

РЕАЛИЗАЦИЯ УГЛУБЛЕННЫХ МЕТОДОВ ДИСПЕРСИОННОГО АНАЛИЗА В ПРОГРАММНОМ КОМПЛЕКСЕ STATISTICA

Борбаць Н.М., Школина Т.В.

Брянский государственный технический университет, Брянск

Ключевые слова: контроль качества, дисперсионный анализ, частично вложенный иерархический план, модель смешанных эффектов, программный комплекс STATISTICA, редактор кода.

Аннотация. Проведение дисперсионного анализа предусмотрено во многих статистических программных комплексах. Однако в большинстве случаев подобная реализация ограничена случаем перекрёстной классификации по модели фиксированных эффектов. В тоже время пользователю иногда требуется выполнить дисперсионный анализ иерархической классификации с одним или большим числом случайных факторов. В статье описан способ проведения подобного анализа с использованием популярного программного комплекса STATISTICA компании StatSoft.

IMPLEMENTATION OF ADVANCED METHOD S ANALYSIS OF VARIANCE IN STATISTICA

Borbats N.M., Shkolina T.V.

Bryansk State Technical University, Bryansk

Keywords: quality control, analysis of variance, partially nested hierarchical plan, mixed effects model, STATISTICA, code editor.

Abstract. Conducting dispersion analysis is provided in many statistical software packages. However, in most cases, such implementation is limited to the case of cross-classification by the fixed effects model. At the same time, the user sometimes needs to perform dispersion analysis of hierarchical classification with one or more random factors. The article describes a method for conducting such an analysis using the popular STATISTICA software package from StatSoft.

Во многих практических ситуациях контроля качества приходится исследовать влияние на отклик двух и более факторов. С этой целью могут использоваться соответствующие процедуры дисперсионного анализа, при этом предполагается, как правило, что данные получены в ходе проведения экспериментов по *факторным планам*. При использовании таких планов в каждом полном опыте или реплике эксперимента исследуются все возможные комбинации уровней (обработок) факторов. При этом, если каждый уровень одного фактора комбинируется с каждым уровнем другого фактора, то такие факторы называются *пересекающимися*, а соответствующий дисперсионный анализ – *анализом перекрёстной классификации*.

Однако в некоторых многофакторных экспериментах уровни одного фактора (например, фактора *B*) подобны, но не идентичны для разных уровней другого фактора (например, *A*). При этом уровни фактора *B* оказываются *вложенными* внутри уровней фактора *A*, то есть каждый уровень фактора *B* комбинируется только с одним уровнем фактора *A*. Это означает, что если фактор *A* имеет *a* уровней, то уровни фактора *B* оказываются распределены по *a*

множествам из $b_1, b_2, \dots, b_a$ уровней соответственно, таких, что i -е множество уровней комбинируется с i -м уровнем фактора A . Подобные планы многофакторных экспериментов называются **иерархическими** или **вложенными планами**. Если внутри каждого уровня фактора A сгруппировано одинаковое число уровней фактора B и для каждой такой комбинации выполняется равное число реплик, то такой иерархический план часто называется **сбалансированным гнездовым планом**. В свою очередь, разновидность дисперсионного анализа, используемая для данных иерархических (гнездовых) планов, соответственно называется **анализом иерархической классификации**.

При наличии только двух исследуемых факторов они могут быть либо только пересекающимися, либо только вложенными. Если же исследуемых факторов больше двух, то одни из них могут быть пересекающимися, а другие – вложенными. Подобные планы эксперимента называют **частично вложенными (иерархическими)** или **вложенными факторными планами**.

Оставив за пределами статьи описание теоретических основ этих видов дисперсионного анализа [1-3], в качестве примера рассмотрим эксперимент, в котором изучается операция сборки электронных компонентов на печатных платах с целью повышения её скорости. Предложено три монтажных приспособления и два варианта планировки рабочих мест. Сборка осуществляется различными операторами, и было решено выбрать случайным образом четырех операторов для каждой комбинации приспособления и варианта планировки. Однако, поскольку участки сборки находятся в разных местах в пределах предприятия, то трудно использовать одних и тех же четырёх операторов для каждой планировки. Следовательно, четыре оператора, выбранных для варианта планировки I, отличаются от четырёх операторов, выбранных для планировки II. Таким образом, в данном эксперименте монтажные приспособления пересекаются с вариантами планировки и операторами, а операторы вложены в варианты приспособления. Кроме того, поскольку уровни одного из исследуемых факторов (операторы) были выбраны случайным образом, в то время как уровни двух других факторов (планировка и приспособления) – фиксированы, то анализ собранных данных должен проводиться по модели смешанных эффектов.

Предположим, в ходе эксперимента были получены приведённые в табл. 1 данные о продолжительности сборки в секундах (значения, помеченные символом *, представляют собой суммы наблюдений y_{ijk} по двум репликам эксперимента для каждой комбинации уровней исследуемых факторов), и необходимо провести их анализ.

Рассмотрим выполнение анализа данных в популярном статистическом программном комплексе STATISTICA [4]. Данные должны быть занесены в таблицу данных обычным образом (рис. 1, приведено только несколько первых значений). Однако в программе отсутствует специальный модуль, позволяющий выполнять анализ данных частично вложенных планов по смешанной модели, поэтому необходимо воспользоваться специальным редактором кода.

Табл. 1. Данные эксперимента

Планировка	Оператор	Приспособление						$y_{ij..}$	$y_{i...}$
		П1		П2		П3			
I	O1	22	46*	30	57*	25	46*	619	
		24		27		21			
	O2	23	47*	29	57*	24	46*		
		24		28		22			
	O3	28	57*	30	62*	27	52*		
		29		32		25			
	O4	25	48*	27	52*	26	49*		
		23		25		23			
	$y_{1.k.}$	198		228		193			
II	O1	26	54*	29	57*	27	52*	633	
		28		28		25			
	O2	27	52*	30	57*	26	50*		
		25		27		24			
	O3	28	53*	24	47*	27	51*		
		25		23		24			
	O4	24	47*	28	58*	28	55*		
		23		30		27			
	$y_{2.k.}$	206		219		208		$y_{....} = 1252$	
$y_{..k.}$	404		447		401				

	1	2	3	4					
	Планировка	Оператор	Присп.	Время					
1	I	O1	П1	22					
2	I	O1	П1	24					
3	I	O2	П1	23					
4	I	O2	П1	24					
5	I	O3	П1	28					
6	I	O3	П1	29					
7	I	O4	П1	25					
8	I	O4	П1	23					
9	I	O1	П2	30					
10	I	O1	П2	27					

Рис. 1. Фрагмент таблицы данных

Для открытия редактора кода необходимо запустить любой модуль анализа, например **Общие линейные модели** из меню **Углубленные методы**, и в открывшемся диалоговом окне в области **Задание анализа:** выбрать пункт **Редактор кода** и нажать кнопку **ОК**. На экране появится диалоговое окно редактора кода (рис. 2), в котором в области **Код анализа** необходимо ввести код выполнения для соответствующего вида анализа.

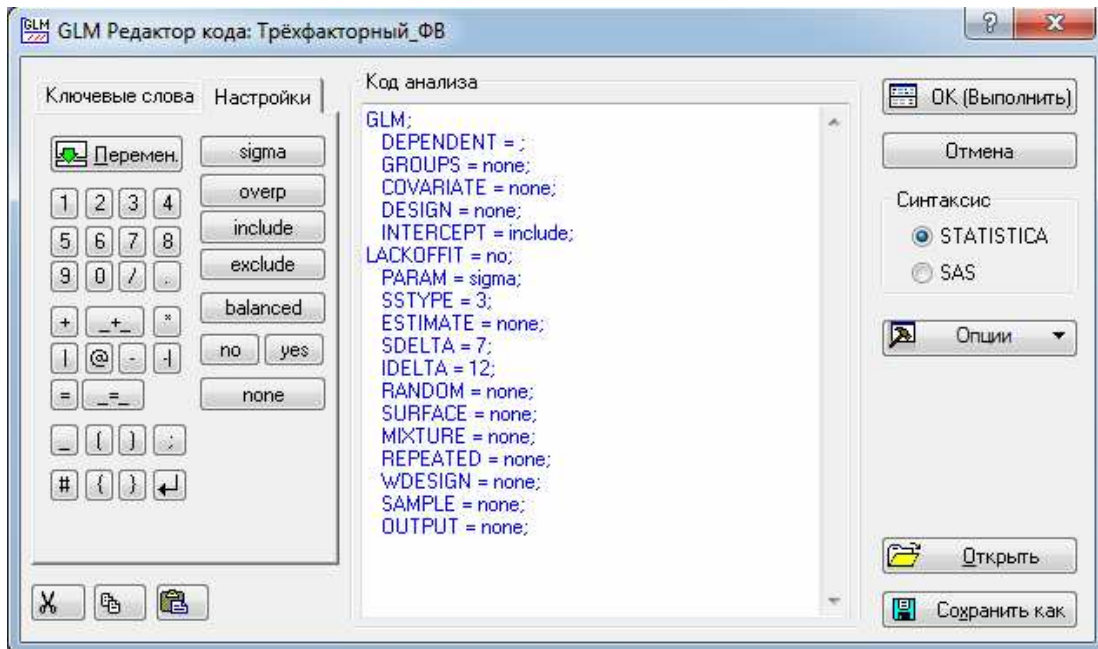


Рис. 2. Диалоговое окно редактора кода

Рассмотрение особенностей синтаксиса кода анализа выходит за рамки данной статьи, поэтому отметим лишь, что для проведения анализа по рассматриваемой модели необходимо ввести следующий код:

```
GLM;
DEPENDENT = "Время";
GROUPS = "Планировка"
        "Оператор"
        "Присп.";
COVARIATE = none;
DESIGN = "Планировка" + "Оператор"("Планировка") + "Присп." +
"Планировка"*"Присп." + "Оператор"*"Присп."("Планировка");
INTERCEPT = include;
LACKOFFIT = no;
PARAM = overp;
SSTYPE = balanced;
ESTIMATE = none;
SDELTA = 7;
IDELTA = 12;
RANDOM = "Оператор";
SURFACE = none;
MIXTURE = none;
REPEATED = none;
WDESIGN = none;
SAMPLE = none;
OUTPUT = none;
```

Для удобства ввода можно использовать кнопки, расположенные в левой части диалогового окна, в частности, указывать имена переменных можно с помощью кнопки **Перемен.**, при нажатии которой открывается диалоговое окно с перечнем имён переменных в таблице данных. После ввода кода анализа для его запуска следует нажать кнопку **ОК (Выполнить)**. На экране появится диалоговое окно результатов анализа (рис. 3), в расширенной версии которого для получения таблицы дисперсионного анализа следует щёлкнуть по кнопке **Проверить все эффекты**.

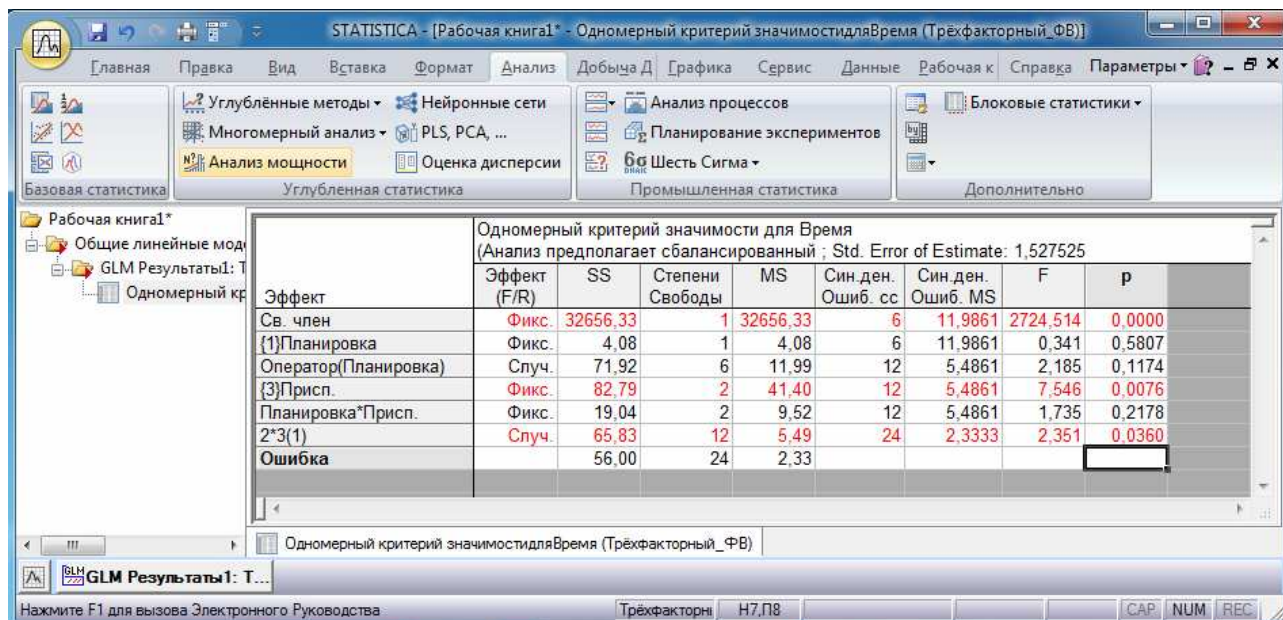


Рис. 3. Таблица дисперсионного анализа

Из результатов анализа на рисунке 3 видно, что эффект планировки рабочего места на продолжительность операции сборки является статистически незначим. В то же время эффект приспособлений является значимым. Также существует значимая взаимосвязь между приспособлениями и операторами внутри варианта планировки, что указывает на различный характер влияния приспособлений на продолжительность сборки для разных операторов.

Отметим, что в программе STATISTICA используется неограниченная модель смешанных эффектов, для которой выражения математических ожиданий средних квадратов отличаются от тех, которые получаются при использовании ограниченной модели. Как следствие, используемые F -статистики также несколько отличаются: в частности, при использовании ограниченной модели в качестве знаменателя F -отношения для проверки гипотезы $H_0: \sigma_B^2 = 0$ об отсутствии изменчивости, обусловленной операторами внутри варианта планировки, должен использоваться средний квадрат ошибки MS_E с 24 степенями свободы, а при неограниченной модели – средний квадрат взаимодействия оператора и приспособления внутри планировки $MS_{BC(A)}$ с 12 степенями свободы. Поскольку $MS_{BC(A)} > MS_E$ и у него меньше число степеней свободы, то эффект оператора внутри планировки оказывается значим только на уровне около 12 процентов. Кроме того, соответствующая оценка компонента дисперсии $\hat{\sigma}_B^2 = 1,083$ меньше, чем в случае ограниченной модели. Однако из-за по-прежнему большого эффекта приспособления и значимого взаимодействия оператора и приспособления внутри планировки эффект оператора всё же стоит учитывать, и поэтому на практические выводы из этого анализа не сильно влияет выбор ограниченной или неограниченной формы смешанной модели.

Таким образом, использование редактора кода позволяет реализовать в программном комплексе STATISTICA углубленные методы дисперсионного анализа, не предусмотренные ни одним из модулей.

Список литературы

1. Montgomery D.C. Design and analysis of experiments, 9th ed. – New Jersey: Jhon Wiley, 2017.
2. Kutner M.H., Nachtsheim C.J., Neter J., Li W. Applied linear statistical models, 5th ed. – New York: McGraw-Hill, 2005
3. Sahai H., Ageel M.I. The analysis of variance. Fixed, random and mixed models. – New York: Springer Science, 2000.
4. Боровков В.П. Популярное введение в современный анализ данных и машинное обучение на STATISTICA. – М.: Горячая линия – Телеком, 2021. – 354 с.

Сведения об авторах:

Борбаць Николай Михайлович – к.т.н., доцент, доцент кафедры «Управление качеством, стандартизация и метрология»;

Школина Татьяна Викторовна – к.т.н., доцент, доцент кафедры «Информатика и программное обеспечение».