УДК 519.233.4

ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ С ИСПОЛЬЗОВАНИЕМ ПАКЕТА СТАТИСТИЧЕСКИХ ПРОГРАММ STATA

© 2014 г. [1]Т. Н. Унгуряну, [1,2]А. М. Гржибовский

[1]Северный государственный медицинский университет, г. Архангельск
[1,2]Норвежский институт общественного здравоохранения, г. Осло, Норвегия

В настоящей статье рассмотрены теоретические принципы применения дисперсионного анализа для сравнения трех и более независимых групп. Представлены примеры расчетов «вручную» и с помощью пакета прикладных статистических программ STATA. Особое внимание уделяется условиям, которые должны соблюдаться для применения данного метода анализа. Даются рекомендации о том, как следует представлять результаты дисперсионного анализа в научных публикациях.

Ключевые слова: дисперсионный анализ, независимые группы, среднее арифметическое, дисперсия

Дисперсионный анализ используется для сравнения средних значений количественного признака при наличии в исследовании трех и более групп.

Дисперсионный анализ предпочтителен по сравнению с использованием множественных сравнений с помощью t-критериев, поскольку риск ошибки первого рода для многократного применения t-критериев больше, чем указанный уровень значимости (т. е. вероятность ошибки первого рода) для каждого t-критерия по отдельности. Такая ситуация называется инфляцией ошибки первого рода и может приводить к получению ложнодостоверных результатов, то есть обнаружению различий там, где их на самом деле нет. Об этом подробно рассказывалось в одном из предыдущих выпусков практикума [2].

Слово «дисперсионный» в названии указывает на то, что в процессе анализа сопоставляются дисперсии изучаемого признака. Общая изменчивость переменной раскладывается на две составляющие – межгрупповую (факторную), обусловленную различием групп (средних значений), и внутригрупповую, обусловленную случайными (неучтенными) причинами. Чем больше частное, полученное в результате деления межгрупповой дисперсии на внутригрупповую дисперсию (F-отношение), тем больше различаются средние значения сравниваемых выборок и тем выше статистическая значимость этого различия [1]. В данной статье рассматривается только однофакторный дисперсионный анализ для независимых групп, который в зарубежной литературе называется One-way analysis of variances или One-way ANOVA. В ходе анализа проверяется нулевая гипотеза (H_0) о равенстве средних значений для трех и более независимых групп.

Перед выполнением дисперсионного анализа необходимо проверить следующие условия его применения [2]:

1. Количественный тип данных.
2. Независимые выборки.
3. Нормальное распределение признака в популяциях, из которых отобраны выборки.
4. Равенство дисперсий изучаемого признака в популяциях, из которых отобраны выборки. Для проверки существенности различий дисперсий используют критерий Levene.
5. Независимые наблюдения в каждой из выборок.

Расчет. Для вычисления статистики критерия для ANOVA – отношения F – необходимо рассчитать средний квадрат отклонений между группами (межгрупповую дисперсию) и средний квадрат отклонений внутри групп (внутригрупповую дисперсию) [1].

Средний квадрат отклонений между группами (межгрупповая дисперсия):

1. Рассчитать разность между средним каждой группы и общим средним по всем группам вместе. Общее среднее нельзя вычислять как среднее групповых средних, так как в группах может быть разное количество наблюдений. Для этого необходимо сложить все значения количественной переменной по всем группам вместе и далее полученную сумму разделить на сумму наблюдений по всем группам.

2. Полученные разности возвести в квадрат.

3. Полученные значения для каждой группы умножить на число наблюдений в данной группе.

4. Выполнив описанные выше процедуры для всех групп, сложить полученные величины по всем группам вместе.

5. Полученную сумму разделить на число степеней свободы m (число групп − 1).

Средний квадрат отклонений внутри групп (внутригрупповая дисперсия):

1. Рассчитать разность каждого отдельного значения от среднего значения в своей группе.

2. Полученные разности возвести в квадрат.

3. Полученные квадраты сложить.

4. Полученную сумму разделить на число степеней свободы n (общее число наблюдений по всем группам вместе − число групп, n − m).

Далее для вычисления F-критерия находят отношение межгрупповой дисперсии к внутригрупповой дисперсии. Эта F-статистика подчиняется F-распределению Фишера − Снедекора с (m − 1, n − m) степенями свободы соответственно в числителе и знаменателе [4]. После расчета F-критерия необходимо сравнить его значение с критическим значением, взятым из таблицы. В случае если рассчитанное значение F равно или превышает критическое значение F для заранее определенного уровня значимости (обычно 0,05), H_0 отвергается и делается вывод о том, что существуют статистически значимые различия между средними значениями в популяциях, из которых извлечены выборки ($p < 0,05$), однако какие из групп различаются между собой − неизвестно.

Поэтому если при выполнении ANOVA получен статистически значимый результат, то далее следует провести попарные апостериорные (post-hoc) сравнения. Апостериорные сравнения представляют собой попарные сравнения изучаемых групп для обнаружения различий между ними. Попарные сравнения выполняются с помощью специальных статистических критериев (Бонферрони, Шеффе, Тьюки и др.), обзор которых подробно представлен в практикуме «Анализ трех и более независимых групп количественных данных» [2]. Если при выполнении ANOVA получен статистически незначимый результат, то продолжать дальнейший анализ с помощью апостериорных сравнений не имеет смысла, так как различия между группами будут отсутствовать.

Пример. В исследовании приняли участие 30 человек, которые случайным образом были разделены на три группы. Каждая группа в течение недели была на разной диете. По окончании исследования у участников измерялся уровень сахара в крови. Результаты представлены в табл. 1. Необходимо определить, есть ли различия между уровнями сахара крови в изучаемых группах.

Таблица 1

Уровень сахара, ммоль/л

Группа 1 (диета 1)	Группа 2 (диета 2)	Группа 3 (диета 3)
4,6	4,3	4,4
5,0	4,4	4,5
5,2	4,9	4,9
5,5	5,1	5,0
4,8	4,5	4,6
5,1	4,5	4,6
5,7	5,4	5,4
5,4	5,1	5,2
5,7	5,1	5,2
5,6	5,3	5,4

Нулевой гипотезой служит гипотеза об отсутствии различий между средними значениями сахара в крови в изучаемых группах. Рассмотрим расчет F-критерия, допуская выполнение всех условий применения однофакторного дисперсионного анализа в данном примере.

Расчет межгрупповой дисперсии. Общая средняя по всем группам вместе = 5,18.

Средние арифметические значения по группам: в 1-й группе $\overline{X}_1 = 5,25$; во 2-й группе $\overline{X}_2 = 5,26$; в 3-й группе $\overline{X}_3 = 4,92$. Сумма квадратов отклонений между группами (межгрупповая вариабельность) = $(5,18 - 5,25)^2 \times 10 + (5,18 - 5,26)^2 \times 10 + (5,18 - 4,92)^2 \times 10 = 1,064$. Межгрупповая дисперсия = $1,064/(3 - 1) = 0,532$.

Расчет внутригрупповой дисперсии представлен в табл. 2.

Таблица 2

Расчет сумм квадратов отклонений значений в группах от групповых средних

Группа 1 (диета 1)		Группа 2 (диета 2)		Группа 3 (диета 3)	
$x_i - \overline{X}_1$	$(x_i - \overline{X}_1)^2$	$x_i - \overline{X}_2$	$(x_i - \overline{X}_2)^2$	$x_i - \overline{X}_3$	$(x_i - \overline{X}_3)^2$
−0,66	0,436	−0,06	0,004	−0,52	0,270
−0,26	0,068	0,04	0,002	−0,42	0,176
−0,06	0,004	0,54	0,292	−0,02	0,000
0,24	0,058	−0,26	0,068	0,08	0,006
−0,46	0,212	0,14	0,020	−0,32	0,102
−0,16	0,026	0,14	0,020	−0,32	0,102
0,44	0,194	0,04	0,002	0,48	0,230
0,14	0,020	−0,26	0,068	0,28	0,078
0,44	0,194	−0,26	0,068	0,28	0,078
0,34	0,116	−0,06	0,004	0,48	0,230
Сумма	1,324	−	0,544	−	1,276

Сумма квадратов отклонений по всем группам (внутригрупповая вариабельность) = 1,324 + 0,544 + 1,276 = 3,144. Внутригрупповая дисперсия = 3,144 / (30 − 3) = 0,116.

Для расчета F-критерия нужно межгрупповую дисперсию разделить на внутригрупповую дисперсию: 0,532 / 0,116 = 4,58. В табл. 2 критическое значение F для степеней свободы 2 (по горизонтали) и 27 (по вертикали) для $\alpha = 0,05$ составляет 3,35. Так как расчетное значение F больше критического значения, то нулевая гипотеза отклоняется, следовательно, в рассматриваемом примере средние значения содержания сахара в крови статистически значимо различаются в трех группах с разными диетами.

ANOVA в STATA. Для выполнения однофакторного дисперсионного анализа в STATA необходимо, чтобы в базе данных значения исследуемой переменной всех групп были размещены в одном столбце, а в соседнем столбце представлены номера групп, которым принадлежат данные. Перед тем как начать проверку с помощью однофакторного дисперсионного анализа, следует проверить, можно ли применять этот критерий в данной ситуации. Уровень сахара в крови является непрерывной величиной, все три группы являются независимыми. Для проверки условия нормальности распределения в каждой из групп в меню Statistics следует выбрать Summaries, tables, and tests → Distributional plots and tests → Shapiro-Wilk normality test. В поле Variables необходимо перенести переменную Glucose, а на вкладке by/if/in отметить галочкой Repeat commands by groups и указать переменную, которая определяет разделение значений на группы Diets (рис. 1). Результаты теста Shapiro-Wilk показали, что значения во всех трех группах имеют нормальное распределение (рис. 2).

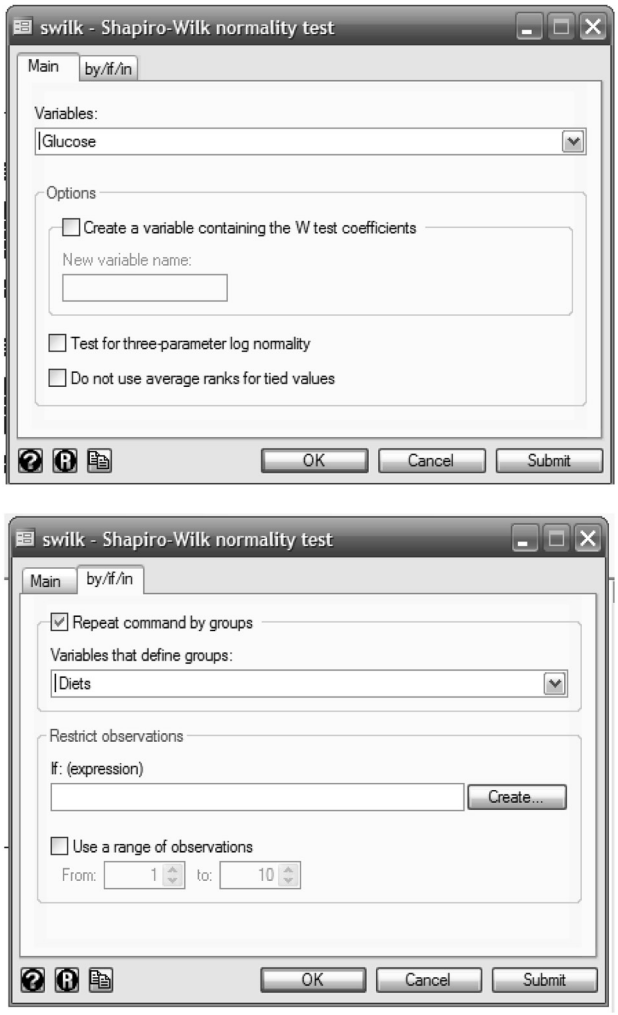

Рис. 1. Диалоговое окно для расчета теста Shapiro-Wilk

```
->  Diets = 1
                Shapiro-Wilk W test for normal data
    Variable      Obs        W           V          z       Prob>z
    Glucose        10    0.96490       0.541     -0.994     0.83993

->  Diets = 2
                Shapiro-Wilk W test for normal data
    Variable      Obs        W           V          z       Prob>z
    Glucose        10    0.90766       1.423      0.627     0.26531

->  Diets = 3
                Shapiro-Wilk W test for normal data
    Variable      Obs        W           V          z       Prob>z
    Glucose        10    0.96653       0.516     -1.067     0.85697
```

Рис. 2. Результаты теста Shapiro-Wilk

Для проверки условия о равенстве дисперсий изучаемого признака в популяциях, из которых отобраны выборки, следует воспользоваться тестом Levene. Для этого в меню Statistics нужно выбрать Summaries, tables, and tests → Classical tests of hypothesis → Robust equal variance test. В поле Variable переносится переменная Glucose, а в поле Variable defining comparison groups указывается группировочная переменная Diet.

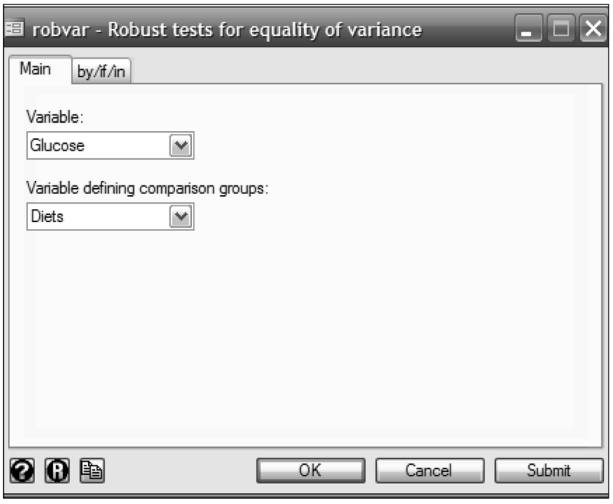

Рис. 3. Диалоговое окно для расчета теста Levene

Как видно из рис. 4, достигнутый уровень значимости (Pr > F) для критерия Levene составил 0,121, что не позволяет отвергнуть нулевую гипотезу о равенстве дисперсий в изучаемых группах. Таким образом, все необходимые условия для применения однофакторного дисперсионного анализа выполняются.

Summary of Glucose			
Diets	Mean	Std. Dev.	Freq.
1	5.25	0.38355061	10
2	5.36	0.24585457	10
3	4.92	0.3765339	10
Total	5.18	0.38092446	30
W0 = 2.2849728	df (2, 27)	Pr > F = 0.1211195	
W50 = 2.2105250	df (2, 27)	Pr > F = 0.12910155	
W10 = 1.9705940	df (2, 27)	Pr > F = 0.15891	

Рис. 4. Результаты теста Levene

Для выполнения однофакторного дисперсионного анализа в STATA [5–8] следует открыть в меню Statistics → Linear models and related → ANOVA / MANOVA → One-way ANOVA (рис. 5). В поле Response variable переносится зависимая переменная, средние значения которой планируется сравнить. В данном примере это переменная Glucose. В поле Factor variable помещается группировочная переменная, то есть переменная, которая используется для разделения всей выборки на группы. В данном примере это переменная Diet. Ниже в Output можно отметить галочкой Produce summary table для получения данных описательной статистики. Запуск анализа осуществляется нажатием на кнопку OK внизу диалогового окна One-way ANOVA.

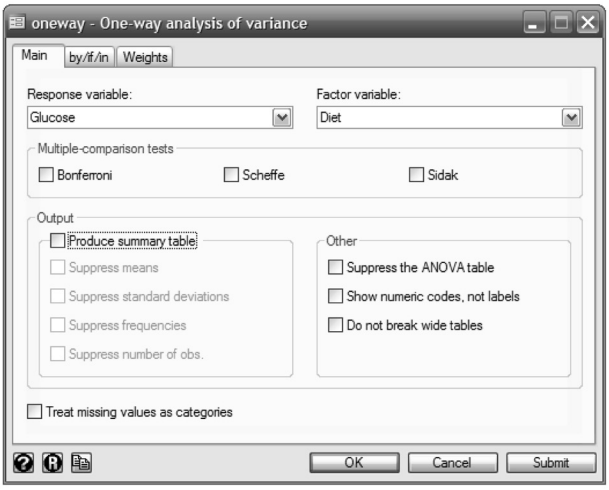

Рис. 5. Диалоговое окно One-way ANOVA

Результаты применения дисперсионного анализа представлены на рис. 6. Во втором столбце представлена общая вариабельность признака (Total Sum of Squares), а также ее составляющие – межгрупповая (Between groups Sum of Squares) и внутригрупповая (Within groups Sum of Squares) вариабельность. В третьем столбце представлено количество степеней свободы (df, degrees of freedom), которое используется для расчета межгрупповой и внутригрупповой дисперсий. В четвертом столбце приведена межгрупповая, внутригрупповая и общая дисперсии. Критерий F получен путем деления 0,5320 / 0,1164 = 4,57. Достигнутый уровень статистической значимости составил 0,0195, что свидетельствует о существовании статистически значимых различий между средними значениями в трех сравниваемых группах.

При описании полученных результатов с применением ANOVA следует указать статистику критерия F, число степеней свободы (ст. св. или df) для межгрупповой и внутригрупповой дисперсий, достигнутую величину статистической значимости – p-значение, округленное до трех знаков после запятой. Необходимо отметить, что наименьшее p-значение, которое следует отразить в отчете, удовлетворяет условию $p < 0{,}001$ [3]. В примере F-критерий может быть описан как $F = 4{,}57$; 2 и 27 ст. св.; $p = 0{,}019$, или $F_{2/27df} = 4{,}57$; $p = 0{,}019$.

Analysis of Variance					
Source	SS	df	MS	F	Prob > F
Between groups	1.06400014	2	0.532000072	4.57	0.0195
Within groups	3.14399984	27	0.116444439		
Total	4.20799998	29	0.145103448		

Рис. 6. Результаты дисперсионного анализа

Так как результаты дисперсионного анализа показали наличие статистически значимых различий между сравниваемыми группами, следующим шагом необходимо выполнить апостериорные сравнения для обнаружения, между какими группами имеются различия. Для апостериорных сравнений STATA в диалоговом окне One-way analysis of variance предлагает три критерия Bonferroni, Scheffe, Sidak (см. рис. 5). Апостериорные сравнения представляют собой попарные сравнения изучаемых групп для обнаружения различий между ними. Наиболее популярным критерием для выполнения попарных сравнений является поправка Bonferroni. Подробные рекомендации по выбору критерия для апостериорных сравнений представлены в практикуме «Анализ трех и более независимых групп количественных данных» [2]. Результаты апостериорных сравнений в STATA выглядят, как на рис. 7. В рассматриваемом примере с помощью поправки Bonferroni установлены статистически значимые различия только между 2-й и 3-й группами ($p = 0{,}023$). Если при выполнении дисперсионного анализа статистически значимые различия между группами не выявлены, то анализ завершается и попарные сравнения не проводятся.

Comparison of Glucose by Diet (Bonferroni)			
Row	Mean-		
Col	Mean	1	2
	2	0.1	
		1.000	
	3	-0.34	-0.44
		0.103	0.023

Рис. 7. Результаты применения критерия Bonferroni для апостериорных сравнений

В настоящей статье мы разобрали основные принципы применения однофакторного дисперсионного анализа для сравнения средних арифметических для трех и более независимых групп. Напомним, что данный метод является параметрическим, а потому может

применяться только при соблюдении ряда условий, рассмотренных выше. Если условия нормальности распределения не соблюдаются, то следует применять непараметрические критерии, например критерий Краскела – Уоллиса, который будет рассмотрен в следующем выпуске практикума.

Список литературы

1. *Банержи А.* Медицинская статистика понятным языком: вводный курс. М. : Практическая медицина, 2007. 287 с.

2. *Гржибовский А. М.* Анализ трех и более независимых групп количественных данных // Экология человека. 2008. № 3. С. 50–58.

3. *Ланг Т. А., Сесик М.* Как описывать статистику в медицине. Аннотированное руководство для авторов, редакторов, рецензентов / пер. с англ. под ред. В. П. Леонова. М. : Практическая медицина, 2011. 480 с.

4. *Петри А., Сэбин К.* Наглядная медицинская статистика. М. : ГЭОТАР-Медиа, 2009. 168 с.

5. *Acock A. C.* Gentle Introduction to Stata. USA, Texas : Stata Press, 2006. 289 p.

6. *Hamilton C.* Statistics with Stata. USA, Belmont, CA : Brooks/Cole, 2006. 409 p.

7. *Kohler U., Kreute F.* Data Analysis Using Stata. USA, Texas : Stata Press, 2005. 378 p.

8. *Rabe-Hesketh S., Everit, Brian.* A Handbook of Statistical Analyses Using Stata. New York : Chapman & Hall, 2007. 352 p.

References

1. Banerjee A. *Meditsinskaya statistika ponyatnym yazykom: vvodnyi kurs* [Medical Statistics Made Clear: Introduction]. Moscow, 2007, 287 p.

2. Grjibovski A. M. Analysis of three and more independent groups of quantitative data. *Ekologiya cheloveka* [Human Ecology]. 2008, 3, pp.50-58. [in Russian]

3. Lang T. A. *Kak opisyvat' statistiku v meditsine* [How to present statistics in medicine]. Moscow, 2011, 480 p.

4. Petrie A., Sabin K. *Naglyadnaya statistika v meditsine* [Medical Statistics at Glance]. Moscow, 2003, 144 p.

5. Acock A. C. *Gentle Introduction to Stata*. USA, Texas, Stata Press, 2006. 289 p.

6. Hamilton C. *Statistics with Stata*. USA, Belmont, CA, Brooks/Cole, 2006. 409 p.

7. Kohler U., Kreute F. *Data Analysis Using Stata*. USA, Texas, Stata Press, 2005. 378 p.

8. Rabe-Hesketh S., Everit, Brian. *A Handbook of Statistical Analyses Using Stata*. New York, Chapman & Hall, 2007. 352 p.

ONE-WAY ANALYSIS OF VARIANCE (ANOVA) IN STATA SOFTWARE

[1]T. N. Unguryanu, [1,2]A. M. Grjibovski

[1]International School of Public Health, Northern State Medical University, Arkhangelsk, Russia
[1,2]Department of International Public Health, Norwegian Institute of Public Health, Oslo, Norway

In the article, we have presented theoretical principles of one-way analysis of variance (ANOVA) for comparisons of three or more independent groups. Examples of the use of ANOVA with manual calculations using formulas have been given as well as algorithms of the use of ANOVA in STATA software. Special consideration has been given to the assumptions which have to be tested as well as the ways to present the results in research papers.

Keywords: one-way ANOVA, independent groups, variance, means

Контактная информация:

Гржибовский Андрей Мечиславович – профессор, доктор медицины, старший советник Норвежского института общественного здоровья, г. Осло, Норвегия; директор Архангельской международной школы общественного здоровья ГБОУ ВПО «Северный государственный медицинский университет» Министерства здравоохранения Российской Федерации, г. Архангельск

Адрес: Nasjonalt folkehelseinstitutt, Pb 4404 Nydalen, 0403 Oslo, Norway

Тел.: +47 22048319, e-mail: angr@fhi.no