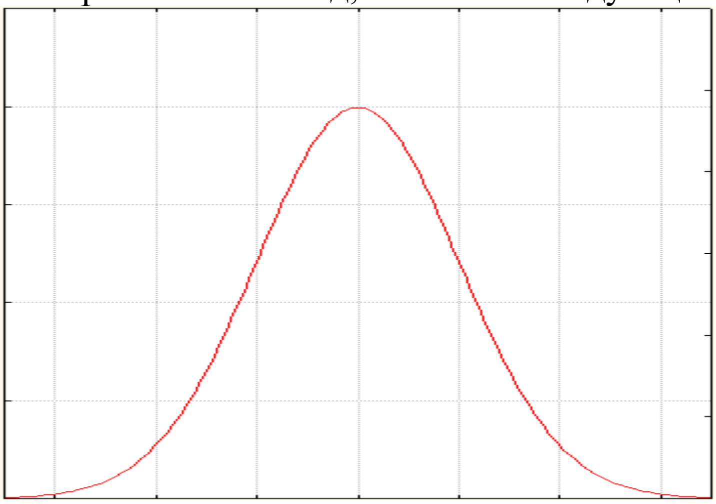
6. НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ

6.1 Назначение непараметрических методов анализа данных

Большинство методов анализа данных, рассмотренных ранее, применимы не во всех случаях, а лишь при соблюдении определенных условий. Строго говоря, они применимы лишь при условии, что анализируемые данные соответствуют нормальному (гауссовскому) закону распределения. Упрощенно говоря, можно считать, что распределение исследуемой величины близко к нормальному закону, если ее гистограмма имеет вид, близкий к следующему:



В других случаях следует применять непараметрические методы анализа данных. Таким образом, непараметрические методы рекомендуется использовать в следующих случаях:

- когда есть основания предполагать, что закон распределения исследуемой величины существенно отличается от нормального;
- при малых объемах выборок (до 30), так как в этом случае, строго говоря, проверка на соответствие нормальному закону невозможна;
- в случаях, когда некоторые из исследуемых величин измеряются в баллах с небольшим диапазоном значений (например, по пятибалльной шкале) или являются нечисловыми.

Непараметрические методы обычно не требуют вычисления параметров исследуемых величин (среднего, стандартного отклонения и т.д.).

Для всех параметрических методов анализа данных, рассмотренных ранее (t-критерий Стьюдента для сравнения независимых и зависимых выборок, корреляционный анализ на основе коэффициента парной корреляции и т.д.) имеются аналогичные непараметрические методы.

6.2 Непараметрические методы оценки различия между независимыми выборками

Параметрический метод оценки различия между независимыми выборками – t-критерий Стьюдента (см. раздел 2.1). Имеется несколько *непараметрических* методов сравнения независимых выборок: U-критерий Манна-Уитни, критерий Вальда-Вольфовица (критерий серий) и т.д.

Рассмотрим применение одного из этих критериев – U-критерия Манна-Уитни.

Пример 6.1 – Сравнивается прочность двух материалов (М1 и М2). Испытано 8 образцов М1 и 7 образцов М2. Характеристики их прочности:

– М1: 570, 620, 620, 560, 560, 620, 630, 580;

– М2: 530, 610, 610, 540, 530, 530, 560.

Выяснить, является ли статистически значимым различие материалов по прочности.

В данном случае объемы выборок очень малы (n_1=8, n_2=7). Поэтому невозможно проверить, соответствуют ли исследуемые величины нормальному закону распределения: данных недостаточно даже для построения гистограммы. Для решения задачи необходимо использовать непараметрический метод.

Решим задачу, используя U-критерий Манна-Уитни.

1 Все значения из обеих выборок записываются по возрастанию. Каждому значению присваивается ранг (номер), т.е. минимальное значение (из обеих выборок) получает номер 1, следующее за ним – номер 2, и т.д. ***Если значения равны, то они получают одинаковый (усредненный) ранг.***

Материал	М2	М2	М2	М2	М1	М1	М2	М1
Прочность	530	530	530	540	560	560	560	570
Ранг	2	2	2	4	6	6	6	8

Материал	М1	М2	М2	М1	М1	М1	М1
Прочность	580	610	610	620	620	620	630
Ранг	9	10,5	10,5	13	13	13	15

Здесь, например, есть три одинаковых значения (530). Можно сказать, что они «делят места с первого по третье». Поэтому им присвоен усредненный ранг, равный 2. Аналогично, есть два одинаковых значения 610. Они делят 10-е и 11-е места. Им присвоен средний ранг 10,5.

2 Вычисляются суммы рангов выборок:

для М1: R_1 = 6+6+8+9+13+13+13+15=83;

для М2: R_2 = 2+2+2+4+6+10,5+10,5=37.

3 Вычисляются вспомогательные величины U_1 и U_2:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1,$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2.$$

В данном примере $U_1 = 8 \cdot 7 + \frac{8 \cdot (8+1)}{2} - 83 = 9$, $U_2 = 8 \cdot 7 + \frac{7 \cdot (7+1)}{2} - 37 = 47$.

4 Вычисляется U-критерий Манна-Уитни:

$U = \min(U_1, U_2) = \min(9; 47) = 9.$

5 Если расчет выполняется вручную, то по значениям n_1 и n_2, а также по заданному уровню значимости α из таблиц U-критерия Манна-Уитни определяется критическое значение $U_{кр}$. Если $U < U_{кр}$, то различие между выборками статистически значимо. В данном примере при $n_1 = 8$, $n_2 = 7$, $\alpha = 0{,}05$, $U_{кр} = 10$. Так как $U = 9$, т.е. $U < U_{кр}$, различие между выборками статистически значимо.

Из рангов видно, что значения прочности для М1 в основном больше, а для М2 – меньше. Можно также вычислить среднюю прочность: для М1 – 595, для М2 – 558,6. Таким образом, М1 значимо прочнее, чем М2.

Если расчет выполняется в программе (например, Statistica), то по значениям U, n_1 и n_2 вычисляется расчетный уровень значимости P. Если $P < \alpha$, то различие между выборками статистически значимо.

Решение задачи в Statistica

1 Присвоить переменным имена MATERIAL и PROCHNOST. Для ввода данных потребуется 15 строк (по общему количеству данных) и два столбца.

2 Ввести данные.

	1 MATERIAL	2 PROCHNOST
1	M1	570
2	M1	620
3	M1	620
4	M1	560
5	M1	560
6	M1	620
7	M1	630
8	M1	580
9	M2	530
10	M2	610
11	M2	610
12	M2	540
13	M2	530
14	M2	530
15	M2	560

3 Из меню Statistics выбрать команду Nonparametrics, затем – команду Comparing two independent samples (groups).

4 В появившемся окне Comparing two groups нажать кнопку Variables для выбора анализируемых переменных. В списке Dependent variable list выбрать переменную PROCHNOST, в списке Indep. (grouping) variable – переменную MATERIAL. Нажать OK.

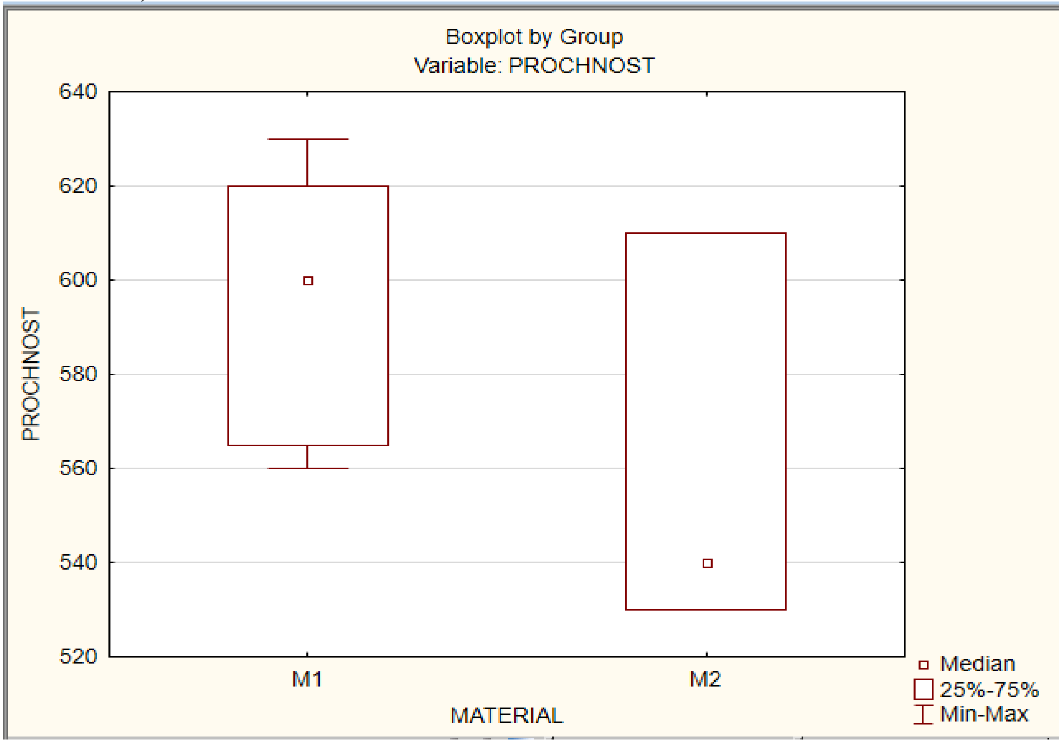
5 В очередном окне Comparing two groups нажать кнопку Mann-Whitney U-test. На экран выводится окно результатов.

Mann-Whitney U Test (w/ continuity correction) (Stat_лекции_непарам)
By variable MATERIAL
Marked tests are significant at p <,05000

variable	**Rank Sum M1**	Rank Sum M2	U	Z	p-value	Z adjusted	p-value	Valid N M1	Valid N M2	2*1sided exact p
PROCHNOST	83,00000	37,00000	9,000000	2,140959	0,032278	2,166251	0,030293	8	7	0,028904

Так как p-value = 0,0323 < 0,05, различие между выборками значимо.

Можно также проиллюстрировать результаты диаграммой размаха (**Box & Whisker**):



Из нее также видно, что M1 прочнее, чем M2.

6.3 Непараметрические методы оценки различия между зависимыми выборками

Параметрический метод оценки различия между зависимыми выборками (на основе *t*-критерия Стьюдента) рассмотрен в разделе 2.2. Имеется несколько непараметрических критериев для сравнения зависимых выборок, например, *T*-критерий Вилкоксона, *Z*-критерий знаков. Рассмотрим применение одного из них – *T*-критерия Вилкоксона.

Пример 6.2 – Анализируется эффективность некоторого мероприятия по снижению вредных выбросов. Имеются данные по выбросам 12 заводов:

Номер завода	1	2	3	4	5	6	7	8	9	10	11	12
До мероприятия	112	98	99	120	95	94	101	115	105	93	108	107
После мероприятия	98	106	105	110	95	85	103	105	103	82	58	103

Здесь, например, на первом заводе выбросы до мероприятия составляли 112 единиц, а после – 98 единиц. На втором заводе мероприятие было неэффективным (выбросы до мероприятия – 98 единиц, после – 106 единиц), и т.д.

Выяснить, привело ли мероприятие к значимому снижению выбросов.

Так как объем выборки небольшой (n=12), требуется применять непараметрический метод сравнения выборок –T-критерий Вилкоксона.

1 Находятся разности значений (в данном примере – разности выбросов до и после мероприятия). Каждой разности присваивается ранг (номер) в соответствии с ее *абсолютным значением (без учета знака)*. Если разности равны, то они получают одинаковый (усредненный) ранг. Нулевые разности *не учитываются* (им не присваиваются ранги).

Номер завода	1	2	3	4	5	6	7	8	9	10	11	12
Разность	14	-8	-6	10	0	9	-2	10	2	11	50	4
Ранг	10	5	4	7,5	–	6	1,5	7,5	1,5	9	11	3

Примечание – Можно было, наоборот, определять разности выбросов после и до мероприятия. Изменились бы только знаки разностей.

В данном случае минимальная ненулевая разность (2 единицы) наблюдалась у седьмого и девятого заводов. Можно сказать, что в ряду разностей они «делят первое и второе места». Поэтому им присвоен усредненный ранг (1,5). Следующая по абсолютной величине разность (4 единицы) – у 12-го завода; ей присвоен ранг 3, и т.д.

2 Вычисляются суммы рангов положительных и отрицательных разностей:

T_1 = 10+7,5+6+7,5+1,5+9+11+3=55,5;

T_2 = 5+4+1,5=10,5.

3 Вычисляется T-критерий Вилкоксона:

$T = \min(T_1, T_2)$.

В данном примере T=10,5.

4 Если расчет выполняется вручную, то по значению n, а также по заданному уровню значимости α из таблиц T-критерия Вилкоксона определяется критическое значение $T_{кр}$. Если $T < T_{кр}$, то различие между выборками статистически значимо. В данном примере при n = 12, α = 0,05, $T_{кр}$ = 17. Так как T = 10,5, т.е. $T < T_{кр}$, то различие между выборками статистически значимо. Из исходных данных (и из расчетов d) видно, что в большинстве случаев выбросы снижаются. Значит, можно считать, что мероприятие привело к значимому *снижению* выбросов.

Если расчет выполняется в специальной программе (например, Statistica), то по значениям T и n определяется расчетный уровень значимости P. Если $P < \alpha$, то различие между выборками статистически значимо.

Решение задачи в программе Statistica.

1 Присвоить переменным имена **Do** и **Posle**. Ввести данные (два столбца).

	1 Do	2 Posle
1	112	98
2	98	106
3	99	105
4	120	110
5	95	95
6	94	85
7	101	103
8	115	105
9	105	103
10	93	82
11	108	58
12	107	103

2 Выбрать Statistics – Nonparametrics – Comparing two dependent samples (variables).

3 В появившемся окне Comparing two variables нажать кнопку Variables для выбора анализируемых переменных. В одном списке переменных выбрать **Do**, в другом - **Posle**. Нажать OK. Выполняется возврат в окно Comparing two variables.

4 Чтобы выполнить анализ данных, нажать кнопку Wilcoxon matched pairs test. На экран выводится окно результатов.

	Wilcoxon Matched Pairs Test (Spreadsheet1) Marked tests are significant at p <,05000			
Pair of Variables	**Valid N**	T	Z	p-value
Do & Posle	11	10,50000	2,000494	0,045448

Оценка значимости различия между выборками выполняется по величине p-value (расчетный уровень значимости). В данном примере P=0,045 < 0,05, т.е. различие выбросов до и после мероприятия статистически значимо.

6.4 Непараметрические методы анализа связи между величинами: ранговые коэффициенты корреляции

Вычисление коэффициентов корреляции, отражающих связь между величинами, рассмотрено в части 3. Для анализа связи между величинами могут также применяться непараметрические (ранговые) коэффициенты корреляции: коэффициенты Спирмена, Кендалла, «гамма».

Пример 6.3 – Исследуется связь некоторого заболевания с воздействием шума. Имеются следующие данные:

Работа в условиях шума	Заболеваемость, %
Никогда	3
Эпизодически	6
До 20% рабочего времени	6
20–40% рабочего времени	5
40–60% рабочего времени	9
60–80% рабочего времени	8
Свыше 80% рабочего времени	8

Эти данные означают, что среди работников, никогда не работающих в условиях шума, заболеваемость исследуемой болезнью составляет 3%; среди тех, кто работает в условиях шума эпизодически, заболеваемость составляет 6%, и т.д.

В данном случае одна из исследуемых величин (время работы в условиях шума) – нечисловая. Кроме того, объем выборки очень мал (n=7). Поэтому необходимо использовать непараметрический метод – ранговый коэффициент корреляции Спирмена.

1 Значениям каждой из анализируемых величин присваиваются ранги: минимальное значение получает номер 1, следующее – номер 2, и т.д. Если варианты равны, то они получают одинаковый (усредненный) ранг.

В данном примере для величины «время работы в условиях шума» ранги следующие: значению «никогда» соответствует ранг 1, значению «эпизодически» – 2, и т.д. Для величины «заболеваемость» ранги следующие: минимальное значение (3) получает ранг 1, следующее за ним (5) – ранг 2, и т.д.

2 Вычисляются разности рангов d_i, i=1,…,n.

Ранги	Время работы в условиях сильного шума	1	2	3	4	5	6	7
	Заболеваемость	1	3,5	3,5	2	7	5,5	5,5
Разности рангов, d_i		0	–1,5	–0,5	2	-2	0,5	1,5

Примечание – При вычислении разностей рангов неважно, какая из величин указана первой, а какая – второй.

3 Вычисляется ранговый коэффициент корреляции Спирмена (здесь n=7):

$$R_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

Примечание – Число 6 в формуле – постоянное, оно не зависит от объема выборки или каких-либо других условий.

В данном примере R_s = 0,7679.

Диапазон возможных значений (от –1 до 1) и смысл рангового коэффициента корреляции Спирмена те же, что и для обычного коэффициента парной корреляции (см. раздел 3).

Проверка статистической значимости рангового коэффициента корреляции Спирмена выполняется точно так же, как и для обычного коэффициента парной корреляции (см. раздел 3). Для этого вычисляется t-критерий Стьюдента:

$$t = \frac{R_s\sqrt{n-2}}{\sqrt{1-R_s^2}}.$$

В данном примере $t = 2{,}68$.

По абсолютному значению t-критерия и числу степеней свободы n–2 из таблиц распределения Стьюдента определяют расчетный уровень значимости P (например, в Excel с помощью функции СТЬЮДЕНТ.ОБР.2Х). Получим P=0,044. Так как $P < 0{,}05$, ранговый коэффициент корреляции Спирмена статистически значим, т.е. есть значимая связь исследуемых величин (работы в условиях шума и заболеваемости).

Решение задачи в программе Statistica.

1 Присвоить переменным имена **Shum** и **Zab**. Для ввода исходных данных потребуется семь строк и два столбца.

2 В столбце переменной **Shum** ввести *ранги*. В столбце **Zab** ввести *значения* заболеваемости.

Примечание – Вводить ранги для переменной **Shum** потребовалось только потому, что время работы в условиях шума представлено в нечисловой форме. Если бы обе анализируемые величины были числовыми, то потребовалось бы ввести только их значения, а не ранги.

Shum	Zab
1	3
2	6
3	6
4	5
5	9
6	8
7	8

3 Выбрать Statistics – Nonparametrics – Correlations (Spearman, Kendall tau, gamma).

4 В появившемся окне Nonparametric correlations установить Compute – Detailed report (чтобы получить подробные результаты).

5 Нажать кнопку Variables. Указать переменные для анализа: в одном из списков выбрать переменную **Shum**, в другом – **Zab**. Нажать OK. Выполняется возврат в окно Nonparametric correlations.

6 Чтобы вычислить ранговый коэффициент корреляции Спирмена, нажать кнопку Spearman rank R. На экран выводится окно результатов.

	Spearman Rank Order Correlations (Stat_непарам) MD pairwise deleted Marked correlations are significant at p <,05000			
Pair of Variables	Valid N	Spearman R	t(N-2)	p-value
Shum & Zab	7	0,763763	2,645751	0,045659

Примечание – В Statistica используется более сложный (и более точный) метод вычисления рангового коэффициента корреляции Спирмена, чем рассмотренный выше. Поэтому возможны некоторые отличия от ручного расчета.

Пример 6.4 – Исследуется зависимость между возрастом обучаемого и результатом некоторого теста (по 10-балльной шкале). Имеются данные для 14 человек:

Возраст	28	47	32	47	24	51	32	49	56	31	27	32	48	52
Результат теста	9	7	9	6	9	5	6	6	5	10	8	8	6	6

Выполнить анализ зависимости между возрастом и результатами теста.

В этой задаче величина «результат теста» не соответствует гауссовскому распределению, так как представляет собой оценки по 10-балльной шкале. Поэтому следует использовать непараметрические методы анализа данных.

Решение в Statistica:

	Vozrast	Result
1	28	9
2	47	7
3	32	9
4	47	6
5	24	9
6	51	5
7	32	6
8	49	6
9	56	5
10	31	10
11	27	8
12	32	8
13	48	6
14	52	6

	Spearman Rank Order Correlations (Spreadsheet1) MD pairwise deleted Marked correlations are significant at p <,05000			
Pair of Variables	**Valid N**	Spearman R	t(N-2)	p-value
Vozrast & Result	14	-0,846658	-5,51156	0,000134

Коэффициент корреляции Спирмена статистически значим, т.е. есть связь (обратная) между возрастом и результатами теста.

6.5 Сравнение нескольких независимых выборок

Ранее был рассмотрен параметрический метод оценки различия между несколькими независимыми выборками (метод однофакторного дисперсионного анализа). Его непараметрические аналоги – критерий Краскала-Уоллиса, медианный критерий.

Пример 6.5 – Известны средние зарплаты на 16 предприятиях трех отраслей:

– отрасль П1: 640, 637, 648, 624, 639;
– отрасль П2: 638, 625, 662, 645, 640, 637;
– отрасль П3: 625, 617, 631, 629, 619.

Это означает, например, что имеются данные по пяти предприятиям отрасли П1. Средняя зарплата на этих предприятиях составляет 640, 637, 648, 624 и 639 ден. ед.

Требуется: а) определить, значимо ли различаются зарплаты в зависимости от отрасли (или, другими словами, является ли отрасль фактором, значимо влияющим на зарплату); б) если окажется, что зарплаты значимо различаются в зависимости от отрасли, определить, какие именно отрасли значимо различаются по зарплате.

В данной задаче анализируются три выборки (группы), т.е. три отрасли: m=3. Объемы выборок (групп) составляют n_1=5, n_2=6, n_3=5. Общий объем выборки составляет n=16.

1 Перечислить все значения из всех групп по возрастанию. Присвоить значениям ранги: ранг 1 назначается первому значению (т.е. минимальному из всех групп), ранг 2 -–следующему значению, и т.д. Если какие-либо значения равны, им присваивается одинаковый (усредненный) ранг, аналогично рассмотренному выше.

Отрасль	П3	П3	П1	П2	П3	П3	П3	П1	П2	П2	П1	П1	П2	П2	П1	П2
Зарплата	617	619	624	625	625	629	631	637	637	638	639	640	640	645	648	662
Ранг	1	2	3	4,5	4,5	6	7	8,5	8,5	10	11	12.5	12.5	14	15	16

2 Для каждой группы вычислить сумму рангов:

R_1 = 3+8,5+11+12.5+15 = 50
R_2 = 4,5+8,5+10+12,5+14+16 = 65,5
R_3 = 1+2+4,5+6+7 = 20,5

3 Вычислить критерий Краскала-Уоллиса (H-критерий):

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{m} \frac{R_i^2}{n_i} - 3(n+1).$$

Для данного примера H=6,31.

4 Для оценки различия между группами воспользоваться одним из следующих способов:

а) по таблицам критерия Краскала-Уоллиса найти критическое значение H_{crit}. Если $H > H_{\text{crit}}$, это означает, что группы различаются значимо. В данном примере для $n_1 = 5$, $n_2 = 6$, $n_3 = 5$, $\alpha = 0,05$ критическое значение $H_{\text{crit}} = 5,729$. Так как $H > H_{\text{crit}}$, группы (т.е. отрасли, точнее – зарплаты по отраслям) различаются значимо;

б) по таблицам распределения «хи-квадрат» для найденного значения H и числа степеней свободы $s = m$-1 найти уровень значимости P. Если $P < \alpha$, это означает, что группы различаются значимо. В данном примере H=6,31, s = 3-1=2. Чтобы найти значение P в MS Excel, воспользоваться функцией **ХИ2.РАСП.ПХ** с параметрами **X**: H; **Степени свободы**: s. В данном случае уровень значимости P = 0,043. Таким образом, $P < \alpha$ (считая α=0,05). Это означает, что группы различаются значимо. Таким образом, отрасль – фактор, значимо влияющий на зарплату, так как зарплаты в отраслях различаются значимо.

Решение задачи в Statistica

1 Присвоить переменным имена, например, **Industry** и **Salary**. Ввести данные. Для ввода данных потребуется 16 строк.

1 Industry	2 Salary
I1	640
I1	637
I1	648
I1	624
I1	639
I2	638
I2	625
I2	662
I2	645
I2	640

2 Выбрать **Statistics – Nonparametrics – Comparing multiple independent samples (groups).**

3 В окне **Kruskal-Wallis ANOVA and Median Test** нажать кнопку **Variables**, чтобы выбрать переменные для анализа. В списке **Dependent variable list** выбрать **Salary**; в списке **Indep. (grouping) variable** выбрать **Industry**. Нажать **OK**.

4 В окне **Kruskal-Wallis ANOVA and Median Test** нажать **Summary: Kruskal-Wallis ANOVA & Median Test**. Вычисляются два критерия – Краскала-Уоллиса и медианный критерий (не рассматривается). Соответственно, открываются два окна. Перейти в окно критерия Краскала-Уоллиса. Отображается уровень значимости (**p**), а также суммы рангов, H-критерий (см. описание алгоритма выше) и еще несколько величин.

	Kruskal-Wallis ANOVA by Ranks; Salary (Lab3_ANOVA) Independent (grouping) variable: Industry Kruskal-Wallis test: H (2, N= 16) =6,340842 p =,0420			
Depend.: Salary	Code	Valid N	Sum of Ranks	Mean Rank
I1	101	5	50,00000	10,00000
I2	102	6	65,50000	10,91667
I3	103	5	20,50000	4,10000

5 Чтобы выявить группы, различающиеся значимо, вернуться в окно **Kruskal-Wallis ANOVA and Median Test** и нажать кнопку **Multiple compari-**

sons of mean ranks for all groups. Вычисляются уровни значимости для попарной оценки значимости различия между группами.

Из полученных результатов видно, что между отраслями I1 и I2 различия практически нет (**p**=1). Различие между I1 и I3 незначимо (**p**=0,15). Только различие между отраслями I2 и I3 имеет тенденцию к значимости (**p**=0,054).

	Multiple Comparisons p values (2-tailed); Salary (Lab3_ANOVA) Independent (grouping) variable: Industry Kruskal-Wallis test: H (2, N= 16) =6,340842 p =,0420		
Depend.: Salary	I1 R:10,000	I2 R:10,917	I3 R:4,1000
I1		1,000000	0,150190
I2	1,000000		0,054161
I3	0,150190	0,054161	

6 Чтобы проиллюстрировать различие между группами (отраслями), построить диаграмму размаха. Для этого снова перейти в окно **Kruskal-Wallis ANOVA** и нажать кнопку **Box & whisker**. В следующем окне выбрать переменную **Salary** и нажать **OK**. В очередном окне выбрать тип диаграммы **Median/Quart/Range** (это означает, что средняя точка на диаграмме будет соответствовать медиане, а в качестве показателей размаха используются квартили и границы выборки). Нажать **OK**. Строится диаграмма размаха. Из нее видно, что зарплата в отрасли I3 значимо *ниже*, чем в I1 и I2 (максимальное различие – между I2 и I3).