PRACTICAL COURSE
OF COMMON BIOSTATISTICS
AND BIOINFORMATICS SOFTWARES

张祥胜◎主编

常用生物统计学与生物信息学软件实用教程

科学出版社

本书得到江苏省重点建设学科——生物学科
江苏省特色专业——生物学
盐城师范学院科研项目
经费资助出版

常用生物统计学与生物信息学软件实用教程

张祥胜　主　编

科学出版社
北京

内 容 简 介

本书以生命科学研究的一般规律为主线，以生物统计学与生物信息学软件为主，深入浅出地介绍生命科学研究中常用的软件，以便学生在本科阶段初步掌握生物统计学与生物信息学相关的软件操作。其包括统计类软件如正交设计助手、Excel、Origin、SPSS、Minitab 和 DPS，生物信息学软件和工具如 NCBI、DNA 序列比对和系统发育树的构建、引物设计、DNAMAN 软件，以及科技写作相关的软件和工具等。

本书适合作为生物类、农林类、食品类等专业的本科教材，也可供相关专业的研究生和科研人员参考。

图书在版编目（CIP）数据

常用生物统计学与生物信息学软件实用教程 / 张祥胜主编. —北京：科学出版社，2015

ISBN 978-7-03-042408-2

Ⅰ. ①常… Ⅱ. ①张… Ⅲ. ①生物统计—统计分析—软件包—高等学校—教材 ②生物信息论—统计分析—软件包—高等学校—教材 Ⅳ. ① Q-332 ② Q811.4-39

中国版本图书馆 CIP 数据核字（2014）第 259032 号

责任编辑：刘 畅 丛 楠 韩书云 / 责任校对：郑金红
责任印制：霍 兵 / 封面设计：铭轩堂

科 学 出 版 社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

文林印务有限公司印刷

科学出版社发行 各地新华书店经销

*

2015 年 1 月第 一 版 开本：787×1092 1/16
2015 年 1 月第一次印刷 印张：14 1/4
字数：337 000

定价：38.00 元

（如有印装质量问题，我社负责调换）

编写委员会

主　编　张祥胜

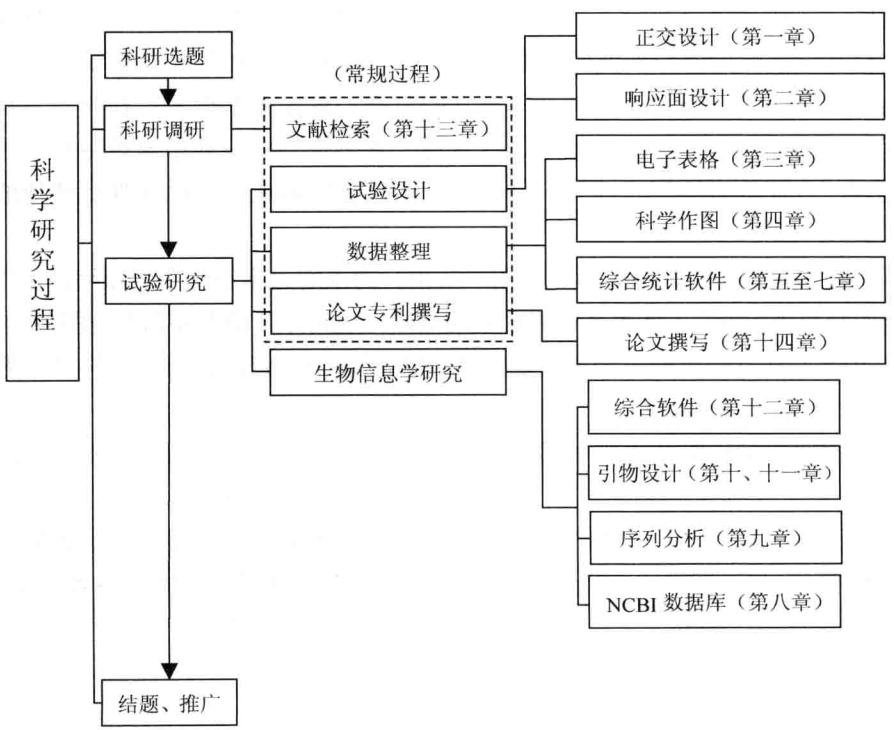
副主编　秦耀国　史正军

参　编　严泽生　胡化广　薛　菲　耿荣庆　王兰萍

前　言

生物统计学与生物信息学软件操作是生物技术、生物工程等专业本科生的专业技能课，对提高学生专业素质、增强专业能力具有重要作用，并能与今后的学习和工作有效接轨。但目前较系统地讲授生物统计学和生物信息学软件操作的教材较少，多数是较深入系统地讲解某一软件的教材，这些均不适合非统计学专业本科生或初学者使用。根据这一现状，笔者与相关的专业课教师发挥各自特长，共同编写了本书，希望对该课程的教学起到推动作用。

生物科学的研究过程大致可分为科研选题、科研调研（资料收集）、试验研究、结题鉴定和成果开发推广等过程。在学生阶段，主要分为文献检索、试验设计、数据整理、科学作图和论文专利撰写等过程。此外，在进行生物信息学和分子生物学研究时，还应包括美国国家生物技术信息中心（NCBI）数据库的使用、序列分析、引物设计等过程。本书按照这一主线进行编写，具体如下图所示。



其中虚线部分为所有自然科学科研调研和试验研究的常规过程，而生命科学的

研究还包括生物信息学研究。本书以生命科学研究的一般规律为主线，以生物统计学与生物信息学软件为主，较为系统地、深入浅出地介绍生命科学研究中常用的软件，以便学生在本科阶段初步掌握生物统计学与生物信息学相关的软件操作，掌握生物学实验（或试验）的设计、数据分析相关的概念和相关软件的使用，逐步提高学生在实验中的观察能力、分析能力、独立思考与解决问题的能力，为科研能力和专业素质的培养奠定基础，为生物统计学与生物信息学软件的教学提供参考，这样就达到了本书编写的目的。

对于以专业能力提升为目的的软件操作学习，在教与学中，应贯彻以下原则：自主学习，以练为主；严格考核，强化训练；不学则已，学则必会；实用为主，会用就行。

建议学习程序如下。

（1）课前准备：安装好相关软件，如在多媒体教室自带笔记本电脑上课，并尽量保证其能上网；仔细阅读教材，并尽量尝试操作一遍；下载或拷贝教师准备的操作素材。

（2）上课时：专心听课，认真操作，不应出现学生要求拷贝教师电脑上的软件等现象。

（3）课余时间：多加练习，直到熟练。根据一般的学习规律，要学好某种知识或技能，需要适当“过度”学习，方可完全掌握，而不是一知半解，生物统计学与生物信息学软件的操作技能的掌握也不例外。同时坚持独立完成操作作业。

值得注意的是，不同的软件之间有功能重叠，这是正常现象，在学习过程中，可根据自己的体会和使用习惯加以灵活运用。本书也有意识地尽量使用相同的数据进行处理，方便读者比较。同时，重点介绍本软件特有或比其他软件更便捷的功能和操作。

本书在编写过程中，尽量做到简明、实用，力求通过本书这个“压缩饼干”，让学生全面吸收“营养”，熟悉并掌握必要的操作，达到提升能力并为专业服务的目标。

本书除各章节分工外，全书由秦耀国副教授通读、修改书稿，张祥胜副教授撰写本书提纲，进行统稿。

非常感谢科学出版社农林与生命科学分社的编辑在本书出版过程中给予的大力支持！

由于编者水平有限，本书内容中涉及的操作较为初步和浅显，不足之处也在所难免，恳请同行和读者提出宝贵意见，使之更加完善，以有利于教学。编者电子信箱：yctu_shengwu@163.com。

张祥胜

2014 年 8 月

目　录

第一篇
生物统计类软件

本篇中的生物统计类软件是指试验设计、数据整理、科学作图和统计分析相关的软件。

试验设计是从事科学研究的设想和计划，是进行科学研究首先必须经过的步骤，也是能否达到研究目的的关键所在，主要包括课题研究设计和试验操作设计。在这里仅提因素和水平的设计，主要应用于优化试验，如发酵培养基配方优化、药用植物有效成分提取参数优化等，除单因素试验设计（简单比较法）外，还有部分因子设计 [如 Plackett-Burman(PB) 设计]、正交设计、均匀设计和响应面设计 [包括 Box-Behnken Design(BBD) 和 Central Composite Design(CCD)]，本书主要涉及 PB 设计、正交设计和 BBD，可以借助于正交设计助手和响应面设计软件完成。

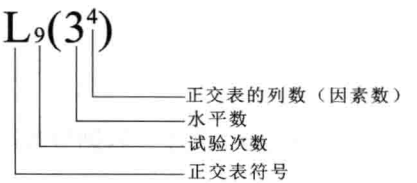
在本书中数据整理和科学作图的软件主要介绍 Excel 和 Origin，前者为电子表格软件，后者为专业科学作图软件。

统计分析主要包括差异显著性分析、作图等，本书主要介绍较为简单易学的 Statistical Product and Service Solutions(SPSS) 和 Data Processing System (DPS)，Minitab 为选学内容，Statistics Analysis System(SAS) 等专业统计软件不介绍。此外，Excel 和 Origin 也可进行数据的统计分析。

第一章　正交设计助手

正交试验设计是研究多因素多水平的一种设计方法，根据正交性从全面试验中选择部分有代表性的点进行试验，这些点具有“均匀分散、齐整可比”的特点，是一种高效、快速、经济的试验设计方法。正交试验法优点有：①试验点代表性强，试验次数少；②不需做重复试验，就可以估计试验误差；③可以分清因素的主次；④可以使用数理统计的方法处理试验结果，归纳出最优组合。

正交表是一整套规则的设计表格，L 表示正交表的符号。例如，$L_9(3^4)$ 的各组成部分的含义如下。



一个正交表中也可以各列的水平数不相等，称为混合型正交表，如 $L_8(4\times2^4)$，此表的 5 列中，有 1 列为 4 水平，4 列为 2 水平。本科阶段一般不要求掌握混合型正交表的设计与试验结果分析。

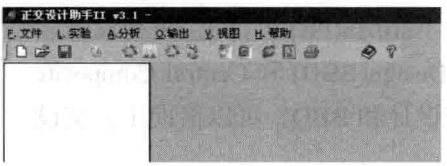


图 1-1

网上共享版本下载解压后，双击“”即可，界面较为简洁，可完成一般的正交试验的因素和水平设计、正交表自动生成、正交试验结果分析等，如图 1-1 所示。

例如，设计一个药物有效成分提取工艺优化试验，对提取温度、提取时间和用碱量进行优化，经单因素试验检验，设计各因素的水平见表 1-1。

表 1-1　因素水平表

水平	温度（A）/℃	时间（B）/min	用碱量（C）	虚拟因素
1	80	90	5%	1
2	85	120	6%	2
3	90	150	7%	3

加上虚拟因素共 4 个因素，3 个水平，标准正交试验次数为水平数的平方，因此共 $3^2=9$ 次试验。其操作步骤如下。

点击菜单“文件”→“新建工程”，再点击菜单“实验”→“新建实验”，即得如图 1-2 所示窗口。

填入“实验名称”，选择“标准正交表”（图 1-3）。

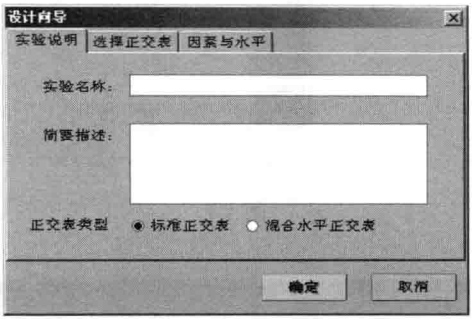


图 1-2

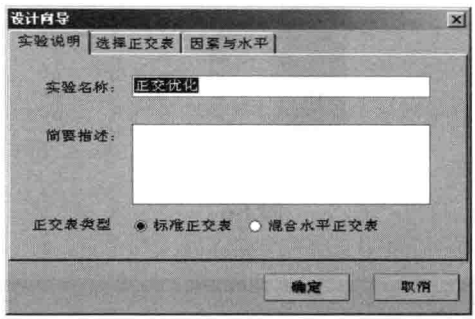


图 1-3

点击“选择正交表”，选择 L9-3-4，即 4 因素 3 水平，共 9 个试验（图 1-4）。

点击“因素与水平”，按预定的设计填入相应参数，如图 1-5 所示。

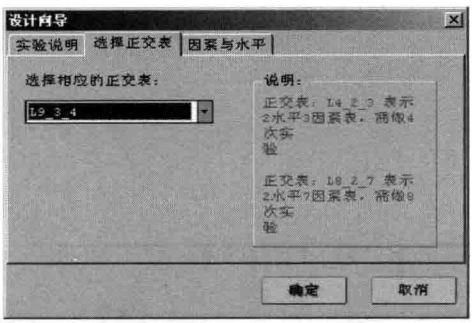


图 1-4

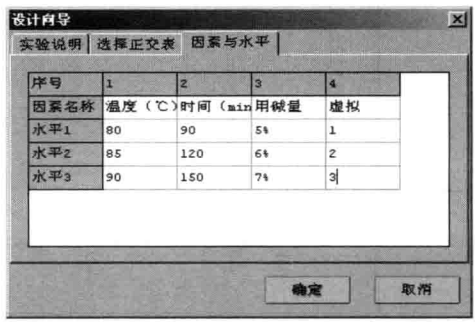


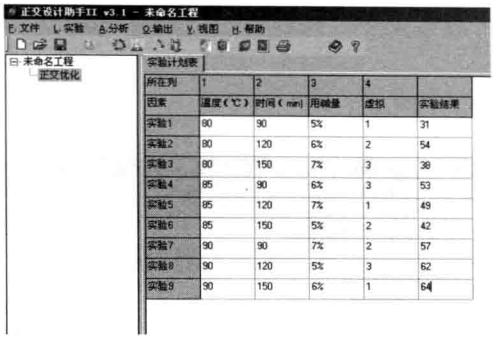
图 1-5

点击“确定”，即获得正交表，根据此正交表实施试验，将试验结果依次填入（图 1-6）。

点击“分析”→“直观分析”，得如图 1-7 所示结果。

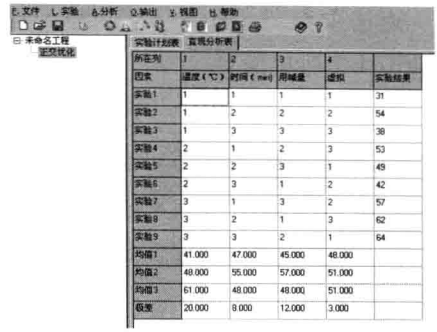
根据直观分析表，可获得各因素的重要性排序和主要因素、各因素的最佳水平，以及各因素水平的最优组合等信息。例如，由极差分析可知，虚拟因素各水平极差最小，说明试验误差在可控范围内，试验结果可靠。3 个因素的重要性依次为：温度 > 用碱量 > 时间，最优组合为：A3B2C1，即温度 90℃，时间 120min，用碱量 5%。

点击“分析”→“因素指标”，并轻轻滚动鼠标滑轮，使折线图居中，得到如图 1-8 所示结果。



所在列	1	2	3	4	
因素	温度(℃)	时间(min)	用碱量	虚拟	实验结果
实验1	80	90	5%	1	31
实验2	80	120	6%	2	54
实验3	80	150	7%	3	38
实验4	85	90	6%	3	53
实验5	85	120	7%	1	49
实验6	85	150	5%	2	42
实验7	90	90	7%	2	57
实验8	90	120	5%	3	62
实验9	90	150	6%	1	64

图 1-6



所在列	1	2	3	4	
因素	温度(℃)	时间(min)	用碱量	虚拟	实验结果
实验1	1	1	1	1	31
实验2	1	2	2	2	54
实验3	1	3	3	3	38
实验4	2	1	2	3	53
实验5	2	2	3	1	49
实验6	2	3	1	2	42
实验7	3	1	3	2	57
实验8	3	2	1	3	62
实验9	3	3	2	1	64
均值1	41.000	47.000	45.000	48.000	
均值2	48.000	55.000	57.000	51.000	
均值3	61.000	48.000	48.000	51.000	
极差	20.000	8.000	12.000	3.000	

图 1-7

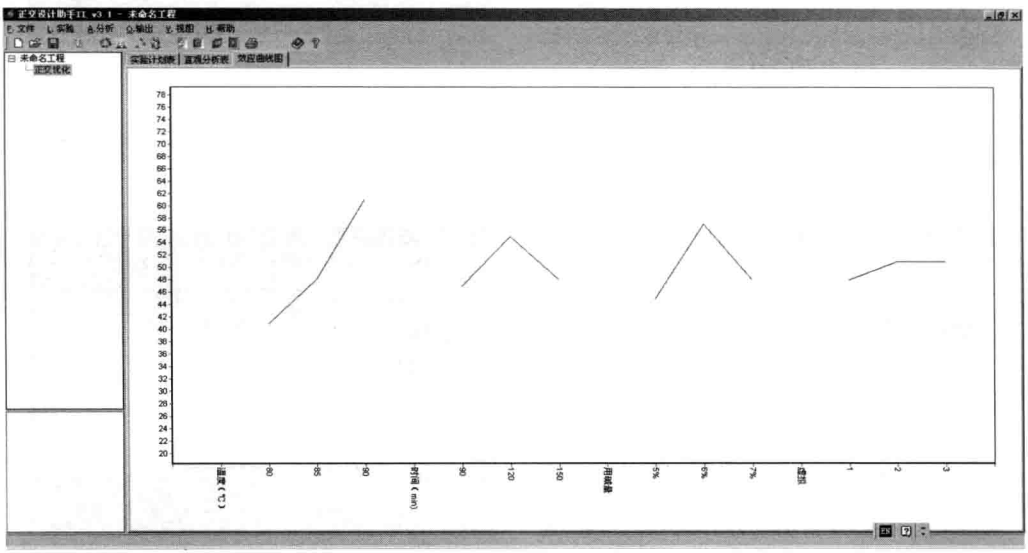


图 1-8

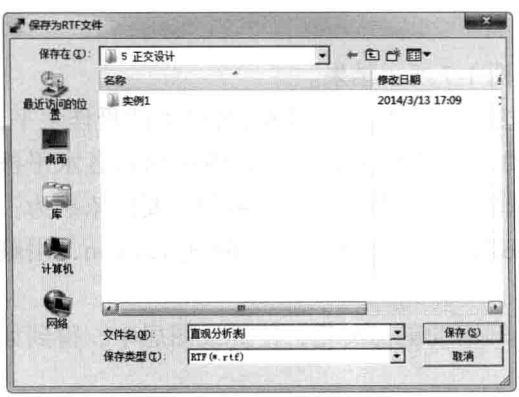


图 1-9

可以继续点击“分析”→“交互作用”，“分析”→“方差分析”，此处从略，本科阶段可不要求。

点击“直观分析表”选项卡，点击菜单“输出”→“保存 RTF”，可打开对话框，点击“保存”即可（图 1-9）。

在 Word 中，可以点击菜单“插入”→“文件”，注意在对话框中选择“所有文件类型”，可插入直观分析表。经设置调整，可获得如表 1-2 所示表格。

表 1-2　直观分析表

因素	1	2	3	4	
	温度 /℃	时间 /min	用碱量	虚拟	试验结果
实验 1	1	1	1	1	31
实验 2	1	2	2	2	54
实验 3	1	3	3	3	38
实验 4	2	1	2	3	53
实验 5	2	2	3	1	49
实验 6	2	3	1	2	42
实验 7	3	1	3	2	57
实验 8	3	2	1	3	62
实验 9	3	3	2	1	64
均值 1	41.000	47.000	45.000	48.000	
均值 2	48.000	55.000	57.000	51.000	
均值 3	61.000	48.000	48.000	51.000	
极差	20.000	8.000	12.000	3.000	

点击“效应曲线图”，点击菜单“输出”→“保存图形”，可保存效应曲线图，在 Word 中点击菜单“插入”→“图形”，即得如图 1-10 所示效应曲线图。

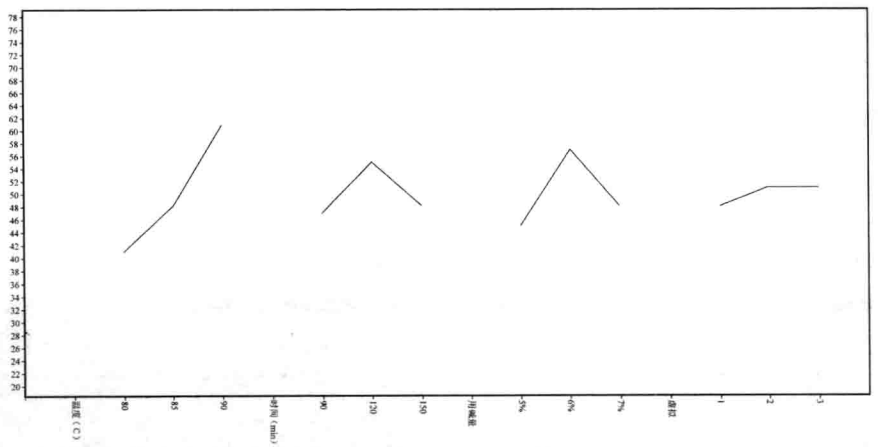


图 1-10

效应曲线图更直观，便于找出各因素的最佳水平。但为避免重复，在提交实验报告时，只用直观分析表即可。如果效应曲线图不够直观，可以利用 Origin 作图（见第四章）。

（张祥胜、胡化广）

第二章　响应面设计软件 Design Expert

Design Expert 是全球顶尖级的试验设计软件。在所有响应面设计相关的软件中，Design Expert 具有易学、操作方便、功能完整、界面亲和力强等优点。在已经发表的有关响应面 (RSM) 优化试验的论文中，Design Expert 是最广泛使用的软件。Plackett-Burman(PB) 设计、Central Composite Design(CCD)、Box-Behnken Design(BBD) 是最常用的试验设计方法。本章以 PB 设计和 BBD 为例展开论述。

1. 安装

目前网上可下载Design Expert绿色版本，也可购买注册版本。该软件界面如图2-1所示。

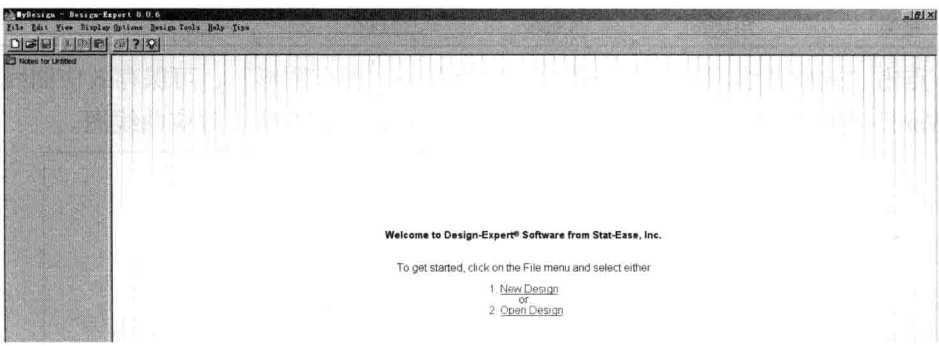


图 2-1

点击“New Design”，则打开界面（图 2-2）。

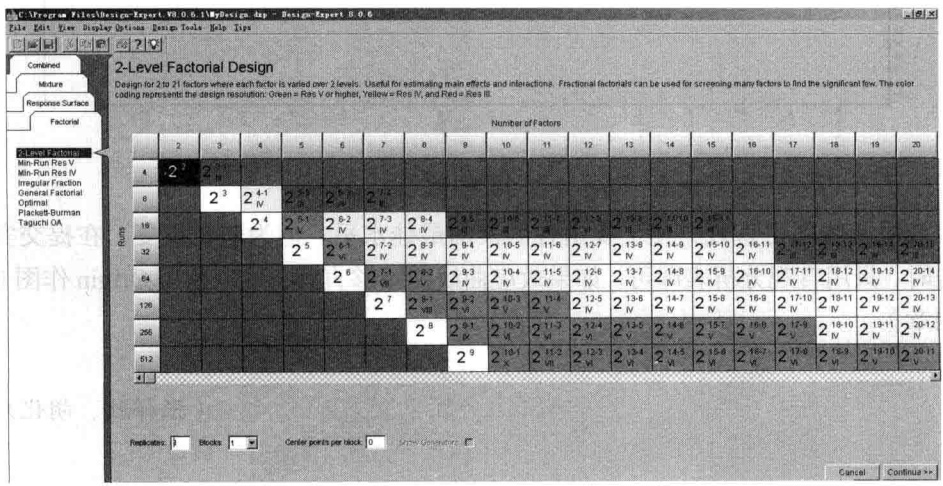


图 2-2

2. PB 设计

PB 设计方法广泛应用于参数初步优选，如微生物发酵工艺关键参数的筛选，通过对试验进行设计和试验结果的分析，筛选出对目标值影响最大的关键因素，可大大减少优化过程中的试验成本，提高试验效率。

例如，在某次发酵培养参数优化试验中，有以下因素需要优化，水平设置如表 2-1 所示（高水平为低水平的 1.25 倍，各因素单位均为 g/L，响应值为发酵产量，单位为 g/L）。

表 2-1

	因素名称	低水平（-1）	高水平（+1）
A	C 源	1	1.25
B	N 源	2	2.5
C	P 源	2	2.5
D	K 源	0.5	0.625
E	酵母提取物	0.5	0.625
F	NaCl	1	1.25
G	$MgSO_4$	0.2	0.25
H	$CaCl_2$	0.02	0.025

打开软件，点击“New Design”，选择“Factorial”→“Plackett-Burman”，然后输入因素名称和水平（图 2-3）。

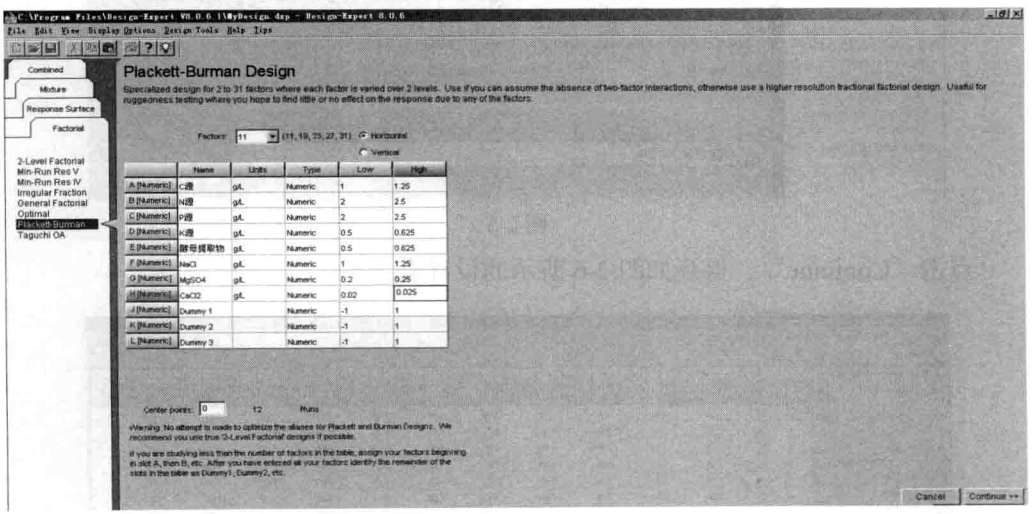


图 2-3

输入因素和水平时应注意以下几点：①操作窗口中表格选定后，可直接从 Excel 或 Word 中拷贝，提高操作效率；②请注意窗口中最后一段话，即剩余因素可设置

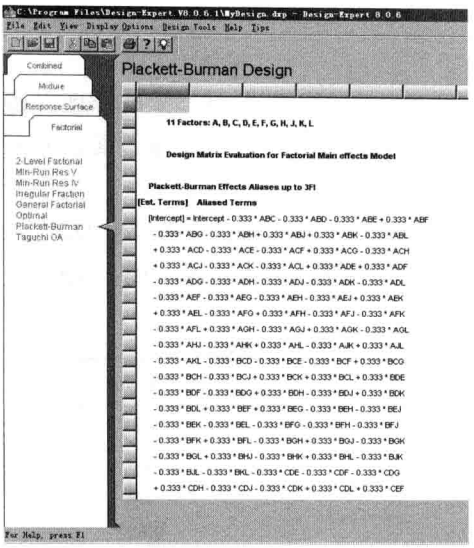


图 2-4

为虚拟因素，本例中共 3 个虚拟因素，依次命名为 Dummy 1、Dummy 2 和 Dummy 3。

点击“Continue”，则如图 2-4 所示。

点击“Continue”，并输入响应值名称和单位，则如图 2-5 所示。

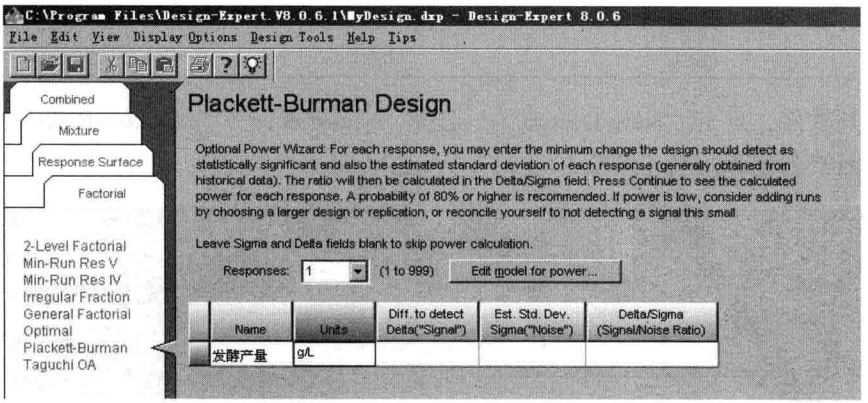


图 2-5

点击“Continue”，得到如图 2-6 所示的设计表。

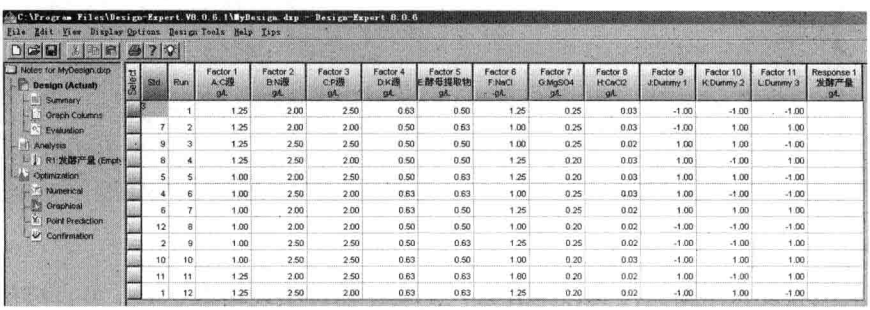


图 2-6

按此表设计试验，并将试验结果填入表中（图 2-7）。