

Математические методы обработки информации

Меркулова Ольга Петровна

доцент кафедры психологии образования и развития

Olga.Merkulova@list.ru
Merkulova-OP@yandex.ru

Тема 2. Описательная статистика

Литература

- Наследов А.Д. Математические методы психологического исследования. Анализ и интерпретация данных. Учебное пособие. СПб.: Речь, 2007. Глава 3 и 4.
- Тюменева Ю.А. Психологическое измерение [Электронный ресурс]: учебное пособие/ Тюменева Ю.А.— Электрон. текстовые данные.— М.: Аспект Пресс, 2007.— 192 с.— Режим доступа: <http://www.iprbookshop.ru/8884>.— ЭБС «IPRbooks», по паролю. Главы 3, 4.
- Борытко Н. М., Моложавенко А.В., Соловцова И.А. Методология и методы психолого-педагогических исследований – М.: Изд. центр "Академия", 2008. С. 250–268.
- Сидоренко Е.В. Методы математической обработки в психологии. СПб.: Речь, 2010. Глава 6.
- Некрасов С.Д. Математические методы в психологии (MS Excel): учеб. пособие. 3-е изд., испр. и доп. Краснодар: Кубанский гос. ун-т, 2014. 147 с. URL: <http://docspace.kubsu.ru/docspace/handle/1/295> .
- Гласс Дж., Стенли Дж. Статистические методы в педагогике и психологии /Пер. с англ. С общ. Ред. Ю.П. Адлера. М.: Прогресс, 1976.

Табличное представление данных исследования

- **Строки таблицы** соответствуют **объектам измерения** (индивид, группа, ситуация)
- Конкретный объект измерения, данные которого представляются в одной строке - **случай**

- **Столбцы таблицы** соответствуют **измеряемым переменным**.
- В каждом столбце записываются данные одной измеренной переменной.

Кодирование случаев	Переменные, относящиеся к первой методике	Переменные, относящиеся ко второй методике	...
---------------------	---	--	-----

№ исп.	Пер_A1	Пер_A2	Пер_A3	Пер_B1	Пер_B2	Пер_B3	Пер_B4	...
1	5	15	2	7	6	7	9	
2	7	26	1	9	8	7	6	
3	9	14	1	7	5	8	6	
4	5	12	3	8	9	9	7	
5	4	23	3	6	7	8	6	
...								

Методы описательной статистики

позволяют представлять количественные данные в обобщенном виде, что необходимо для их анализа:

- выявления типичных тенденций,
- формулировки гипотез о различиях и взаимосвязях

Повторя-
емость

- Группировка
- Подсчет частот

Нагляд-
ность

- Построение графиков

Сравни-
мость

- Расчет параметров распределения

Табличное и графическое представление распределения признака

- **Признак** – отдельное свойство, измеренное определенным способом. Для признака должна быть известна шкала, в которой он измерен.
- **Выборка** – совокупность объектов, включенных в исследование. Для каждого объекта выборки должны быть известны значения всех измеренных признаков.
- **Группировка** – разбиение всего множества возможных значений признака на интервалы (обычно равные) и отнесение каждого значения к одному из интервалов. Дальнейшая обработка идет не по значениям, а по интервалам. Каждый интервал задается верхней и нижней границей.

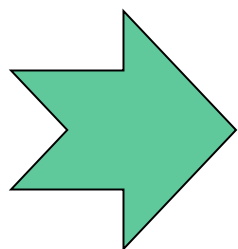
- **Частота** значения или интервала группировки (n_i , где i – номер значения или интервала) – число объектов, имеющих данное значение или отнесенных к данному интервалу группировки.
- **Объем выборки** (N) – общее число объектов в выборке.
- **Относительная частота** (частость) – отношение частоты к объему выборки. Может выражаться в долях от единицы (n_i / N) или процентах. В последнем случае полученная доля домножается на 100 %.
- **Накопленная частота** – число объектов в выборке, значения которых не превосходят данное (при группировке – верхнюю границу данного интервала). Имеет смысл для данных, начиная с порядковых. Накопленная частота минимального значения совпадает с его частотой, максимального – с объемом выборки.
- **Накопленная относительная частота** – отношение накопленной частоты соответствующего значения (интервала) к объему выборки. **Накопленный процент** – накопленная относительная частота * 100%.

Распределение признака

показывает как связаны значения признака с их повторяемостью на выборке

Пример

- 5, 7, 4, 5,
10, 3, 5, 6,
7, 8, 5, 3,
4, 6, 8, 4,
6, 6, 7, 9.



			x_i	n_i	n_i'
3	1	1	3	2	2
3	2	2			
4	1	3	4	3	5
4	2	4			
4	3	5			
5	1	6	5	4	9
5	2	7			
5	3	8			
5	4	9			
6	1	10	6	4	13
6	2	11			
6	3	12			
6	4	13			
7	1	14	7	3	16
7	2	15			
7	3	16			
8	1	17	8	2	18
8	2	18			
9	1	19	9	1	19
10	1	20	10	1	20

			x_i	n_i	n_i'
3	1	1	3	2	2
3	2	2			
4	1	3	4	3	5
4	2	4			
4	3	5			
5	1	6	5	4	9
5	2	7			
5	3	8			
5	4	9			
6	1	10	6	4	13
6	2	11			
6	3	12			
6	4	13			
7	1	14	7	3	16
7	2	15			
7	3	16			
8	1	17	8	2	18
8	2	18			
9	1	19	9	1	19
10	1	20	10	1	20

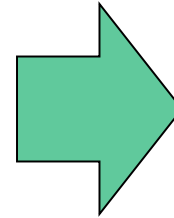


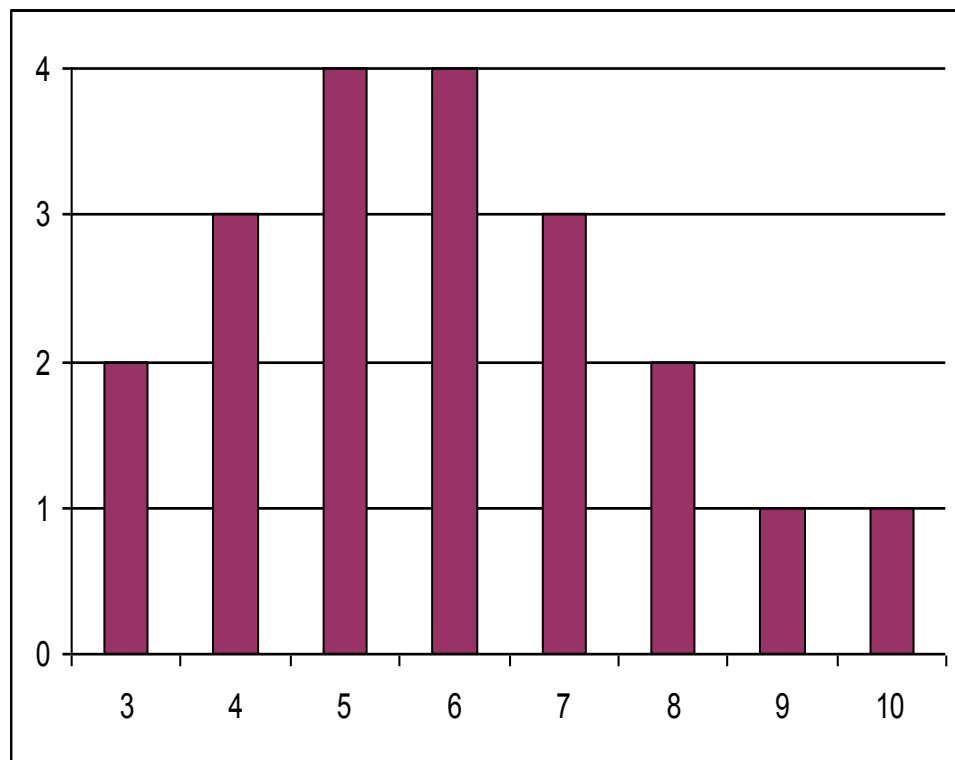
Таблица
распределения
признака

Значения x_i	Частота n_i	Накопл. частота n_i'
3	2	2
4	3	5
5	4	9
6	4	13
7	3	16
8	2	18
9	1	19
10	1	20

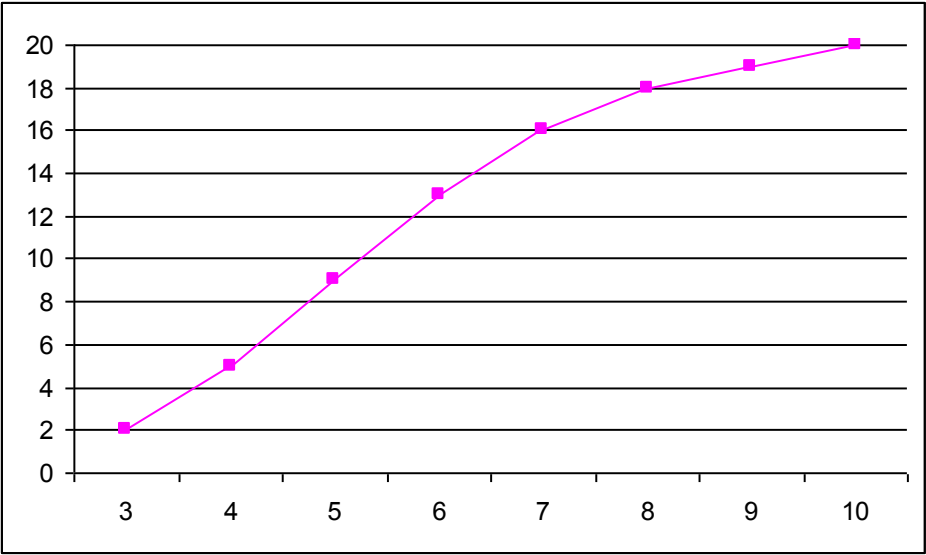
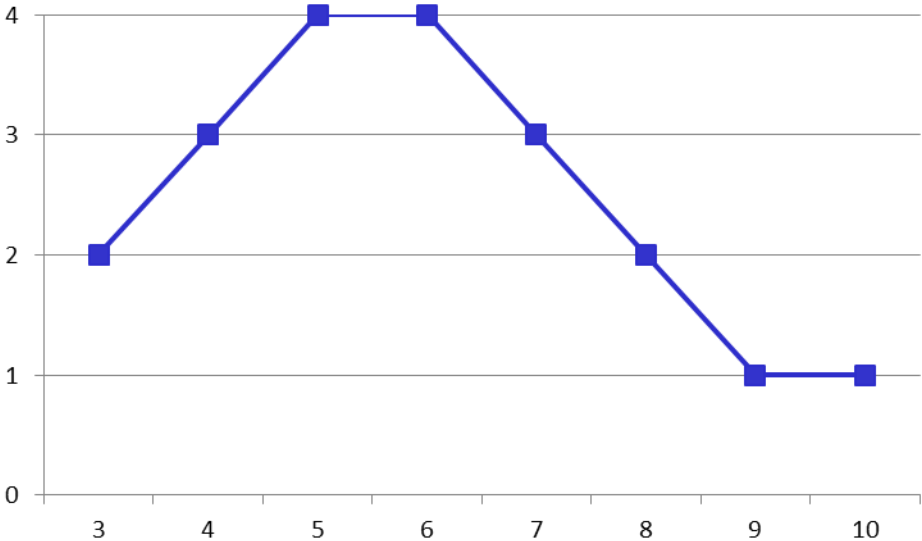
- Эмпирическое распределение признака (вариационный ряд)**
 – двойной числовой ряд, показывающий, каким образом численные значения изучаемого признака связаны с их повторяемостью в выборке. Состоит из упорядоченных по возрастанию неповторяющихся значений признака (или интервалов группировки) и сопоставленных им частот. Дополнительные характеристики, используемые в таблицах распределений: относительные и накопленные частоты.

Значения x_i	Частота n_i	Накопл. частота n_i'	Относит. частота	Накопл. относит. частота
3	2	2	0,10	0,10
4	3	5	0,15	0,25
5	4	9	0,20	0,45
6	4	13	0,20	0,65
7	3	16	0,15	0,80
8	2	18	0,10	0,90
9	1	19	0,05	0,95
10	1	20	0,05	1,00
Сумма	20		1	

- **Столбчатая диаграмма** – графическое представление распределения, в котором на горизонтальной оси отложены отдельные значения признака, на вертикальной – абсолютные или относительные частоты. Частота каждого значения показана столбиком, основание которого соответствует значению, а высота – частоте. Используется для дискретных признаков, между столбиками, соответствующими отдельным значениям, остается зазор.



- **Полигон частот** – графическое представление распределения, в котором на горизонтальной оси отложены отдельные значения признака, на вертикальной – абсолютные или относительные частоты. Частота каждого значения показана точкой, координаты которой определяются парой чисел (значение или середина интервала группировки; частота). Точки соединены ломанной линией.
- **Полигон накопленных частот** – графическое представление распределения. Строится аналогично *полигону частот*, но по значениям накопленных частот. Представляется из себя неубывающую ломанную линию.



Параметр распределения

число, которое
характеризует
какое-либо
свойства
распределения в
целом



Меры средней (центральной) тенденции

Показывают наиболее общую, типичную с какой-либо точки зрения, т.е. среднюю или центральную тенденцию, характеризующую всю выборку в целом.

- **Мода (M_o)** – наиболее часто встречающееся значение признака, т.е. значение (или интервал группировки), частота которого максимальна.
 - Если наибольшая частота характерна для двух смежных значений, модой считается среднее арифметическое этих значений.
 - Если наибольшая частота характерна для двух несмежных значений, модами считаются оба эти значения, а распределение – бимодальным.
 - Если наибольшая частота характерна для трех и более несмежных значений, принято считать, что распределение является полимодальным или не имеет моды. В статистических пакетах для бимодального или полимодального распределения может быть показано какое-то одно (например, наименьшее) значение моды.

			x_i	n_i	n_i'
3	1	1	3	2	2
3	2	2			
4	1	3	4	3	5
4	2	4			
4	3	5			
5	1	6	5	4	9
5	2	7			
5	3	8			
5	4	9			
6	1	10	6	4	13
6	2	11			
6	3	12			
6	4	13			
7	1	14	7	3	16
7	2	15			
7	3	16			
8	1	17	8	2	18
8	2	18			
9	1	19	9	1	19
10	1	20	10	1	20

Пример расчета моды и медианы

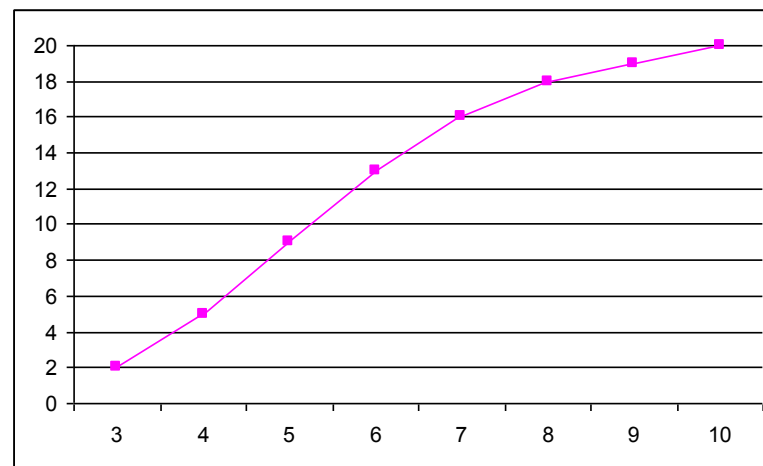
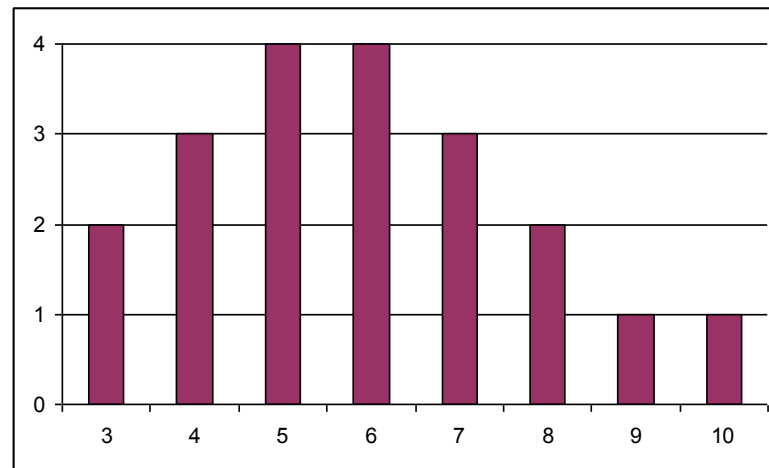
Мода. $M_o = 5,5$.

Медиана.

$N = 20$ (четное).

Медианные значения
– на 10 и 11
позициях.

$Me = 6$



Меры средней (центральной) тенденции

Медиана (Me , Md) – такое значение в ряду данных, что половина всех значений больше него, а половина – меньше; значение, которое делит ряд данных пополам. $Md = Q_2 = D_5 = P_{50}$. Имеет смысл для данных, измеренных в шкале, начиная с порядковой.

- Определение медианы по несгруппированным данным:
 - Упорядочить значения выборки по возрастанию – переписать их, начиная с наименьшего. Пронумеровать полученный ряд – наименьшему значению присвоить номер 1, следующему – 2, и т.д. Последнее значение в ряду (являющееся максимальным), должно получить номер, равный *объему выборки* (N).
 - Если *объем выборки* является нечетным числом, следует найти значение, которое располагается в ряду данных под номером $(N+1)/2$. Это значение является медианой.
 - Если *объем выборки* является четным числом, следует найти два значения, которые располагаются в ряду данных под номерами $N/2$ и $(N/2)+1$. Медианой является среднее арифметическое этих значений (рассчитывать среднее надо только в том случае, если найденные значения не совпадают).
- Для нахождения медианы по сгруппированным данным следует воспользоваться значениями *накопленных частот* таблицы *распределения*.

Меры средней (центральной) тенденции

- **Среднее арифметическое** (\bar{x} , M_x) – значение, которое обладает таким свойством, что сумма отклонений всех значений ряда от него равна 0; сумма всех значений, деленная их количество (т.е. на *объем выборки*).

Нахождение для несгруппированных данных:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

- где N – объем выборки, x_i – «пробегают» все значения в выборке.
- Если уже построена таблица *распределения*, для нахождения среднего арифметического следует предварительно найти произведения каждого неповторяющегося значения на соответствующую ему *частоту*, затем сложить эти произведения – получим сумму всех значений исходного ряда данных:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^k n_i x_i$$

- где N – объем выборки, k – число разрядов, т.е. неповторяющихся значений в таблице распределения, x_i – «пробегают» все неповторяющиеся значения, n_i – соответствующее каждому x_i значение *частоты*.

Пример расчета среднего значения по сгруппированным данным

x_i	n_i	$x_i n_i$
(1)	(2)	(3)
3	2	6
4	3	12
5	4	20
6	4	24
7	3	21
8	2	16
9	1	9
10	1	10
Σ	20	118

$$\bar{x} = \frac{1}{N} \sum_{i=1}^k n_i x_i = \frac{1}{20} 118 = \frac{118}{20} = 5,9$$

Какая мера средней тенденции лучше?

- Пример 1. Доходы бедняка - 2 т.р.
представителя «среднего класса» – 30 т.р.
олигарха – 500 т.р.
- Усреднять не имеет смысла!
 - среднее = 178,3 т.р.
 - медиана = 30 т.р.
 - моды нет
- Пример 2. Доходы сотрудников организации (т.р.)
- 5, 10, 10, 10, 15, 15, 20, 20, 300
 - среднее = 45
 - медиана = 15
 - мода = 10
-

Недостаточность использования только мер средней тенденции

- Пример. В трех группах учащихся получены оценки:
 - 3 3 3 3 3 3 $M = 3, M_o = 3, M_d = 3$
 - 1 2 3 3 4 5 $M = 3, M_o = 3, M_d = 3$
 - 1 1 1 5 5 5 $M = 3, M_o \text{ нет}, M_d = 3$

Меры варитивности

- Показывают разброс, т.е. вариативность, изменчивость значений в выборке. В сочетании с мерами средней тенденции являются достаточно информативными для описания распределения в целом.
- **Размах вариации** (R) – расстояние (разность) между максимальным и минимальным значением в выборке; широта диапазона в котором варьируют значения. Расчет:
 - $R = x_{max} - x_{min}$,
 - где x_{max} и x_{min} – соответственно максимальное и минимальное значения в выборке.

Меры варитивности

- **Дисперсия** (S^2) – средний квадрат отклонений от среднего арифметического. Одна из основных мер вариативности, используемая во многих методах статистического анализа. Расчет для негруппированных данных:

$$S^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

- где N – объем выборки, \bar{x} – среднее арифметическое, x_i – «пробегают» все значения в выборке.

- Если уже построена таблица распределения, аналогично среднему арифметическому расчет может выполняться по формуле:

$$S^2 = \frac{1}{N} \sum_{i=1}^k (x_i - \bar{x})^2 n_i$$

- где N – объем выборки, \bar{x} – среднее арифметическое, k – число разрядов, т.е. неповторяющихся значений в таблице распределения, x_i – «пробегают» все неповторяющиеся значения, n_i – соответствующее каждому x_i значение частоты.

Меры варитивности

- **Стандартное отклонение** (S или σ) – положительный квадратный корень из дисперсии. В отличие от дисперсии выражается в тех же единицах измерения, что и исходные данные, поэтому чаще используется для оценки широты разброса значений на оси данных. Расчет:

$$S = \sqrt{S^2}$$

– где S^2 – дисперсия.

Пример расчета дисперсии и стандартного отклонения

x_i	n_i	$x_i \cdot n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot n_i$
(1)	(2)	(3)	(4)	(5)	(6)
3	2	6	-2,9	8,41	16,82
4	3	12	-1,9	3,61	10,83
5	4	20	-0,9	0,81	3,24
6	4	24	0,1	0,01	0,04
7	3	21	1,1	1,21	3,63
8	2	16	2,1	4,41	8,82
9	1	9	3,1	9,61	9,61
10	1	10	4,1	16,81	16,81
Σ	20	118			69,8

x_i – значения признака
 n_i – частоты
 \bar{x} – среднее арифметическое значение

$$S^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 \cdot n_i = \frac{1}{20} \cdot 69,8 = 3,49$$

$$S = \sqrt{S^2} = \sqrt{3,49} \approx 1,87$$

Меры вариативности для примера недостаточности мер средней тенденции

- 3 3 3 3 3 3 $R = 0, Q = 0, S = 0$
- 1 2 3 3 4 5 $R = 4, Q = 1, S = 1,41$
- 1 1 1 5 5 5 $R = 4, Q = 2, S = 2,19$

Интерпретация основных параметров нормального распределения

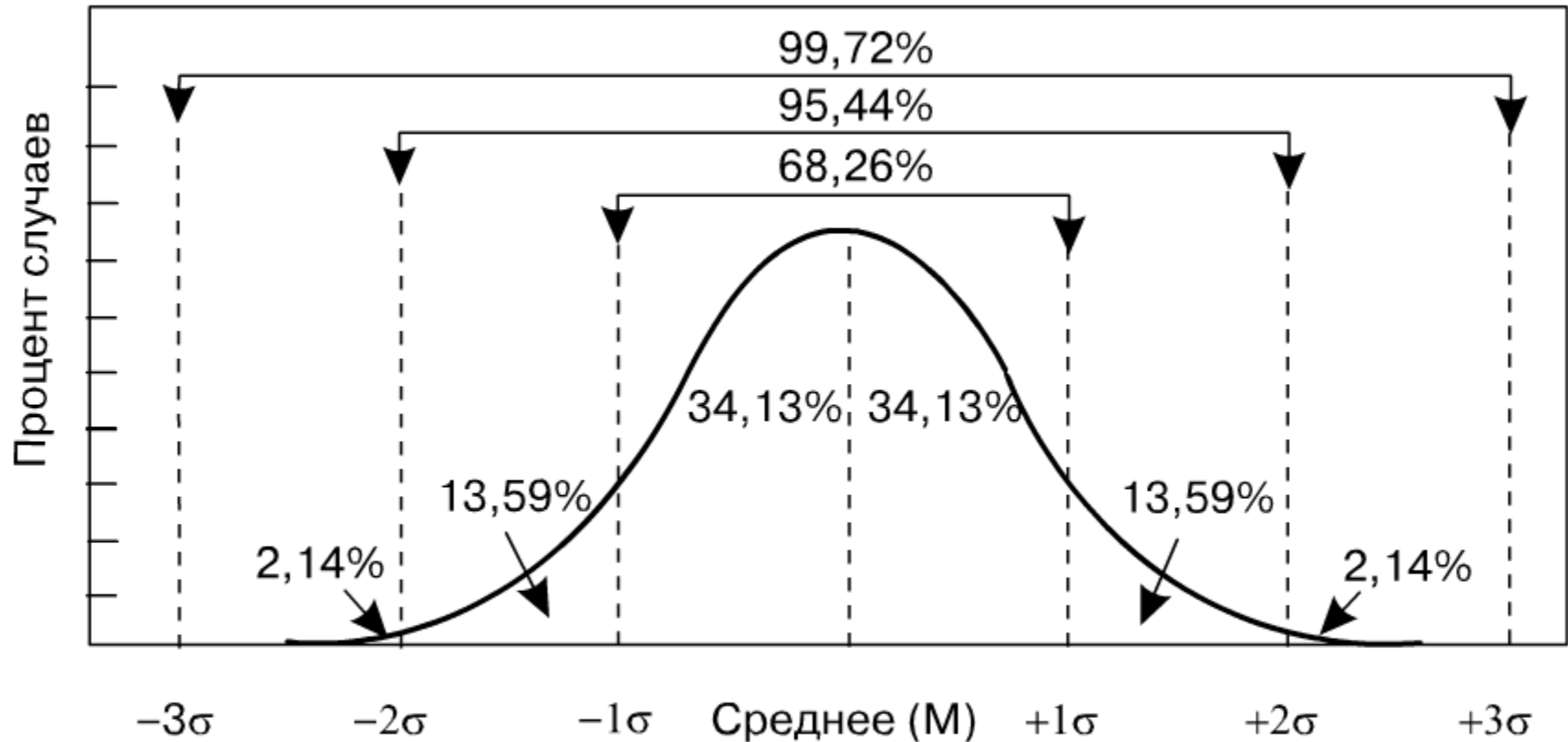
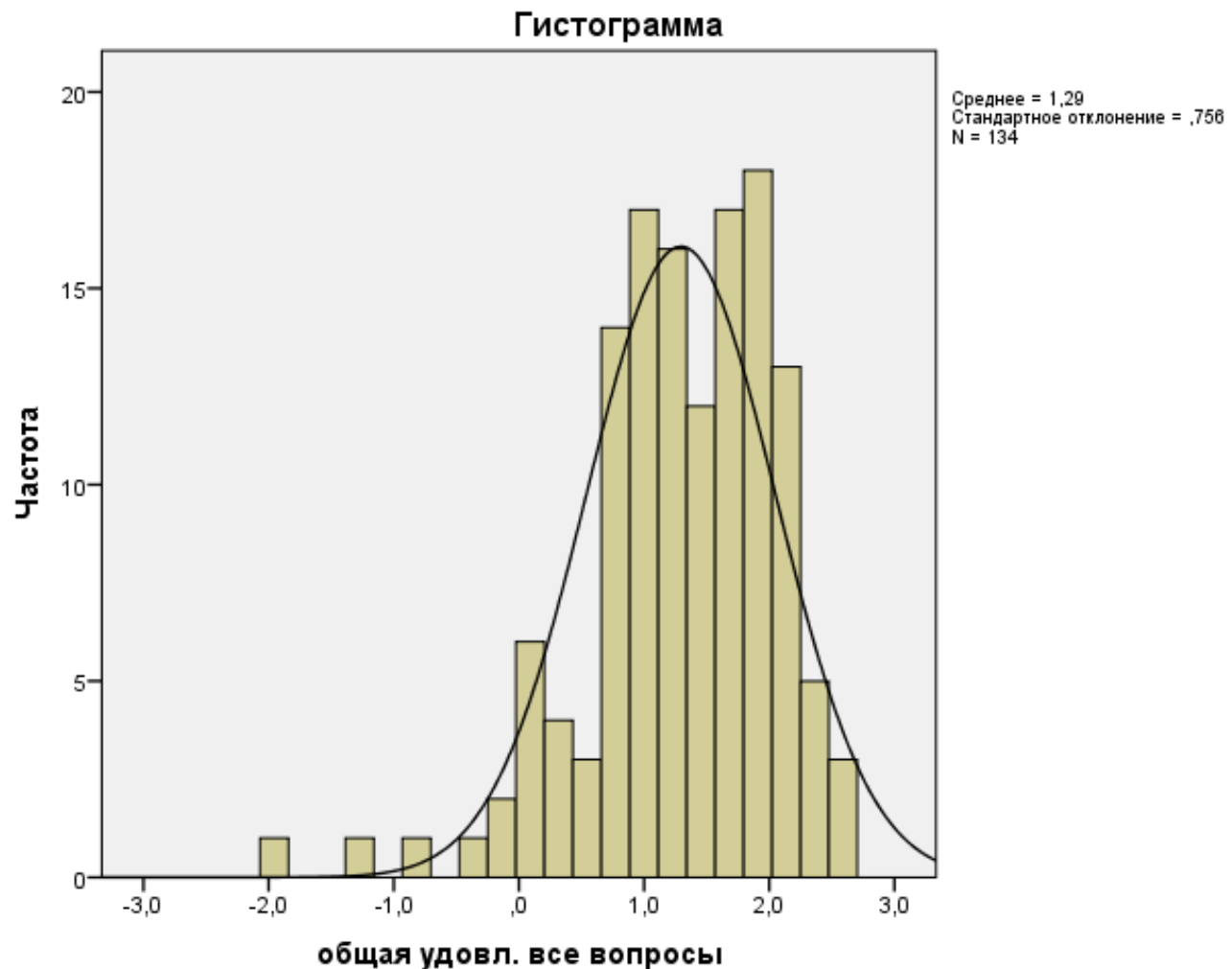


Рис. 17. Процентное распределение случаев под нормальной кривой
(*Источник:* Анастаси, Урбина, 2003)

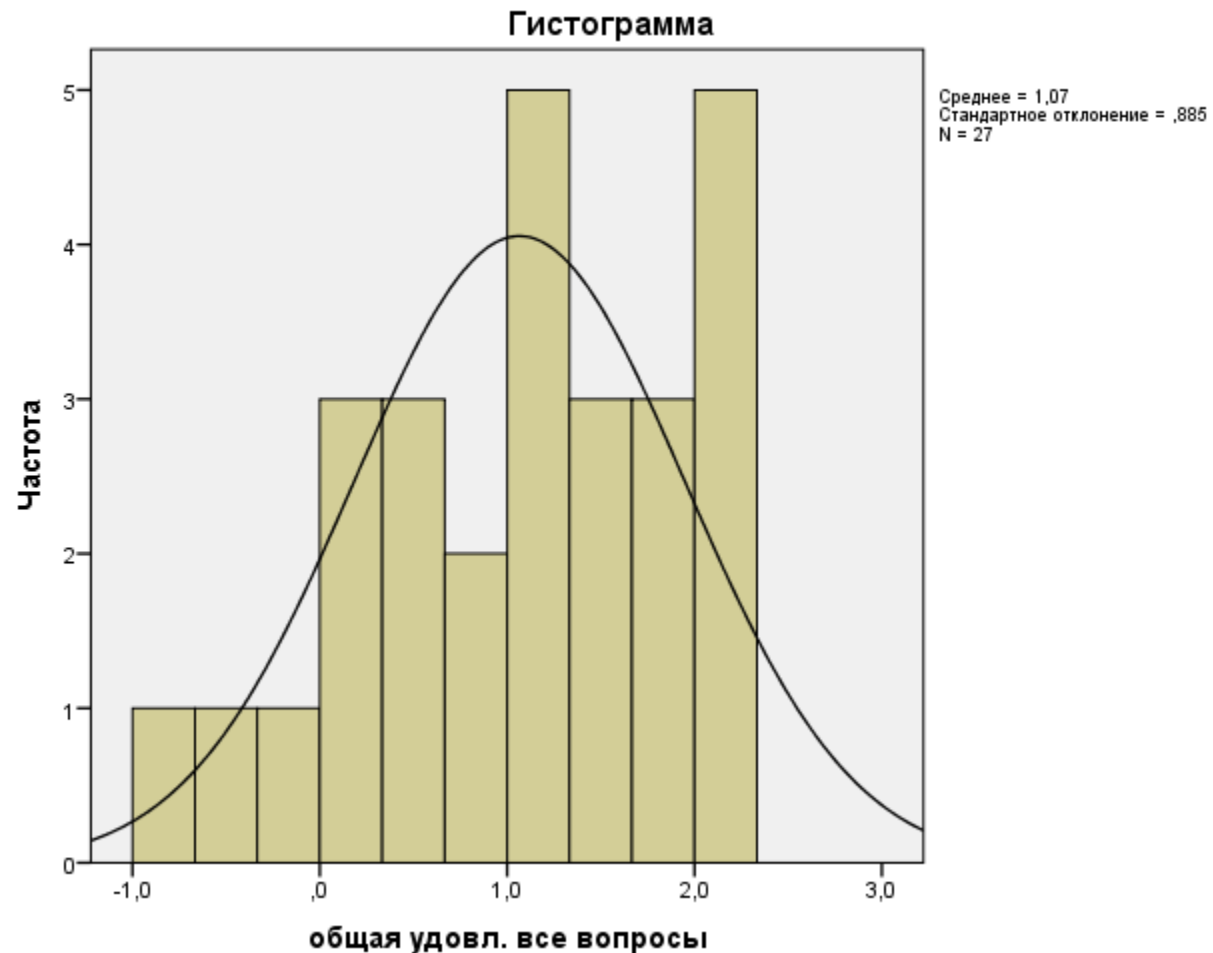
Пример. Обобщающий показатель опросника. Группа 1

Статистики		
общая удовл. все вопросы		
N	Валидные	134
	Пропущенные	0
Среднее		1,294
Медиана		1,364
Мода		1,0 ^a
Стд. отклонение		,7562
Дисперсия		,572
Размах		4,6
Минимум		-2,0
Максимум		2,6
а. Имеется несколько мод. Показана наименьшая.		



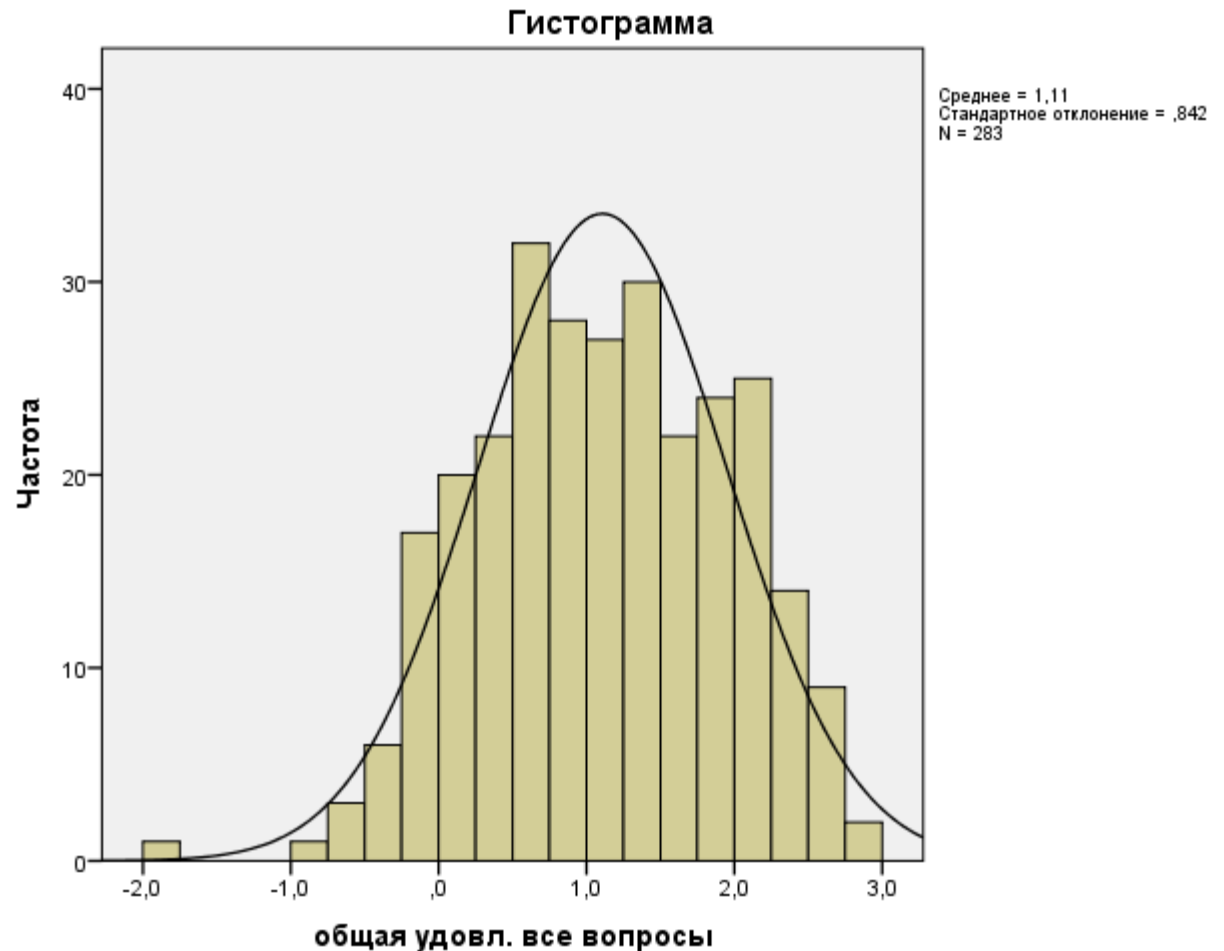
Пример. Обобщающий показатель опросника. Группа 2

Статистики		
общая удовл. все вопросы		
N	Валидные	27
	Пропущенные	0
Среднее		1,066
Медиана		1,182
Мода		,2 ^a
Стд. отклонение		,8854
Дисперсия		,784
Размах		3,2
Минимум		-,9
Максимум		2,3
а. Имеется несколько мод. Показана наименьшая.		



Пример. Обобщающий показатель опросника. Группа 3

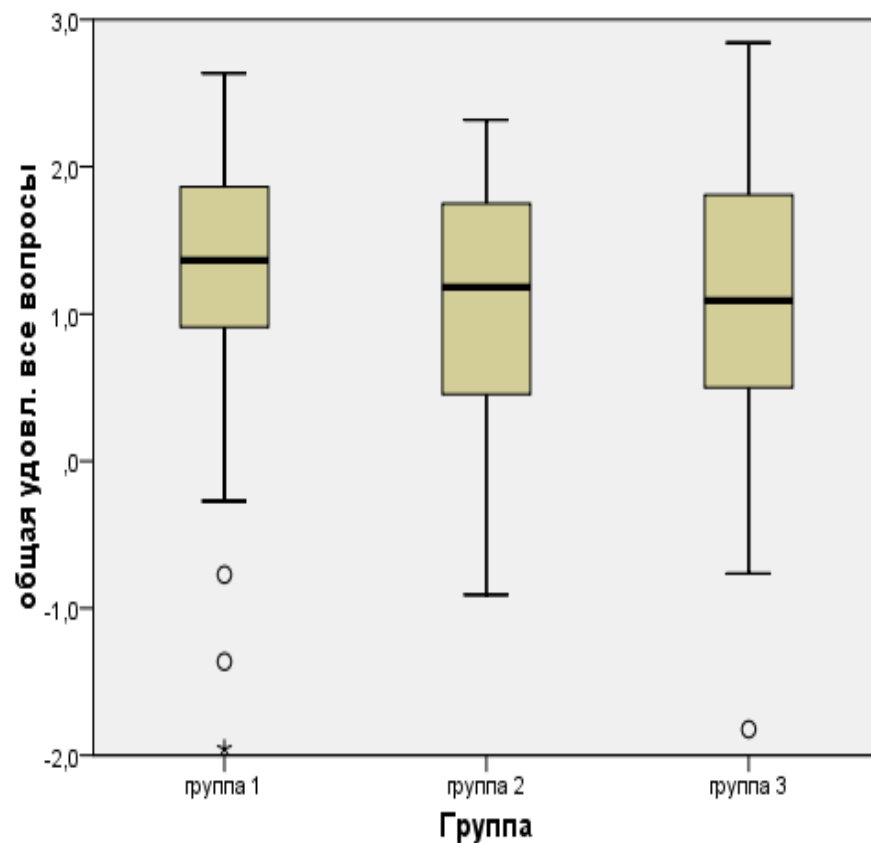
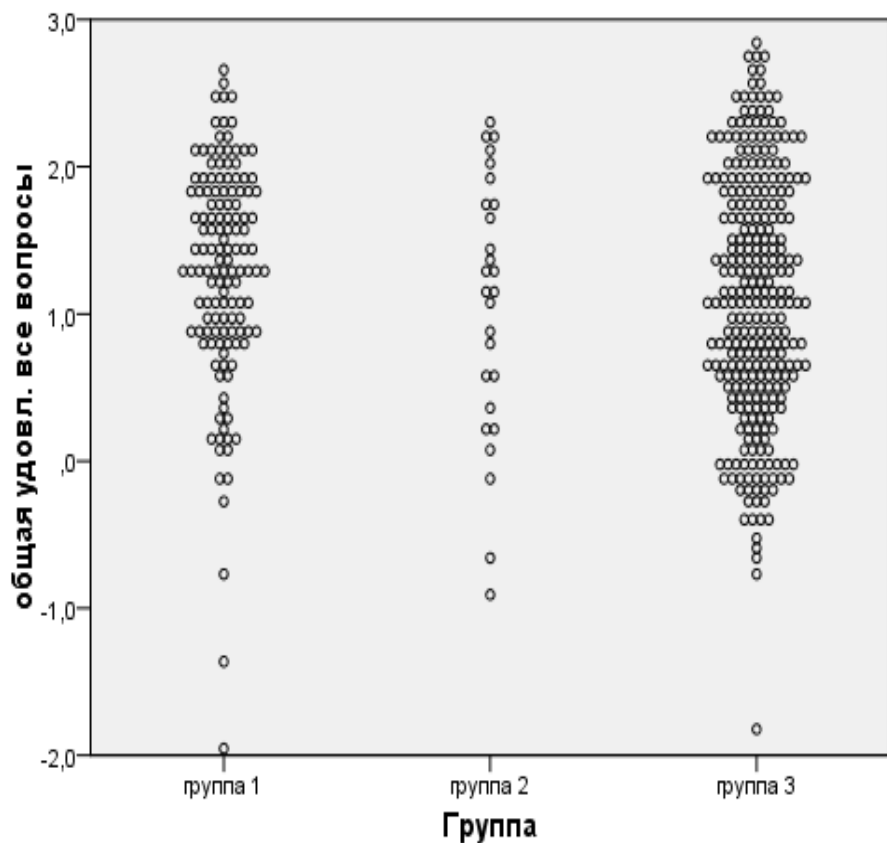
Статистики		
общая удовл. все вопросы		
N	Валидные	283
	Пропущенные	13
Среднее		1,107
Медиана		1,091
Мода		2,0
Стд. отклонение		,8417
Дисперсия		,708
Размах		4,7
Минимум		-1,8
Максимум		2,8



Пример. Сопоставление 3 групп

общая удовл. все вопросы

Группа	Среднее	Стд. Отклонение	Медиана	Минимум	Максимум	N
группа 1	1,294	,7562	1,364	-2,0	2,6	134
группа 2	1,066	,8854	1,182	-,9	2,3	27
группа 3	1,107	,8417	1,091	-1,8	2,8	283
Итого	1,161	,8225	1,227	-2,0	2,8	444



Двумерное распределение

- **Двумерное распределение** – совместное распределение двух признаков, т.е. частота встречаемости в выборке всех возможных пар значений первого и второго признака. Строится на основе данных об измерении двух признаков на одной выборке объектов.
- **Диаграмма рассеяния** – графическое представление двумерного распределения. По осям откладываются значения первого и второго признаков. Каждый объект в выборке обозначается точкой, координаты которой определяются значениями первого и второго признаков, полученными для этого объекта.
 - Диаграмма рассеяния является удобной формой представления двумерного распределения, если значения обоих признаков имеют низкую повторяемость. В противном случае несколько объектов оказываются изображенными одной точкой.
- **Таблица сопряженности** – таблица, в которой заголовки строк образуют значения одного признака, заголовки столбцов – значения второго признака, в ячейках указаны частоты встречаемости в выборке каждой пары значений.
 - Таблица сопряженности является удобной формой представления двумерного распределения, если значения как первого, так и второго признака имеют достаточно высокую повторяемость.
- **Задачи** анализа двумерных распределений связаны с изучением **взаимосвязей** между признаками

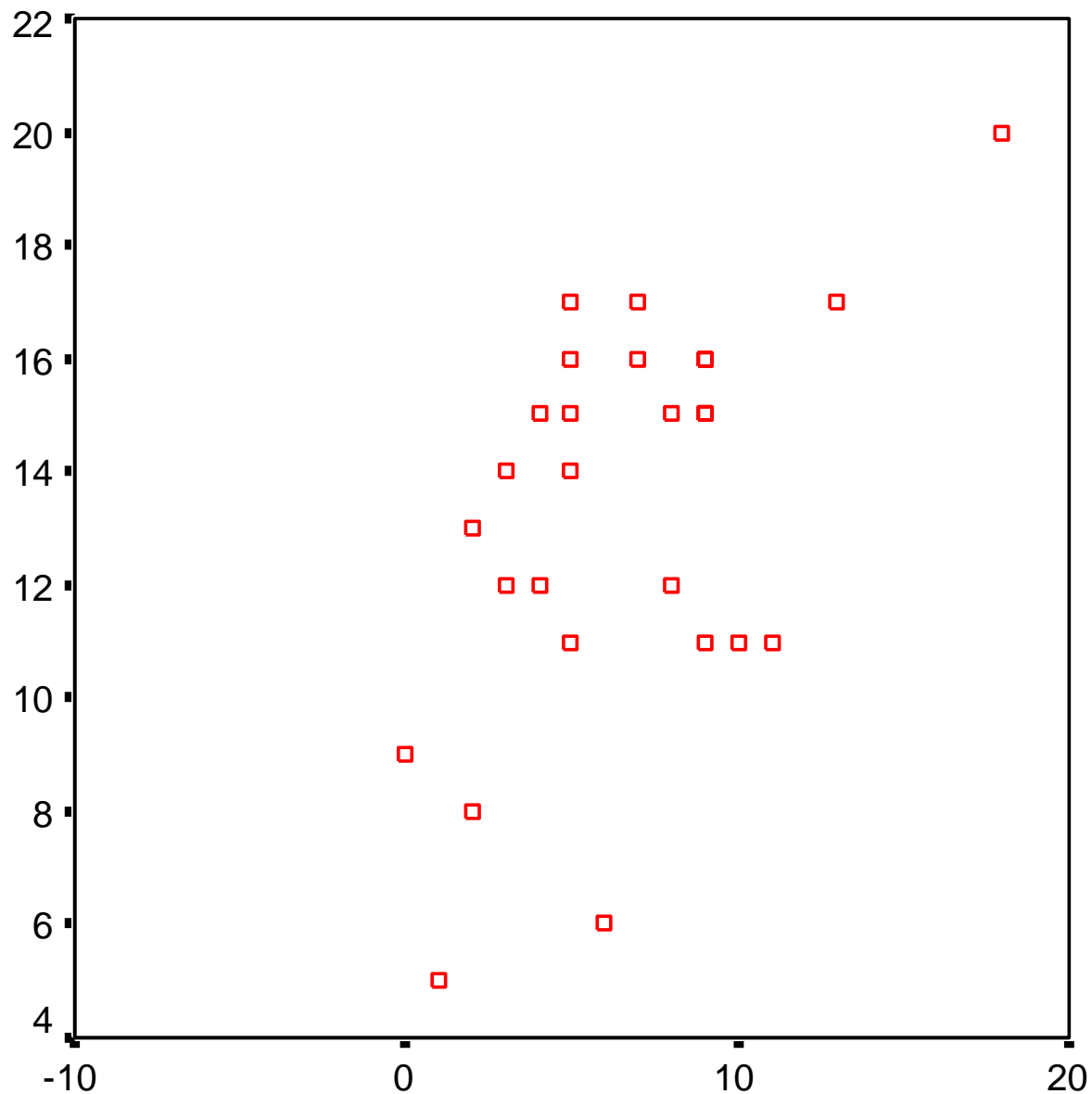
Пример таблицы двумерного распределения

Высшее образование является хорошим, если оно... * СТАТУС

Count

		СТАТУС		Total
		родитель студента	студент	
Высшее образование является хорошим, если оно...	является интересным, увлекательным для студента	2	7	9
	дает хорошую теоретическую базу	3	3	6
	позволяет устроиться на хорошо оплачиваемую работу	11	12	23
	практические навыки для будущей профессии	25	22	47
	получено в государственном вузе	7	3	10
	преподается на высоком уровне квалификации	8	9	17
	Total	56	56	112

Пример графика
двумерного
распределения



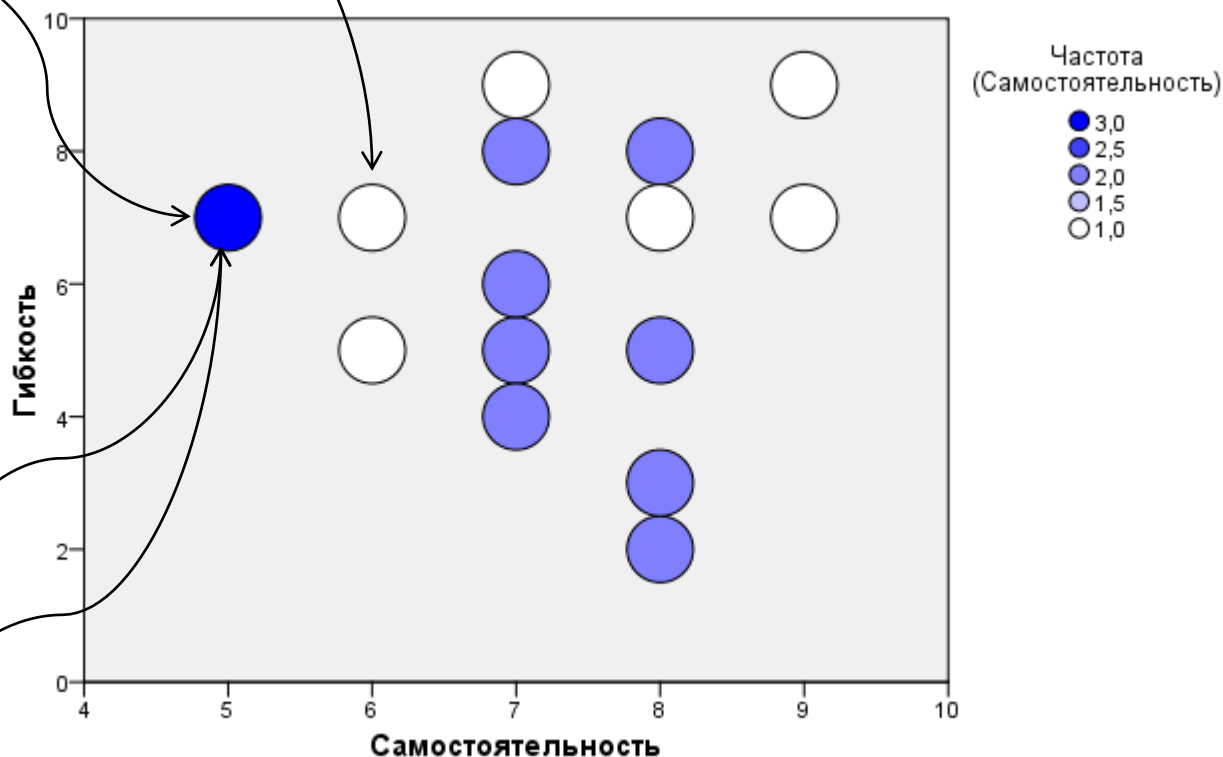
контрольная работа 2

Пример двумерного распределения Таблица исходных данных и диаграмма рассеяния

Опросник стилевой саморегуляции поведения (Моросанова).
Шкалы гибкости и самостоятельности

$N=25, r = -0,151$

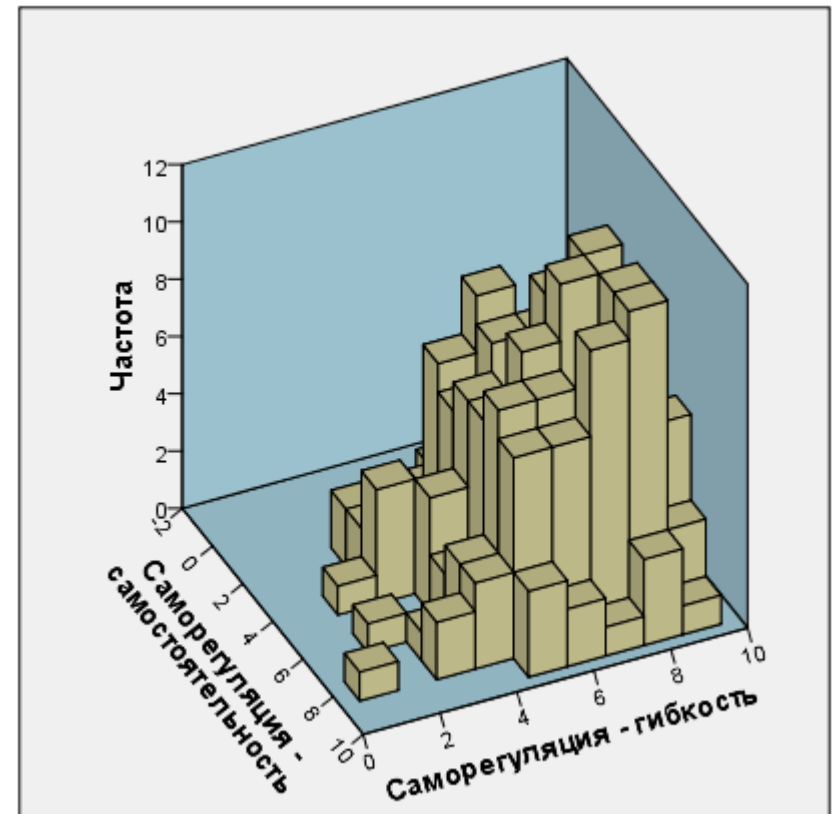
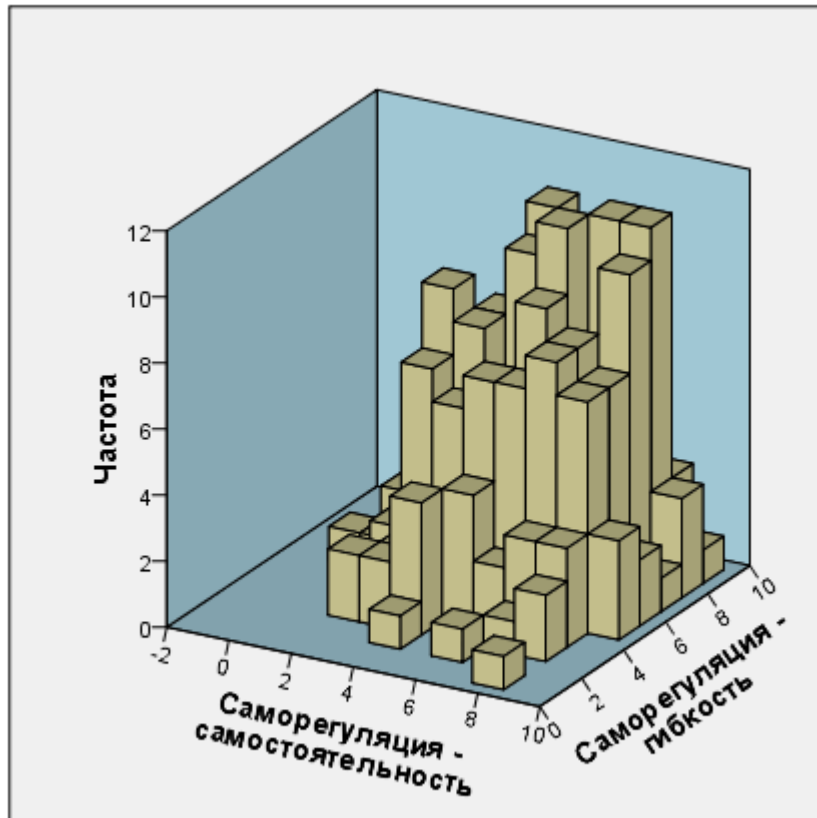
№ исп.	Самостоя- тельность	Гибкость
1	6	7
2	8	2
3	8	3
4	8	2
5	8	5
6	7	5
7	7	6
8	5	7
9	7	4
10	6	5
11	9	9
12	7	8
13	7	6
14	7	9
15	7	8
16	8	5
17	8	3
18	8	8
19	8	7
20	9	7
21	7	4
22	8	8
23	5	7
24	7	5
25	5	7



Пример двумерного распределения
Трехмерная гистограмма рассеяния

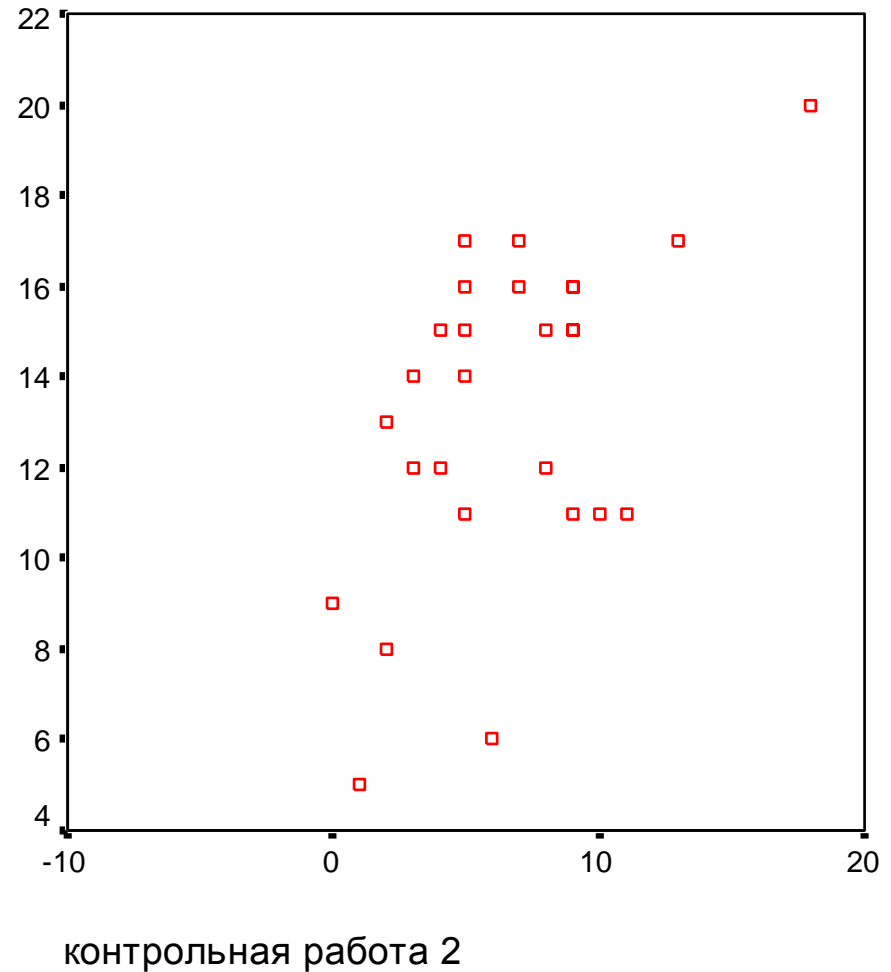
Опросник стилевой саморегуляции поведения (Моросанова).
Шкалы гибкости и самостоятельности

$N=265$, $r = -0,151$



Взаимосвязь признаков двумерного распределения

- Признаки являются **взаимосвязанными**, если определенные значения одного признака чаще встречаются вместе с определенными значениями другого признака.
- **Коэффициент корреляции** – число, показывающее меру взаимосвязи между двумя признаками. Изменяется в пределах от -1 до $+1$.
- Пример: $r = 0,541$



Обоснование формулы коэффициента корреляции

Пример

x	y
1	2
2	2
3	4
5	4
4	6
7	9
6	8
5	6
6	7
5	7
8	7
8	8

$$M_x = 5$$

$$M_y = 5,8$$

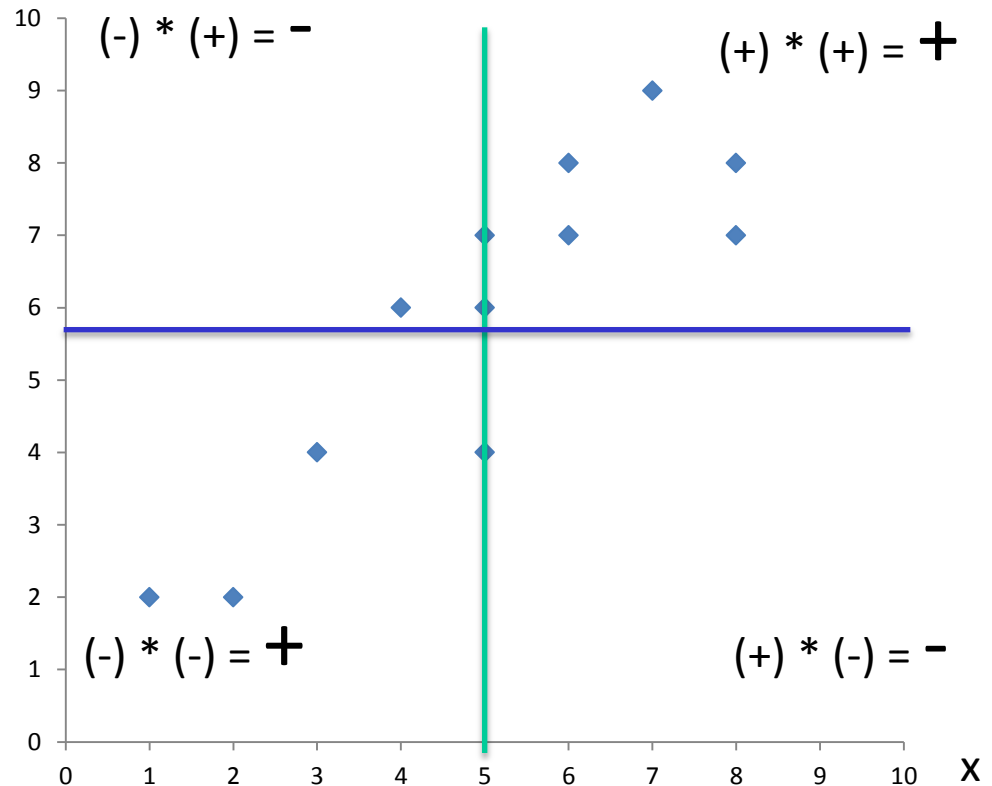
$$r = 0,88$$

Для каждой точки
вычисляется
момент:
 $M_i = (x_i - M_x) (y_i - M_y)$

знак зависит от
области по
отношению к точке
(M_x, M_y)

абсолютная
величина – от
удаленности от
этой точки

y



$$\sum M_i$$

зависит от знаков и абсолютных величин моментов,
числа объектов

$$\sum M_i / (n-1)$$

зависит от знаков и абсолютных величин моментов
(конфигурации данных и «размерности» измерений)

- коэффициент корреляции завит от конфигурации
данных

$$r = \frac{\sum M_i}{(n-1)S_x S_y}$$

Корреляции

- **Коэффициент линейной корреляции Пирсона** используется для интервальных данных, вычисляется по формуле:

$$r = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 \sum_{i=1}^N (y_i - \bar{y})^2}}$$

- **Коэффициент ранговой корреляции Спирмена** является частным случаем коэффициента линейной корреляции Пирсона, вычисляется по формуле:

$$r_s = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)}$$

Интерпретация коэффициента корреляции

- **Форма корреляции** отражает вдоль какой линии группируются точки (объекты выборки) на диаграмме рассеяния. При **линейной** корреляции точки, соответствующие объектам выборки располагаются близко к прямой линии, при **нелинейной** – группируются вокруг какой-либо кривой.

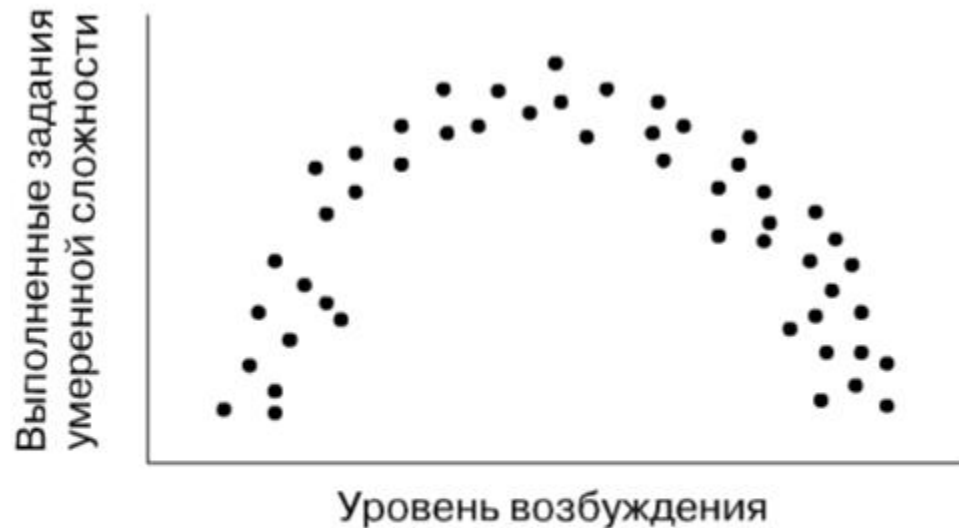


Рис. 39. График зависимости уровня возбуждения и успешности в выполнении заданий разной сложности

Интерпретация коэффициента корреляции

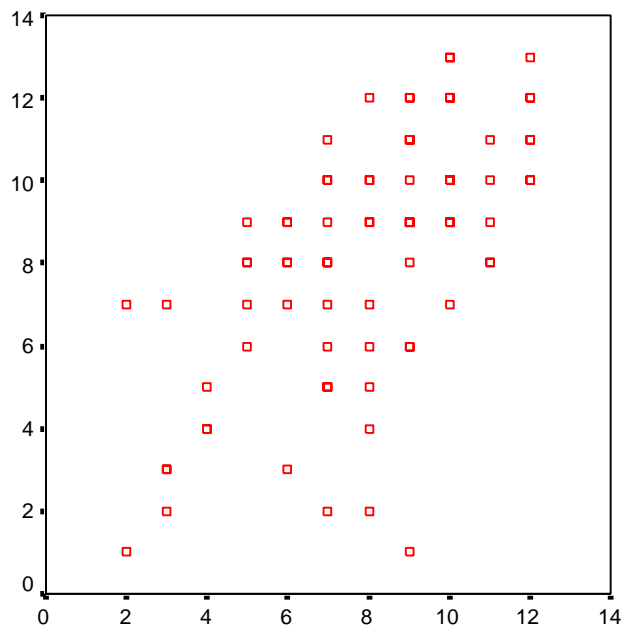
- **Направление корреляции** показывает характер взаимосвязи. По направлению выделяют прямую и обратную корреляции.
 - **Прямая** или **положительная** корреляция отражает взаимосвязь, при которой высоким значениям первого признака чаще соответствуют высокие значения другого признака, а низким – низкие. $r > 0$.
 - **Обратная** или **отрицательная** корреляция отражает взаимосвязь, при которой высоким значениям первого признака чаще соответствуют низкие значения другого признака, и наоборот. $r < 0$.
 - Два признака являются **независимыми**, если все сочетания значений первого и второго признаков являются равновероятными, т.е. встречаются примерно одинаково часто. $r \approx 0$.

Интерпретация коэффициента корреляции

- **Сила или теснота корреляции** (взаимосвязи) показывает насколько явно выражена взаимосвязь. Чем ближе значение коэффициента корреляции по абсолютной величине к 1, тем теснее взаимосвязь.

интервал при обратной связи	интервал при прямой связи	интерпретация
$r < -0,70$	$r > 0,70$	сильная или тесная
$-0,69 < r < -0,50$	$0,50 < r < 0,69$	средняя
$-0,49 < r < -0,30$	$0,30 < r < 0,49$	умеренная
$-0,29 < r < -0,20$	$0,20 < r < 0,29$	слабая
$-0,19 < r < 0,19$		очень слабая

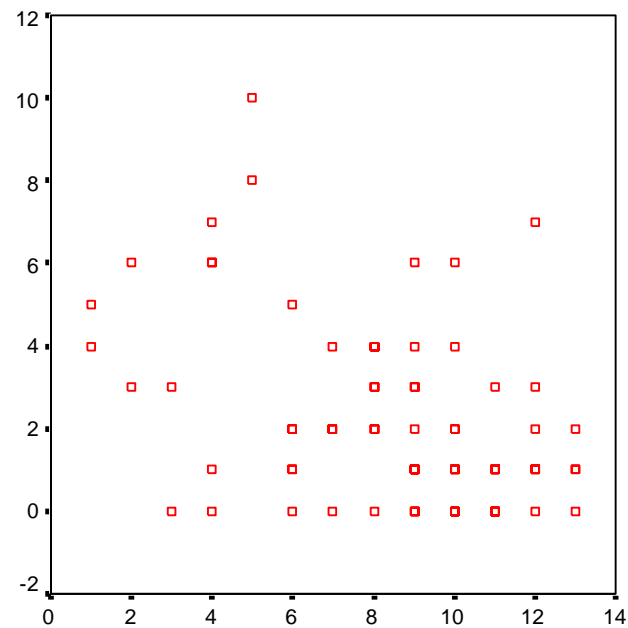
Сидоренко Е.В. Методы математической обработки в психологии. СПб.: Речь, 2010. С. 204.



SC10_1

$r = 0,636$

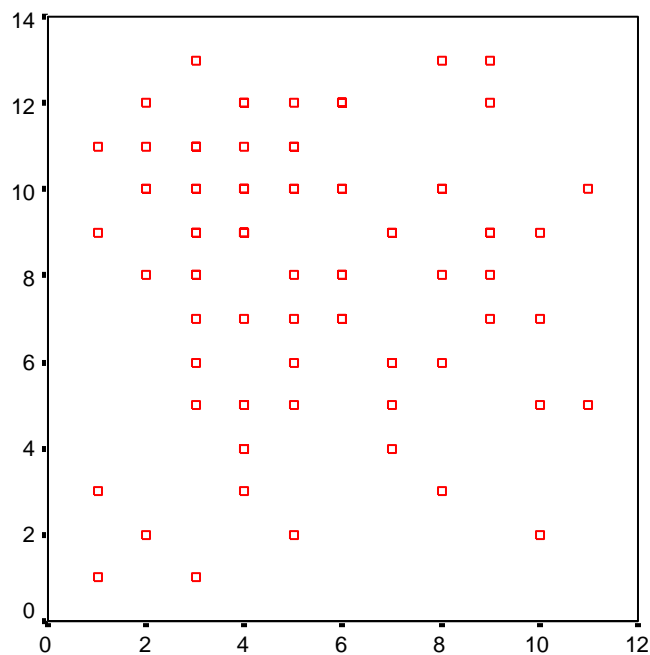
прямая,
средняя



C16_1

$r = -0,473$

обратная,
умеренная



SC6_2

$r = -0,011$

очень слабая