УДК 314.144(481)

АНАЛИЗ ТРЕХ И БОЛЕЕ НЕЗАВИСИМЫХ ГРУПП КОЛИЧЕСТВЕННЫХ ДАННЫХ

© 2008 г. **А. М. Гржибовский**

Национальный институт общественного здоровья, г. Осло, Норвегия

В статье рассматривается сравнение трех и более независимых групп количественных данных с помощью одномерного (однофакторного) дисперсионного анализа (One-Way ANOVA) и критерия Краскела-Уоллиса (Kruskal-Wallis test) с использованием пакета прикладных статистических программ SPSS. Особое внимание уделяется рассмотрению проблемы множественных сравнений и проверке необходимых условий для применения дисперсионного анализа. Кроме того, рассматриваются критерии для апостериорных сравнений при обнаружении статистически значимых различий в ходе дисперсионного анализа. Изложенный материал дает общие сведения о статистических критериях, применяемых для проверки гипотез о равенстве средних трех и более независимых групп, и призван вызвать интерес читателей журнала к прочтению специализированной литературы перед началом работы над будущими публикациями.

Ключевые слова: распределение, дисперсия, однофакторный дисперсионный анализ, критерий Краскела-Уоллиса, апостериорные сравнения.

В предыдущих выпусках журнала отмечалось, что выбор статистического критерия для проверки гипотез определяется типом и распределением данных [3, 4]. Особо подчеркивалось, что непарный критерий Стьюдента предназначен только для сравнения двух независимых групп при выполнении необходимых условий для применения параметрических критериев. В настоящей статье будет рассмотрен дисперсионный анализ, предназначенный для проверки статистических гипотез о равенстве средних для трех и более независимых групп количественных данных. Как и критерий Стьюдента, дисперсионный анализ — параметрический метод, поэтому будут рассмотрены необходимые условия для его применения. Также будет рассмотрен критерий Краскела-Уоллиса, который применяется в тех случаях, когда использовать дисперсионный анализ невозможно из-за несоблюдения условий для его применения.

К сожалению, в отечественной литературе до сих пор часто встречаются примеры применения непарного критерия Стьюдента для сравнения трех, четырех и даже пяти и более независимых групп. Причем сравниваются либо группы попарно, либо все группы с контрольной, в результате чего публикации пестрят большим количеством выражений типа «$p < 0,05_{1\text{-}2}$, $p < 0,05_{2\text{-}3}$, $p < 0,05_{1\text{-}3}$» или звездочками, обозначающими наличие статистически значимых различий между сравниваемыми группами. Такое представление данных не приветствуется по причине малой информативности выражений типа «$p < 0,05$» (вместо которых всегда нужно указывать абсолютные значения достигнутых уровней значимости (p), а еще лучше — доверительные интервалы для выявленных различий). Больше того, оно указывает на использование ошибочно завышенного критического уровня значимости при проведении проверки нескольких статистических гипотез и тем самым увеличивает вероятность случайного обнаружения статистически значимых различий там, где их в действительности нет. Данная проблема называется проблемой множественных сравнений, причем встречается она не только в работах начинающих исследователей, но и в статьях известных ученых. Опасность этой проблемы заключается в вероятности обнаружения ложнодостоверных различий, что приводит к декларированию наличия эффекта от нового лечения в случае, когда его нет, или в случае обнаружения вредоносного действия изучаемого фактора даже в ситуации, когда фактор в действительности никакого влияния на изучаемый исход не оказывает. Представление данных в виде «$p < 0,05$» лишь усугубляет ситуацию, не позволяя грамотным читателям самим принимать решение о принятии или отвержении нулевой гипотезы на основании достигнутых уровней значимости.

В чем же суть проблемы множественных сравнений? В биомедицинской литературе принято считать, что нулевая гипотеза об отсутствии различий между сравниваемыми группами может быть отвергнута, если достигнутый уровень значимости (p) < 0,05. Это означает, что мы в 5 % случаев готовы отвергнуть верную нулевую гипотезу, то есть принять решение о наличии различий там, где их на самом деле нет, что еще называется ошибкой 1 типа. Если изначально допустить, что истинных различий между сравниваемыми группами нет, то величина p покажет, с какой вероятностью мы можем обнаружить выявленные или еще более существенные различия в исследованиях с аналогичными объемами выборки.

Если мы принимаем традиционные 0,05 за критический уровень значимости, то вероятность ошибки 1 типа составляет 5 %, значит, вероятность отсутствия этой ошибки составит 0,95, или 95 %. Если мы проводим три сравнения (сравниваем попарно три группы, проверяем три статистические гипотезы), то вероятность отсутствия ошибки 1 типа в любом из сравнений составит $0{,}95^n$, то есть $0{,}95^3 = 0{,}857$, или 85,7 %, а значит, вероятность сделать хотя бы одну ошибку 1 типа будет равна $1 - 0{,}95^n = 1 - 0{,}857 = 0{,}142$, или 14,2 % вместо декларируемых 5 %. В такой ситуации необходимо использовать меньший критический уровень значимости, который рассчитывается по формуле: $p^* = 1 - 0{,}95^{1/n}$, где n — количество производимых сравнений. Для данного примера $p^* = 1 - 0{,}95^{1/3} = 0{,}0170$, то есть различия между группами можно считать статистически значимыми, только если $p < 0{,}0170$. Из этого следует, что в публикациях, где встречается «$p < 0{,}05_{1\text{-}2}$, $p > 0{,}05_{2\text{-}3}$, $p > 0{,}05_{1\text{-}3}$», совершенно невозможно сделать вывод о статистической значимости различий между группами 1 и 2, а потому результаты должны интерпретироваться читателем минимум как сомнительные.

Для ситуации с тремя сравниваемыми группами количество возможных попарных сравнений равно количеству изучаемых групп (таблица). Если групп больше, то количество возможных попарных сравнений можно рассчитать по формуле: $n = 0{,}5N(N - 1)$, где N — количество изучаемых групп. Например, если имеется 12 групп (при попарных сравнениях среднемесячных значений тех или иных показателей), то максимальное количество возможных сравнений составит $n = 0{,}5 \cdot 12 \cdot (12 - 1) = 66$. Если оставить критический уровень значимости без изменений (0,05), то верятность случайного обнаружения статистически значимых различий составит $1 - 0{,}95^{66} = 0{,}966$, или 96,6 %. Критический уровень значимости для данного примера при проведении всех 66 сравнений должен быть установлен на уровне $1 - 0{,}95^{1/66} = 0{,}00078$, то есть статистически значимыми могут считаться только те различия, для которых $p < 0{,}00078$.

Читатель может встретиться с проблемой множественных сравнений в следующих случаях:

1. Наличие нескольких сравниваемых групп (например, сравнение средних уровней артериального давления у врачей, учителей и чиновников).

2. Проверка нескольких независимых статистических гипотез на основании данных одной выборки (например, изучение взаимосвязи между употреблением витаминов А, В, С и Е и раком молочной железы).

3. Анализ подгрупп (например, сравнение результатов двух видов лечения для группы испытуемых с последующим анализом в подгруппах пациентов с разными степенями тяжести заболевания).

Количество возможных сравнений, вероятность ошибки 1 типа и уровни значимости для наиболее часто встречающегося в литературе количества сравниваемых групп

	Количество сравниваемых групп			
	2	3	4	5
Количество попарных сравнений	1	3	6	10
Вероятность случайного выявления статистически значимых различий (ошибка 1 типа) для множественных попарных сравнений, %	5	14	26	40
Критический уровень значимости	0,0500	0,0170	0,0085	0,0051
Количество сравнений с контрольной группой	1	2	3	4
Вероятность случайного выявления статистически значимых различий (ошибка 1 типа) для множественных сравнений с контрольной группой, %	5	10	14	19
Критический уровень значимости	0,0500	0,0253	0,0170	0,0127

Во всех приведенных примерах исследователи должны принимать во внимание проблему множественных сравнений и рассчитывать новые критические уровни значимости. Для большей убедительности в необходимости изменения критического уровня значимости можно привести пример, опубликованный для демонстрации важности проблемы в США еще в 1980 году [7]. Исследователи провели симуляцию изучения эффективности двух различных методов лечения ишемической болезни сердца. Они случайным образом разбили всех пациентов на две равные группы, но несмотря на то, что все пациенты получали одно и то же лечение, данные были обработаны так, как будто бы одной группе назначалось лечение А, а другой — лечение Б. При сравнении эффективности «двух видов лечения» различий обнаружено не было, что неудивительно, так как все пациенты получали одно и то же лечение. Затем исследователи разбили каждую из групп пациентов еще на 6 по количеству пораженных коронарных артерий (1, 2 или 3 сосуда) и сократительной способности миокарда левого желудочка (выше или ниже определенного критического уровня). Анализ выявил, что результаты лечения не различались в пяти подгруппах, а в подгруппе пациентов с наиболее тяжелой формой заболевания лечение А было более эффективно ($p = 0{,}025$). Если бы исследование было настоящим, то исследователи могли бы предположить, что лечение А эффективнее лечения Б для наиболее тяжелых случаев заболевания, и сделать соответствующие практические выводы. Но в действительности-то обе группы получали одно и то же лечение! Разбивка групп на лечение А и Б было искусственным и использовалось только для проведении статистического анализа. Пример наглядно демонстрирует, что при делении выборки на подгруппы и проведении множественных сравнений мы значительно увеличиваем вероятность ошибки 1 типа, то есть обнаружения различий там, где их на самом деле нет. При проведении 6 сравнений (см. таблицу) вероятность ошибки 1 типа возрастает до 26 %! Если изменить критический уровень значимости до 0,0085, как это следует делать при проверке 6 гипотез, то различия между лечением А и Б, полученные в подгруппе пациентов с наиболее тяжелой формой заболевания ($p = 0{,}025$) не будут статистически значимыми, что позволит сделать вывод об отсутствии различий между двумя видами лечения, что будет верно, так как в действительности все пациенты получали одно и то же лечение.

Если читатель знает о количестве проведенных сравнений в исследовании, то сделать заключение о правомочности выводов авторов не составит труда. Однако часто встречаются ситуации, когда исследователи проверяют огромное количество гипотез или сравнивают «все со всем, авось что найдется», а к публикации представляют только те результаты, для которых были получены статистически значимые различия. Порочность такой практики опять же следует из вышеприведенных формул: если провести 100 сравнений, то вероятность получить статистически значимые различия ($p < 0{,}05$) хотя бы в одном из них в результате чистой случайности составляет 99,4 %. Если представить результаты только одного сравнения, то читателю остается только поверить автору о наличии статистически значимых различий. Если бы исследователь сообщил, что всего сравнений было 100, то читатель мог бы самостоятельно рассчитать необходимый для такой ситуации критический уровень значимости (0,0005). Чтобы избежать подобных ситуаций, исследователи должны придерживаться простых правил:

1. Планировать детальный анализ исследования до начала сбора данных.
2. Представлять план анализа в письменном виде руководителю проекта.
3. При проведении анализа данных строго следовать плану исследования.
4. Докладывать результаты проверки всех статистических гипотез, а не только тех, где нулевая гипотеза была отвергнута.

Эти правила давно стали рутинной практикой на Западе. Например, протоколы всех проводимых рандомизированных контролируемых испытаний (РКИ) подлежат обязательной регистрации, а результаты, полученые в ходе этих испытаний, оцениваются с учетом соответствия фактически применяемых методов анализа с методами, которые исследователи изначально планировали использовать. В настоящее время практически невозможно опубликовать результаты РКИ в международных рецензируемых журналах, если протокол исследования не был предварительно зарегистрирован.

Чтобы избежать проблемы множественных сравнений при анализе средних трех и более групп, следует применять дисперсионный анализ. Основы дисперсионного анализа были разработаны в 20-е годы XX столетия английским биологом и генетиком сэром Рональдом Фишером (1890—1962), который по праву считается одним из основателей современной статистической науки. Существует много видов дисперсионного анализа, детальное описание которых представлено в литературе. В данной статье рассматривается самый простой вариант — одномерный (однофакторный) дисперсионный анализ для независимых групп (One-way ANalysis Of VAriance, ANOVA) в ходе которого проверяется нулевая гипотеза о равенстве средних для трех и более независимых групп. Как следует из названия, основным элементом анализа является дисперсия. Теоретические знания о том, как «работает» данный критерий изложены в большинстве пособий по статистике [1, 2], поэтому здесь будет представлено лишь практическое применение дисперсионного анализа с помощью SPSS, однако перед тем, как приступить к анализу данных, необходимо проверить необходимые условия для его применения:

1. Количественный тип данных, причем желательны непрерывные, а не дискретные данные.
2. Независимые выборки.
3. Нормальное распределение изучаемого признака в популяциях, из которых отобраны выборки.
4. Равенство дисперсий изучаемого признака в популяциях, из которых отобраны выборки.
5. Независимые наблюдения в каждой из выборок.

Рассмотрим одномоментное поперечное исследование, в ходе которого сравнивались средние значения систолического артериального давления в трех профессиональных группах (врачи, учителя, чиновники). В качестве нулевой служит гипотеза об отсутствии различий между средними значениями артериального давления в изучаемых группах. Файл с данными (Human_Ecology_3_2008) доступен на сайте журнала: http://www.nsmu.ru/nauka_sgmu/rio/eco_human/. Перед тем как начать проверку гипотезы с помощью однофакторного дисперсионного анализа, следует проверить, можно ли применять этот критерий в данной ситуации. Артериальное давление является непрерывной количественной величиной. Группы являются независимыми, так как один и тот же человек в один момент времени не может быть отнесен к более чем одной категории рода занятий.

Для проверки условия нормальности распределения в каждой из групп с помощью описательной статистики, графических методов и статистических критериев, в меню Analyze следует выбрать «Descriptive statistics», затем «Explore». В открывшемся диалоговом окне слева будет список переменных, из которых следует выбрать те, для которых планируется провести проверку распределения (в данном случае переменную AD). Кроме того, чтобы изучить распределение в обеих группах, следует в окно «Factor List» поместить группировочную переменную «RZ» (рис. 1). После выбора меню Plots лучше в «Descriptive» убрать флажок «Stem and leaf» и отметить гистограмму (Histogram), как показано на рис. 2. В меню Boxplot можно отметить «Factor level together» для получения «ящичных диаграмм» и поставить флажок на «Normality plots with tests». В меню Spread vs. Level with Levene Test надо отметить «Untransformed» для проверки условия равенства дисперсий.

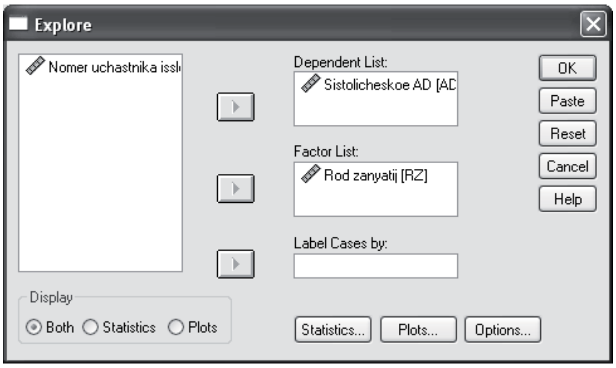

Рис. 1. Диалоговое окно «Explore» для определения зависимых переменных (Dependent List) и группировочных переменных (Factor List)

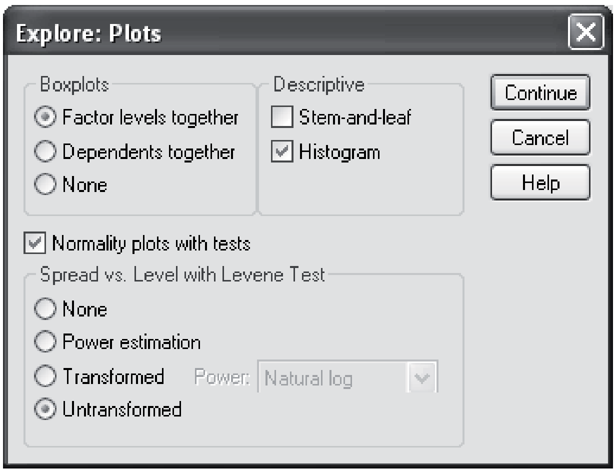

Рис. 2. Диалоговое окно «Plots» для оценки распределения данных с помощью графиков и статистических критериев, а также для проверки равенства дисперсий

Общие сведения о группах (количество наблюдений и пропущенные значения) представлены на рис. 3. Полученные данные описательной статистики для каждой из групп представлены на рис. 4. Исходя из данных асимметрии и эксцесса видим, что распределения лишь незначительно смещены, однако для окончательного решения о возможности применения дисперсионного анализа следует оценить гистограммы и квантильные диаграммы, а также результаты применения статистических критериев для проверки распределения (рис. 5).

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	Rod zanyatij	N	Percent	N	Percent	N	Percent
Sistolicheskoe AD	Vrachi	35	100,0%	0	,0%	35	100,0%
	Uchitelya	35	100,0%	0	,0%	35	100,0%
	Chinovniki	35	100,0%	0	,0%	35	100,0%

Рис. 3. Данные об общем количестве наблюдений и количестве пропущенных величин

Поскольку результаты применения критерия Shapiro-Wilk показывают, что нулевую гипотезу об отсутствии различий между распределением в каждой из групп и нормальным распределением отвергнуть нельзя ($p = 0{,}927$; $p = 0{,}797$ и $p = 0{,}881$), что не противоречит результатам графической оценки (графики не представлены), можно считать, что данные в обеих группах подчиняются закону нормального распределения. Кроме того, достигнутый уровень значимости (величина p) для критерия Levene (рис. 6) составил 0,390, что не позволяет отвергнуть нулевую гипотезу о равенстве дисперсий в изучаемых группах. Таким образом, все необходимые условия для применения однофакторного дисперсионного анализа выполняются.

Для применения однофакторного дисперсионного анализа следует открыть диалоговое окно «One-Way ANOVA», которое открывается при помощи меню Analyze → Compare Means → One-Way ANOVA (рис. 7). В область «Dependent List» переносится зависимая переменная, средние значения которой планируется сравнить. В данном примере это переменная «AD». В область «Factor» помещается группировочная переменная, то есть переменная, которая будет использоваться для разделения всей выборки на группы. В данном примере это переменная «RZ». В диалоговом окне «Options» следует отметить «Descriptive» для получения данных описательной статистики, «Homogeneity of variance test» для проверки условия равенства дисперсий, а также «Means plot» для графического изображения средних арифметических для каждой из групп (рис. 8). Запуск анализа осуществляется нажатием на кнопку «OK» в правом верхнем углу диалогового окна «One-Way ANOVA» (см. рис. 7).

Результаты применения дисперсионного анализа представлены на рис. 9 и 10: на рис. 9 — общее количество наблюдений в каждой из групп (N), средние арифметические значения (Mean), стандартные отклонения (Std. Deviation), стандартные ошибки средних арифметических (Std. Error), 95 % доверительные интервалы для средних (95 % Confidence Interval for Means), а также минимальные (Minimum) и максимальные (Maximum) значения; на рис. 10

– результаты проверки равенства дисперсий с помощью критерия Levene.

Descriptives

	Rod zanyatij			Statistic	Std. Error
Sistolicheskoe AD	Vrachi	Mean		127,9034	2,00123
		95% Confidence Interval for Mean	Lower Bound	123,8364	
			Upper Bound	131,9703	
		5% Trimmed Mean		127,7575	
		Median		128,4264	
		Variance		140,172	
		Std. Deviation		11,83942	
		Minimum		103,64	
		Maximum		154,23	
		Range		50,59	
		Interquartile Range		14,04	
		Skewness		,155	,398
		Kurtosis		-,002	,778
	Uchitelya	Mean		131,8018	1,51750
		95% Confidence Interval for Mean	Lower Bound	128,7179	
			Upper Bound	134,8858	
		5% Trimmed Mean		131,7477	
		Median		131,8680	
		Variance		80,599	
		Std. Deviation		8,97768	
		Minimum		112,44	
		Maximum		150,07	
		Range		37,63	
		Interquartile Range		15,46	
		Skewness		,066	,398
		Kurtosis		-,602	,778
	Chinovniki	Mean		133,0380	1,53580
		95% Confidence Interval for Mean	Lower Bound	129,9169	
			Upper Bound	136,1591	
		5% Trimmed Mean		133,1195	
		Median		133,7043	
		Variance		82,553	
		Std. Deviation		9,08589	
		Minimum		112,74	
		Maximum		151,11	
		Range		38,37	
		Interquartile Range		12,27	
		Skewness		-,009	,398
		Kurtosis		-,408	,778

Рис. 4. Описательная статистика для переменной «AD» в изучаемых группах

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Rod zanyatij	Statistic	df	Sig.	Statistic	df	Sig.
Sistolicheskoe AD	Vrachi	,094	35	,200*	,986	35	,927
	Uchitelya	,097	35	,200*	,981	35	,797
	Chinovniki	,074	35	,200*	,984	35	,881

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Рис. 5. Результаты проверки нормальности распределения данных в каждой из групп с помощью критериев Kolmogorov-Smirnov и Shapiro-Wilk

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Sistolicheskoe AD	Based on Mean	,951	2	102	,390
	Based on Median	,939	2	102	,394
	Based on Median and with adjusted df	,939	2	89,201	,395
	Based on trimmed mean	,958	2	102	,387

Рис. 6. Результаты проверки равенства дисперсий с помощью критерия Levene

Результаты проверки гипотезы о равенстве средних представлены на рис. 11. Во втором столбце представлены общая вариабельность признака (Total Sum of Squares), а также ее составляющие – внутригрупповая (Within Groups Sum of Squares) и межгрупповая (Between Groups Sum of Squares) вариабельность.

В данном примере 4,6 % всей вариабельности артериального давления обусловлено межгрупповыми различиями (502,7 / 10 815,7 = 0,046). В третьем столбце представлено количество степеней свободы, которое используется для рассчета межгрупповой и внутригрупповой дисперсии. Разделив первое на второе, получим число F (названное в честь Фишера), которое равно 1, если верна нулевая гипотеза об отсутствии межгрупповых различий. Таким образом, буквальный смысл дисперсионного анализа заключается в сравнении межгрупповой и внутригрупповой дисперсии признака и при их равенстве делается вывод об отсутствии межгрупповых различий между средними. В данном случае F = 251,4 / 101,1 = 2,486.

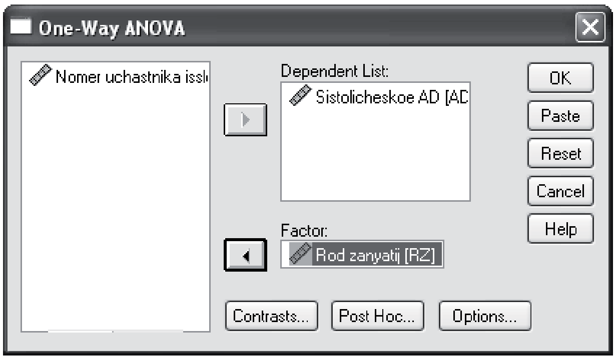

Рис. 7. Диалоговое окно «One-Way ANOVA»

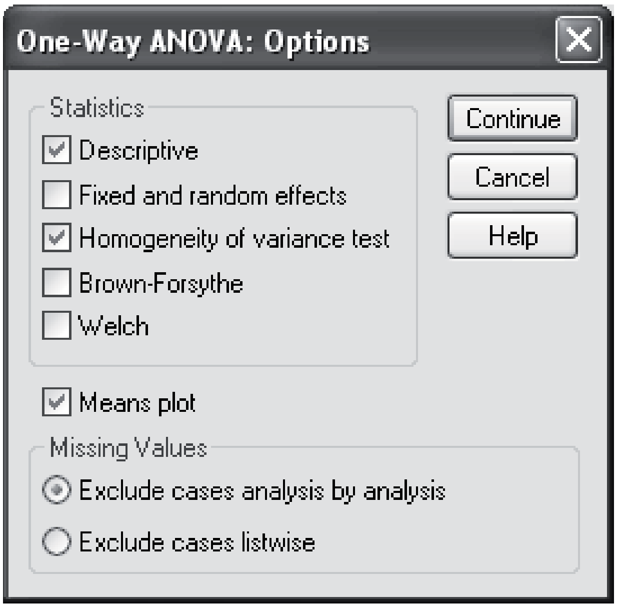

Рис. 8. Диалоговое окно «Options»

Descriptives

Sistolicheskoe AD

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Vrachi	35	127,9034	11,83942	2,00123	123,8364	131,9703	103,64	154,23
Uchitelya	35	131,8018	8,97768	1,51750	128,7179	134,8858	112,44	150,07
Chinovniki	35	133,0380	9,08589	1,53580	129,9169	136,1591	112,74	151,11
Total	105	130,9144	10,19791	,99521	128,9408	132,8879	103,64	154,23

Рис. 9. Описательная статистика для сравниваемых групп

При сравнении числа F с табличными значениями для имеющегося количества степеней свободы рассчитывается вероятность получения выявленных различий между дисперсиями, если нулевая гипотеза верна (величина p). В данном примере p = 0,088, то есть нулевую гипотезу об отсутствии различий между групповыми средними отвергнуть нельзя. Значит, можно сделать вывод об отсутствии статистически значимых различий между средними значениями артериального давления в изучаемых группах. Представляя результаты дисперсионного анализа, рекомендуется указывать значение F с указанием количества степеней свободы и достигнутый уровень значимости (для данного примера $F_{2,102} = 2,486$; p = 0,088).

Test of Homogeneity of Variances

Sistolicheskoe AD

Levene Statistic	df1	df2	Sig.
,951	2	102	,390

Рис. 10. Результаты проверки равенства дисперсий с помощью критерия Levene

ANOVA

Sistolicheskoe AD

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	502,720	2	251,360	2,486	,088
Within Groups	10313,015	102	101,108		
Total	10815,736	104			

Рис. 11. Результаты дисперсионного анализа

Получили бы мы такие же результаты при попарном сравнении групп? При проведении трех сравнений с использованием критерия Стьюдента были получены три величины p: 0,569; 0,126 и 0,046. Последнее значение было получено при сравнении артериального давления врачей и чиновников. Если бы мы проводили только одно сравнение, то можно было бы считать различия статистически значимыми, но поскольку всего проводилось три сравнения, то критический уровень значимости должен быть не 0,05, а 0,017. Достигнутый уровень значимости (0,046) превышает новый критический уровень (0,017), значит, различия не могут считаться статистически значимыми, что не противоречит результатам дисперсионного анализа.

При интерпретации результатов дисперсионного анализа всегда следует помнить, что, во-первых, выявление статистически значимых различий говорит только о том, что различия между средними существуют, но не говорит о том, какие из групп различаются между собой. Во-вторых, несмотря на название метода, результаты дисперсионного анализа не говорят о различиях между дисперсиями в изучаемых группах. Это проверяется с помощью критерия Levene, причем равенство дисперсий является одним из необходимых условий применения дисперсионного анализа.

Что делать, если условие равенства дисперсий не выполняется (результаты применения критерия Levene показывают наличие статистически значимых различий между дисперсиями в изучаемых группах)? Для таких случаев существуют критерии Brown-Forsythe и Welch, которые можно выбрать в диалоговом окне «Options», (см. рис. 8). Интерпретация результатов аналогична таковой для дисперсионного анализа.

Что делать, если дисперсионный анализ покажет наличие статистически значимых различий между средними трех или более групп? Следующим шагом будет проведение апостериорных сравнений для обнаружения, между какими группами имеются различия. Для апостериорных сравнений SPSS предлагает 18 критериев (рис. 12). Какой из них выбрать?

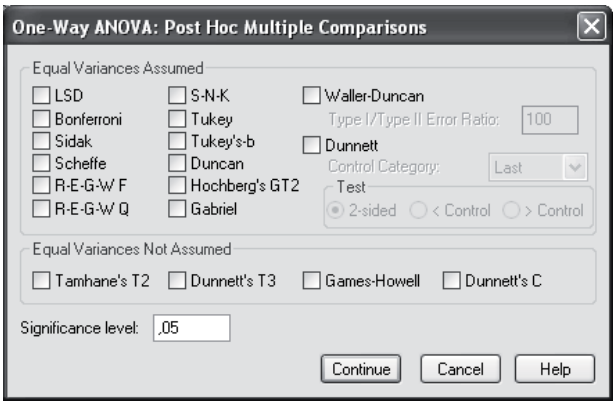

Рис. 12. Диалоговое окно для выбора апостериорных сравнений

Апостериорные сравнения представляют собой попарные сравнения изучаемых групп для обнаружения различий между ними. Подобные сравнения могут быть проведены с помощью критерия Стьюдента для независимых выборок, что выглядит по меньшей мере странно, учитывая все сказанное выше о проблеме множественных сравнений. Однако отличие от простых попарных сравнений заключается в том, что при проведении апостериорных сравнений рассчитываются новые критические уровни значимости для удержания ошибки 1 типа в пределах 5 % как показано в таблице. Наиболее простым и наиболее популярным способом коррекции ошибки 1 типа является поправка Бонферрони (Bonferroni), при проведении которой традиционный уровень ошибки 1 типа делится на количество сравнений для получения нового критического уровня значимости. Так, если имеется 3 сравнения, то новый критический уровень должен быть 0,05 / 3 = 0,017. Поправка Бонферрони хорошо контролирует ошибку 1 типа, но вместе с тем является очень консервативной и приводит к уменьшению статистической мощности критерия и повышению вероятности ошибки 2 типа, то есть вероятности принятия решения об отсутствии различий там, где они на самом деле есть. Либеральные критерии, в свою очередь, завышают вероятность ошибки 1 типа, то есть вероятность принятия решения о наличии различий там, где их нет. Таким образом, при выборе статистического критерия для апостериорных сравнений необходимо принимать во внимание, как критерии контролируют ошибки 1 и 2 типов и как они работают при несоблюдении необходимых условий применения дисперсионного анализа.

Критерий LSD (Least Significant Difference), или критерий наименьших значимых различий, совсем не контролирует ошибку 1 типа и поэтому для проведения адекватных сравнений непригоден. Критерий S-N-K (Studentized-Neuman-Keuls) также слишком либерален. Поправка Бонферрони дает хорошие результаты при небольшом (до 5) количестве сравнений. При проведении большего числа сравнений лучше пользоваться критерием Тьюки (Tukey). Критерии Данна (Dunn) и Шеффе (Scheffe) обладают несколько меньшей статистической мощностью, чем критерий Тьюки. Наилучшее сочетание мощности и контроля за ошибкой 1 типа предлагает критерий REGWQ (Ryan, Einot, Gabriel, Welsh Q-критерий), который можно рекомендовать как критерий выбора при необходимости сравнить большое количество групп, но только если объемы групп и дисперсии изучаемого признака в группах равны.

Если количество наблюдений в сравниваемых группах отличается незначительно (например, n_1 = 35, n_2 = 39, n_3 = 32), то рекомендуется применять критерий Габриэля (Gabriel). Если же объемы групп различаются более существенно, то тогда лучше использовать GT-2 критерий Хохберга (Hochberg's GT-2 test). Несмотря на то, что эти критерии допускают разные объемы групп, дисперсии изучаемого признака в группах не должны отличаться. Если же в результате применения критерия Levene обнаружены различия дисперсий, то выводы о различиях между средними следует делать только по результатам применения критерия Welch или критерия Brown-Forsythe (см. рис. 8). Апостериорные сравнения в таких ситуациях рекомендуется проводить с использоваенным критерия Games-Howell, однако следует помнить, что этот критерий может быть слишком либеральным при малых группах, а также когда группы неравны по объему. Если необходимо проводить сравнения нескольких групп с контрольной группой, то для этого SPSS предлагает критерий Даннетта (Dunnett's test). Подробнее о выборе критерия для апостериорных сравнений можно прочитать в пособии Toothaker [8].

Результаты апостериорных сравнений в SPSS выглядят как на рис. 13. В рассматриваемом примере проведение апостериорных сравнений не является необходимостью, так как значимых различий между средними значениями артериального давления между группами выявлено не было, однако рис. 13 убедительно показывает, насколько разные результаты можно получить при применении разных критериев. Так, согласно критерию LSD, имеются статистически значимые различия между врачами и чиновниками (p = 0,035), в то время как, согласно критерию Бонферрони, нулевую гипотезу об отсутствии различий отвергнуть нельзя, что соответствует результатам дисперсионного анализа.

Помимо апостериорных попарных сравнений всех групп или сравнения всех групп с контрольной SPSS дает возможность проводить плановые сравнения нескольких групп. Данная функция особенно удобна, когда необходимо провести анализ трендов (если группы можно расположить в логической последовательности) или сравнить лишь некоторые из групп, а не каждую с каждой. Представим на момент, что артериальное давление в данном примере изучалось не в трех профессиональных, а в трех возрастных группах, где первая группа включала в себя самых молодых, а третья — самых старших участников исследования. В такой ситуации можно оценить тренд или ответить на вопрос, имеется ли зависимость между возрастом и артериальным давлением.

Multiple Comparisons

Dependent Variable: Sistolicheskoe AD

	(I) Rod zanyati	(J) Rod zanyati	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
LSD	Vrachi	Uchitelya	-3,89847	2,40366	,108	-8,6661	,8692
		Chinovniki	-5,13461*	2,40366	,035	-9,9023	-,3670
	Uchitelya	Vrachi	3,89847	2,40366	,108	-,8692	8,6661
		Chinovniki	-1,23613	2,40366	,608	-6,0038	3,5315
	Chinovniki	Vrachi	5,13461*	2,40366	,035	,3670	9,9023
		Uchitelya	1,23613	2,40366	,608	-3,5315	6,0038
Bonferroni	Vrachi	Uchitelya	-3,89847	2,40366	,324	-9,7492	1,9523
		Chinovniki	-5,13461	2,40366	,105	-10,9854	,7162
	Uchitelya	Vrachi	3,89847	2,40366	,324	-1,9523	9,7492
		Chinovniki	-1,23613	2,40366	1,000	-7,0869	4,6146
	Chinovniki	Vrachi	5,13461	2,40366	,105	-,7162	10,9854
		Uchitelya	1,23613	2,40366	1,000	-4,6146	7,0869

*. The mean difference is significant at the .05 level.

Рис. 13. Результаты применения критериев LSD и Bonferroni для апостериорных сравнений

Для анализа трендов и проведения плановых сравнений нужно открыть диалоговое окно «Contrasts» (рис. 14) в основном окне дисперсионного анализа (см. рис. 7). Для оценки тренда следует отметить «Polynomial» и выбрать линейный тренд или тренды второго (Quadratic), третьего (Cubic), четвертого (4th) или пятого (5th) порядков.

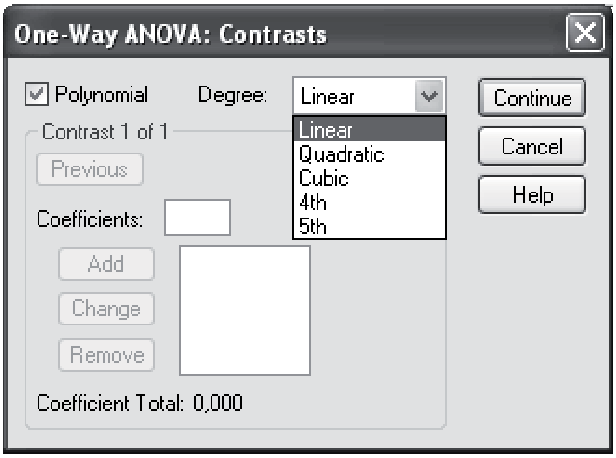

Рис. 14. Диалоговое окно «Contrasts»

Результаты дисперсионного анализа вместе с анализом тренда представлены на рис. 15. Помимо уже известных результатов об отсутствии статистически значимых различий между групповыми средними (p = 0,088) во второй строке представлен результат анализа тренда, который говорит о том, что можно отвергнуть нулевую гипотезу об отсутствии линейного тренда (p = 0,035). Таким образом, можно было бы сделать вывод о наличии линейного тренда между возрастом и артериальным давлением. Поскольку в нашем примере в качестве группирующей переменной используется номинальная переменная «Род занятий», выводы о наличии или отсутствии тренда смысла не имеют. Оценка плановых сравнений в деталях описана в литературе [6] и здесь не рассматривается. Многомерный дисперсионный анализ, при котором анализируется более одного фактора, а также дисперсионный анализ для повторных наблюдений, который следует применять, если группы не являются независимыми, будут рассматриваться в последующих выпусках «Экологии человека».

ANOVA

Sistolicheskoe AD

			Sum of Squares	df	Mean Square	F	Sig.
Between Groups	(Combined)		502,720	2	251,360	2,486	,088
	Linear Term	Contrast	461,373	1	461,373	4,563	,035
		Deviation	41,347	1	41,347	,409	,524
Within Groups			10313,015	102	101,108		
Total			10815,736	104			

Рис. 15. Результаты дисперсионного анализа с оценкой тренда

Что делать, если надо сравнить три или более независимые группы, в которых данные не подчиняются закону нормального распределения? Такие ситуации весьма нередки в медицинских исследованиях и часто сочетаются с малыми объемами выборок. В таких ситуациях следует либо трансформировать имеющиеся данные с помощью различных арифметических преобразований до достижения нормальности распределения [4], после чего можно будет применить дисперсионный анализ, либо применять критерий Краскела-Уоллиса (Kruskal-Wallis H-test), иногда еще называемый непараметрическим дисперсионным анализом. Критерий Краскела-Уоллиса рассчитывается с использованием не фактических значений переменных, а их рангов, поэтому является методом выбора при сильно скошенных распределениях. Так же, как и дисперсионный анализ, критерий Краскела-Уоллиса поможет выяснить, имеются ли различия между группами, но не сможет показать, между какими из групп эти различия существуют. При обнаружении статистически значимых различий между группами с помощью критерия Краскела-Уоллиса далее следует проводить апостериорные сравнения с помощью критерия Манна-Уитни, рассмотренного в предыдущем выпуске журнала [3]. Следует помнить, что, поскольку SPSS не дает возможности автоматически проводить апостериорные сравнения с помощью непараметрических методов статистики, исследователям самим необходимо рассчитывать новые критические уровни значимости исходя из представленных выше формул или как показано в таблице.

Для использования критерия Краскела-Уоллиса в SPSS необходимо открыть диалоговое окно «Tests for Several Independent Samples», которое открывается при помощи меню Analyze → Nonparametric Tests → K Independent Samples (рис. 16). В поле «Test Variable List» помещается изучаемая переменная (AD). В поле «Grouping Variable» помещается группировочная переменная (RZ). Для определения сравниваемых групп следует открыть диалоговое окно «Define Range» (рис. 17) и задать минимальное и максимальное значения, с помощью которых кодируются сравниваемые группы (в данном случае «1» и «3»). Далее в меню Options можно выбрать «Descriptive» для получения данных описательной статистики (рис. 18).

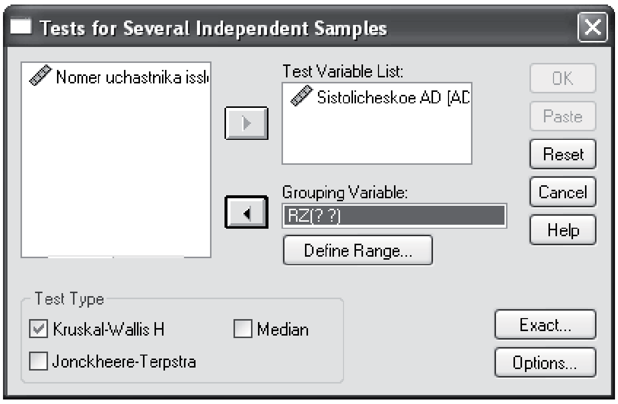

Рис. 16. Диалоговое окно «Tests for Several Independent Samples»

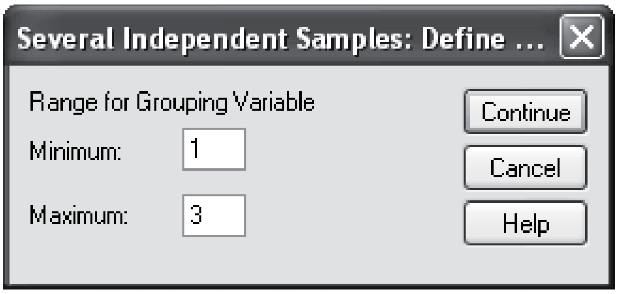

Рис 17. Диалоговое окно «Several Independent Samples: Define Groups»

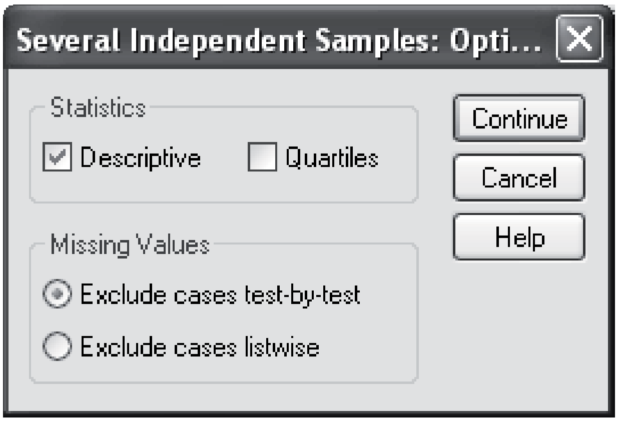

Рис.18. Диалоговое окно «Several Independent Samples: Options»

Результаты сравнения групп с помощью критерия Краскела-Уоллиса представлены на рис. 19. В нижней таблице представлены значения критерия Краскела-Уоллиса, обозначенные в таблице как Chi-Square, количество степеней свободы (df) и достигнутый уровень значимости различий (Asymp. Sig.). Результаты показывают, что, хотя средний ранг значений артериального давления в группе чиновников был выше, чем в остальных группах, статистически значимых различий между группами нет. Если бы различия были обнаружены, следовало бы проводить попарные сравнения групп при помощи критерия Манна-Уитни с новым критическим уровнем значимости: 0,05 / 3 = 0,017.

Ranks

	Rod zanyatij	N	Mean Rank
Sistolicheskoe AD	Vrachi	35	44,23
	Uchitelya	35	55,34
	Chinovniki	35	59,43
	Total	105	

Test Statistics[a,b]

	Sistolicheskoe AD
Chi-Square	4,670
df	2
Asymp. Sig.	,097

a. Kruskal Wallis Test
b. Grouping Variable: Rod zanyatij

Рис. 19. Результаты сравнения артериального давления в группах врачей, учителей и чиновников с помощью критерия Краскела-Уоллиса

Можно ли было применять непараметрический критерий Краскела-Уоллиса в этой ситуации (данные подчиняются закону нормального распределения)? Да, но, как видно из достигнутого уровня значимости, критерий Краскела-Уоллиса имеет несколько меньшую статистическую мощность, чем дисперсионный анализ, поэтому при нормальном распределении и выполнении прочих условий дисперсионный анализ является методом выбора. Некоторые исследователи не рекомендуют применять параметрические методы (в том числе и дисперсионный анализ), если объем каждой из групп составляет менее 30 наблюдений, даже если выборочные данные имеют нормальное распределение [5]. Можно ли использовать дисперсионный анализ при отклонении распределения от нормального? При наличии больших выборок с равными дисперсиями дисперсионный анализ достаточно устойчив к небольшим отклонениям распределения от нормального, особенно при равных объемах выборок. При малых выборках применение дисперсионного анализа для скошенных распределений может привести к сильно искаженным результатам, поэтому рекомендуется в такой ситуации применять критерий Краскела-Уоллиса.

В следующем выпуске будут рассмотрены статистические критерии для анализа парных наблюдений.

Список литературы

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

2. *Гланц С.* Медико-биологическая статистика / С. Гланц. – М. : Практика, 1998. – 460 с.

3. *Гржибовский А. М.* Анализ количественных данных для двух независимых групп / А. М. Гржибовский // Экология человека. – 2008. – № 2. – С. 54–61.

4. *Гржибовский А. М.* Типы данных, проверка распределения и описательная статистика / А. М. Гржибовский // Экология человека. – 2008. – № 1. – С. 52–58.

5. *Chang Y. H.* Biostatistics 101: Data presentation / Y. H. Chang // Singapore Medical Journal. – 2003. – N 6. – P. 280–285.

6. *Field A.* Discovering statistics using SPSS / A. Field. – SAGE Publications, 2005. – 779 p.

7. *Lee K. L. et al.* Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease / K. K. Lee, J. F. McNeer, C. F. Starmer et al. // Circulation. – 1980. – Vol. 61. – N 3. – P. 508–515.

8. *Toothaker L. E.* Multiple comparison procedures. Sage University paper series on quantitative applications in the social sciences, 07-089 / L.E. Toothaker. – SAGE Publications, 1993. – 104 p.

ANALYSIS OF THREE AND MORE INDEPENDENT GROUPS OF QUANTITATIVE DATA

A. M. Grjibovsky

National Institute of Public Health, Oslo, Norway

In the article, a comparison of three and more independent groups of quantitative data has been considered with the help of the one-dimension (one-factor) dispersion analysis (One-Way ANOVA) and the Kruskal-Wallis test with the use of the package of applied statistical programs SPSS. Special attention has been paid to the problem of multiple comparisons and check of necessary conditions for application of the dispersion analysis. Besides, the criteria for aposteriori comparisons in cases of detection of statistically significant differences during the dispersion analysis have been considered. The stated facts have given general information about statistical criteria used for check of hypotheses about equality of three and more independent groups, and should arouse interest of journals' readers to reading of single-purpose literature before beginning of work with future publications.

Key words: distribution, dispersion, one-factor dispersion analysis, the Kruskal-Wallis criterion, aposteriori comparisons.

Контактная информация:

Гржибовский Андрей Мечиславович – старший советник Национального института общественного здоровья, г. Осло, Норвегия

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

Тел.: +47 22042392, +47 45268913; e-mail: angr@fhi.no

Статья поступила 06.02.2008 г.