）| 天津市武清区创业总部基地 B07 栋（邮编：310700）

总体分布不同。由于秩和检验对两总体分布形状的差别不敏感，对位置相同、形状不同但类似的两总体分布，推断不出两总体分布(形状)有差别，故在实际应用中，H_0可写作两总体分布位置相同，也可简化为两总体中位数相等。与两独立样本 t 检验相对应，wilcoxon 秩和检验也是扩增子分析中应用最多非参数检验的类型。两组独立样本比较的 Wilcoxon 秩和检验对应于 R 语言中的 wilcox.test(x,y,paired=FALSE)函数，x 和 y 分别是两组样品的数值。

完全随机设计多个样本比较 Kruskal-Wallis H 检验用于推断计量资料或等级资料的多个独立的样本所来自的多个总体分布是否有差别。在比较两个以上的总体时，广泛使用 KW 秩和检验，它是两样本的 Wilcoxon 方法在多于两个样本时的推广。H 检验的假设 H_0应为多个总体分布的位置相同，H_1为多个总体分布的位置不同。Kruskal-Wallis 秩和检验的 R 函数是 kruskal.test（x~g）,其中 x 是由各个分组数据构成的向量或列表，g 是由因子构成的向量。当 Kruskal-Wallis H 检验的结果拒绝 H_0，接受 H_1，认为多个总体分布位置不全相同。当需要进一步判定哪两个分组之间具有差异时，还需要使用 Wilcoxon 秩和检验，对任意两样本进行两两比较，然后再进行假设检验的校正。

FriedmanM 检验用于推断随机区组设计的多个相关样本所来自的总体分布是否有差别。FriedmanM 检验和 Kruskal-Wallis H 检验区别主要取决于实验的目的，如果目的在于比较分组间有没有差别，就用 Kruskal-Wallis H 检验，如果是比较每组自身内部多次实验间有没有差别，就用 Friedman M 检验。Friedman M 检验会在分组内部进行排秩，而 Kruskal-Wallis H 检验是在多个分组之间进行排秩。Friedman M 检验秩和检验的 R 函数是 friedman.test()。如果检验结果判断分组之间有差别，还需要使用 Wilcoxon 秩和检验，对任意两样本进行两两比较，然后再进行假设检验的校正。

6 方差分析

方差分析（ANOVA）由英国统计学家 R.A.Fisher 首创，为纪念 Fisher，方差分析又称 F 检验（F test）。在实际工作中，影响一件事的因素是很多的，例如：不同的生产厂家，不同的原材料，不同的操作规程，以及不同的技术指标对产品的质量都会有影响。方差分析就是研究一种或多种因素对试验结果的观测值是否具有显著影响。与 T 检验用于两个样本平均数的假设检验不同，方法差分析是用于两个以上样本平均数的假设检验方法，它通常用于推断多个总体均数有无差异。与 Kruskal-Wallis H 非参数检验类似，ANOVA 分析并不能详细得出哪几个分组存在差异，如果想得到存在差异的分组还需要通过多重假设检验来验证。

人们在试验中所考察到的数量指标如产量，性能等称为观测值，影响观测值的条件称为因素。因素的不同状态称为水平，一个因素可以采用多个水平。在一项试验中，可以得出一系列不同的观测值。引起观测值不同的原因是多方面的，有的是处理方式不同或者条件不同引起的，称为因素效应（或者称处理效应，条件变异）。有的是试验过程中偶然性因素的干扰或者观测误差所导致的，称为试验误差。方差分析的原理就是将测量数据的总变异按照变异原因的不同分解为因素效应和试验误差，并对其做出数量分析，比较各种原因在总变异中所占的重要程度，作为统计推断的依据，由此确定进一步的工作方向[7]。

方差分析按照研究因素的数量的不同分为单因素方差分析，多因素方差分析。其中多因素方差分析通常还要考虑多个因素之间的相互作用，较为复杂，多数只是研究双因素方差分析，本文档以单因素方差分析为例做简单介绍。单因素方差分析的步骤和一般的假设检验的步骤相同，也是分为提出假设，构造检验统计量，根据显著水平，确定临界值和拒绝域，做出检验决策这几个步骤。

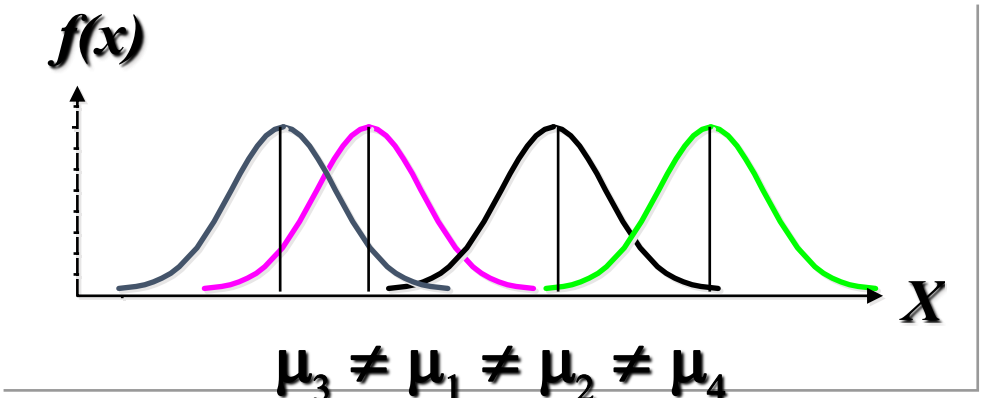
（1）提出假设

北京诺禾致源生物信息科技有限公司 网址：www.novogene.com 邮箱：support@novogene.com 电话：010-8283 7801
地址：北京市海淀区金码大厦 B 座 17 层（邮编：100083）| 天津市武清区创业总部基地 B07 栋（邮编：310700）

方差分析的 H_0 原假设为比较因素 A 的 k 个水平之间没有差异，Ha 备选假设为比较因素 A 的 k 个水平不全相等。

$$Ho: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \ldots = \mu_k$$

$$Ha: \text{ At least one } \mu_k \text{ is different}$$



（2） 构造检验统计量

构造检验统计量需要计算各个水平的均值$\bar{y}j$，全部观察值的总均值$\bar{\bar{y}}$,离差平方和 SS 均方 *MS*。最后根据计算的检验统计量 F 比和 F 分布确定方差分析的检验结果。表 1 列出了单因素方差分析的相关参数，表的下方列出了总变差，因素变差，误差变差计算的公式。

表 1 单因素方差分析表

方差来源	SS（平方和）	df(自由度)	均方	F 比	P 值
因素 A	SS（Factor）效应平方和	g - 1	MS_A=SS（Factor）/ df factor	F= MSA/MSE	p
误差	SS（Error）误差平方和	$\sum_{j=1}^{g}(n_j-1)$	MS_E = SE/df error		
总和	SS（Total）总平方和	$\left(\sum_{j=1}^{g} n_j\right)-1$			

g 为水平的个数，n_j是各个水平所含的样本数

$$\sum_{j=1}^{g}\sum_{i=1}^{n_j}(y_{ij}-\bar{\bar{y}})^2 = \sum_{j=1}^{g} n_j(\bar{y}_j-\bar{\bar{y}})^2 + \sum_{j=1}^{g}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2$$

$$SS(Total) = SS(Factor) + SS(Error)$$

SS(总) = 实验总平方和（个别 - 总均值）

SS(因素) = 因素平方和（群体均值 - 总均值）

SS(误差) = 群体中的平方和（个别 - 群体均值）

SS（Total）称为总离差平方和（或称为总变差），它是所有观测值 yij 与总均值$\bar{y}$差的平方和，描绘了所有观测数据的离散程度。

SS（Error）称为误差平方和或组内平方和，它是所有观测值 yij 与各个水平均值$\bar{y}_j$差的平方和，表示了随机误差的影响。

SS（Factor）称为效应平方和或组间平方和，它是各个水平下的均值$\bar{y}_j$和总均值$\bar{y}$差的平方和，反应了因素 A 的各个水平差异的影响。

判断因素的水平是否对其观察值有影响，实际上就是比较组间方差与组内方差之间差异的大小。组间平方和 SS（Factor）除以自由度后的均方与组内平方和 SS（Error）除以自由度后的均方差异就不会太大；如果组间均方显著地大于组内均方，说明各水平(总体)之间的差异受因素的影响较大，该因素各个水平之间具有差异。

和 T 检验一样，方差分析也需要满足正态性，独立性和方差齐性的条件。ANOVA 方差分析的 R 函数是 aov[8]（x~g）其中 x 是由各个分组数据构成的向量或列表，g 是由因子构成的向量。方差分析得到单因素各个水平之间有差异的结果后，可用多重比较分析，例如 Tukey 检验的方法对各组均值差异的成对检验。扩增子报告的 Alpha、Beta 多样性指数组间差异分析的 Tukey 分析，就是先多个分组之间进行方差分析然后再两两分组之间进行 Tukey 检验。

7 参考文献

[1] 贾俊平，统计学基础[M].中国人民大学出版社,2010.

[2] 郭祖超，洪立基，杨琦.有关非参数检验应用的若干问题．中国卫生统计.1987.4.2

[3] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 65–70.

[4] Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**, 289–300.

[5] Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proceedings of the National Academy of Science 100, 9440-9445.

[6] Myles Hollander and Douglas A. Wolfe (1973), *Nonparametric Statistical Methods.* New York: John Wiley & Sons. Pages 115–120.

[7] 薛毅，陈立萍．统计建模与 R 软件[M]．清华大学出版社, 2007.

[8] Chambers, J. M., Freeny, A and Heiberger, R. M. (1992) *Analysis of variance; designed experiments.* Chapter 5 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

北京诺禾致源生物信息科技有限公司 网址：www.novogene.com 邮箱：support@novogene.com 电话：010-8283 7801
地址：北京市海淀区金码大厦 B 座 17 层（邮编：100083）| 天津市武清区创业总部基地 B07 栋（邮编：310700）