

**Тартуский университет
Нарвский колледж**

Informaatika

P2NC.00.450

Игорь Костюкевич



**Нарва
2009**

Оглавление

Основные понятия статистики	4
Введение.....	4
Типы вопросов анкеты.....	4
Понятие вариационного ряда. Виды вариационных рядов.....	7
Числовые характеристики вариационного ряда.....	9
Этапы принятия статистического решения	11
Классификация психологических задач, решаемых с помощью статистических методов ..	11
Нормальное распределение	13
Корреляционная связь.....	14
Назначение факторного и кластерного анализа.....	16
Проверка статистических гипотез	16
Задание по теме «Основы статистики»	21
Статистические функции MS Excel.....	22
Подсчет количества наблюдений	22
Средние значения и порядковые статистики.....	24
Характеристики разброса	26
Перекодирование полученных значений	27
Вычисление корреляций между переменными	29
Проверка нормальности распределения	30
Коэффициент корреляции рангов (Спирмена).....	31
Критерий согласия распределений χ^2 –Пирсона.....	31
Задание по теме «Статистические функции MS Excel»	33
Сводные таблицы – PivotTables	34
Создание сводной таблицы	34
Параметры сводных таблиц	37
Примеры. Группирование элементов сводной таблицы	39
Пакет анализа (Analysis ToolPack).....	42
Описательная статистика (Descriptive Statistics).....	43
Корреляционный анализ.....	44
Гистограмма.....	45
Ранг и перцентиль.....	47
t-тест	47
Двухвыборочный z-тест для средних.....	49
Задания по теме «Пакет анализа MS Excel»	50
Графики и диаграммы в MS Excel.....	52
Введение.....	52
Типы диаграмм в MS Excel	52
Процедура построения диаграммы с помощью мастера Chart Wizard	54
Построение графиков в MS Excel (задание I).....	55
Построение диаграмм и графиков с использованием MS Excel (задание II)	57
Основы работы с программой SPSS.....	62
Основные окна программы	62
Меню программы SPSS	63
Описание переменных	63
Допустимые имена переменных	64
Типы переменных.....	65
Модификация данных.....	67

Вычисление новых переменных путем использования различных арифметических выражений (математических формул)	67
Compute Variable: If Cases (Вычислить переменную: Отбор наблюдений).....	70
Перекодирование переменных.....	71
Сортировка наблюдений.....	72
Отбор наблюдений.....	73
Расщепление	75
Частотные таблицы и описательная статистика.....	76
Таблицы сопряженности	80
Статистики, вычисляемые для таблиц сопряженности	83
Проверка гипотезы о нормальности распределения.....	86
Сравнение средних.....	87
Непараметрические тесты сравнения выборок	89
Задание по теме «Основы работы с файлами и переменными в SPSS»	91
Корреляционный анализ	93
Задание по SPSS (вычисление переменных, анализ таблиц сопряженности, проверка гипотез, корреляционный анализ)	96
Кластерный анализ.....	98
Задание по теме «Корреляционный и кластерный анализ средствами SPSS».....	99
Приложения.....	101
Описание файла данных	101
Методика диагностики самооценки психических состояний (по Г. Айзенку)	101
Критические точки распределения Стьюдента	103
Критические значения линейного коэффициента корреляции Пирсона.....	104
Критические значения выборочного коэффициента корреляции рангов (по Спирмену)...	105
Критические значения критерия χ^2 для уровней статистической значимости $\alpha < 0,05$ и $\alpha < 0,01$ при разном числе степеней свободы	105

Основные понятия статистики

Введение

Основной задачей математической статистики является разработка методов получения научно обоснованных выводов о массовых явлениях и процессах из данных наблюдений и экспериментов. Эти выводы и заключения относятся не к отдельным испытаниям, из повторения которых складывается данное массовое явление, а представляют собой утверждения об общих вероятностных характеристиках данного процесса, то есть о вероятностях, законах распределения, математических ожиданиях, дисперсиях и т. д. Такое использование фактических данных как раз и является отличительной чертой статистического метода.

Пусть мы располагаем сведениями (обычно довольно ограниченными), например, о характеристиках некоторой совокупности объектов (респондентов). Собранные нами данные могут представлять непосредственный интерес в смысле информации о данной конкретной группе. Статистические же проблемы возникают тогда, когда мы на основе той же информации начинаем делать выводы относительно более широкого круга явлений. Так, например, нас может интересовать качество технологического процесса, для чего мы оцениваем вероятность получения в нем дефектного изделия или среднюю долговечность изделия. В этом случае мы рассматриваем собранный материал не ради его самого, а лишь как некую пробную группу или выборку, представляющую только серии из возможных результатов, которые мы могли бы встретить при продолжении наблюдений массового процесса в данной обстановке. Выводы и оценки, основанные на материале наблюдений, отражают случайный состав пробной группы и поэтому считаются приблизительными оценками вероятностного характера. Во многих случаях теория указывает, как наилучшим способом использовать имеющуюся информацию для получения по возможности более точных и надежных характеристик, указывая при этом степень надежности выводов, обусловленную ограниченностью запаса сведений.

В математической статистике рассматриваются две основные категории задач: оценивание и статистическая проверка гипотез. Первая задача разделяется на точечное оценивание и интервальное оценивание параметров распределения. Например, может возникнуть необходимость по наблюдениям получить точечные оценки среднего значения и дисперсии. Если мы хотим получить некоторый интервал, с той или иной степенью достоверности содержащий истинное значение параметра, то это задача интервального оценивания.

Вторая задача – проверка гипотез – заключается в том, что мы делаем предположение о распределении вероятностей случайной величины (например, о значении одного или нескольких параметров функции распределения) и решаем, согласуются ли в некотором смысле эти значения параметров с полученными результатами наблюдений.

Одним из основных методов получения информации об объектах исследования в психологии является опрос (анкетирование).

Типы вопросов анкеты

Каждой поставленной задаче соответствует блок вопросов, которые подразделяются на:

- закрытые;
- открытые;
- полужакрытые.

Закрытые вопросы предполагают выбор ответов из полного набора вариантов, приводимых в анкете. Основное преимущество использования данного типа вопросов состоит в том, что они позволяют использовать машинную обработку данных. Однако при этом в них скрывается

существенный недостаток: выбор ответа респондента (анкетируемого) ограничен. Поэтому обязательным при использовании закрытых вопросов является требование, чтобы значимые аспекты исследуемой проблемы были отражены в предлагаемых респонденту вариантах ответов.

Существуют два типа закрытых вопросов:

- альтернативный (дихотомический);
- с выборочным ответом (многовариантный).

Альтернативный вопрос предполагает выбор из двух вариантов ответа.

Вопрос с выборочным ответом предполагает выбор из трех и более вариантов ответов. Для постановки таких вопросов используются шкалы измерений, являющиеся инструментом приведения разнородных качественных признаков к сопоставимым количественным:

- номинальная;
- порядковая;
- интервальная;
- отношений.

Номинальная шкала (шкала наименований) обладает только характеристикой описания и представляет собой простое перечисление вариантов ответа без всякого намека на упорядочение и сопоставление.

Порядковая шкала (шкала рангов) состоит из категорий, отличающихся друг от друга условными понятиями или качественными признаками, в которых подразумевается наличие упорядоченности в каком-либо виде. Для того чтобы шкальные оценки отличались от чисел в быденном понимании, их на порядковом уровне называют рангами.

Имеется большое количество вариантов шкал, построенных на основе изложенных выше базовых принципах. Наиболее часто для формулировки закрытых вопросов применяются:

- Шкала Лайкерта;
- Семантический дифференциал;
- Шкала Стэпела;
- Мнемоническая шкала.

Шкала Лайкерта является разновидностью порядковой. При ее использовании разрабатывается ряд утверждений, которые относятся к проблеме или объекту исследования. Респондентов просят указать свою степень согласия или несогласия с каждым утверждением.

На практике широко используется пятиразрядная (пятиуровневая) шкала Лайкерта. В дополнение к двум крайним («полностью согласен», «полностью не согласен») утверждениям и нейтральному вводятся два промежуточных ответа («согласен», «не согласен»). Часто применяется цифровое обозначение утверждений, величина которого соответствует степени оценки отношения: абсолютное согласие («полностью согласен») оценивается значением 5, а радикальное несогласие («полностью не согласен») соответствует оценке 1. суммируя цифровые показатели, можно получить среднее значение степени согласия с приведенными утверждениями, а также построить статистические распределения. Можно также построить профиль разных групп респондентов.

Семантический дифференциал представляет собой серию биполярных (противоположных) характеристик, определяющих свойства изучаемого объекта. Семантическая шкала состоит из большого количества антонимов: «плохой – хороший», «удобный – неудобный», «полезный – бесполезный» и т.п. На такой шкале наносятся граничные определения, а все

пространство между ними разбивается на 7 диапазонов, характеризующих степень приближенности мнения респондента к тому или иному определению. Для удобства обработки анкет применяются шкалы с цифровыми обозначениями. Респондент должен отметить цифру, соответствующую его отношению к исследуемой проблеме.

Шкала Стэнела является модификацией шкалы семантического дифференциала. Отличие состоит в следующем: 1) прилагательные как биполярные пары тестируются отдельно, а не одновременно; 2) выделяются 10 отдельных положений на шкале, а не 7. Использование этой шкалы освобождает исследователя от разработки биполярных пар, позволяет использовать более детальную дифференциацию в измеряемых отношениях. Тем не менее, эта шкала используется реже, чем семантический дифференциал.

Мнемонические (рисованные) шкалы используются как для облегчения формулировок вариантов ответов на вопросы, так и для упрощения выбора ответа респондентами (например, детьми).

Такого рода рисунки четко и однозначно показывают реакцию опрашиваемых на тот или иной вопрос.

Открытые вопросы в отличие от закрытых не содержат подсказок, не «навязывают» тот или иной вариант ответа, а рассчитаны на получение неформализованного (нестандартного) мнения. Открытые вопросы используются в случае, когда заранее неизвестен перечень возможных ответов респондентов. После сбора информации затем обычно классифицируют, группируют полученные ответы на открытые вопросы, переводя их в номинативную (чаще) или порядковую (реже) шкалы.

Виды открытых вопросов:

- неструктурированный – допускает любую словесную форму ответа;
- подбор словесных ассоциаций – опрашиваемому называют отдельные слова с целью выяснить возникающие у него ассоциации;
- завершение предложения – предлагается завершить незаконченное предложение;
- завершение рассказа – предлагается завершить незаконченный рассказ;
- завершение рисунка – опрашиваемого просят представить себя на месте одного из героев, как правило, шутового рисунка и от его имени написать на рисунке собственное мнение;
- тематический апперцепционный тест – предлагается придумать рассказ по предложенной картинке.

В анкетах также часто присутствуют смешанные вопросы, состоящие из вопросов, относящихся к разным типам.

Обработка ответов на открытые вопросы весьма затруднительна и практически не поддается автоматизации. Обычно исследователь пытается все-таки выделить доминирующие типы ответов с целью их классификации и дальнейшего анализа методами статистики. Также весьма популярным методом анализа ответов на открытые вопросы (и вообще – информации в виде текста) является метод контент-анализа (применяется в области политологии, рекламы, социологии, экономики, психологии, культурологи; в данном курсе не рассматривается).

Интервальная шкала - это шкала, классифицирующая по принципу "больше на определенное количество единиц - меньше на определенное количество единиц". Каждое из возможных значений признака отстоит от другого на равном расстоянии.

Шкала равных отношений - это шкала, классифицирующая объекты или субъектов пропорционально степени выраженности измеряемого свойства. В шкалах отношений классы обозначаются числами, которые пропорциональны друг другу: 2 так относится к 4,

как 4 к 8. Это предполагает наличие абсолютной нулевой точки отсчета.

От того, в какой шкале «измерен» тот или иной показатель, зависит применимость разных статистических методов. Для шкалы равных отношений применимы практически все методики (с учетом наличия нормального распределения). На практике обычно не делают различий между переменными, относящимися к интервальной шкале и шкале равных отношений (так сделано и в программе SPSS). Номинативные переменные используются только для классификации. Порядковые переменные анализируются (главным образом) с применением непараметрических критериев, основанных на частотах и рангах.

Понятие вариационного ряда. Виды вариационных рядов

Дадим определение другим понятиям, используемым в статистике. Так, совокупность предметов или явлений, объединенных каким-либо общим признаком или свойством качественного или количественного характера, называется объектом наблюдения.

Всякий объект статистического наблюдения состоит из отдельных элементов — единиц наблюдения.

Результаты статистического наблюдения представляют собой числовую информацию — данные. Статистические данные — это сведения о том, какие значения принял интересующий исследователя признак в статистической совокупности. Признаки бывают количественными и качественными.

Количественным называется признак, значения которого выражаются числами.

Качественным называется признак, характеризующийся некоторым свойством или состоянием элементов совокупности.

Статистическая совокупность называется генеральной, если исследованию подлежат все элементы совокупности (сплошное наблюдение).

Часть элементов генеральной совокупности, подлежащая исследованию, называется выборочной совокупностью (выборкой). Она извлекается из генеральной совокупности случайно, так чтобы каждый из n элементов выборки имел равные шансы быть отобранным.

Значения признака, которые при переходе от одного элемента совокупности к другому изменяются (варьируют), называются вариантами и обычно обозначаются малыми латинскими буквами x, y, z .

Порядковый номер варианта (значения признака) называется рангом: x_1 — 1-й вариант (1-е значение признака), x_2 — 2-й вариант (2-е значение признака), x_i — i -й вариант (i -е значение признака).

Ряд значений признака (вариантов), расположенных в порядке возрастания или убывания с соответствующими им весами, называется вариационным рядом (рядом распределения).

В качестве весов выступают частоты или частоты.

Частота (T) показывает, сколько раз встречается тот или иной вариант (значение признака) в статистической совокупности.

Частость или относительная частота (ω_i) показывает, какая часть единиц совокупности имеет тот или иной вариант. Частость рассчитывается как отношение частоты того или иного варианта к сумме всех частот ряда.

Сумма всех частостей равна 1.

Вариационные ряды бывают дискретными и интервальными.

Дискретные вариационные ряды строят обычно в том случае, если значения изучаемого признака могут отличаться друг от друга не менее чем на некоторую конечную величину. В дискретных вариационных рядах задаются точечные значения признака.

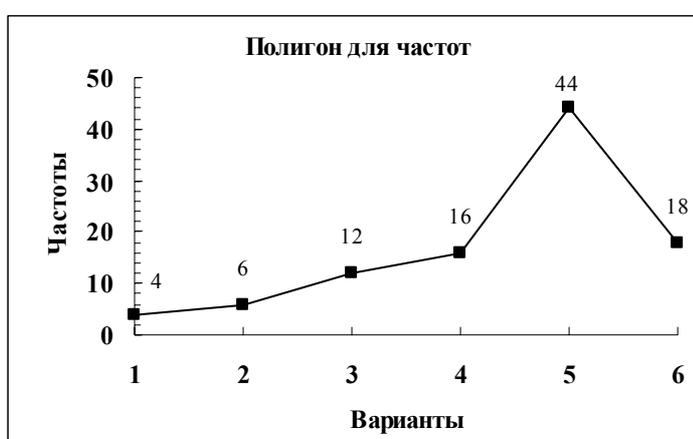
Интервальные вариационные ряды строят обычно в том случае, если значения изучаемого признака могут отличаться друг от друга на сколь угодно малую величину. Значения признака в них задаются в виде интервалов.

В интервальных вариационных рядах в каждом интервале выделяют верхнюю и нижнюю границы. Разность между верхней и нижней границами интервала называется интервальной разностью или длиной (величиной) интервала.

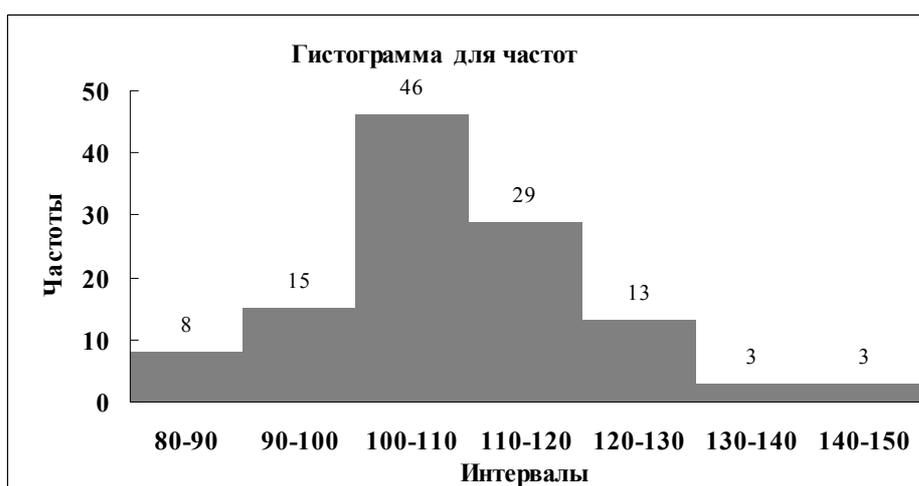
Кроме этого в интервальном вариационном ряде могут встречаться интервалы разной длины. Если интервалы в вариационном ряде имеют одинаковую длину (интервальную разность), их называют равновеликими, в противном случае — неравновеликими.

Для характеристики вариационного ряда наряду с частотами и частостями используются накопленные частоты и частости. Накопленные частоты (частости) показывают, сколько единиц совокупности (какая их часть) не превышает заданного значения (варианта) x .

Дискретный вариационный ряд графически можно представить с помощью полигона распределения частот или частостей (рис.6.1).



Интервальные вариационные ряды графически можно представить с помощью гистограммы, т. е, столбчатой диаграммы.



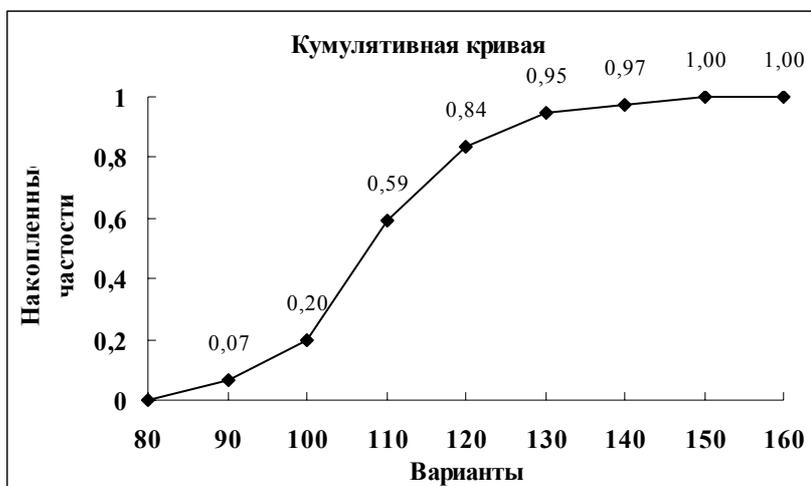
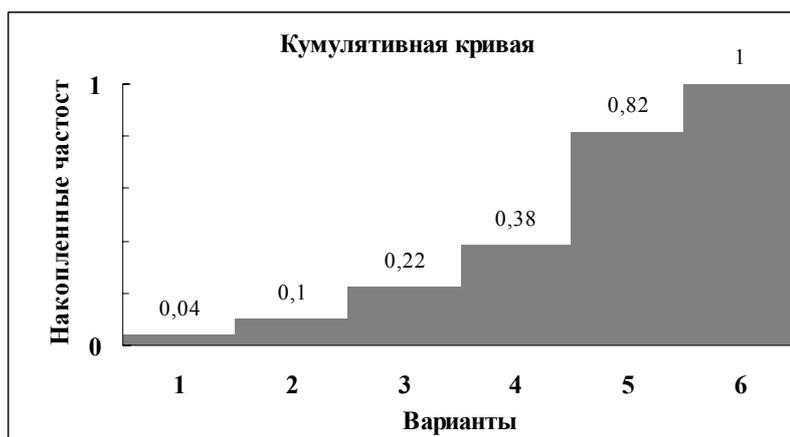
При ее построении по оси абсцисс откладываются значения изучаемого признака (границы интервалов). В том случае, если интервалы одинаковой величины, по оси ординат можно откладывать частоты или частости. Если же интервалы имеют разную величину, по оси ординат необходимо откладывать значения абсолютной или относительной плотности распределения. Абсолютная плотность — отношение частости интервала к его величине.

Абсолютная плотность показывает, сколько единиц совокупности приходится на единицу интервала.

Относительная плотность — отношение частости интервала к его величине.

Относительная плотность показывает, какая часть единиц совокупности приходится на единицу интервала.

И дискретные, и интервальные вариационные ряды графически можно представить в виде кумуляты. При построении первой по данным дискретного ряда по оси абсцисс откладываются значения признака (варианты), а по оси ординат — накопленные частоты или частоты. На пересечении значений признака (вариантов) и соответствующих им накопленных частот (частостей) строятся точки, которые в свою очередь соединяются отрезками или кривой. Получающаяся таким образом ломаная (кривая) называется кумулятой (кумулятивной кривой). Абсциссами ее точек являются верхние границы интервалов. Ординаты образуют накопленные частоты (частости) соответствующих интервалов. Часто добавляют еще одну точку, абсциссой которой является нижняя граница первого интервала, а ордината равна нулю. Соединяя точки отрезками или кривой, получаем кумуляту.



Числовые характеристики вариационного ряда

Одной из основных числовых характеристик ряда распределения (вариационного ряда) является средняя арифметическая.

Существует две формулы расчета средней арифметической: простая и взвешенная. Простую среднюю арифметическую обычно используют, когда данные наблюдения не сведены в вариационный ряд либо все частоты равны единице или одинаковы

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

где x_i — i -е значение признака; n — объем ряда (число наблюдений; число значений

признака).

Если частоты отличны друг от друга, расчет производится по формуле средней арифметической взвешенной

$$\bar{x} = \frac{\sum_i x m_i}{\sum_i m_i},$$

где x_i — i -е значение признака; t_i — частота i -го значения признака; k — число его значений (вариантов).

Перечислим некоторые другие часто используемые статистические показатели.

- Медиана. Значение, выше и ниже которого попадает по половине наблюдений, иначе 50-й перцентиль. Если число наблюдений четно, медиана есть арифметическое среднее двух находящихся в середине значений, если выборку упорядочить по убыванию или по возрастанию. Медиана представляет собой меру центральной тенденции, которая нечувствительна к выбросам, в отличие от среднего значения, которое могут исказить несколько экстремально больших или малых значений.
- Мода. Чаще всего встречающееся значение. Если таких значений несколько, каждое из них является модой. Процедура Частоты выдает только наименьшее из этих значений.
- Стандартное отклонение. Мера разброса вокруг среднего. При нормальном распределении 68% наблюдений укладываются в одно стандартное отклонение от среднего, и 95% - в два стандартных отклонения.
- Дисперсия. Мера разброса относительно среднего значения. Равна сумме квадратов отклонений от среднего, деленной на число, на единицу меньше числа наблюдений. Дисперсия измеряется в единицах, которые равны квадратам единиц измерения самой переменной.
- Диапазон (размах вариации). Разность между наибольшим и наименьшим значениями числовой переменной; максимум минус минимум.
- Минимум. Наименьшее значение числовой переменной.
- Максимум. Наибольшее значение числовой переменной.
- Стандартная ошибка среднего. Мера того, как сильно может отличаться значение среднего от выборки к выборке, извлекаемое из одного и того же распределения. Можно применять для грубого сравнения наблюдаемого среднего с гипотетическим значением (то есть можно заключить, что два значения различаются, если отношение их разности к стандартному отклонению меньше -2 или больше +2).
- Асимметрия. Мера асимметрии распределения. Нормальное распределение симметрично, и для него асимметрия равна 0. Распределение со значимой положительной асимметрией имеет длинный хвост справа. Распределение со значимой отрицательной асимметрией имеет длинный хвост слева. В качестве грубого правила можно сказать, что значение асимметрии, более чем вдвое превышающее ее стандартную ошибку, указывает на наличие асимметрии распределения.
- Эксцесс. Мера сгруппированности наблюдений вокруг центральной точки. Для нормального распределения значение эксцесса равно 0. Положительный эксцесс означает, что наблюдения более плотно группируются вокруг центра, и их хвосты длиннее, чем у нормального распределения, а отрицательный - что наблюдения группируются менее плотно, и хвосты короче.

Возможности вычисления этих статистических характеристики имеются в любом программном обеспечении, ориентированном на работу с электронными таблицами, в том числе в MS Excel и SPSS.

Этапы принятия статистического решения

Принятие статистического решения разбивается на этапы или шаги:

1. Формулировка нулевой и альтернативной гипотез.
2. Определение объема выборки N .
3. Выбор соответствующего уровня значимости или вероятности отклонения нулевой гипотезы. Это может быть величина меньшая или равная 0,05 (5% уровень значимости). В зависимости от важности исследования можно выбрать уровень значимости в 0,1% или даже в 0,001%.
4. Выбор статистического метода, который зависит от типа решаемой психологической задачи.
5. Вычисление соответствующего эмпирического значения по экспериментальным данным согласно выбранному статистическому методу.
6. Нахождение для выбранного статистического метода критических значений соответствующих уровню значимости (обычно для $p = 0,05$ и для $p = 0,01$).
7. Принятие решения (выбор между нулевой и альтернативной гипотезами – см. следующий раздел курса).

Классификация психологических задач, решаемых с помощью статистических методов

Подчеркнем еще раз, что, прежде чем выполнить любой психологический эксперимент, необходимо четко сформулировать его задачи, определить экспериментальную гипотезу и все этапы ее статистической проверки, а также выбрать соответствующий статистический метод, наиболее эффективный для решения поставленных в исследовании задач.

Подавляющее большинство задач, решаемых психологом в эксперименте, предполагает те или иные сопоставления. Это могут быть сопоставления одних и тех же показателей в разных группах испытуемых или, напротив, разных показателей в одной и той же группе. Для определения степени эффективности каких-либо воздействий (обучение, тренировка, тренинг, инструктаж и т.п.) сравниваются показатели «до» и «после» этих воздействий. Например, сравниваются показатели уровня агрессивности у подростков до и после психотренинга, что позволяет определить его эффективность. Или сопоставляются результаты у одних и тех же испытуемых по одним и тем же методикам, но в разном возрасте, это позволяет выявить временную динамику анализируемых показателей. Иногда возникает задача сравнить индивидуальные показатели, полученные при различных внешних условиях, для выявления связи между показателями и факторов, объединяющих эти связи.

Два выборочных распределения сравниваются между собой или с теоретическим законом распределения, чтобы выявить различия или, напротив, сходство в типах распределений. Например, сравнение распределений времени решения простой и сложных задач позволит построить классификацию задач и типологию испытуемых

Психологические задачи, решаемые с помощью методов математической статистики, условно можно разделить на несколько групп:

- задачи, требующие установления сходства или различия;
- задачи, требующие группировки и классификации данных;
- задачи, ставящие целью анализ источников вариативности получаемых психологических признаков;
- задачи, предполагающие возможность прогноза на основе имеющихся данных.

Более полная сводка типов задач и методов их решения дана в таблице.

<i>Задачи</i>	<i>Условия</i>	<i>Методы</i>
Выявление различий в уровне исследуемого признака	2 выборки	критерий Макнамары Q критерий Розенбаума U критерий Манна-Уитни φ критерий (угловое преобразование Фишера)
	3 и больше выборок	S критерий Джонкира H критерий Крускала-Уоллиса
Оценка сдвига значений исследуемого признака	2 замера на одной и той же выборке	T критерий Вилкоксона G критерий знаков φ критерий (угловое преобразование Фишера) t критерий Стьюдента
	3 и более замеров на одной и той же выборке	$\chi^2_{\text{фр}}$ критерий Фридмана L критерий тенденций Пейджа t критерий Стьюдента
Выявление различий в распределении признака	при сопоставлении эмпирического распределения с теоретическим	χ^2 критерий Пирсона λ критерий Колмогорова-Смирнова t критерий Стьюдента
	при сопоставлении двух эмпирических распределений	χ^2 критерий Пирсона λ критерий Колмогорова-Смирнова φ критерий (угловое преобразование Фишера)
Выявление степени согласованности изменений	двух признаков	φ коэффициент корреляции Пирсона τ коэффициент корреляции Кендала R - бисериальный коэффициент корреляции η корреляционное соотношение Пирсона ρ коэффициент ранговой корреляции Спирмена г коэффициент корреляции Пирсона линейная и криволинейная регрессии
	трех или большего числа признаков	ρ коэффициент ранговой корреляции Спирмена

		г коэффициент корреляции Пирсона множественная и частная корреляции линейная, криволинейная и множественная регрессии
Анализ изменений признака под влиянием контролируемых условий	под влиянием одного фактора	S критерий Джонкира L критерий тенденций Пейджа однофакторный дисперсионный анализ критерий Линка и Уоллиса критерий Немени множественное сравнение независимых выборок
	под влиянием двух факторов одновременно	двухфакторный дисперсионный анализ

Часто при анализе данных используются несколько методик для сравнения получаемых статистических выводов.

Нормальное распределение

Распределением признака называется закономерность встречаемости разных его значений.

В психологических исследованиях чаще всего ссылаются на нормальное распределение.

Нормальное распределение характеризуется тем, что крайние значения признака в нем встречаются достаточно редко, а значения, близкие к средней величине – достаточно часто. Нормальным такое распределение называется потому, что оно очень часто встречалось в естественнонаучных исследованиях и казалось "нормой" всякого массового случайного проявления признаков. Это распределение следует закону, открытому Муавром, Гауссом и Лапласом. График нормального распределения представляет собой так называемую колоколообразную кривую.

На рисунке ниже приведено изображение дискретного вариационного ряда и аппроксимирующей его кривой нормального распределения. Один из критериев проверки гипотезы о нормальном характере распределения исследуемого признака (см. раздел о статистических функциях MS Excel) подтверждает незначимость отклонений эмпирического распределения от теоретического.

Некоторые методы проверки гипотез (их называют параметрическими) для корректности своего использования требуют соответствия анализируемых данных нормальному распределению. К таким методам относятся, например, критерий Стьюдента и корреляционный анализ по Пирсону (линейный коэффициент корреляции). Свободными от «привязки» к нормальному распределению являются непараметрические методы, использующие получаемые в эксперименте (наблюдении) частоты встречаемости значений признаков и их ранги.



Корреляционная связь

Психолога нередко интересует, как связаны между собой две или большее количество переменных в одной или нескольких изучаемых группах. Например, могут ли учащиеся с высоким уровнем тревожности демонстрировать стабильные академические достижения или связана ли продолжительность работы психологов в школе с размером их заработной платы или с чем больше связан уровень умственного развития учащихся — с их успеваемостью по математике или по литературе и т.п.

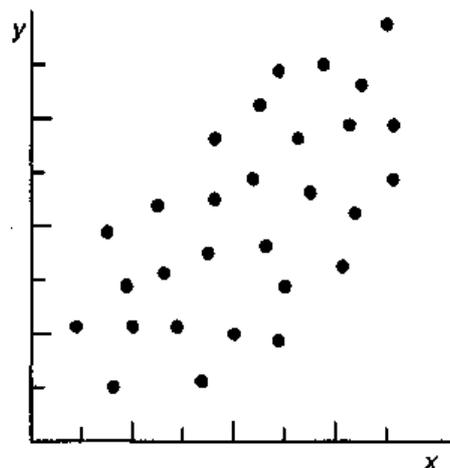
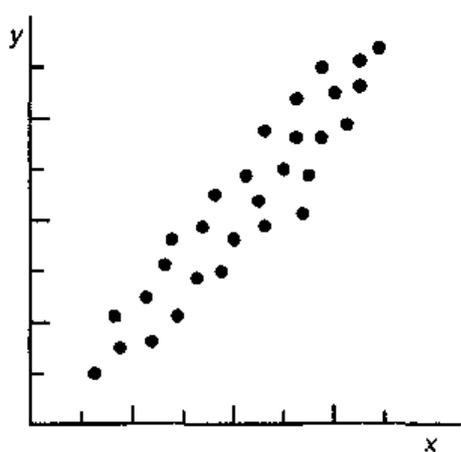
В математике для описания связей между переменными величинами используют понятие функции, которая ставит в соответствие каждому определенному значению независимой переменной определенное значение зависимой переменной. Но подобные однозначные или функциональные связи между переменными величинами встречаются далеко не всегда.

Известно, например, что в среднем между ростом людей и их весом наблюдается положительная связь, и такая, что чем больше рост, тем больше вес человека. Однако из этого правила имеются исключения, когда относительно низкие люди имеют избыточный вес, и, наоборот, астеники, при высоком росте имеют малый вес. Причиной подобных исключений является то, что каждый биологический, физиологический или психологический признак определяется воздействием многих факторов средовых, генетических, социальных, экологических и т.д. Поэтому связи между психологическими признаками имеют не функциональный, а статистический характер, когда в среднем определенному значению одного признака, рассматриваемому в качестве аргумента, соответствует не одно какое-либо значение, а целый спектр распределяющихся в вариационный ряд числовых значений. Такого рода зависимость между переменными величинами называется корреляционной, или корреляцией. Корреляционная связь — это согласованное изменение двух признаков, отражающее тот факт, что изменчивость одного признака находится в соответствии с изменчивостью другого.

Функциональные связи легко обнаружить и измерить на единичных и групповых объектах, однако этого нельзя проделать с корреляционными связями, которые можно изучать только на представительных выборках методами математической статистики. Корреляционные связи — это вероятностные изменения.

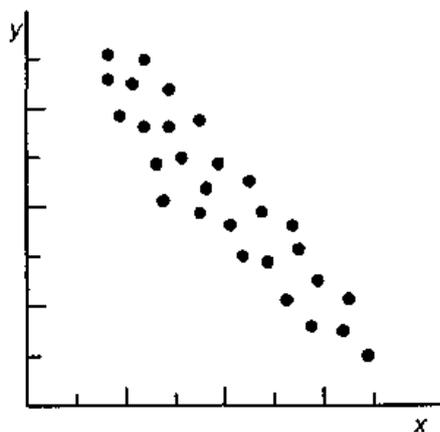
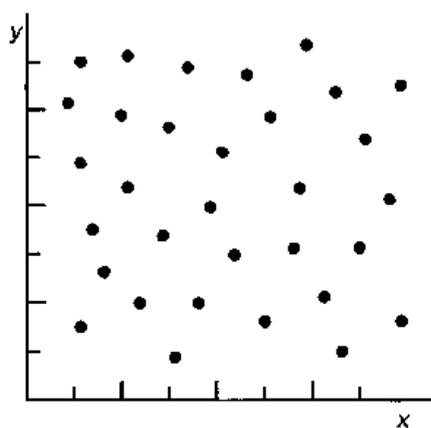
Согласованные изменения признаков и отражающая это корреляционная связь между ними может свидетельствовать не о зависимости этих признаков между собой, а о зависимости обоих этих признаков от какого-то третьего признака или сочетания признаков, не рассматриваемых в исследовании. Зависимость подразумевает влияние, связь — любые согласованные изменения, которые могут объясняться сотнями причин. Корреляционные связи не могут рассматриваться как свидетельство причинно следственной зависимости они свидетельствуют лишь о том что изменениям одного признака, как правило, сопутствуют определенные изменения другого но находится ли причина изменения в одном из признаков или она оказывается за пределами исследуемой пары признаков нам неизвестно.

Виды корреляционных связей между измеренными признаками могут быть различны: корреляция бывает линейной и нелинейной, положительной и отрицательной. Она линейна, если с увеличением или уменьшением одной переменной вторая переменная в среднем либо также растет, либо убывает. Она нелинейна, если при увеличении одной величины характер изменения второй нелинейен, а описывается другими законами.



Сильная положительная корреляция

Слабая положительная корреляция



Корреляция отсутствует

Сильная отрицательная корреляция

Корреляция будет положительной, если с увеличением переменной X переменная Y в среднем также увеличивается, а если с увеличением X переменная Y имеет в среднем тенденцию к уменьшению, то говорят о наличии отрицательной корреляции. Возможна ситуация когда между переменными невозможно установить какую-либо зависимость. В этом случае говорят об отсутствии корреляционной связи. Нередко встречаются задачи, в которых традиционная и наиболее часто встречающаяся в психологических исследованиях линейная корреляционная связь отсутствует, в то время как имеется криволинейная связь, например, полиномиальная или гиперболическая.

Задача корреляционного анализа сводится к установлению направления (положительное или отрицательное) и формы (линейная, нелинейная) связи между варьирующими признаками,

измерению ее тесноты, и, наконец, к проверке уровня значимости полученных коэффициентов корреляции.

Переменные X и Y могут быть измерены в разных шкалах, именно это определяет выбор соответствующего коэффициента корреляции.

Тип шкалы		Мера связи
Переменная X	Переменная Y	
интервальная или отношений	интервальная или отношений	коэффициент Пирсона γ
интервальная или отношений	ранговая, интервальная или отношений	коэффициент Спирмена ρ
ранговая	ранговая	коэффициент τ Кендалла
дихотомическая	дихотомическая	коэффициент ϕ
дихотомическая	ранговая	рангово-бисериальный коэффициент
дихотомическая	интервальная или отношений	бисериальный коэффициент
интервальная	ранговая	не разработан

Назначение факторного и кластерного анализа

Факторный анализ - это процедура, с помощью которой большое число переменных, относящихся к имеющимся наблюдениям сводит к меньшему количеству независимых влияющих величин, называемых факторами. При этом в один фактор объединяются переменные, сильно коррелирующие между собой. Переменные из разных факторов слабо коррелируют между собой. Таким образом, целью факторного анализа является нахождение таких комплексных факторов, которые как можно более полно объясняют наблюдаемые связи между переменными, имеющимися в наличии.

В результате *кластерного анализа* при помощи предварительно заданных переменных формируются группы наблюдений. Под наблюдениями здесь понимаются отдельные личности (респонденты) или любые другие объекты. Члены одной группы (одного кластера) должны обладать схожими проявлениями переменных, а члены разных групп различными.

Оба метода основываются на сложных математических алгоритмах. К сожалению, в программе MS Excel они не реализованы. Ниже мы рассмотрим примеры проведения факторного и кластерного анализов с помощью программы SPSS.

Наряду с кластеризацией наблюдений в SPSS предусмотрена кластеризация переменных. Здесь на основе заданных наблюдений образуются группы переменных (в принципе то же самое делает и факторный анализ).

Одной из проблем кластерного анализа является определение количества кластеров. В настоящий момент отсутствуют общепринятые методики решения этой проблемы. Решение обычно принимает исследователь исходя из теоретических соображений.

Проверка статистических гипотез

В процессе статистического анализа иногда бывает необходимо сформулировать и проверить предположения (гипотезы) относительно величины независимых параметров или закона распределения изучаемой генеральной совокупности (совокупностей). Например, исследователь выдвигает гипотезу о том, что «выборка извлечена из нормальной генеральной совокупности» или «генеральные средние двух анализируемых совокупностей

равны». Такие предположения называются статистическими гипотезами.

Сопоставление высказанной гипотезы относительно генеральной совокупности с имеющимися выборочными данными, сопровождаемое количественной оценкой степени достоверности получаемого вывода и осуществляемое с помощью того или иного статистического критерия, называется проверкой статистических гипотез.

Выдвинутая гипотеза называется нулевой (основной). Ее принято обозначать H_0 .

По отношению к высказанной (основной) гипотезе всегда можно сформулировать альтернативную (конкурирующую), противоречащую ей. Альтернативную (конкурирующую) гипотезу принято обозначать H_1 .

Цель статистической проверки гипотез состоит в том, чтобы на основании выборочных данных принять решение о справедливости основной гипотезы H_0 . Если выдвигаемая гипотеза сводится к утверждению о том, что значение некоторого неизвестного параметра генеральной совокупности в точности равно заданной величине, то эта гипотеза называется простой. В других случаях гипотеза называется сложной.

В качестве нулевой гипотезы H_0 принято выдвигать простую гипотезу, так как обычно бывает удобнее проверять более строгое утверждение.

По своему содержанию статистические гипотезы можно подразделить на несколько основных типов:

- гипотезы о виде закона распределения исследуемой случайной величины;
- гипотезы о числовых значениях параметров исследуемой генеральной совокупности;
- гипотезы об однородности двух или нескольких выборок или некоторых характеристик анализируемых совокупностей;
- гипотезы об общем виде модели, описывающей статистическую зависимость между признаками; и др.

Так как проверка статистических гипотез осуществляется на основании выборочных данных, т. е. ограниченного ряда наблюдений, решения относительно нулевой гипотезы H_0 имеют вероятностный характер. Другими словами, такое решение неизбежно сопровождается некоторой, хотя возможно и очень малой, вероятностью ошибочного заключения как в ту, так и в другую сторону.

Так, в какой-то небольшой доле случаев α нулевая гипотеза H_0 может оказаться отвергнутой, в то время как в действительности в генеральной совокупности она является справедливой. Такую ошибку называют ошибкой 1-го рода, а ее вероятность - α уровнем значимости и обозначают α .

Наоборот, в какой-то небольшой доле случаев β нулевая гипотеза H_0 принимается, в то время как на самом деле в генеральной совокупности она ошибочна, а справедлива альтернативная гипотеза H_1 .

Такую ошибку называют ошибкой 2-го рода. Вероятность ошибки 2-го рода обозначается как β .

Вероятность $1 - \beta$ называют мощностью критерия.

При фиксированном объеме выборки можно выбрать по своему усмотрению величину вероятности только одной из ошибок α или β . Увеличение вероятности одной из них приводит к снижению другой.

Принято задавать вероятность ошибки 1-го рода α — уровень значимости. Как правило, пользуются некоторыми стандартными значениями уровня значимости α : 0,1; 0,05; 0,025; 0,01; 0,005; 0,001. Тогда, очевидно, из двух критериев, характеризующихся одной и той же вероятностью α (отклонить правильную в действительности гипотезу H_0), следует принять тот, которому соответствует меньшая ошибка 2-го рода β , т.е. большая мощность. Снижения

вероятностей обеих ошибок α и β можно добиться путем увеличения объема выборки.

Правильное решение относительно нулевой гипотезы H_0 также может быть двух видов:

- будет принята нулевая гипотеза H_0 , когда в генеральной совокупности верна нулевая гипотеза H_0 ; вероятность такого решения $1 - \alpha$;
- нулевая гипотеза H_0 будет отклонена в пользу альтернативной H_1 , когда в генеральной совокупности нулевая гипотеза H_0 отклоняется в пользу альтернативной H_1 , вероятность такого решения $1 - \beta$ — мощность критерия.

Результаты решения относительно нулевой гипотезы можно проиллюстрировать с помощью таблицы.

Нулевая гипотеза H_0	Результаты решения относительно нулевой гипотезы	
	Отклонена	Принята
Верна	Ошибка 1-го рода, ее вероятность $P(H_1/H_0)=\alpha$	Правильное решение, его вероятность $P(H_0/H_0)=1-\alpha$
Неверна	Правильное решение, его вероятность $P(H_1/H_1)=1-\beta$	Ошибка 2-го рода, ее вероятность $P(H_0/H_1)=\beta$

Проверка статистических гипотез осуществляется с помощью статистического критерия (назовем его в общем виде K), являющего функцией от результатов наблюдения.

Статистический критерий — это правило (формула), по которому определяется мера расхождения результатов выборочного наблюдения с высказанной гипотезой H_0 .

Статистический критерий, как и всякая функция от результатов наблюдения, является случайной величиной и в предположении справедливости нулевой гипотезы H_0 подчинен некоторому хорошо изученному (и затабулированному) теоретическому закону распределения с плотностью распределения $f(k)$.

Выбор критерия для проверки статистических гипотез может быть осуществлен на основании различных принципов. Чаще всего для этого пользуются принципом отношения правдоподобия, который позволяет построить критерий, наиболее мощный среди всех возможных критериев. Суть его сводится к выбору такого критерия K с известной функцией плотности $f(k)$ при условии справедливости гипотезы H_0 , чтобы при заданном уровне значимости α можно было бы найти критическую точку K распределения $f(k)$, которая разделила бы область значений критерия на две части: область допустимых значений, в которой результаты выборочного наблюдения выглядят наиболее правдоподобными, и критическую область, в которой результаты выборочного наблюдения выглядят менее правдоподобными в отношении нулевой гипотезы H_0 .

Если такой критерий K выбран, и известна плотность его распределения, то задача проверки статистической гипотезы сводится к тому, чтобы при заданном уровне значимости α рассчитать по выборочным данным наблюдаемое значение критерия $K_{\text{набл}}$ определить, является ли оно наиболее или наименее правдоподобным в отношении нулевой гипотезы H_0 .

Проверка каждого типа статистических гипотез осуществляется с помощью соответствующего критерия, являющегося наиболее мощным в каждом конкретном случае. Например, проверка гипотезы о виде закона распределения случайной величины может быть осуществлена с помощью критерия согласия Пирсона χ^2 ; проверка гипотезы о равенстве неизвестных значений дисперсий двух генеральных совокупностей — с помощью критерия Фишера F ; ряд гипотез о неизвестных значениях параметров генеральных совокупностей проверяется с помощью критерия Z — нормальной распределенной случайной величины и критерия t -Стьюдента и т. д.

Значение критерия, рассчитываемое по специальным правилам на основании выборочных данных, называется наблюдаемым значением критерия ($K_{\text{набл}}$).

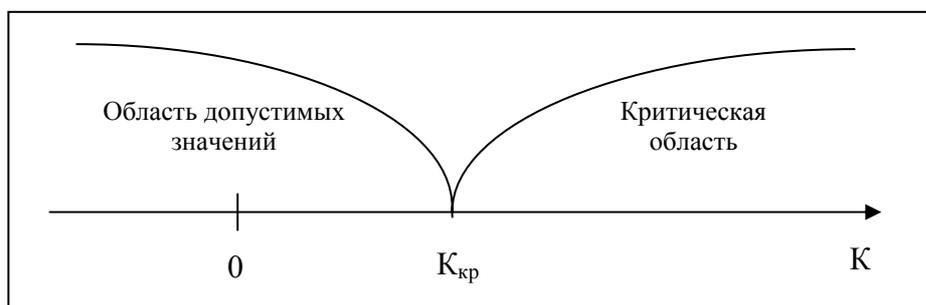
Значения критерия, разделяющие совокупность значений критерия на область допустимых значений (наиболее правдоподобных в отношении нулевой гипотезы H_0) и критическую область (область значений, менее правдоподобных в отношении нулевой гипотезы H_0), определяемые на заданном уровне значимости α по таблицам распределения случайной величины K , выбранной в качестве критерия, называются критическими точками ($K_{кр}$).

Областью допустимых значений (областью принятия нулевой гипотезы H_0) называют совокупность значений критерия K , при которых нулевая гипотеза H_0 не отклоняется.

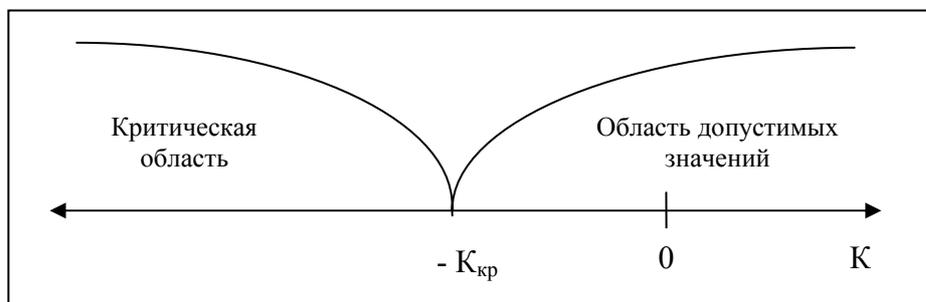
Критической областью называют совокупность значений критерия K , при которых нулевая гипотеза H_0 отклоняется в пользу конкурирующей H_1 .

Различают одностороннюю (правостороннюю или левостороннюю) и двустороннюю критические области.

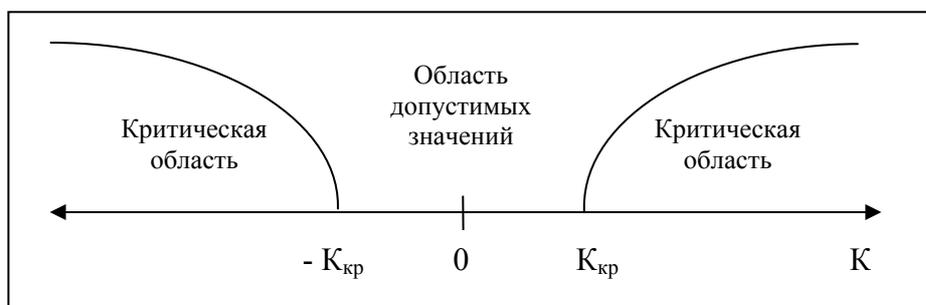
Если конкурирующая гипотеза — правосторонняя, например, $H_1: a > a_0$, то и критическая область — правосторонняя. При правосторонней конкурирующей гипотезе критическая точка ($K_{кр.п}$) принимает положительные значения.



Если конкурирующая гипотеза — левосторонняя, например, $H_1: a < a_0$, то и критическая область — левосторонняя. При левосторонней конкурирующей гипотезе критическая точка принимает отрицательные значения ($K_{кр.л}$).



Если конкурирующая гипотеза — двусторонняя, например, $H_1: a \neq a_0$, то и критическая область — двусторонняя. При двусторонней конкурирующей гипотезе определяются 2 критические точки ($K_{кр.л}$ и $K_{кр.п}$).



Основной принцип проверки статистических гипотез состоит в следующем:

- если наблюдаемое значение критерия ($K_{набл}$) принадлежит критической области, то нулевая гипотеза H_0 отклоняется в пользу конкурирующей H_1 ;

- если наблюдаемое значение критерия ($K_{\text{набл}}$) принадлежит области допустимых значений, то нулевую гипотезу H_0 нельзя отклонить.

Можно принять решение относительно нулевой гипотезы H_0 путем сравнения наблюдаемого ($K_{\text{набл}}$) и критического значений критерия ($K_{\text{кр}}$).

При правосторонней конкурирующей гипотезе:

- если $K_{\text{набл}} \leq K_{\text{кр}}$, то нулевую гипотезу H_0 нельзя отклонить;
- если $K_{\text{набл}} > K_{\text{кр}}$, то нулевая гипотеза H_0 отклоняется в пользу конкурирующей H_1 .

При левосторонней конкурирующей гипотезе:

- если $K_{\text{набл}} \geq -K_{\text{кр}}$, то нулевую гипотезу H_0 нельзя отклонить;
- если $K_{\text{набл}} < -K_{\text{кр}}$, то нулевая гипотеза H_0 отклоняется в пользу конкурирующей H_1 .

При двусторонней конкурирующей гипотезе:

- если $-K_{\text{кр}} \leq K_{\text{набл}} \leq K_{\text{кр}}$, то нулевую гипотезу H_0 нельзя отклонить;
- если $K_{\text{набл}} > K_{\text{кр}}$ или $K_{\text{набл}} < -K_{\text{кр}}$, то нулевая гипотеза H_0 отклоняется в пользу конкурирующей H_1 .

Алгоритм проверки статистических гипотез сводится к следующему:

- 1) сформулировать нулевую H_0 и альтернативную H_1 гипотезы;
- 2) выбрать уровень значимости α ;
- 3) в соответствии с видом выдвигаемой нулевой гипотезы H_0 выбрать статистический критерий для ее проверки, т.е. — специально подобранную случайную величину K , точное или приближенное распределение которой заранее известно;
- 4) по таблицам распределения случайной величины K , выбранной в качестве статистического критерия, найти критическое значение $K_{\text{кр}}$ (критическую точку или точки);
- 5) на основании выборочных данных по специальному алгоритму вычислить наблюдаемое значение критерия $K_{\text{набл}}$;
- 6) по виду конкурирующей гипотезы H_1 определить тип критической области;
- 7) определить, в какую область (допустимых значений или критическую) попадает наблюдаемое значение критерия $K_{\text{набл}}$, и в зависимости от этого — принять решение относительно нулевой гипотезы H_0 .

Следует заметить, что даже в том случае, если нулевую гипотезу H_0 нельзя отклонить, это не означает, что высказанное предположение о генеральной совокупности является единственно подходящим: просто ему не противоречат имеющиеся выборочные данные, однако таким же свойством наряду с высказанной могут обладать и другие гипотезы.

Можно интерпретировать результаты проверки нулевой гипотезы следующим образом:

- если в результате проверки нулевую гипотезу H_0 нельзя отклонить, то это означает, что имеющиеся выборочные данные не позволяют с достаточной уверенностью отклонить нулевую гипотезу H_0 , вероятность нулевой гипотезы H_0 больше α , а конкурирующей H_1 — меньше $1 - \alpha$;
- если в результате проверки нулевая гипотеза H_0 отклоняется в пользу конкурирующей H_1 , то имеющиеся выборочные данные не позволяют с достаточной уверенностью принять нулевую гипотезу H_0 , вероятность нулевой гипотезы H_0 меньше α , а конкурирующей H_1 — больше $1 - \alpha$.

Задание по теме «Основы статистики»

Для всех вопросов представленной ниже анкеты укажите его и шкалу измерения.

Вопрос	Тип	Шкала
Ваш возраст: 1) до 22 лет 2) 22-34 года 3) 35 – пенс.		
Ваше образование:		
Пол: муж. - жен. (нужное подчеркните)		
Ваша национальность (впишите)		
Национальность отца (впишите)		
Национальность матери (впишите)		
Ваше семейное положение: женат (замужем), холост, разведен		
Если женат (замужем), укажите национальность супруга (и)		
Имеете ли вы детей: да нет		
Количество детей в семье: 1) 1 2) 2-3, 3) свыше 4-х		
Курс обучения: 1 курс, 2 курс, 3 курс, 4 курс, 1 маг., 2 маг.		
Специальность, на которую обучаетесь:		
Дополнительная специальность, которую приобретаете:		
Имеете ли Вы стаж педагогической деятельности: да , нет.		
Стаж педагогической деятельности: 1) до 3 лет, 2) 4-10, 3) 11-20, 4) выше 20		
Работаете ли сейчас в системе образования: да нет		
Если работаете, укажите должность		
Имели ли Вы педагогическую практику: да нет		
Количество разных этносов и наций в Вашей педагогической работе (практике) (если она имела место быть) 1) 1 2) 2-3 3) 4-5 4) 6 и более		

Отметьте, какие из ниже перечисленных курсов Вы прослушали:

1) «Семейная психология»		
2) «Работа с семьёй»		
3) «Межкультурное общение»		
4) «Педагогическое общение»		
5) «Основы консультирования»		
6) «Мультикультурное образование»		

Статистические функции MS Excel

Ниже кратко описываются некоторые статистические функции MS Excel. Для наиболее сложных функций приводится также описание аргументов и примеры использования.

Примеры использования функций находятся в демонстрационном файле данных (excel_functions.xls), а также отражены на иллюстрациях.

Подсчет количества наблюдений

При первичной статистической обработке данных одной из начальных задач является подсчет количества наблюдений, в которых исследуемая характеристика приняла то или иное значение или ее значение попало в заданный интервал.

Для выполнения таких подсчетов можно использовать некоторые функции MS Excel – функции «семейства» COUNT и функцию FREQUENCY.

COUNT

Подсчитывает количество чисел в списке аргументов. Используется для получения количества числовых ячеек в интервалах или массивах ячеек.

В качестве аргумента этой функции выступает обычно диапазон ячеек, содержащий анализируемую информацию. Количество таких диапазонов (или отдельных ячеек) ограничено - не более 30.

COUNTA

Подсчитывает количество непустых значений в списке аргументов. Используется для подсчета количества ячеек с данными в интервале или массиве.

COUNTBLANK

Подсчитывает число пустых ячеек внутри интервала.

COUNTIF

Подсчитывает количество ячеек внутри интервала, удовлетворяющих заданному критерию. С ее помощью можно не только подсчитать количество ячеек, соответствующих критерию, но и построить частотное распределение результатов наблюдений. К сожалению, функция не допускает использования сложных критериев.

Синтаксис - COUNTIF(интервал;критерий)

Интервал - это интервал, в котором нужно подсчитать ячейки.

Критерий - это критерий в форме числа, выражения или текста, который определяет, какие ячейки надо подсчитывать. Например, критерий может быть выражен следующим образом: 4, "4", ">=4", "naine".

Использование этой функции для подсчета количества студентов разных форм обучения показано на рисунке (в строках 2-54 столбца В таблицы находится информация о форме обучения респондентов).

57	количество студентов		
58	стационар (1)	21	=COUNTIF(B2:B54;1)
59	открытый университет (2)	32	=COUNTIF(B2:B54;2)

Подсчитаем с помощью этой функции количество наблюдений, попавших в интервалы [0;5], (5;10], (10;15], (15;20]. Предварительно вычислим количество наблюдений, не превышающих 5, 10, 15 и 20, а затем простым вычитанием получим искомые значения.

	A	B	C	D	E	F
62			тревожность	фрустрация	агрессивность	ригидность
63		5	20	17	11	3
64		10	45	46	40	40
65		15	52	53	50	50
66		20	53	53	53	53
67						
68						
69			тревожность	фрустрация	агрессивность	ригидность
70		[0;5]	20	17	11	3
71		(5;10]	25	29	29	37
72		(10;15]	7	7	10	10
73		(15;20]	1	0	3	3

	A	B	C	D	E	F
62			тревожность	фрустрация	агрессивность	ригидность
63		5	=COUNTIF(C\$2:C\$54;"<=5")	=COUNTIF(D\$2:D\$54;"<=5")	=COUNTIF(E\$2:E\$54;"<=5")	=COUNTIF(F\$2:F\$54;"<=5")
64		10	=COUNTIF(C\$2:C\$54;"<=10")	=COUNTIF(D\$2:D\$54;"<=10")	=COUNTIF(E\$2:E\$54;"<=10")	=COUNTIF(F\$2:F\$54;"<=10")
65		15	=COUNTIF(C\$2:C\$54;"<=15")	=COUNTIF(D\$2:D\$54;"<=15")	=COUNTIF(E\$2:E\$54;"<=15")	=COUNTIF(F\$2:F\$54;"<=15")
66		20	=COUNTIF(C\$2:C\$54;"<=20")	=COUNTIF(D\$2:D\$54;"<=20")	=COUNTIF(E\$2:E\$54;"<=20")	=COUNTIF(F\$2:F\$54;"<=20")
67						
68						
69			тревожность	фрустрация	агрессивность	ригидность
70		[0;5]	=C63	=D63	=E63	=F63
71		(5;10]	=C64-C63	=D64-D63	=E64-E63	=F64-F63
72		(10;15]	=C65-C64	=D65-D64	=E65-E64	=F65-F64
73		(15;20]	=C66-C65	=D66-D65	=E66-E65	=F66-F65

Существенно проще эта задача решается с использованием следующей функции.

FREQUENCY

Вычисляет частоту появления значений в интервале значений и возвращает массив цифр. Может использоваться, например, для подсчета количества результатов, попадающих в заданные интервалы. Поскольку данная функция возвращает массив, она должна задаваться в качестве формулы массива.

Синтаксис - FREQUENCY(массив_данных;массив_карманов)

Массив_данных - это массив или ссылка на множество данных, для которых вычисляются частоты. Если массив_данных не содержит значений, то функция ЧАСТОТА возвращает массив нулей.

Массив_карманов - это массив или ссылка на множество интервалов, в которые группируются значения аргумента массив_данных. Если массив_карманов не содержит значений, то функция возвращает количество элементов в аргументе массив_данных.

- FREQUENCY вводится как формула массива после выделения интервала смежных ячеек, в которые нужно вернуть полученный массив распределения.
- Количество элементов в возвращаемом массиве на единицу больше количества элементов в аргументе массив_карманов.
- FREQUENCY игнорирует пустые ячейки и тексты.
- Формулы, которые возвращают массивы, должны быть введены как формулы массивов (ввод такой функции завершается нажатием комбинации клавиш Ctrl+Shift+Enter).
- Массив карманов можно задавать перечислением значений в фигурных скобках (например, {5;10;15;20}).

Вычислим частоту интервалов с помощью этой функции. Предварительно подготовим таблицу для вывода результатов. Выделим ячейки C77:C80 (мы знаем, что значений, превышающих 20, нет, поэтому нет необходимости выделять 4+1=5 ячеек, на одну больше, чем количество ячеек в диапазоне карманов), с помощью мастера введем функцию (используем смешанную и абсолютную адресацию ячеек во избежание необходимости многократного ввода формулы). Завершаем ввод выражения нажатием комбинации клавиш Ctrl+Shift+Enter. Затем копируем содержимое ячеек C77:C80. ниже представлена часть полученной таблицы (отображены формулы). Результат же, естественно, совпадает с полученным выше.

	A	B	C
76			тревожность
77		5	=FREQUENCY(C\$2:C\$54;\$B\$77:\$B\$80) :
78		10	=FREQUENCY(C\$2:C\$54;\$B\$77:\$B\$80) :
79		15	=FREQUENCY(C\$2:C\$54;\$B\$77:\$B\$80) :
80		20	=FREQUENCY(C\$2:C\$54;\$B\$77:\$B\$80) :

Средние значения и порядковые статистики

Важной группой являются функции, позволяющие определить вокруг какого значения (значений) группируются результаты наблюдений или место (порядок, ранг) наблюдения в массиве данных.

AVERAGE

Возвращает среднее (арифметическое) своих аргументов.

PERCENTILE

Возвращает k-ую перцентиль для значений из интервала. Эта функция используется для определения порога приемлемости. Например, можно принять решение экзаменовать только тех кандидатов, которые набрали баллов более чем 90-ая перцентиль. При вычислении используется линейная интерполяция между ближайшими к перцентили значениями переменной.

QUARTILE

Возвращает квартиль множества данных. Квартиль часто используются при анализе продаж, чтобы разбить генеральную совокупность на группы. Например, функцию можно использовать, чтобы найти 25 процентов наиболее доходных предприятий среди всех.

Синтаксис – QUARTILE (массив;часть)

Массив - это массив или интервал ячеек с числовыми значениями, для которых определяется значения квартилей.

Часть - это значение, которое нужно вернуть.

Если часть равна 0, то функция возвращает минимальное значение, 1 – первую квартиль (25-ую перцентиль), 2 – значение медианы (50-ую перцентиль), 3 – третью квартиль (75-ую перцентиль), 4 – максимальное значение.

MEDIAN

Возвращает медиану заданных чисел. Медиана - это число, которое является серединой множества чисел, то есть половина чисел имеют значения большие, чем медиана, а половина

чисел имеют значения меньшие, чем медиана.

MODE

Возвращает наиболее часто встречающееся или повторяющееся значение в массиве или интервале данных. Является мерой взаимного расположения значений.

В таблице представлен пример вычисления рассмотренных статистических характеристик.

среднее	6,79	=AVERAGE(A2:A54)
10% перцентиль	3,2	=PERCENTILE(A2:A54;0,1)
90% перцентиль	12	=PERCENTILE(A2:A54;0,9)
0 квартиль	1	=QUARTILE(A2:A54;0)
1 квартиль	4	=QUARTILE(A2:A54;1)
2 квартиль	6	=QUARTILE(A2:A54;2)
3 квартиль	8	=QUARTILE(A2:A54;3)
4 квартиль	18	=QUARTILE(A2:A54;4)
медиана	6	=MEDIAN(A2:A54)
мода	6	=MODE(A2:A54)

RANK

Возвращает ранг числа в списке чисел. Ранг числа - это его величина относительно других значений в списке. (Если список отсортировать, то ранг числа будет его позицией).

RANK присваивает повторяющимся числам одинаковый ранг. Однако наличие повторяющихся чисел влияет на ранг последующих чисел. Например, для списка целых, если число 10 появляется дважды и имеет ранг 5, то 11 будет иметь ранг 7 (и никакое число не будет иметь ранг 6).

Замечание. При использовании непараметрических методов статистики ранг повторяющихся чисел вычисляется как среднее арифметическое их мест (в примере выше числа 10 занимают места 5-ое и 6-ое, поэтому их ранг будет $(5+6)/2=5,5$).

PERCENTRANK

Возвращает категорию значения в наборе данных как процентное содержание в наборе данных. Эта функция используется для оценки относительного положения точки данных во множестве данных.

Характеристики разброса

MIN

Возвращает наименьшее значение в списке аргументов.

MAX

Возвращает наибольшее значение из набора значений.

CONFIDENCE

Возвращает доверительный интервал для среднего генеральной совокупности. Доверительный интервал - это интервал с обеих сторон от среднего выборки, в который с заданной вероятностью попадает среднее значение генеральной совокупности (на самом деле, один из аргументов этой функции задает вероятность непопадания в искомый интервал; обычно эта вероятность выбирается малой – 0,05 или 0,01).

При этом предполагается, что наблюдения подчиняются нормальному распределению.

SKEW

Возвращает асимметрию распределения. Асимметрия характеризует степень несимметричности распределения относительно его среднего.

KURT

Возвращает эксцесс множества данных. Эксцесс характеризует относительную остроконечность или сглаженность распределения по сравнению с нормальным распределением (для последнего эксцесс равен нулю). Положительный эксцесс обозначает относительно остроконечное распределение. Отрицательный эксцесс обозначает относительно сглаженное распределение.

VAR

Оценивает дисперсию по выборке.

STDEV

Оценивает стандартное отклонение по выборке. Стандартное отклонение - это мера того, насколько широко разбросаны точки данных относительно их среднего.

VARP

Вычисляет дисперсию для генеральной совокупности.

STDEVP

Вычисляет стандартное отклонение по генеральной совокупности. Стандартное отклонение - это мера того, насколько широко разбросаны точки данных относительно их среднего.

В таблице представлен пример вычисления характеристик разброса.

минимум	1	=MIN(A2:A54)
максимум	18	=MAX(A2:A54)
CONFIDENCE	0,89	=CONFIDENCE(0,05;G21;53)
асимметрия	1,05	=SKEW(A2:A54)
эксцесс	1,44	=KURT(A2:A54)
выборочная дисперсия	10,86	=VAR(A2:A54)
стандартное отклонение по выборке	3,30	=STDEV(A2:A54)
генеральная дисперсия	10,66	=VARP(A2:A54)
стандартное отклонение по генеральной совокупности	3,26	=STDEVP(A2:A54)

Перекодирование полученных значений

Под перекодированием переменных подразумевается замена величин, полученных непосредственно в результате исследования, на другие значения. Это бывает необходимо, когда первоначальное разнообразие исходных данных не нужно для последующего анализа. Часто такая необходимость обусловлена изменением шкалы измерений, например, при переходе от «сырых» баллов к процентилям.

К сожалению, в MS Excel отсутствует инструментарий для выполнения перекодирования. Однако здесь могут помочь встроенные функции

VLOOKUP (HLOOKUP)

Ищет значение в крайнем левом столбце таблицы и возвращает значение в той же строке из указанного столбца таблицы.

Синтаксис

VLOOKUP(искомое_значение;инфо_таблица;номер_столбца;интервальный_просмотр)

Искомое_значение - это значение, которое должно быть найдено в первом столбце массива. Искомое_значение может быть значением, ссылкой или текстовой строкой.

Инфо_таблица - это таблица с информацией, в которой ищутся данные.

- Если интервальный_просмотр имеет значение ИСТИНА, то значения в первом столбце аргумента инфо_таблица должны быть расположены в возрастающем порядке: ..., -2, -1, 0, 1, 2, ..., A-Z, ЛОЖЬ, ИСТИНА; в противном случае функция может выдать неправильный результат. Если интервальный_просмотр имеет значение ЛОЖЬ, то инфо_таблица не обязана быть сортированной.
- Данные можно упорядочить следующим образом: в меню Data выбрать команду Sort и установить переключатель Ascending.
- Значения в первом столбце аргумента инфо_таблица могут быть текстами, числами или логическими значениями.
- Регистр не учитывается (т. е. строчные и заглавные буквы не различаются).

Номер_столбца - это номер столбца в массиве инфо_таблица, в котором должно быть

найденое соответствующее значение. Если номер_столбца равен 1, то возвращается значение из первого столбца аргумента инфо_таблица; если номер_столбца равен 2, то возвращается значение из второго столбца аргумента инфо_таблица и так далее.

Интервальный_просмотр - это логическое значение, которое определяет, нужно ли, чтобы искала точное или приближенное соответствие. Если этот аргумент имеет значение ИСТИНА или опущен, то возвращается приблизительно соответствующее значение; другими словами, если точное соответствие не найдено, то возвращается наибольшее значение, которое меньше, чем искомое_значение. Если этот аргумент имеет значение ЛОЖЬ, то функция ВПР ищет точное соответствие.

Функция HLOOKUP полностью идентична функции VLOOKUP, но для ее корректной работы необходимо, чтобы инфо_таблица имела горизонтальное расположение (функция HLOOKUP ищет значения в первой строчке информационной таблицы).

Продемонстрируем применение этой функции на примерах.

Пример 1

В тесте Айзенка полученные баллы (от 0 до 20) разбивают на три диапазона: от 0 до 7 – низкий уровень, от 8 до 14 – средний, от 15 до 20 – высокий.

В этом случае инфо_таблица будет иметь следующую структуру:

«старое» значение	«новое» значение
0	низкий
7	низкий
8	средний
14	средний
15	высокий
20	высокий

Формат нашей функции (вводим ее в соседний с исходными данными столбик, например, в ячейку B2, если она, конечно, пустая), осуществляющей перекодирование, будет иметь вид:

=VLOOKUP(A2;\$E\$2:\$F\$7;2;1)

\$E\$2:\$F\$7– адрес инфо_таблицы, ссылки используем абсолютные, поскольку будем формулу копировать вниз.

Последний аргумент (1) означает, что ищется приближенное соответствие (так, например, 5 баллов будет соответствовать низкому уровню, поскольку ближайшее к 5 значение снизу равно 0). Напомним, что в этом случае инфо_таблица должна быть отсортирована по возрастанию.

Пример 2

В тесте «Диагностика типологий психологической защиты» «сырые» баллы переводятся в проценты. В инструкции по обработке данных этого теста приводится фактически соответствующая инфо_таблица (приведем ее часть):

«сырые» баллы	Отрицание, процентиля
0	3
1	13
2	27
3	39
4	50
5	61
6	79
7	84
8	90
9	97
10	98
11	99

В этом случае отсутствуют пропуски значений в левом столбике, и мы можем последний аргумент рассматриваемой функции приравнять к нулю.

Вычисление корреляций между переменными

CORREL

Возвращает коэффициент (линейной) корреляции Пирсона между двумя интервалами ячеек. Интервалы должны содержать равное количество значений.

Например, коэффициент корреляции для показателей «тревожность» и «фрустрация»
 $=\text{CORREL}(B2:B54;C2:C54) = 0,440$.

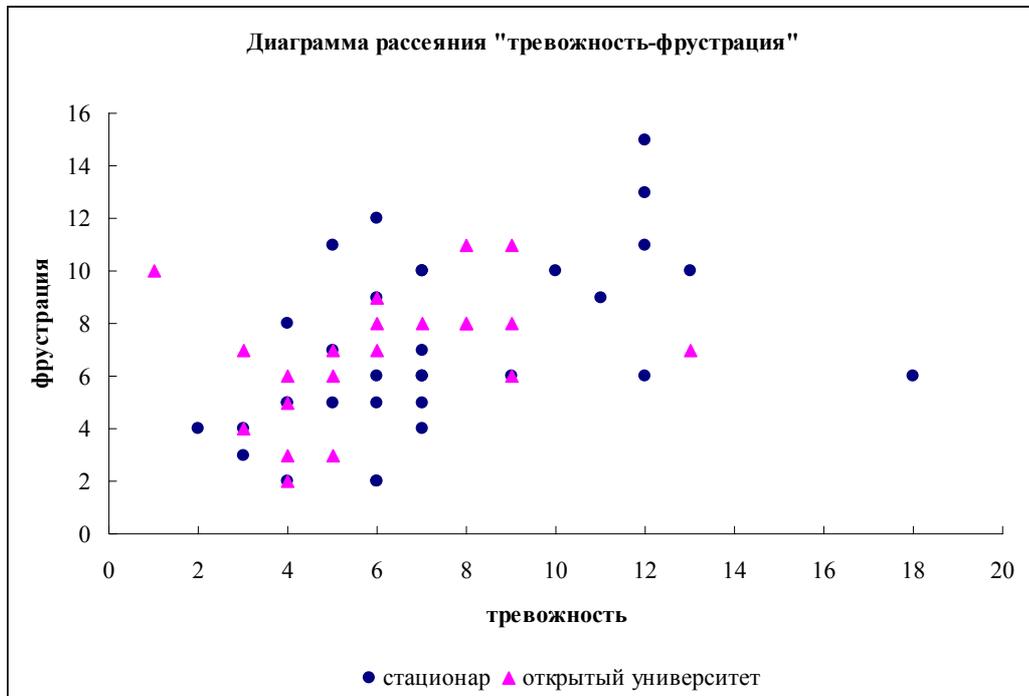
Для проверки значимости коэффициента корреляции можно воспользоваться соответствующими статистическими таблицами. Можно воспользоваться также следующим методом. Вычислим величину

$$t_s = |r| \cdot \sqrt{n-3} = 0,440 \cdot \sqrt{53-3} \approx 3,11.$$

Она распределена по закону Стьюдента. Из сравнения полученного значения t_s с критическими $t_{кр} = 1,96$ ($p=0,05$) и $t_{кр} = 2,58$ ($p=0,01$) следует, что корреляция между переменными значима с уровнем значимости $p < 0,01$.

Наличие корреляционной связи между переменными принято иллюстрировать с помощью диаграммы рассеяния (XY (Scatter)).

Для рассматриваемых данных она представлена на рисунке.



Проверка нормальности распределения

Проверка гипотезы о нормальном характере распределения исследуемых признаков - важный этап статистического анализа данных. Существенная часть статистических методов применима только в случае, когда распределения показателей подчинены нормальному закону.

К сожалению, в программе MS Excel отсутствуют встроенные процедуры проверки этой гипотезы. Выйти из этого затруднения можно применив один из простых методов – метод, основанный на формулах Е.И. Пустыльника.

По этой методике рассчитываются критические значения для показателей асимметрии (А) и эксцесса (Е):

$$A_{кр} = 3 \cdot \sqrt{\frac{6(n-1)}{(n+1)(n+3)}}$$

$$E_{кр} = 5 \cdot \sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}$$

$$A_3 = \frac{\sum_i (x_i - x_{срел})^3 \cdot m_i}{n \cdot \sigma^3}$$

$$E_3 = \frac{\sum_i (x_i - x_{срел})^4 \cdot m_i}{n \cdot \sigma^4} - 3,$$

$$\sigma = \sqrt{\frac{\sum_i (x_i - x_{срел})^2 \cdot m_i}{n-1}}, \quad x_{срел} = \frac{\sum_i x_i \cdot m_i}{n}$$

где n – общее количество наблюдений, m_i – частота (если по наблюдениям не построен вариационный ряд, то m_i = 0).

Если выполняются оба неравенства Аэ Акр и Еэ Екр, то распределение изучаемого признака не отличается от нормального распределения.

В файле excel_stat_hypothesis.xls представлены результаты применения этой методике для

некоторых переменных. Также для визуального сравнения экспериментального и теоретического распределений построен соответствующий график (использовалась функция NORMDIST).

Распределение параметра «фрустрация» близко к нормальному, а параметра «тревожность» - существенно отличается от нормального.

При отклонении от нормальности параметрические критерии (например, критерий Стьюдента, коэффициент корреляции Пирсона) неприменимы.

Покажем пример вычисления коэффициента корреляции рангов (Спирмена) – одной из альтернатив коэффициентам Пирсона в случае отсутствия нормального распределения у признаков.

Коэффициент корреляции рангов (Спирмена)

Коэффициент корреляции рангов используется для определения тесноты связи между признаками в случае их количественного представления. Применим и тогда, когда признаки не распределены по нормальному закону.

Для определения рангового коэффициента корреляции ранжируют все значения обеих переменных. Для связанных рангов проводят пересчет.

Пример. Четыре наблюдения дали одинаковый результат. Только три наблюдения превышают рассматриваемые, то есть эта четверка занимает места с 4-го по 7-ое включительно. Тогда им приписывается одинаковый ранг, равный средней арифметической этих мест, то есть $(4+5+6+7)/4=5,5$.

Коэффициент корреляции рангов вычисляется по формуле

$$\rho = 1 - \frac{6 \cdot \sum d_i^2}{n(n^2 - 1)},$$

где n – число коррелируемых пар значений признаков X и Y ;

$d_i = X_i - Y_i$ – разность рангов i -го испытуемого в обоих ранговых рядах.

Коэффициент корреляции рангов принимает значения от -1 до 1. если никакой связи между признаками не существует, то коэффициент равен 0. При положительной корреляции при увеличении числовых значений одной из переменных соответственно увеличиваются значения другой. При отрицательной корреляции увеличению значений одной переменной соответствует уменьшение значений второй.

В файле excel_stat_hypothesis.xls представлены результаты вычисления коэффициента ранговой корреляции признаков. Для пересчета рангов по MS Excel в ранги по Спирмену используется вспомогательная таблица.

Полученное значение коэффициента корреляции (0,513) превышает критическое (0,40) для уровня значимости 0,01 (см. справочные таблицы). Следовательно, между показателями существует статистически значимая корреляционная связь.

Критерий согласия распределений χ^2 –Пирсона

Критерий χ^2 применяется в двух случаях.

1. Для сопоставления эмпирического распределения признака с теоретическим.
2. Для сопоставления двух и более эмпирических распределений одного и того же признака друг с другом.

Рассмотрим использование метода для сравнения сопоставления двух эмпирических

распределений.

Попытаемся сравнить распределение студентов обеих форм обучения по уровню агрессивности.

Эмпирические частоты

	низкий	средний	высокий	всего
стационар	6	14	1	21
открытый университет	16	12	4	32
всего	22	26	5	53

Для этого вычислим теоретические частоты распределения в предположении, что при отсутствии различий в распределениях респондентов по уровням количество наблюдений было бы пропорционально доле респондентов в группе.

Теоретические частоты

	низкий	средний	высокий	всего
стационар	$21 \cdot 22 / 53 = 8,72$	$21 \cdot 26 / 53 = 10,30$	$21 \cdot 5 / 53 = 1,98$	21
открытый университет	$32 \cdot 22 / 53 = 13,28$	$32 \cdot 26 / 53 = 15,70$	$32 \cdot 5 / 53 = 3,02$	32
всего	22	26	5	53

Затем вычисляется величина

$$\chi^2_{\text{эмп}} = \sum_{i=1}^k \frac{(m_{\text{эмп } i} - m_{\text{теор } i})^2}{m_{\text{теор } i}},$$

где $m_{\text{эмп } i}$ – эмпирическая (экспериментальная) частота по i -му разряду признака;

$m_{\text{теор } i}$ – теоретическая частота по i -му разряду признака;

i – порядковый номер разряда;

k – количество разрядов признака (в нашем примере $k = 3$).

В файле excel_stat_hypothesis.xls представлены результаты вычисления.

Получено значение 2,66. Оно сравнивается с табличными $t_{\text{кр}} = 5,99$ ($p = 0,05$) и $t_{\text{кр}} = 9,21$ ($p = 0,01$) для степеней свободы, равной $3-1=2$. Полученное значение меньше критических. Поэтому нет основания отвергать нулевую гипотезу, и распределения респондентов по уровням агрессивности не отличаются для обеих сравниваемых групп.

Критерий имеет ряд ограничений:

1. Объем выборки должен быть достаточно большим ($n \geq 30$). При малых объемах критерий дает очень приближенные значения. При увеличении объема точность критерия повышается.
2. Теоретическая частота для каждой ячейки таблицы не должна быть меньше 5 (в нашем иллюстративном примере это ограничение не выполняется). При меньших частотах часто прибегают к укрупнению интервалов.
3. Выбранные разряды должны охватывать весь диапазон вариативности признака. Группировка на разряды должна быть одинаковой во всех сопоставляемых распределениях.
4. Необходимо вносить поправку на «непрерывность» при сопоставлении распределений

признаков, которые принимают всего два значения. В этом случае в числителе приведенной выше формулы должно присутствовать следующее выражение:

$$\left(m_{\text{эксп } i} - m_{\text{теор } i} \right)^2.$$

5. Сумма наблюдений по разрядам всегда должна быть равна общему количеству наблюдений.

Задание по теме «Статистические функции MS Excel»

Для указанного набора данных:

- 1) вычислить количество наблюдений;
- 2) вычислить основные статистические характеристики (среднее, минимум, максимум, выборочную дисперсию и стандартное отклонение, моду, медиану, асимметрию и эксцесс);
- 3) построить интервальный вариационный ряд (нижняя граница первого интервала – 200, ширина интервала – 30); использовать функцию COUNT;
- 4) выполнить задание 3 с использованием функции FREQUENCY;
- 5) полученный вариационный ряд изобразить графически;
- 6) методом Е.И. Пустыльника проверить гипотезу о нормальном характере распределения данных;
- 7) сравнить (с помощью графика) вариационный ряд и кривую нормального распределения (использовать выборочное среднее и дисперсию).

При выполнении задания использовать только статистические функции MS Excel.

200	271	295	225	268	245	275	248	250	270
310	315	345	350	270	270	295	360	300	285
270	260	210	274	300	275	300	260	260	260
298	281	284	291	280	235	289	240	280	230
300	290	289	292	360	300	365	290	330	290
327	295	250	337	249	350	271	298	300	345
232	235	248	273	237	256	255	238	220	220
300	275	315	300	300	261	265	262	273	355
325	335	320	300	310	310	300	330	268	300
280	340	280	260	320	345	350	279	258	260

Сводные таблицы – PivotTables

Сводные таблицы – одно из наиболее мощных аналитических средств Excel. Сводная таблица мгновенно преобразует огромное количество информации в небольшую аккуратную таблицу.

Сводная таблица – это динамическая итоговая таблица, обобщающая результаты базы данных, т.е. других таблиц, которые могут быть расположены как в этом же рабочем листе, так и во внешних файлах.

Сводная таблица способна генерировать различные обобщения исходных данных, например распределение частот или итоговые суммы нескольких различных измерений данных. Кроме того, с помощью сводной таблицы можно выводить промежуточные обобщающие данные с любым уровнем детализации. Важнейшее свойство сводных таблиц — их интерактивность. После создания сводной таблицы информацию можно переупорядочивать самыми разными способами. Можно даже создавать пользовательские группы обобщающих данных (например, объединять итоговые данные по северным и восточным регионам).

Сводная таблица названа так (pivot — стержень) потому, что заголовки строк и столбцов таблицы можно как бы вращать вокруг основной области данных, предоставляя обобщенные данные под различными углами зрения.

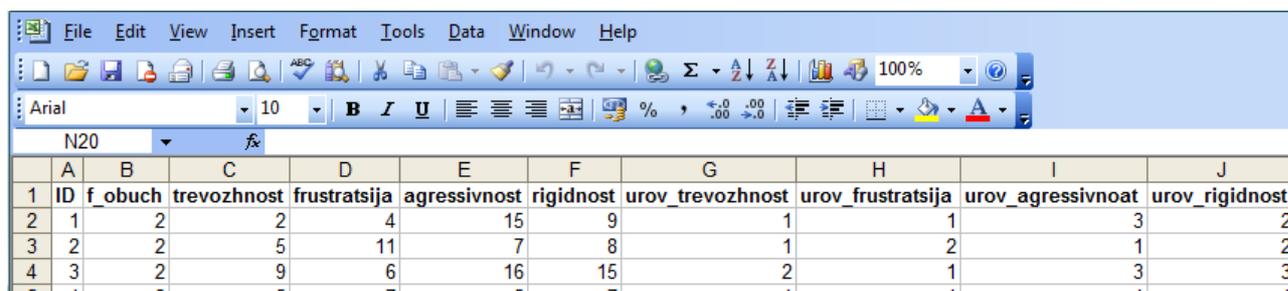
Один из недостатков использования сводных таблиц вместо обобщающих отчетов, основанных на формулах, состоит в том, что сводная таблица не обновляется автоматически при изменении исходных данных. Однако этот недостаток не представляет серьезной проблемы, поскольку для обновления сводной таблицы достаточно одного щелчка на кнопке Refresh Data, расположенной на панели инструментов PivotTables.

С точки зрения возможностей статистической обработки данных психолого-педагогических исследований сводные таблицы в основном применимы для получения данных описательного характера (частоты, средние, дисперсия, наименьшее и наибольшее значения и т.п.). Представляется весьма полезной предоставляемая сводными таблицами возможность получения этой информации для всевозможных подгрупп исследуемых респондентов.

Демонстрация основных процедур построения и модификации сводной таблицы представлено в видеофрагментах. Источник данных – файл pivot_demo.xls, содержащий данные обработки результатов теста Айзенка.

Создание сводной таблицы

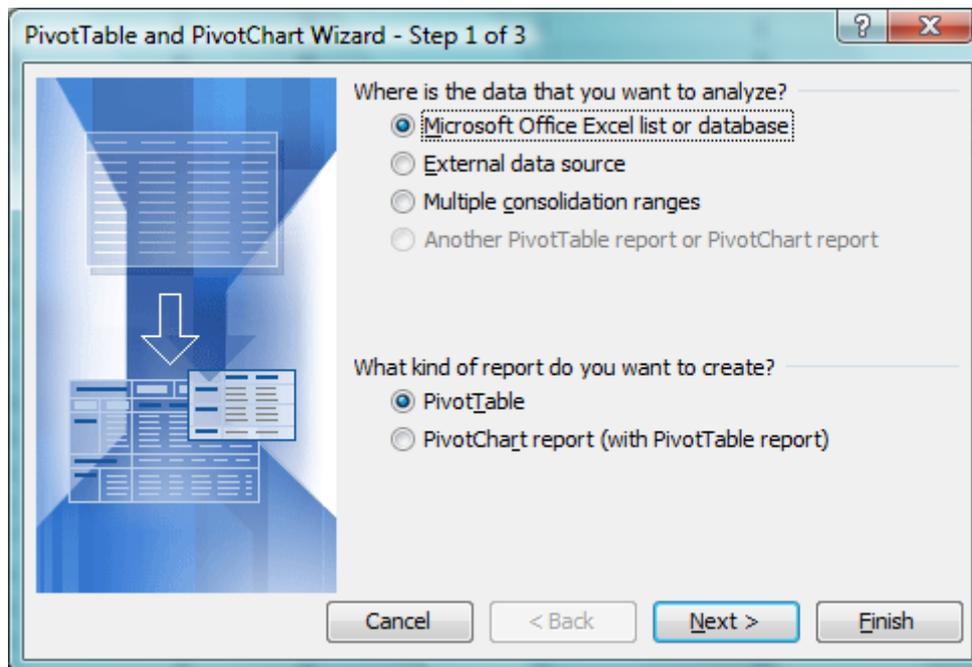
Сводную таблицу можно создать с помощью мастера сводных таблиц. Запустите мастер сводных таблиц, выбрав команду Data -> PivotTable. Затем выполняйте шаги, описанные ниже.



	A	B	C	D	E	F	G	H	I	J			
1	ID	f_obuch	trevozhnost	frustratsija	agressivnost	rigidnost	urov_trevozhnost	urov_frustratsija	urov_agressivnoat	urov_rigidnost			
2	1	2	2	4	15	9	1	1		3	2		
3	2	2		5	11	7	8	1	2		1	2	
4	3	2		9	6	16	15	2		1		3	3

Шаг 1: определение расположения исходных данных

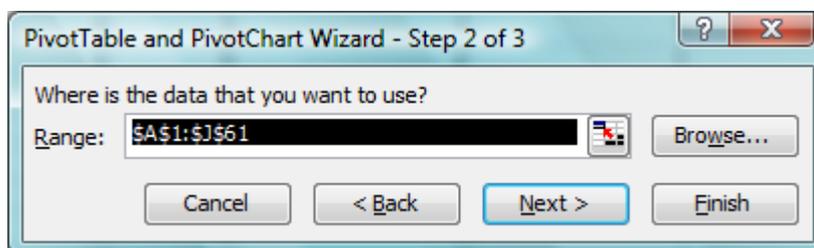
Активизируйте лист Данные рабочей книги, показанной на рисунке. Расположите указатель ячейки в любом месте таблицы исходных данных и выберите команду Data -> PivotTable and PivotChart Report; появится диалоговое окно первого шага мастера сводных таблиц.



На этом шаге задаются исходные данные. Excel предоставляет гибкие возможности задания исходных данных сводной таблицы. В нашем примере используется база данных, расположенная в рабочем листе.

Шаг 2: определение данных

Чтобы перейти к следующему шагу мастера, щелкните на кнопке Next. Появится диалоговое окно второго шага мастера. Помните: внешний вид диалогового окна зависит от параметров, установленных в первом диалоговом окне мастера. На рисунке показано окно, выводимое, если на шаге 1 был установлен переключатель в списке или базе данных Microsoft Excel.

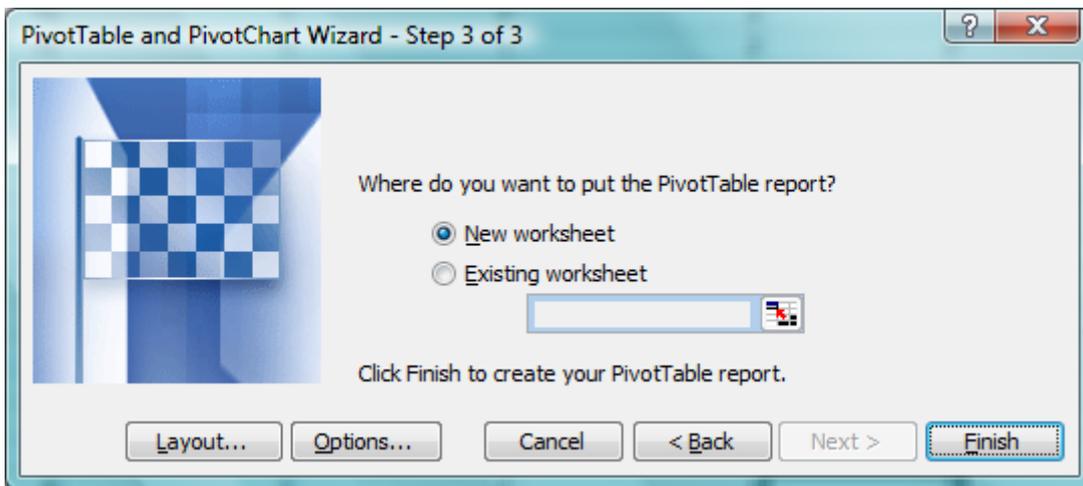


Если при выборе команды Data - > PivotTable указатель ячейки был расположен в любом месте базы данных рабочего листа, то мастер на втором шаге автоматически идентифицирует диапазон базы данных. С помощью кнопки Обзор можно открыть любую другую рабочую книгу и выбрать любой диапазон.

Перейдем к шагу 3 мастера (Next).

Шаг 3: завершение мастера сводных таблиц

На рисунке показано диалоговое окно последнего шага мастера. В нем задается расположение сводной таблицы. Если установить переключатель новый лист, то мастер создает новый рабочий лист и вставляет в него сводную таблицу. Если же установить переключатель существующий лист, то сводная таблица создается в текущем рабочем листе; в этом случае можно задать также расположение начальной ячейки сводной таблицы.

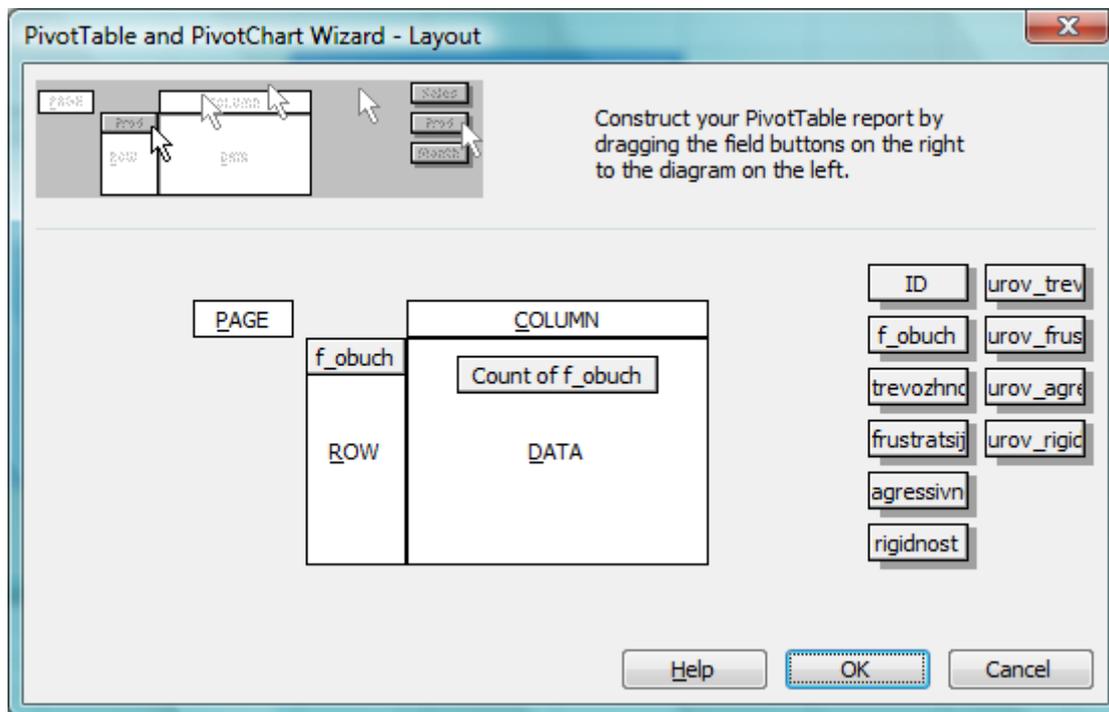


На шаге 3 можно, щелкнув на кнопке Параметры, определить внешний вид сводной таблицы. Эти же параметры можно установить в любое время после создания сводной таблицы, поэтому пока что оставьте их по умолчанию.

Фактическую раскладку сводной таблицы можно определить двумя способами.

- Щелкнуть на кнопке Layout (Макет), расположенной в диалоговом окне третьего шага мастера сводных таблиц. Появляющееся при этом диалоговое окно предоставляет средства определения раскладки сводной таблицы.
- Щелкнув на кнопке Finish, создать пустую сводную таблицу. Затем сводную таблицу можно заполнить с помощью панели инструментов PivotTable.

Будем использовать первый способ.



На рисунке показано диалоговое окно, выводимое в результате щелчка на кнопке Макет, расположенной в последнем окне мастера. Поля базы данных представлены кнопками вдоль правой стороны диалогового окна. Для создания полей сводной таблицы нужно просто перетащить кнопки в соответствующие области макета сводной таблицы, расположенного в центре диалогового окна.

Макет сводной таблицы имеет четыре области.

- Page (Страница). Значения в этом поле выводятся в сводной таблице как элементы страницы.

- Row (Строка). Значения в этом поле выводятся в сводной таблице как элементы строки.
- Data (Данные). Обобщенное поле сводной таблицы.
- Column (Столбец). Значения в этом поле выводятся в сводной таблице как элементы столбца.

В любую из этих областей можно перетащить произвольное количество кнопок полей. Использовать все поля не обязательно. Неиспользуемые поля просто не выводятся в сводной таблице.

При перетаскивании кнопки поля в область Data мастер применяет к полю функцию SUM, если оно содержит числовые значения, и функцию COUNT — если нечисловые.

В процессе определения раскладки сводной таблицы можно дважды щелкнуть на кнопке поля и настроить его параметры (Options). Например, можно задать характер обобщения поля: суммирование, подсчет или любую другую функцию. Можно также задать, какие элементы поля должны быть скрыты или опущены. Если вы перетащили кнопку поля в неправильную область, то просто перетащите ее за пределы макета, удалив таким образом поле. Поля можно настраивать в любое время после создания сводной таблицы.

Показанная на рисунке выше сводная таблица (вернее, пока только структура таблицы) будет выводить информацию о количестве студентов на разных формах обучения (напомним, что мы закодировали эту информацию: 1- стационар, 2 – открытый университет).

	A	B
1	Drop Page Fields Here	
2		
3	Count of f obuch	
4	f obuch	Total
5		1 24
6		2 36
7	Grand Total	60

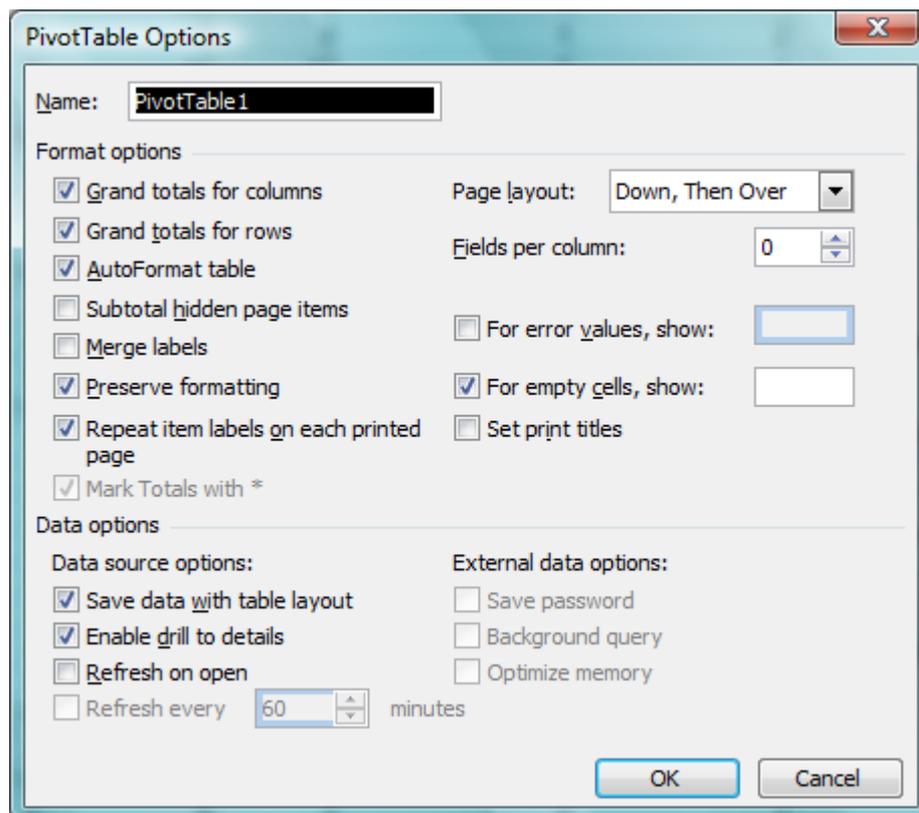
Параметры сводных таблиц

Пользователь может задавать различные параметры, определяющие внешний вид и принцип действия сводной таблицы. Чтобы получить доступ к параметрам сводной таблицы, щелкните на кнопке Options, расположенной в последнем диалоговом окне мастера сводных таблиц. При этом активизируется диалоговое окно PivotTable Options.

Активизировать это же диалоговое окно можно и после создания сводной таблицы. Для этого щелкните правой кнопкой мыши на любой ячейке сводной таблицы и в контекстном меню выберите команду **Параметры таблицы**. Ниже приведен список параметров, задаваемых в окне PivotTable Options.

- Имя. В этом поле можно задать имя сводной таблицы;
- общая сумма по столбцам. Если установить этот флажок, то вычисляются общие суммы элементов, выводимых в столбцах.
- общая сумма по строкам. Если установить этот флажок, то вычисляются общие суммы элементов, выводимых в строках.
- автоформат. Установите этот флажок, если хотите применить к сводной таблице один из автоформатов. Примененный автоформат сохраняется, даже если переупорядочивается раскладка таблицы.
- включать скрытые значения. Если установлен этот флажок, то скрытые элементы включаются в промежуточные итоги полей страницы.

- объединять ячейки заголовков. Если установить этот флажок, то будут объединены ячейки заголовков строк и столбцов.
- сохранять форматирование. При обновлениях сводной таблицы сохраняются все примененные до этого параметры форматирования.
- повторять подписи на каждой странице печати. При распечатке сводной таблицы заголовки строк выводятся на каждой странице.
- помечать итоги*. Этот флажок доступен, только если сводная таблица сгенерирована на основе исходных данных OLAP (On-line Analytical Processing — аналитическая интерактивная обработка в реальном времени). Если флажок установлен, то звездочка выводится после каждого промежуточного и общего итога, учитывающего значения скрытых (и, естественно, видимых) элементов.



- макет страницы. В этом раскрывающемся списке можно задать последовательность вывода полей страницы.
- число полей в столбце. В этом поле можно задать количество полей страницы, выводимых перед началом другой строки полей страницы.
- для ошибок отображать. Задание значения, выводимого в ошибочных ячейках сводной таблицы.
- для пустых полей отображать. Задание значения, выводимого в пустых ячейках сводной таблицы.
- печать заголовков. В распечатке отчета сводной таблицы заголовки столбцов выводятся вверху каждой страницы.
- сохранить данные вместе с таблицей. Если установлен этот флажок, то Excel сохраняет дополнительную копию данных (так называемый кэш сводной таблицы), что позволяет ускорить вычисление сводной таблицы при изменении раскладки. Если в компьютере ощущается недостаток памяти, то флажок следует снять (что немного замедлит обновление таблицы).
- развертывание разрешено. Если установлен этот флажок, то, дважды щелкнув на ячейке

сводной таблицы, можно просмотреть записи, использованные при вычислении обобщенного значения.

- обновить при открытии. Сводная таблица обновляется при каждом открытии рабочей книги.
- обновлять каждые ... мин.. Если компьютер соединен с внешней базой данных, то в этом поле можно задать временной интервал обновления сводной таблицы при открытой рабочей книге.
- сохранить пароль. Если используется внешняя база данных, требующая пароль, то в Excel он сохраняется таким образом, что его не нужно вводить повторно вручную.
- фоновый запрос. Если этот флажок установлен, то запрос к внешней базе данных выполняется в фоновом режиме. Пользователь в это время может продолжать работать со сводной таблицей.
- оптимизировать память. Уменьшение необходимого объема памяти при обновлении сводной таблицы путем выполнения запросов к внешней базе данных.

Примеры. Группирование элементов сводной таблицы

Одно из наиболее полезных свойств сводных таблиц — возможность объединять элементы в группы. Чтобы сгруппировать элементы сводной таблицы, нужно выделить их, щелкнуть на них правой кнопкой мыши и выбрать в контекстном меню команду Группа и структура'=>Группировать.

Этой возможностью удобно пользоваться для формирования интервальных вариационных рядов.

Но сначала покажем, как сформировать дискретный ряд.

Drop Page Fields Here	
Count of trevozhnost	
trevozhnost	Total
1	1
2	1
3	4
4	8
5	9
6	10
7	8
8	5
9	4
10	1
11	1
12	5
13	2
18	1
Grand Total	60

PivotTable and PivotChart Wizard - Layout

Construct your layout by dragging the fields to the diagram.

PAGE: trevozhnost

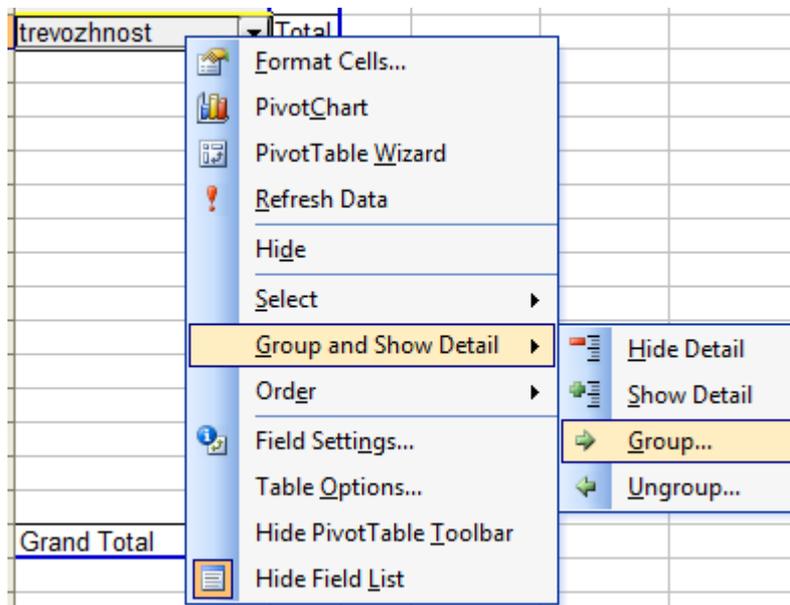
COLUMN: Count of trevozhnost

ROW: trevozhnost

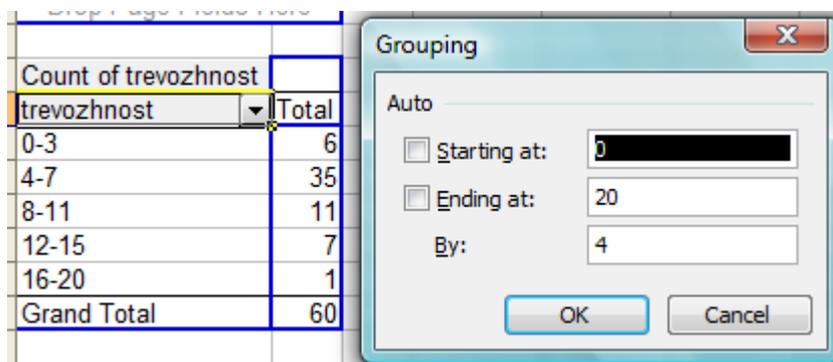
DATA: Count of trevozhnost

На рисунке показан требуемый макет и сама сводная таблица.

Затем выделим заголовок левого столбика и выполним команду Group....- > Group...

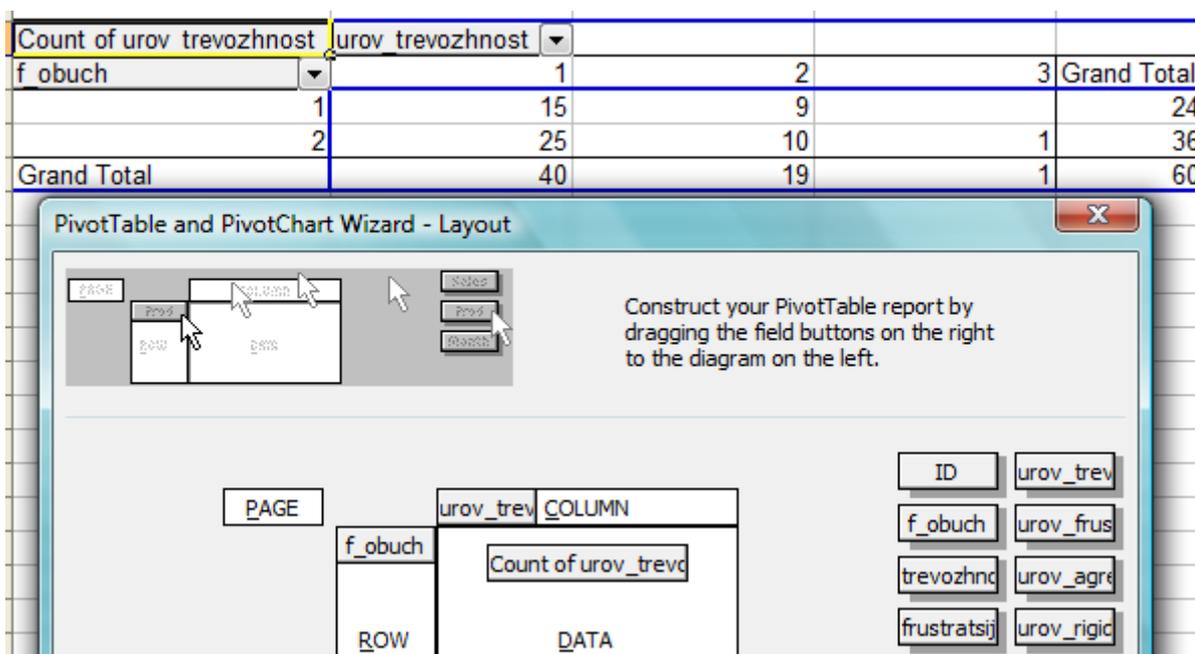


В появившемся диалоговом окне зададим параметры группировки:

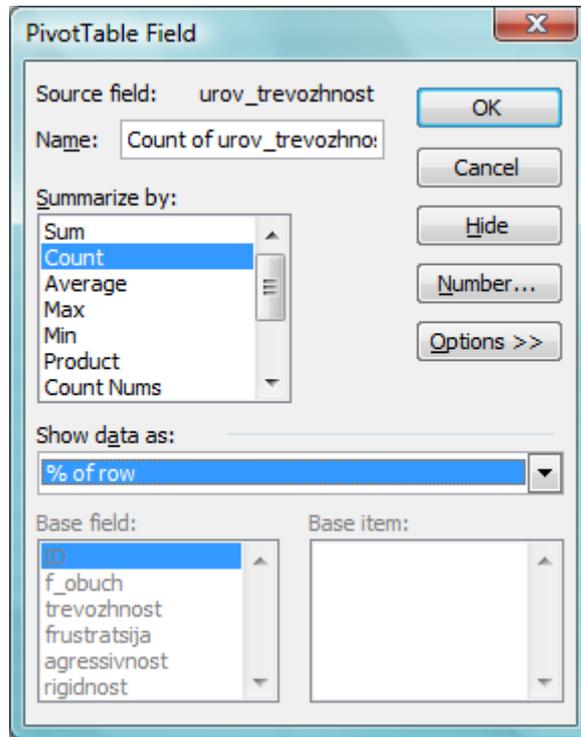


На рисунке показаны настройки группировки и сам ее результат.

Покажем (на рисунке) пример создания таблиц сопряженности (в терминологии SPSS).



Для изменения формы представления информации можно воспользоваться настройками поля сводной таблицы. Надо дважды щелкнуть по полю Count of urov_trevozhnost и в диалоговых окнах (не забудьте нажать кнопку Options) установить указанные на рисунках параметры.



Мы получим распределение студентов по уровням тревожности, выраженное в относительной мере (в процентах).

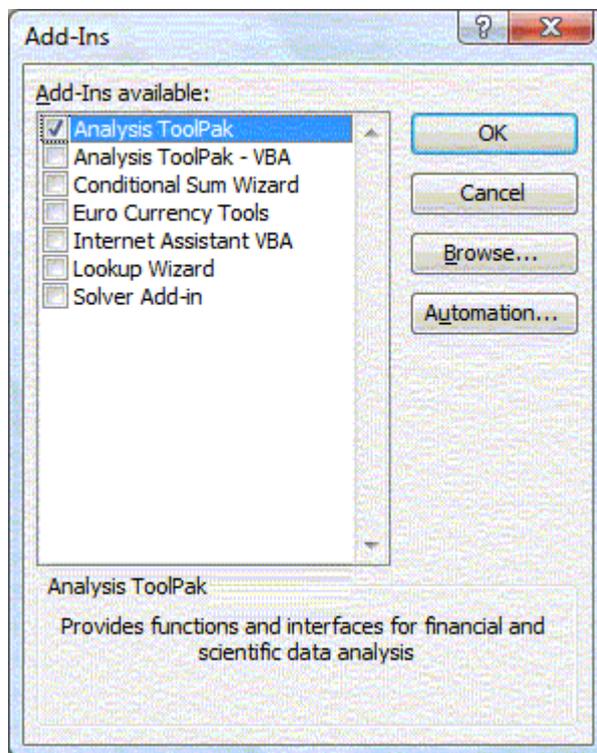
Count of urov trevozhnost	urov trevozhnost			
f_obuch	1	2	3	Grand Total
1	62,50%	37,50%	0,00%	100,00%
2	69,44%	27,78%	2,78%	100,00%
Grand Total	66,67%	31,67%	1,67%	100,00%

С помощью сводных таблиц легко получить данные описательной статистики.

	f_obuch		
Data	1	2	Grand Total
Average of trevozhnost	6,208333333	7,222222222	6,816666667
Min of trevozhnost	1	2	1
Max of trevozhnost	13	18	18
StdDevp of trevozhnost	2,798499102	3,342579768	3,175382742

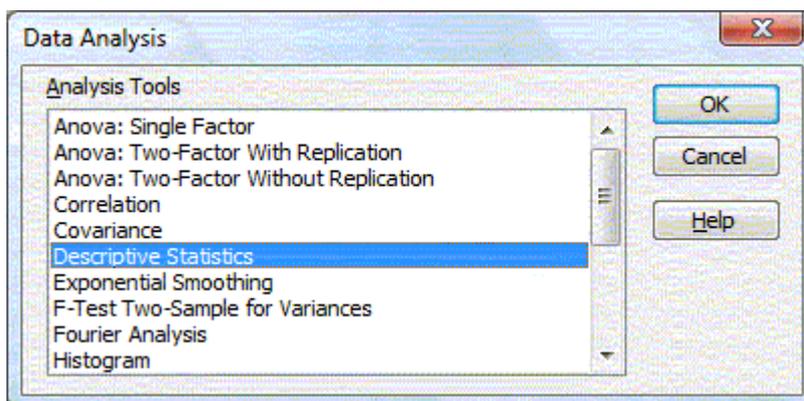
Пакет анализа (Analysis ToolPack)

Для того чтобы можно было использовать инструменты пакета анализа (Analysis ToolPack), необходимо подключить соответствующую надстройку. Для этого выполните следующую последовательность команд: Tools → Add-Ins... и в появившемся диалоговом окне отметьте надстройку Analysis ToolPack.



После этого в перечне команд меню Tools появится команда Data Analysis, а также расширится перечень доступных для использования встроенных (финансовых и инженерных, а также некоторых статистических) функций MS Excel.

Запуск требуемого инструмента пакета анализа осуществляется последовательным выполнением команд меню MS Excel: Tools → Data Analysis с последующим выбором инструмента из списка.



Настройка параметров любого инструмента анализа осуществляется с помощью диалогового окна.

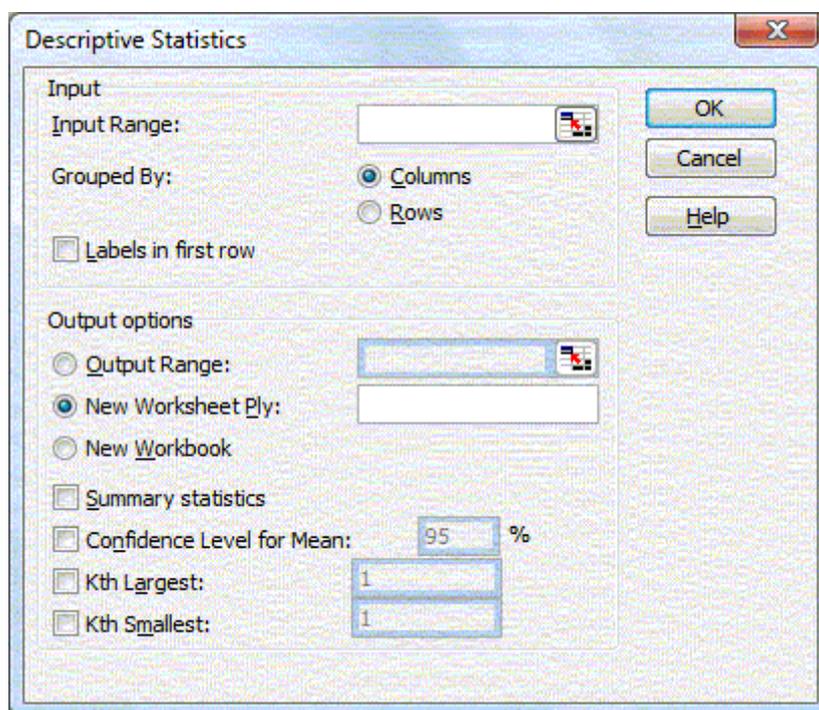
В пакет анализа входят почти 20 инструментов:

- дисперсионный анализ (три вида);
- корреляционный анализ;
- ковариационный анализ;
- описательная статистика;

- экспоненциальное сглаживание;
- F-тест;
- анализ Фурье;
- гистограмма;
- скользящее среднее;
- генерация случайных чисел;
- ранг и перцентиль;
- регрессия;
- выборка;
- t-тест (три вида);
- z-тест.

Мы рассмотрим использование некоторых из них.

Описательная статистика (Descriptive Statistics)



Вверху диалогового окна находятся настройки ввода данных.

Input Range – Входной диапазон

Указывается диапазон рабочего листа, в котором находятся исследуемые данные. Если диапазон содержит метки строк или столбцов, то необходимо указать на это, установив флажок *Labels in first row/column* (Метки в первой строке/столбце).

Grouped By – Группирование

Указывается способ размещения исходных данных – в столбцах или строках.

Остальные элементы диалогового окна используются для настроек вывода информации (*Output Range/New Worksheet Ply/New Workbook – Выходной интервал/Новый рабочий лист/Новая рабочая книга*).

Summary Statistics – Итоговая статистика

Если установить этот флажок, то одновременно будут вычислены параметры описательной статистики для каждого из столбцов (каждой строки) данных.

Confidence Level for Mean – Уровень надежности

Используется для вычисления доверительного интервала для среднего значения (см. описание функции CONFIDENCE).

Kth Largest/Kth Smallest – *K-й наибольший/K-й наименьший*

Возвращает k-ое наибольшее (и/или наименьшее) значение из множества данных.

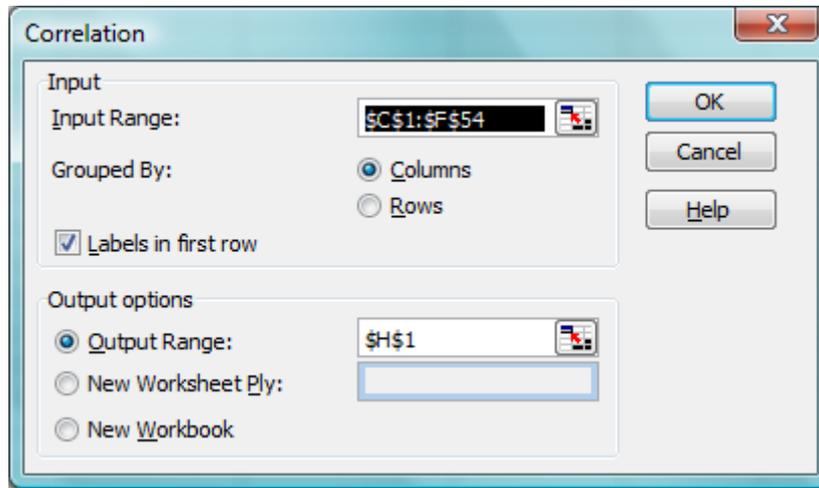
Аналогичный результат можно получить, применяя функцию LARGE (SMALL).

Ниже представлены результаты вычислений для переменной «тревожность» (округлены).

<i>тревожность</i>	
Mean	6,79
Standard Error	0,45
Median	6
Mode	6
Standard Deviation	3,30
Sample Variance	10,86
Kurtosis	1,44
Skewness	1,04
Range	17
Minimum	1
Maximum	18
Sum	360
Count	53
Largest(5)	12
Smallest(3)	3
Confidence Level(95,0%)	0,91

Корреляционный анализ

Коэффициент корреляции широко используется в статистике для оценки степени зависимости двух наборов данных. Например, если наивысшие значения в одном наборе данных связаны с наивысшими значениями во втором наборе, у этих наборов положительная корреляция. Степень корреляции выражается коэффициентом, который лежит в интервале от -1,0 (полная отрицательная корреляция) до +1,0 (полная положительная корреляция). Нулевой коэффициент корреляции означает, что два набора некоррелированы (это часто принимается как подтверждение того, что наборы данных независимы).



Требуется определить входной диапазон ячеек, который может состоять из любого количества переменных, сгруппированных по строкам или столбцам.

Результат представлен в виде корреляционной матрицы, в которой содержатся коэффициенты корреляции для каждой пары переменных.

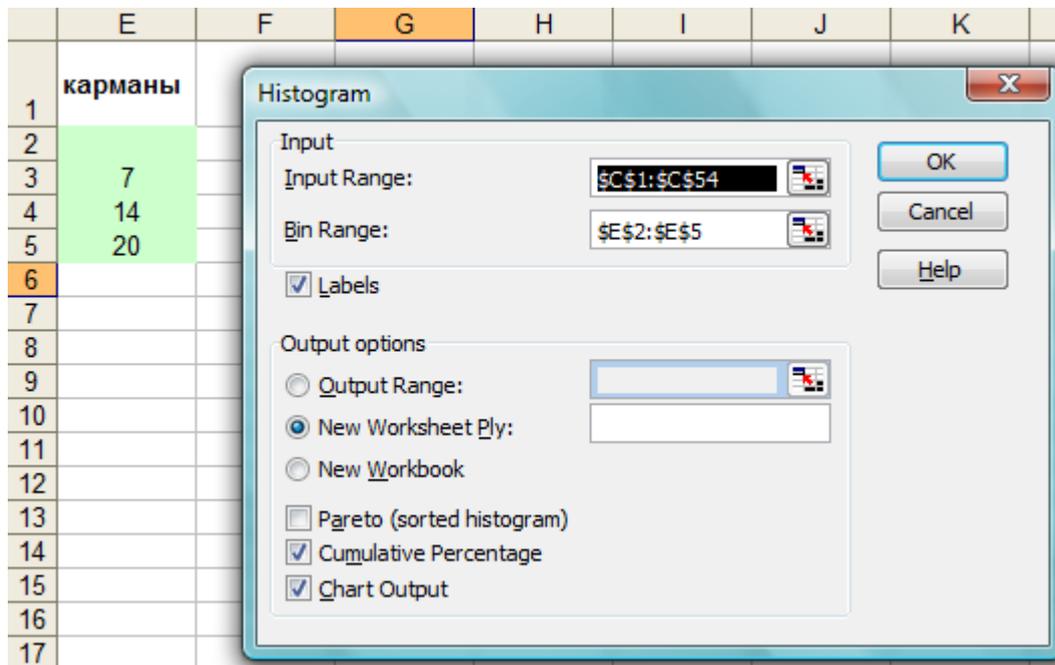
	<i>тревожность</i>	<i>фрустрация</i>	<i>агрессивность</i>	<i>ригидность</i>
<i>тревожность</i>	1			
<i>фрустрация</i>	0,440	1		
<i>агрессивность</i>	0,161	0,041	1	
<i>ригидность</i>	0,190	0,262	0,296	1

Ячейки выше диагонали не заполнены, поскольку корреляционная матрица симметрична.

В результирующей корреляционной матрице не используются формулы для вычисления результатов. Поэтому, если данные изменились, корреляционная матрица останется неизменной. Для создания корреляционной матрицы, которая автоматически обновляется при изменении исходных данных, можно использовать функцию CORREL.

Гистограмма

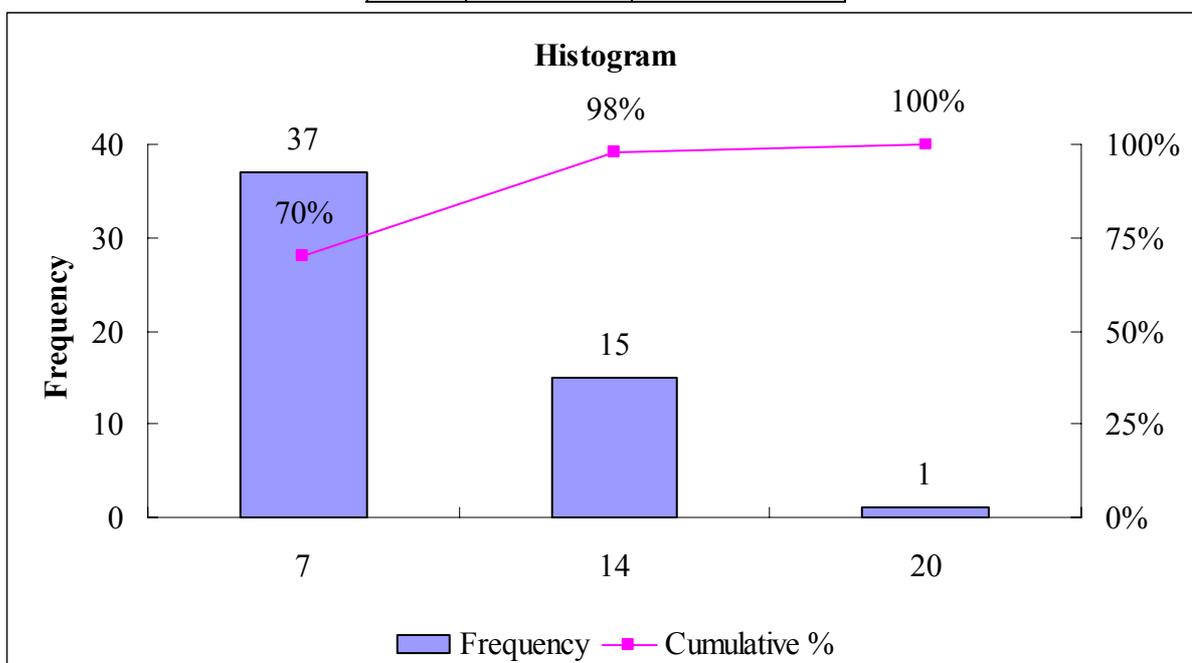
Эта процедура полезна для анализа распределения данных и построения гистограмм (графика эмпирической плотности вероятности). В качестве исходных значений нужно указать входной диапазон и интервалы значений. Интервал значений определяет границы для каждого столбца гистограммы. Если не указывать интервалы значений, Excel автоматически создаст равномерно распределенные интервалы.



В качестве дополнительной возможности предусмотрена сортировка результирующей диаграммы по частоте попадания данных в каждый интервал.

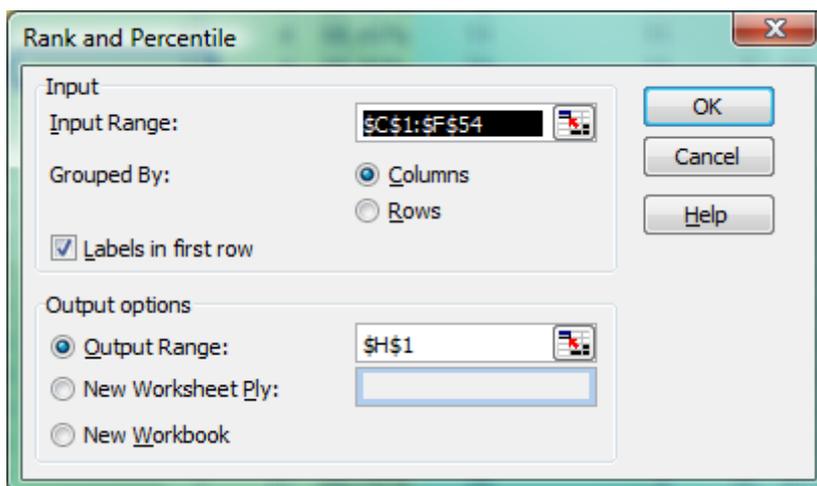
Полученные результаты представлены в таблице и на рисунке (диаграмма отредактирована).

	<i>Frequency</i>	<i>Cumulative %</i>
7	37	69,81%
14	15	98,11%
20	1	100,00%
More	0	100,00%



Ранг и перцентиль

С помощью этого средства создается таблица, показывающая порядковый и процентный ранги для каждого значения из заданного диапазона ячеек.



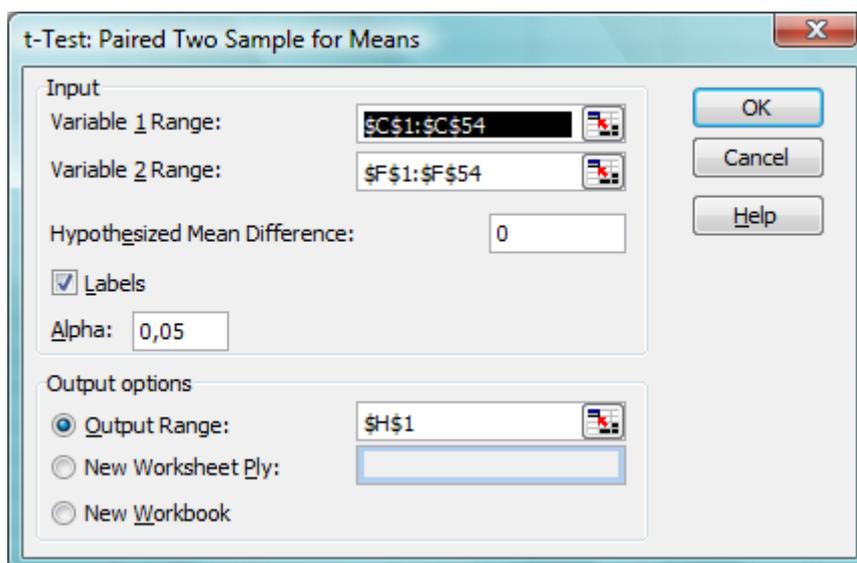
Результат применения этого инструмента аналогичен получаемому при непосредственном применении функций RANK и PERCENTILE. Отличие состоит в том, что все наборы данных сортируются в порядке убывания ранга (наивысший ранг равен 1).

t-тест

Процедура t-тест используется для определения факта статистически значимого различия между двумя небольшими выборками. Пакет анализа может выполнять тест трех типов.

Парный двухвыборочный t-тест для средних

Используется для парных выборок, в которых содержатся данные о двух наблюдениях за одним и тем же объектом. Выборки должны быть одинакового размера.



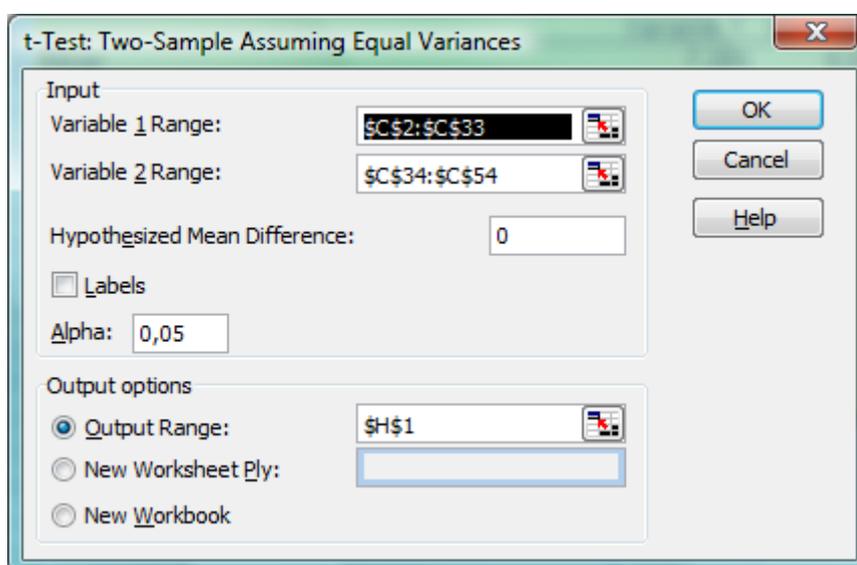
	<i>тревожность</i>	<i>ригидность</i>
Mean	6,79	9,11
Variance	10,86	9,53
Observations	53	53
Pearson Correlation	0,19	

Hypothesized Mean Difference	0	
df	52	
t Stat	-4,16	
P(T<=t) one-tail	6,07E-05	
t Critical one-tail	1,67	
P(T<=t) two-tail	1,21E-04	
t Critical two-tail	2,01	

Вычисленные вероятности (односторонняя и двухсторонняя) существенно меньше 0,05, поэтому нулевая гипотеза (об отсутствии различий средних значений) отвергается в пользу альтернативной – различия статистически значимы.

Двухвыборочный t-тест с одинаковыми дисперсиями

Используется для независимых, а не парных выборок. Предполагает равенство дисперсий этих выборок.



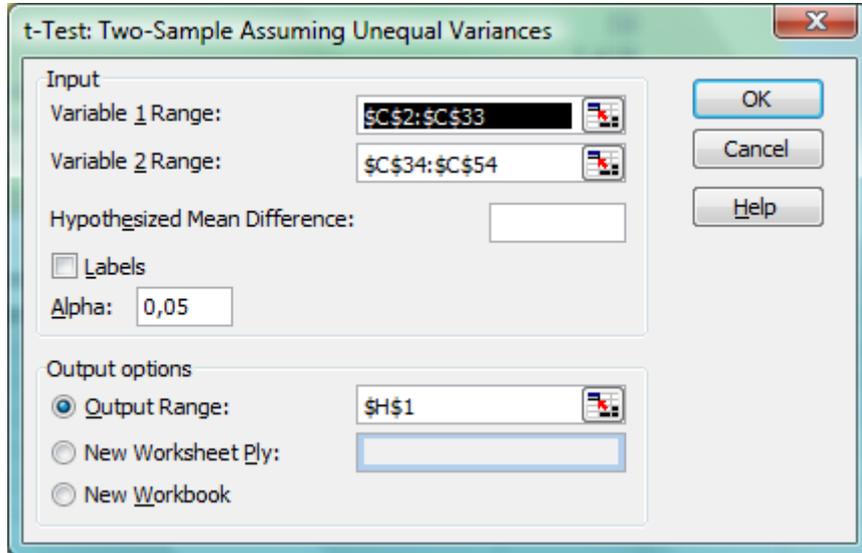
	<i>Variable 1</i>	<i>Variable 2</i>
Mean	7,281	6,048
Variance	12,725	7,548
Observations	32	21
Pooled Variance	10,695	
Hypothesized Mean Difference	0	
df	51	
t Stat	1,343	
P(T<=t) one-tail	0,093	
t Critical one-tail	1,675	
P(T<=t) two-tail	0,185	
t Critical two-tail	2,008	

Вычисленные вероятности (односторонняя и двухсторонняя) превышают значение 0,05, поэтому нулевая гипотеза (об отсутствии различий средних значений) не отвергается.

Двухвыборочный t-тест с разными дисперсиями

Используется для независимых выборок. Предполагает неравные дисперсии выборок.

В диалоговом окне нужно задать уровень надежности — параметр Alpha и гипотетическую среднюю разность между двумя средними (обычно проверяют гипотезу об отсутствии различий – в этом случае этот параметр приравнивают к нулю или просто не указывают).



Результаты вычислений представлены в таблице.

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	7,281	6,048
Variance	12,725	7,548
Observations	32	21
Hypothesized Mean Difference	0	
df	50	
t Stat	1,418	
P(T<=t) one-tail	0,081	
t Critical one-tail	1,676	
P(T<=t) two-tail	0,162	
t Critical two-tail	2,009	

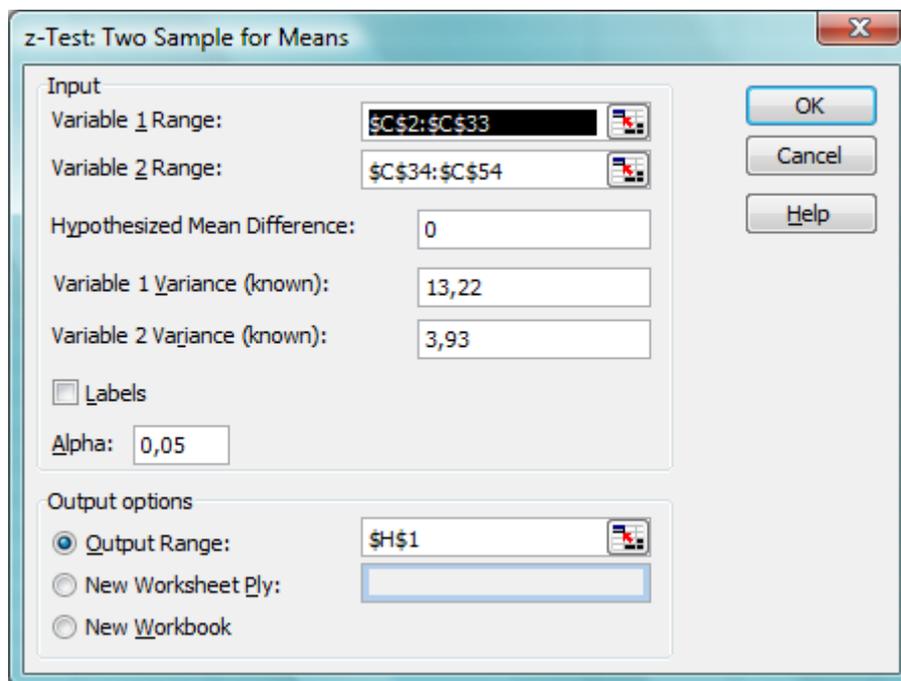
Вычисленные вероятности (односторонняя и двухсторонняя) превышают значение 0,05, поэтому нулевая гипотеза (об отсутствии различий средних значений) не отвергается.

Двухвыборочный z-тест для средних

Тогда как t-тесты используются для небольших выборок, z-тест предназначен для больших выборок или совокупностей. Должны быть известны дисперсии для обоих входных интервалов.

Применим для данных, подчиняющихся нормальному распределению.

Предварительно вычислим выборочные дисперсии показателя ригидности для студентов открытого университета (форма обучения – 2) и стационара (1).



	<i>Variable 1</i>	<i>Variable 2</i>
Mean	9,406	8,667
Known Variance	13,22	3,93
Observations	32	21
Hypothesized Mean Difference	0	
z	0,955	
P(Z<=z) one-tail	0,170	
z Critical one-tail	1,645	
P(Z<=z) two-tail	0,340	
z Critical two-tail	1,960	

Вычисленные вероятности (односторонняя и двухсторонняя) превышают значение 0,05, поэтому нулевая гипотеза (об отсутствии различий средних значений) не отвергается – различия в уровне выраженности исследуемого параметра у двух групп статистически незначимы.

Задания по теме «Пакет анализа MS Excel»

Задание 1

Для указанного набора данных с помощью инструментов анализа данных:

- 1) вычислить основные статистические характеристики (среднее, минимум, максимум, выборочную дисперсию и стандартное отклонение, моду, медиану, асимметрию и эксцесс);
- 2) построить интервальный вариационный ряд (нижняя граница первого интервала – 200, ширина интервала – 30); использовать функцию COUNT;
- 3) полученный вариационный ряд изобразить графически;

- 4) с помощью t-теста проверить гипотезу о равенстве средних значений двух выборок (независимых), образованных верхней и нижней половиной указанной таблицы данных;
- 5) выполнить такую же проверку с помощью z-теста.

200	271	295	225	268	245	275	248	250	270
310	315	345	350	270	270	295	360	300	285
270	260	210	274	300	275	300	260	260	260
298	281	284	291	280	235	289	240	280	230
300	290	289	292	360	300	365	290	330	290
327	295	250	337	249	350	271	298	300	345
232	235	248	273	237	256	255	238	220	220
300	275	315	300	300	261	265	262	273	355
325	335	320	300	310	310	300	330	268	300
280	340	280	260	320	345	350	279	258	260

Задание 2

Перед сдачей экзаменов в конце семестра в 20 группах студентов университета был проведен опрос о том, какую оценку по сдаваемым в сессию курсам они ожидают получить. После сессии средние полученные оценки были сопоставлены со средними ожидаемыми. Результаты приведены в таблице:

Ожидаемая	3,4	3,1	3,0	2,8	3,7	3,5	2,9	3,7	3,5	3,2
Полученная	4,1	3,4	3,3	3,0	4,7	4,6	3,0	4,6	4,6	3,6
Ожидаемая	3,0	3,5	3,3	3,1	3,3	3,9	2,9	3,2	3,4	3,4
Полученная	3,5	4,0	3,6	3,1	3,3	4,5	2,8	3,7	3,8	3,9

- 1) Рассчитайте линейный коэффициент корреляции Пирсона, оцените его значимость при $\alpha=0,05$.
- 2) Проверьте гипотезу о равенстве средних значений этих двух наборов данных (ожидаемая и получаемая оценки) с помощью t-критерия Стьюдента для связанных выборок.

Графики и диаграммы в MS Excel

Введение

Диаграмма – это визуальное представление числовых значений. представление числовых данных в форме наглядных диаграмм делает эти данные более понятными и доступными. Поскольку диаграммы являются рисунками, они особенно полезны для обзорных представлений данных и для иллюстрации взаимоотношений числовых рядов. Диаграмма позволяет визуально заметить тенденцию или повторяемость данных, которые в таблице остались бы незамеченными.

Обычно диаграмма создается для того, чтобы продемонстрировать какую-либо закономерность.

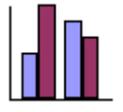
Часто иллюстрируемая закономерность описывается в названии диаграммы или в текстовом окне, расположенном в диаграмме. Сама диаграмма приводится для визуальной поддержки иллюстрации.

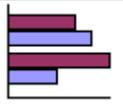
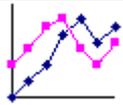
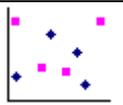
Выбор правильного типа диаграммы – ключевой фактор, влияющий на убедительность иллюстрации. Поэтому не жалейте времени на эксперименты с различными типами диаграмм, только так вы сможете выбрать наиболее эффективный способ представления ваших данных.

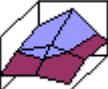
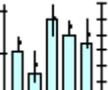
Почти в каждом случае иллюстрация заложенной в данных закономерности базируется на сравнении. Ниже приведены некоторые наиболее общие виды сравнений.

- Сравнение одних показателей с другими. Например, в диаграмме могут сравниваться продажи компании в разных регионах.
- Изменение данных во времени. В этом случае сравниваются данные в разные моменты или периоды времени.
- Относительные сравнения. Примером относительного сравнения служит круговая диаграмма, изображающая относительные значения данных в виде секторов.
- Взаимоотношения данных. Идеальный пример иллюстрации взаимоотношений – точечная диаграмма.
- Сравнение частот (относительных частот). Пример сравнения частот – гистограмма, изображающая количество (или процент) школьников с различной успеваемостью.
- Иллюстрация (или выявление) посторонних точек. Диаграмма может служить средством быстрого визуального выявления ошибочных точек данных, значения которых существенно искажены случайными факторами или ошибками ввода.

Типы диаграмм в MS Excel

Тип диаграммы	Краткое описание
	<p>Column Гистограмма (столбиковая)</p> <p>Гистограмма – один из наиболее популярных типов диаграмм. В гистограмме каждая точка данных выводится в виде отдельного вертикального столбика, высота которого отображает значение. Шкала значений выводится на вертикальной оси, располагаемой обычно на левой стороне диаграммы. В гистограмме может быть отображено произвольное количество рядов данных. Столбики, отображающие разные ряды, могут быть расположены или рядом, или один на другом. В последнем случае диаграмма называется гистограммой с накоплением. Обычно разные ряды</p>

		<p>данных выводятся разным цветом или заполняются разными узорами.</p> <p>Гистограммы часто используются для сравнения дискретных данных. Они могут иллюстрировать как отличия данных одного ряда между собой, так и отличия между разными рядами.</p>
	<p>Bar</p> <p>Линейчатая (полосовая)</p>	<p>Линейчатая диаграмма – это гистограмма, повернутая по часовой стрелке на угол 90°. Ее главное преимущество состоит в том, что подписи ее категорий значительно легче располагать и читать.</p>
	<p>Line</p> <p>График (линейная)</p>	<p>Графики часто используются для иллюстрации непрерывных данных. Они полезны для выявления тенденций.</p>
	<p>Pie</p> <p>Круговая (секторная)</p>	<p>Круговая диаграмма удобна для иллюстрации отношения или вклада отдельных значений. С ее помощью можно вывести только один ряд данных. Круговые диаграммы наиболее эффективны при небольшом количестве точек. Обычно в круговой диаграмме выводят не более пяти или шести точек данных (секторов); если количество точек больше шести, то круговую диаграмму трудно интерпретировать.</p> <p>Все значения круговой диаграммы должны быть положительными числами. Если в ряде данных встречаются отрицательные значения, Excel автоматически преобразует их в положительные.</p> <p>Чтобы привлечь внимание к одному или нескольким секторам, их можно выделить. Для этого нужно активизировать сектор и перетащить его с помощью мыши в направлении от центра.</p> <p>С помощью вторичной круговой диаграммы или круговой диаграммы с вторичной гистограммой можно вывести в той же области построения еще одну диаграмму, детализирующую значения одного или нескольких значений первичной диаграммы.</p>
	<p>XY (Scatter)</p> <p>Точечная (диаграмма рассеяния)</p>	<p>У точечных диаграмм нет оси категорий: по обеим осям выводятся значения. Каждая точка данных выводится только на основе двух значений.</p> <p>Точечные диаграммы часто используются для иллюстрации зависимостей между двумя переменными.</p> <p>Добавление в точечную диаграмму линии тренда позволит лучше проиллюстрировать взаимосвязь переменных. Возможно также добавить к измерениям полосы погрешностей.</p>
	<p>Area</p> <p>С областями (слоевая)</p>	<p>Диаграмма с областями — это график, в котором область под линией закрашена.</p>
	<p>Doughnut</p> <p>Кольцевая</p>	<p>Подобна диаграмме типа Pie (круговой), но может отображать несколько рядов данных. Кольцевые диаграммы используются редко. Во многих случаях вместо кольцевой диаграммы лучше подходит нормированная гистограмма с накоплением.</p> <p>Удобнее всего использовать кольцевую диаграмму вместо</p>

		круговой для иллюстрации одного ряда данных.
	Radar Лепестковая (радиальная)	В лепестковой диаграмме отдельная ось определена для каждой категории. Оси направлены наружу от центра диаграммы. Значение каждой точки данных отмечается на соответствующей оси.
	Surface Поверхность	Отображает изменение данных по двум измерениям в виде поверхности.
	Bubble Пузырьковая	Отображает на плоскости наборы данных из трех значений. Подобна точечной диаграмме, но третья (дополнительная) величина отображается с помощью пузырьков разного размера. Обе оси являются осями значений, осей категорий нет (как и у точечной диаграммы).
	Stock Биржевая	Предназначены для вывода биржевых данных. Отображает наборы данных из трех значений (самый высокий курс, самый низкий курс, курс закрытия).
	Cylinder Cone Pyramid	Практически не отличаются от диаграмм типа Column (гистограмма). Отличие состоит в замене прямоугольников (параллелепипедов) на цилиндры (конусы, пирамиды).

Для визуализации данных педагогических и психологических исследований чаще всего применяются диаграммы типа Column, Bar, Line, Pie, Scatter, Doughnut и Radar.

Процедура построения диаграммы с помощью мастера Chart Wizard

Наиболее распространенный (и наиболее гибкий) способ построения диаграммы состоит в использовании мастера диаграмм. Выполните следующие действия.

1. Выделите ячейки данных, на основе которых необходимо создать диаграмму (этот этап не обязателен).
2. Выберите команду Insert -> Chart (Вставка -> Диаграмма) (или щелкните на пиктограмме Chart Wizard (Мастер диаграмм), расположенной на стандартной панели инструментов).
3. На каждом шаге мастера диаграмм (от 1 до 4) задавайте необходимые параметры диаграммы и щелкайте на кнопке Next (Далее).
4. Щелкнув на кнопке Finish (Готово), создайте диаграмму. Вы можете щелкнуть на этой кнопке на любом из шагов мастера, если сочтете, что диаграмма уже приобрела нужный вид.

Шаги построения диаграмм с помощью Chart Wizard подробно отражены в видеорегистраторах. Таким же способом представлена информация об основах форматирования элементов диаграмм.

Построение графиков в MS Excel (задание I)

По указанным данным построить представленные ниже графики.

Исходные данные и образцы графиков находятся в файле excel_grafiki_task_1.xls.

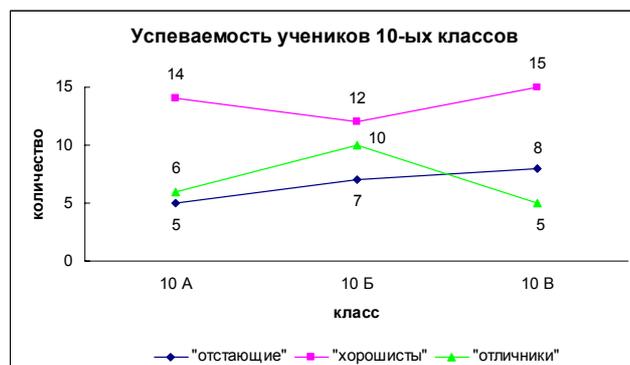
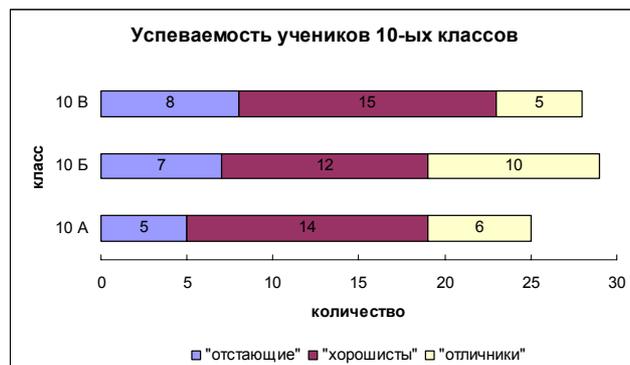
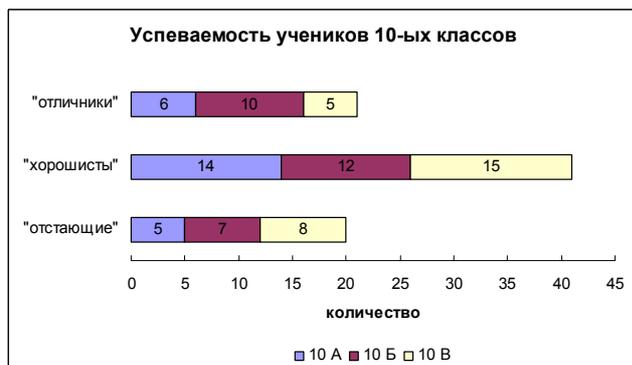
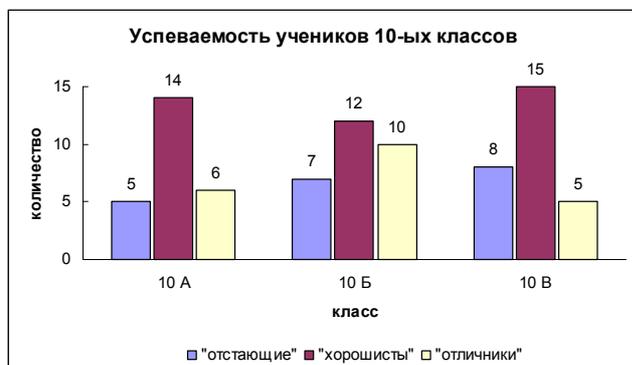
Каждый график или группа графиков строятся на отдельных листах.

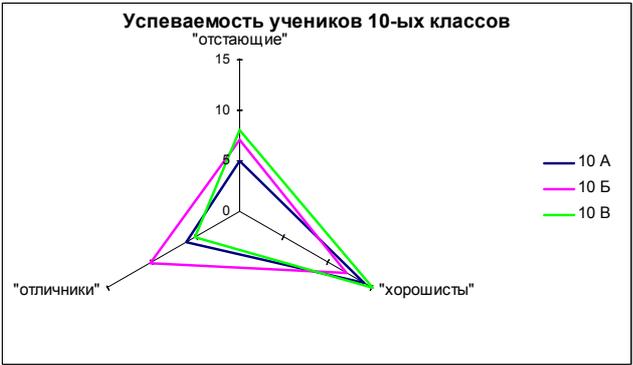
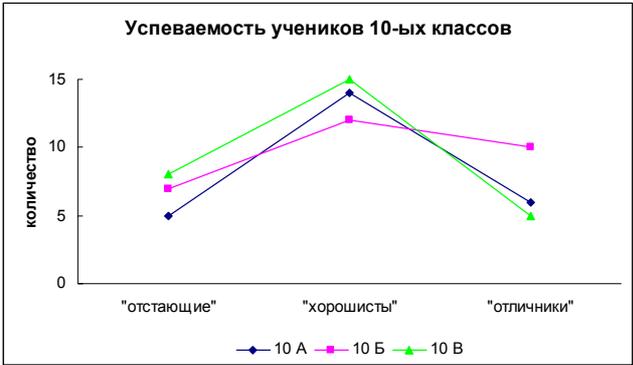
Результаты работы сохранить под именем Familija_Imja_excel_grafiki_task_1.xls (указав, естественно, свою фамилию и имя) и прислать преподавателю по электронной почте.

Исходные данные

	10 А	10 Б	10 В
"отстающие"	5	7	8
"хорошисты"	14	12	15
"отличники"	6	10	5

Образцы графиков





Построение диаграмм и графиков с использованием MS Excel (задание II)

По указанным данным построить диаграммы, полностью соответствующие образцу.

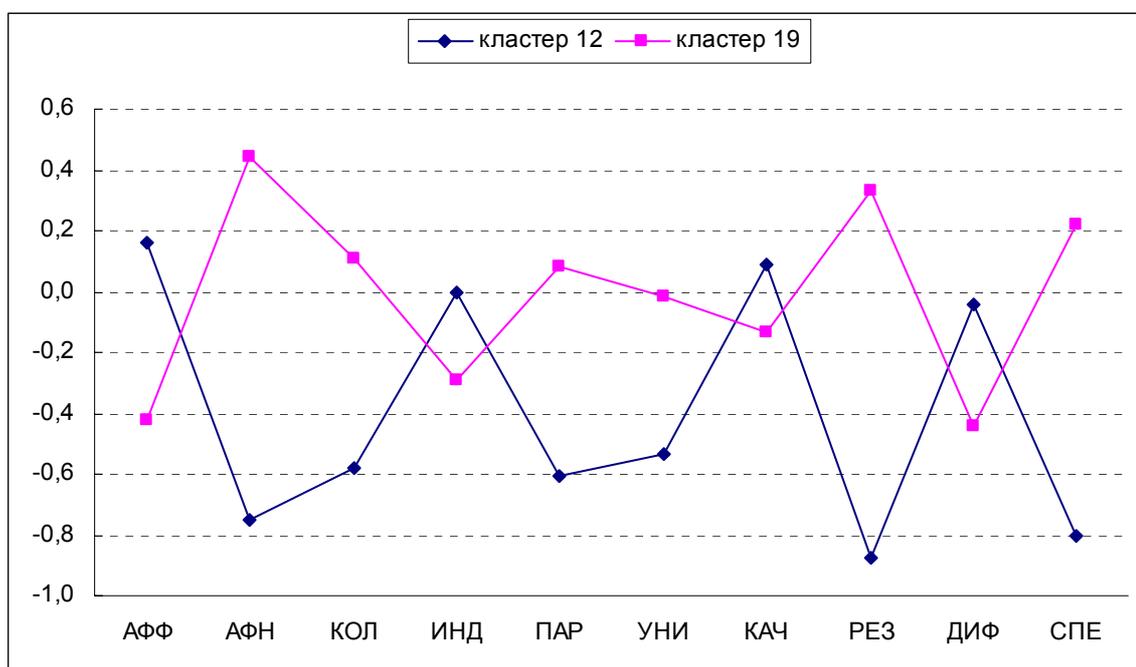
Исходные данные и образцы графиков находятся в файле excel_grafiki_task_2.xls.

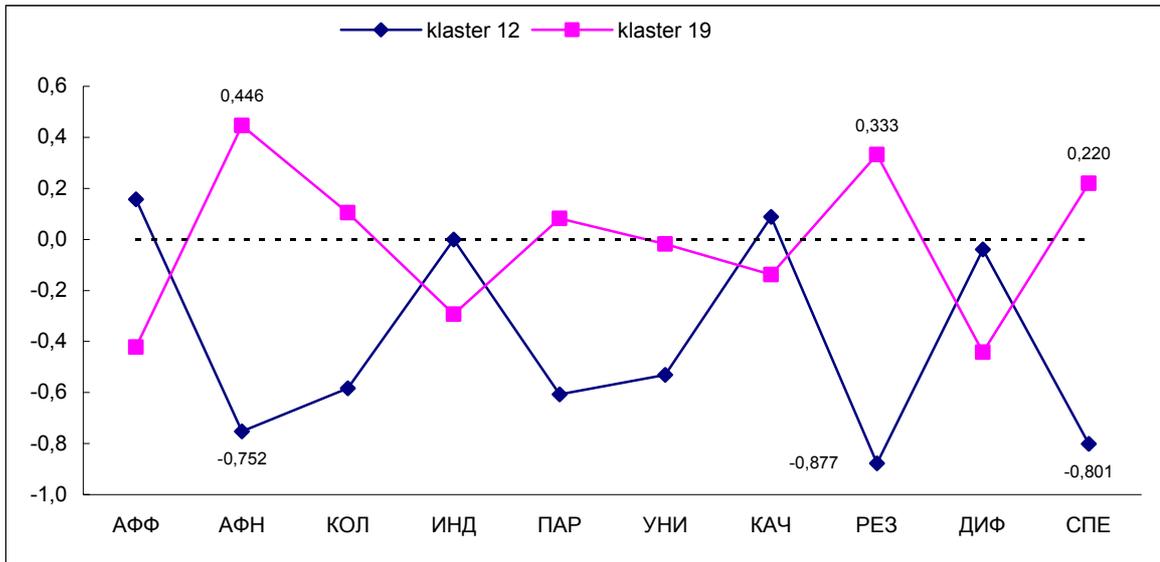
Каждый график или группа графиков строятся на отдельных листах.

Результаты работы сохранить под именем Familija_Imja_excel_grafiki_task_2.xls (указав, естественно, свою фамилию и имя) и прислать преподавателю по электронной почте.

Задание 1

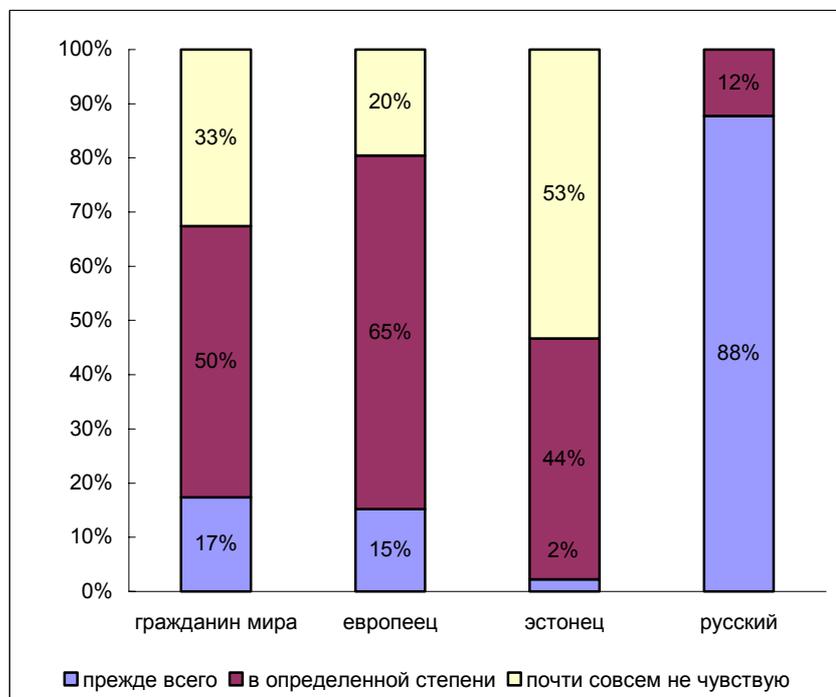
	АФФ	АФН	КОЛ	ИНД	ПАР	УНИ	КАЧ	РЕЗ	ДИФ	СПЕ
Валидные	12	12	12	12	12	12	12	12	12	12
Среднее	0,157	-0,752	-0,583	0,000	-0,607	-0,531	0,088	-0,877	-0,040	-0,801
Стд. ошибка среднего	0,180	0,230	0,334	0,295	0,254	0,200	0,307	0,200	0,298	0,214
Валидные	19	19	19	19	19	19	19	19	19	19
Среднее	-0,422	0,446	0,105	-0,293	0,082	-0,018	-0,138	0,333	-0,442	0,220
Стд. ошибка среднего	0,161	0,155	0,172	0,173	0,215	0,179	0,164	0,185	0,121	0,162





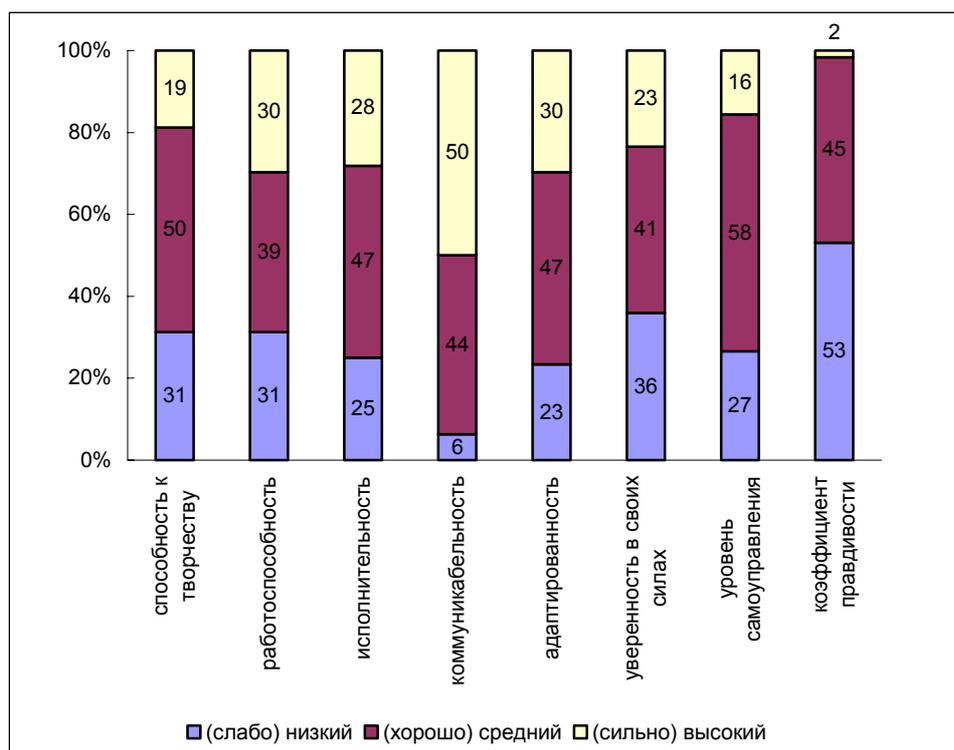
Задание 2

	гражданин мира	европеец	эстонец	русский
прежде всего	8	7	1	43
в определенной степени	23	30	20	6
почти совсем не чувствую	15	9	24	0



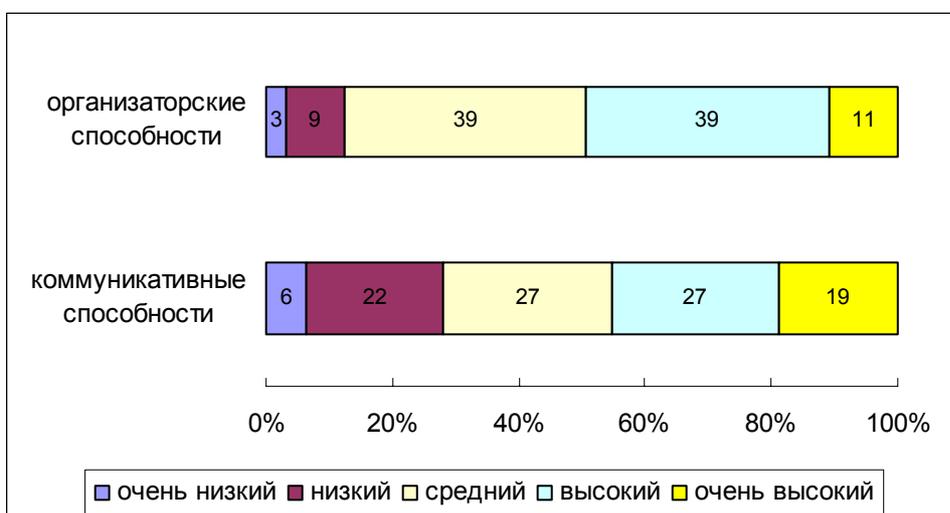
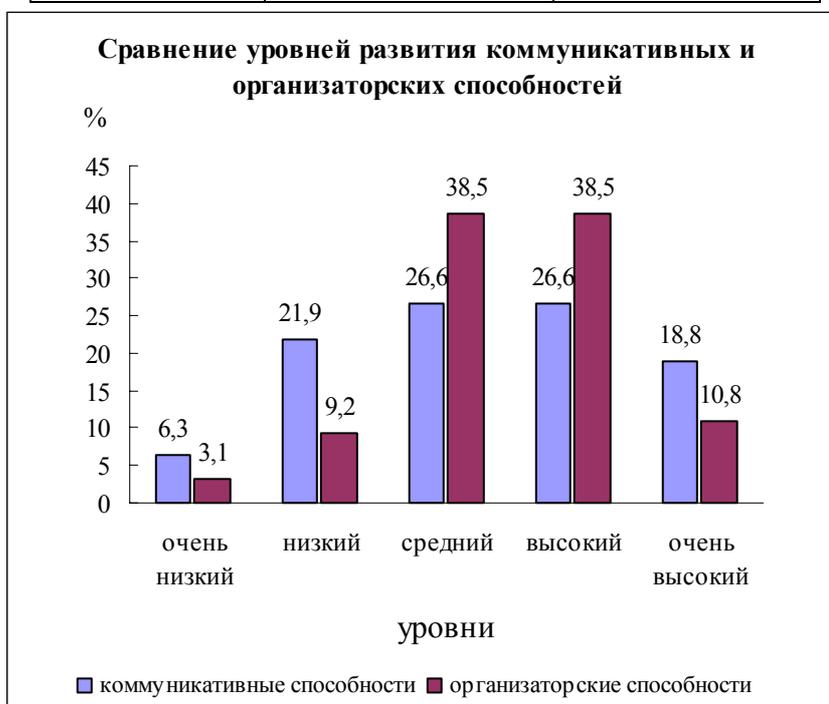
Задание 3

	способность к творчеству	работоспособность	исполнительность	коммуникабельность	адаптированность	уверенность в своих силах	уровень самоуправления	коэффициент правдивости
(слабо) низкий	31,3	31,3	25	6,3	23,4	35,9	26,6	53,1
(хорошо) средний	50	39,1	46,9	43,8	46,9	40,6	57,8	45,3
(сильно) высокий	18,8	29,7	28,1	50	29,7	23,4	15,6	1,6



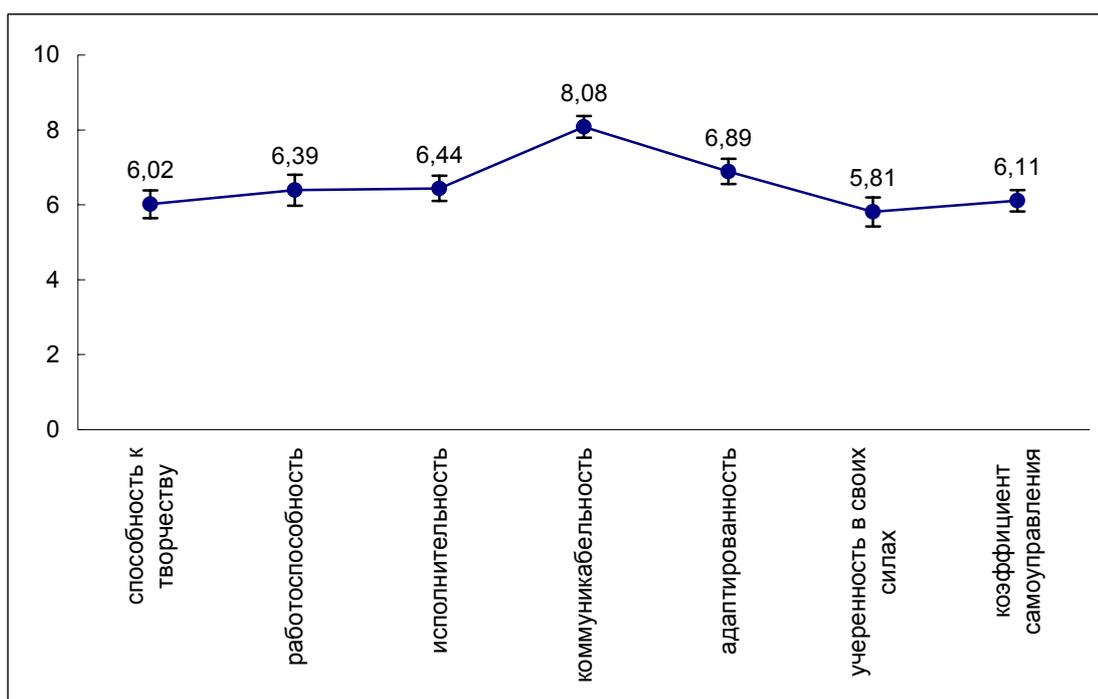
Задание 4

	Коммуникативные способности	Организаторские способности
очень низкий	6,3	3,1
низкий	21,9	9,2
средний	26,6	38,5
высокий	26,6	38,5
очень высокий	18,8	10,8



Задание 5

	Среднее	Стд. ошибка среднего	Стд. отклонение
способность к творчеству	6,02	0,37	2,95
работоспособность	6,39	0,41	3,29
исполнительность	6,44	0,34	2,69
коммуникабельность	8,08	0,29	2,32
адаптированность	6,89	0,34	2,70
уверенность в своих силах	5,81	0,39	3,11
коэффициент самоуправления	6,11	0,28	2,26

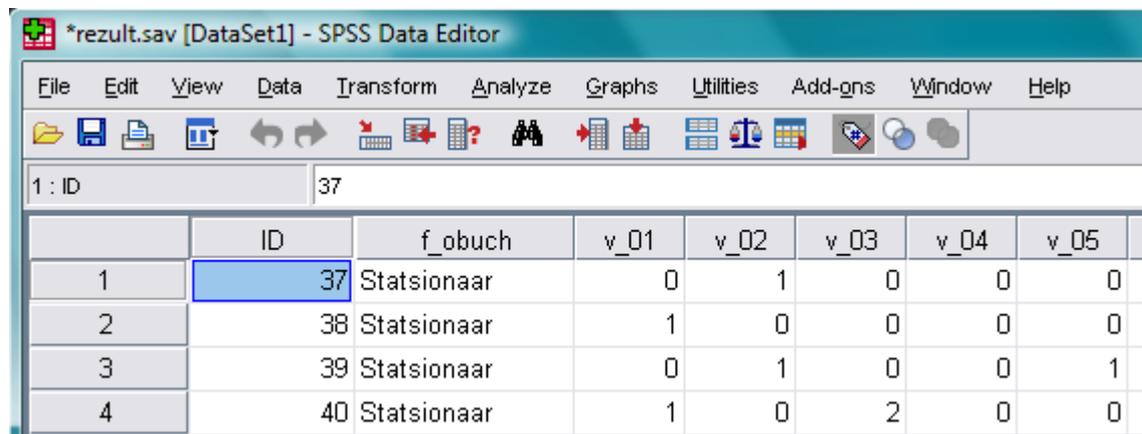


Основы работы с программой SPSS

Основные окна программы

Окна редактора данных

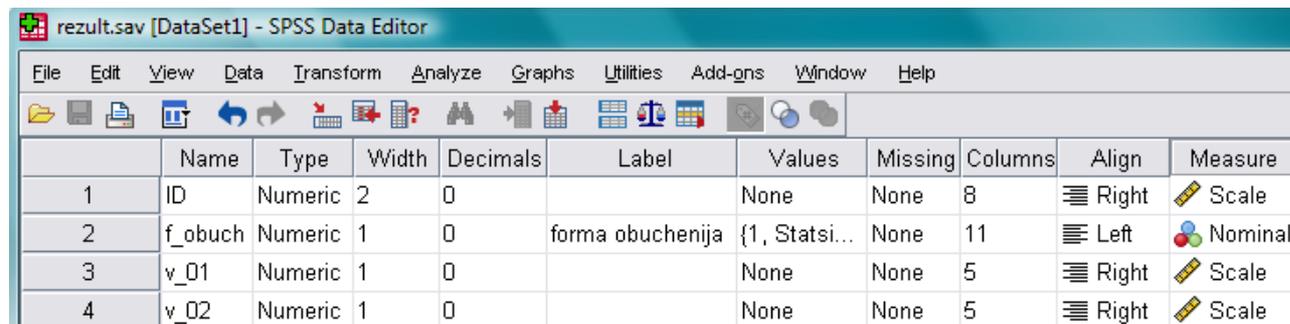
Окно Data View



*result.sav [DataSet1] - SPSS Data Editor

	ID	f_obuch	v_01	v_02	v_03	v_04	v_05
1	37	Stationsaar	0	1	0	0	0
2	38	Stationsaar	1	0	0	0	0
3	39	Stationsaar	0	1	0	0	1
4	40	Stationsaar	1	0	2	0	0

Окно Variable View



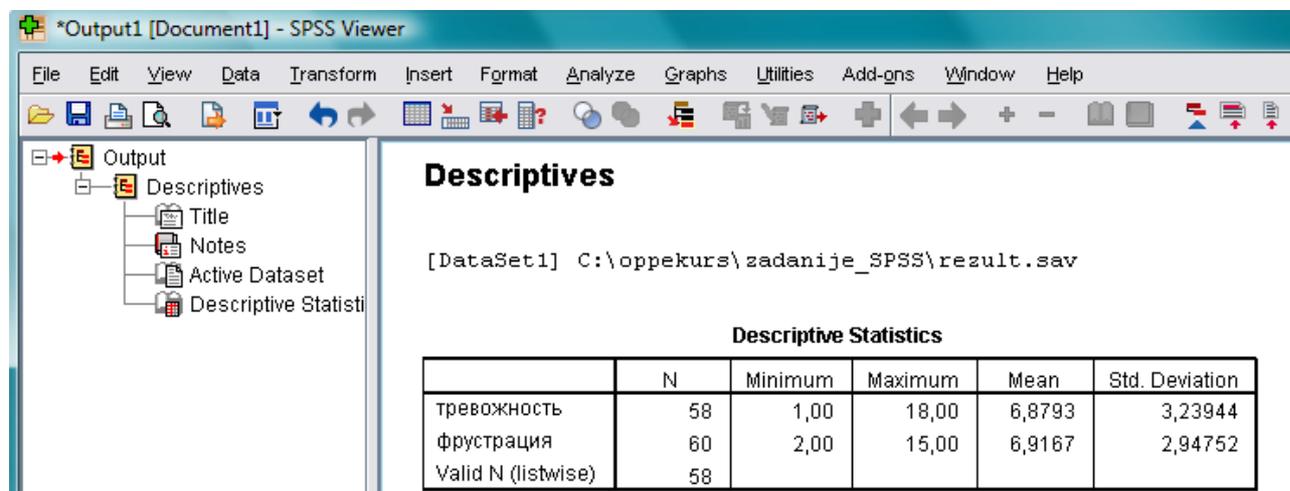
result.sav [DataSet1] - SPSS Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	ID	Numeric	2	0		None	None	8	Right	Scale
2	f_obuch	Numeric	1	0	forma obuchenija	{1, Statsi...	None	11	Left	Nominal
3	v_01	Numeric	1	0		None	None	5	Right	Scale
4	v_02	Numeric	1	0		None	None	5	Right	Scale

Переключение между этими окнами (закладки слева внизу окна редактора данных)



Окно отчета SPSS Viewer



*Output1 [Document1] - SPSS Viewer

Descriptives

[DataSet1] C:\oppekurs\zadanije_SPSS\result.sav

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
тревожность	58	1,00	18,00	6,8793	3,23944
фрустрация	60	2,00	15,00	6,9167	2,94752
Valid N (listwise)	58				

Обратите внимание, что меню редактора данных и окна просмотра результатов (и окна синтаксиса – см. ниже) одинаковые. Следовательно, все команды (диалоговые окна) можно

запускать из любого окна программы SPSS.

Меню программы SPSS



В строке меню главного окна SPSS находится 10 опций:

- File/Файл содержит стандартные возможности по созданию новых файлов, экспорту и импорту в различные форматы, выводу на печать и т.п. Например, открытие ранее созданной базы данных осуществляется так: Файл – Открыть – Данные, после чего появится стандартное окно для открытия файлов.
- Edit/Правка – содержит стандартные возможности редактирования (вырезать, копировать, вставить, отменить действие).
- View/Вид – определяет внешний вид главного окна, содержит возможности настройки строки состояния, панели инструментов, шрифтов, и т.п.
- Data/Данные – содержит основные операции по созданию и редактированию данных SPSS (создание новых переменных, создание шаблона ввода данных, сортировка данных и др.)
- Transform/Трансформация – содержит операции по изменению данных (пересчет данных, перекодирование, замена пропущенных и ошибочных данных).
- Analyze/Анализ – здесь находятся все статистические операции, которые может осуществлять SPSS.
- Graphs/Графики – операции по созданию графиков.
- Utilities/Сервис – дополнительные сервисные возможности по просмотру свойств переменных, файлов данных и др.
- Window/Окно – стандартные возможности по управлению несколькими окнами программы.
- Help/Помощь – различная справочная информация (к сожалению, не переведенная на русский язык).

Описание переменных

Закладка Variable View содержит описания свойств переменных, содержащихся в файле данных. На вкладке Variable View:

- строки - это переменные;
- столбцы - это свойства переменных.

Вы можете не только добавлять или удалять переменные, но и изменять свойства переменных, включая следующие атрибуты:

- имя переменной;
- тип переменной;
- число цифр или символов в переменной;
- число десятичных знаков;
- описательные метки переменных и значений;

- пользовательские пропущенные значения;
- ширину столбца;
- уровень измерения.

Все эти свойства сохраняются вместе с сохраняемым файлом данных.

Допустимые имена переменных

Для имен переменных должны выполняться следующие правила:

- Имя каждой переменной должно быть уникальным; дублирующиеся имена не допускаются.
- Имена переменных могут иметь длиной до 64 байт (символов), первый символ в имени переменной должен быть буквой либо одним из символов @, #, или \$. Последующие символы могут быть любой комбинацией букв, чисел, точек (.) и не пунктуационных символов. Шестьдесят четыре байта обычно означают 64 символа в однобайтовых языках (например, английский).
- Имена переменных не могут содержать пробелов.
- Если имя переменной начинается с символа # - эта переменная является служебной. Создавать служебные переменные можно только при помощи синтаксиса. Не возможно задать # в качестве первого символа имени переменной в диалоговых окнах, в которых создаются новые переменные.
- Если имя переменной начинается с символа \$ - эта переменная является системной. В задаваемых пользователем переменных нельзя задать символ \$ в качестве первого символа в имени переменной.
- В именах переменных можно использовать точку, знак подчеркивания, а также символы \$, # и @. Например, A_ \$#1 - это допустимое имя переменной.
- Следует избегать задания имен переменных, заканчивающихся точкой, поскольку точка в таких случаях может быть воспринята как символ окончания команды синтаксиса. Создавать переменных, имена которых заканчиваются точкой, можно только при помощи синтаксиса. Не возможно создать переменную, имя которой заканчивается точкой в диалоговых окнах, в которых создаются новые переменные.
- Следует избегать имен переменных, заканчивающихся символом подчеркивания, поскольку возможен конфликт с именами, создаваемыми автоматически командами и процедурами.
- В именах переменных не могут использоваться зарезервированные ключевые слова. Зарезервированными словами являются ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO и WITH.
- Имена переменных могут состоять из любого сочетания символов в верхнем и в нижнем регистрах. Регистр сохраняется при отображении имен переменных.
- Если длинное имя переменной нужно перенести в выводе на несколько строк, строка разрывается на символах подчеркивания, на точках и там, где регистр символов изменяется с нижнего на верхний.

Типы переменных

В диалоговом окне Тип переменной задается тип данных для каждой переменной. По умолчанию все новые переменные - числовые. Вы можете использовать диалоговое окно Тип переменной для изменения типа данных. Содержимое диалогового окна Тип переменной зависит от выбранного типа данных. Для некоторых типов данных появляются поля для ввода ширины переменной и числа знаков после запятой, для других вы можете просто выбрать формат из списка с форматами.

Доступны следующие типы данных:

Numeric Числовая	Переменная, значения которой являются числами. Значения отображаются в стандартном числовом формате. При вводе данных Редактор данных принимает числовые значения в стандартном формате или в научной записи.
Comma Запятая	Числовая переменная, значения которой отображаются с запятыми, разделяющими каждые три разряда, а для отделения дробной части используется точка. При вводе данных типа Запятая, Редактор данных принимает числа с запятыми или без запятых, или в форме научной записи. В значениях не могут содержаться запятые справа от десятичного разделителя.
Dot Точка	Числовая переменная, значения которой отображаются с точками, разделяющими каждые три разряда, а для отделения дробной части используется запятая. При вводе данных типа с разделением групп цифр точками, Редактор данных принимает числа с точками или без точек, или в форме научной записи. В значениях не могут содержаться точки справа от десятичного разделителя.
Scientific notation Научная запись	Позволяет задать числовую переменную, значения которой выводятся с показателем степени, представленным буквой E, за которой идет знак и величина степени десятки. Редактор данных принимает в качестве таких переменных числовые значения, как со степенью, так и без. Показателю степени может предшествовать E или D, а также необязательный знак или только знак, например, 123, 1.23E2, 1.23D2, 1.23E+2, 1.23+2.
Date Дата	Числовая переменная, значения которой отображаются в одном из нескольких форматов календарной даты или времени. Формат выбирается из списка. Разделителями могут быть слэши, дефисы, точки, запятые или пробелы. Диапазон столетия при двузначном отображении года определяется установками в диалоговом окне Параметры (меню Правка, Параметры, вкладка Данные).
Dollar Доллар	Числовая переменная отображается со значком доллара вначале (\$), точками, отделяющими группы по три разряда, и точкой в качестве десятичного разделителя. Значения данных можно вводить как со знаком доллара вначале, так и без него.
Custom currency Выбираемая валюта	Числовая переменная, значения которой выводятся в одном из денежных форматов, заданного пользователем на вкладке Валюта диалогового окна Параметры. Заданные символы валюты нельзя использовать при вводе данных, однако они выводятся в Редакторе Данных.
String Текстовая	Переменная, значения которой не являются числовыми, не может использоваться в вычислениях. Текстовая переменная может содержать любые символы, однако их число не должно превышать заданную величину. Заглавные и строчные буквы считаются разными символами. Такие значения называют также буквенно-цифровыми.

Остальные параметры описания переменной представлены в таблице.

<p>Label Метка переменных</p>	<p>Вы можете задать метку переменной длиной до 256 символов (в двухбайтных языках - до 128 символов). Метки переменных могут содержать пробелы и зарезервированные символы, которые не допускается применять в именах переменных.</p>
<p>Values Метки значений</p>	<p>Каждому значению переменной можно присвоить содержательную метку. Метки особенно полезны, если в файле данных для представления нечисловых категорий используются числовые коды (например, коды 1 и 2 для обозначения пола мужской и женский, соответственно).</p> <p>Метки значений сохраняются с файлом данных. Вам не нужно повторно определять метки значений при следующем открытии файла данных.</p> <p>Метки значений могут быть длиной до 120 символов.</p>
<p>Missing Пропущенные значения</p>	<p>В диалоговом окне Пропущенные значения определенные значения задаются как пользовательские пропущенные. Например, Вы хотите отделить данные, пропущенные из-за отказа респондента отвечать, от данных, пропущенных из-за того, что вопрос не относится к респонденту. Значения данных, обозначенные как пользовательские пропущенные, помечаются для специальной обработки и исключаются из большинства вычислений.</p> <p>Пользовательские пропущенные значения сохраняются в файле данных. Вам не нужно повторно задавать пользовательские пропущенные значения при следующем открытии файла данных.</p> <p>Вы можете ввести до трех дискретных (отдельных) пропущенных значений, диапазон пропущенных значений или диапазон плюс одно дискретное значение.</p> <p>Диапазоны пропущенных значений могут быть заданы только для числовых переменных.</p>
<p>Columns Ширина столбца</p>	<p>Вы можете определить количество символов, задающих ширину столбца. Ширину столбца можно также изменить в Редакторе данных на вкладке Данные, перетащив мышью правую границу столбца.</p> <p>Ширина столбцов для пропорциональных шрифтов определяется средней шириной символов. В зависимости от используемых в значении символов в заданной ширине может отображаться меньше или больше символов.</p> <p>Ширина столбца влияет лишь на представление значений в Редакторе данных. Изменение ширины столбца не изменяет заданной ширины переменной.</p>
<p>Align Выравнивание переменной</p>	<p>Выравнивание управляет выводом значений данных и/или меток значений в Редакторе данных. По умолчанию числовые переменные выровнены по правому краю, а текстовые переменные - по левому. Выравнивание влияет только на представление (вывод на экран) данных в Редакторе данных.</p>
<p>Measure Шкала измерения переменной</p>	<p>Вы можете задать шкалу измерения переменной: количественную (числовые данные с интервальным или абсолютным уровнем измерения), порядковую, или номинальную. Номинальные и порядковые данные могут быть текстовыми (алфавитно-цифровыми) или числовыми.</p> <ul style="list-style-type: none"> • Номинальные. Переменную можно рассматривать как номинальную, когда ее значения представляют категории без естественного упорядочения, например, подразделение компании, где работает

	<p>наемный сотрудник. Примеры номинальных переменных включают регион, почтовый индекс или религиозную конфессию.</p> <ul style="list-style-type: none"> • Порядковые. Переменную можно рассматривать как порядковую, когда ее значения представляют категории с некоторым естественным для них упорядочением, например, уровни удовлетворенности обслуживанием от крайней неудовлетворенности до крайней удовлетворенности. Примеры порядковых переменных включают баллы, представляющие степень удовлетворенности или уверенности, или баллы, оценивающие предпочтение. • Шкала. Переменную можно рассматривать как количественную, когда ее значения представляют упорядоченные категории с осмысленной метрикой, так что уместно сравнивать расстояния между значениями. Примеры количественной переменной включают возраст в годах и доход в тысячах долларов.
--	--

Модификация данных

Для проведения анализа часто бывает необходимо выполнить преобразование данных. На основе первоначально собранных данных можно создать новые переменные и изменить кодирование. Подобные преобразования называются модификацией данных.

В SPSS существует много возможностей для модификации данных. К важнейшим из них относятся:

- вычисление новых переменных путем использования различных арифметических выражений (математических формул);
- подсчет частоты появлений определенных значений;
- перекодирование значений;
- вычисление новых переменных при выполнении определенного условия;
- агрегирование данных;
- ранговые преобразования
- вычисление весов наблюдений.

Рассмотрим некоторые из этих возможностей на примере учебного файла данных (тест Айзенка).

Вычисление новых переменных путем использования различных арифметических выражений (математических формул)

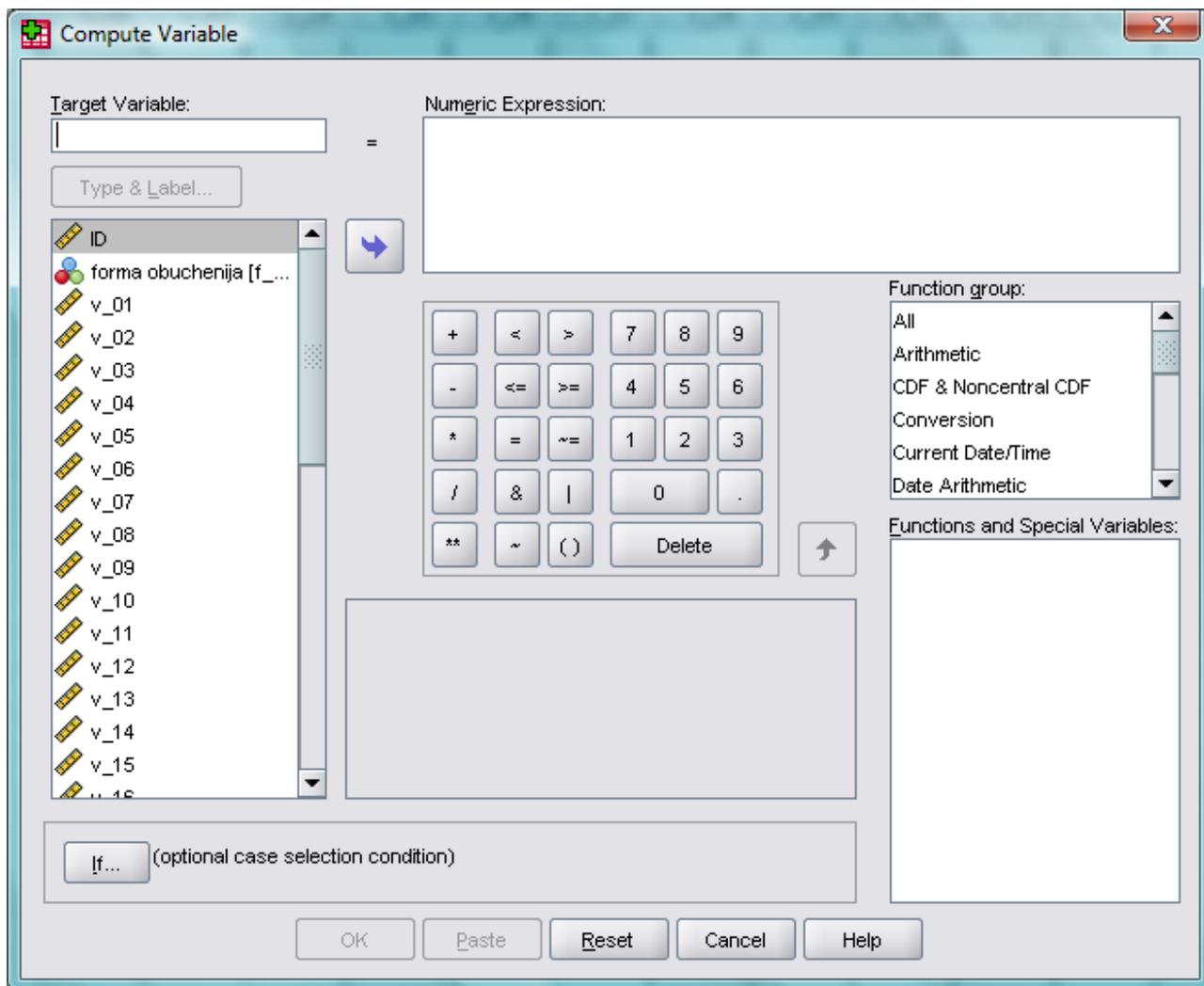
Выберите в меню: Transform -> Compute variable

Диалоговое окно Compute variable используется для вычисления значений переменных на основе значений других переменных.

- Вы можете вычислять значения числовых и текстовых переменных.
- Вы можете создавать новые переменные или изменять значения существующих. Создавая новую переменную, Вы можете задать ее тип и метку.
- Вы можете вычислять значения для подмножества наблюдений, удовлетворяющих логическим условиям.
- Вы можете использовать более 70 встроенных функций, включая арифметические, статистические, текстовые функции и функции распределений.

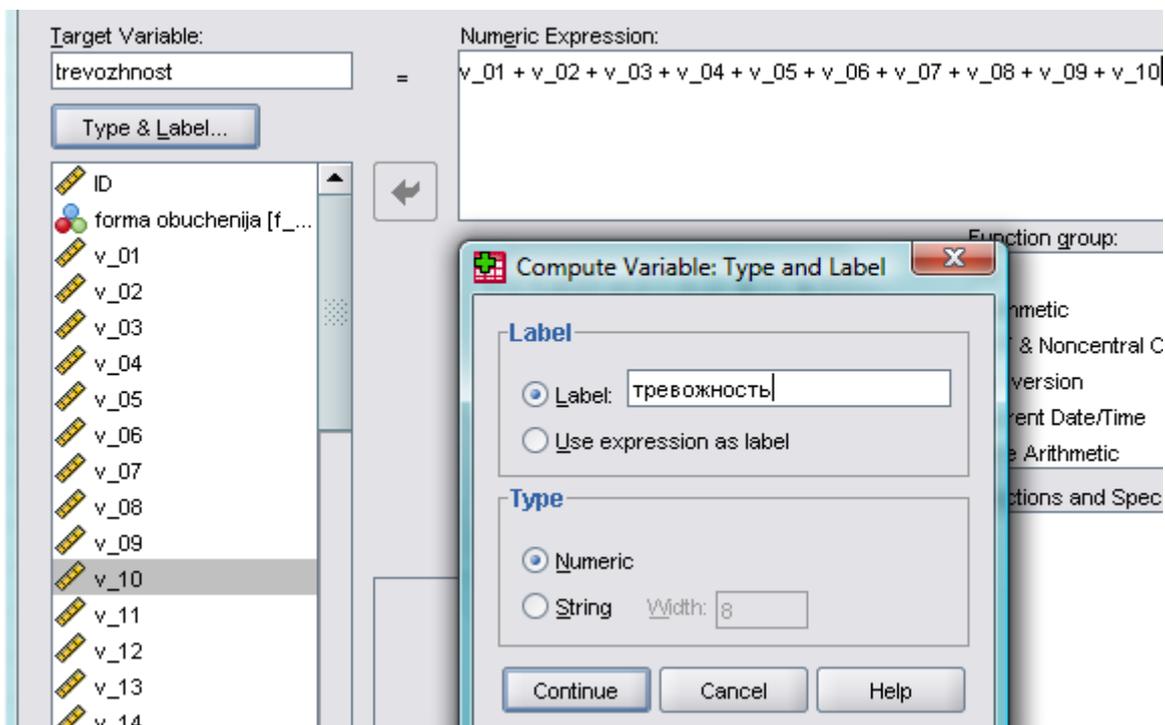
Введите имя вычисляемой переменной. Ей может быть существующая переменная или новая переменная, которую нужно добавить в активный набор данных.

Для создания выражения можно вставлять компоненты в поле выражения или ввести условие вручную в поле выражения.



- Вы можете вставлять функции или часто используемые системные переменные, выбрав группу функций из списка Function group и дважды щелкнув затем на функции или переменной в списке Functions and Special variables (или выбрав функцию или переменную и замет щелкнув на кнопке со стрелкой). Введите все параметры, отмеченные знаками вопроса (применимо только к функциям). Выбор группы All обеспечивает вывод списка всех доступных функций и системных переменных. В специально выделенной области диалогового окна отображается краткое описание выбранной функции или переменной.
- Текстовые константы должны быть заключены в апострофы.
- В значениях с десятичными знаками в качестве десятичного разделителя должна использоваться точка (.).
- Для новых текстовых переменных Вам придется нажать на кнопку Type & Label, чтобы задать тип переменной.

Пример. Вычислим показатель «Тревожность» по данным теста Айзенка. Для этого нам необходимо просуммировать результаты ответов на первые 10 вопросов теста. Необходимые действия отражены на рисунке.



После вычислений в файле данных появится новая переменная (и ее вычисленные значения) с указанными свойствами.

Аналогичные действия необходимо провести для вычисления и остальных трех показателей.

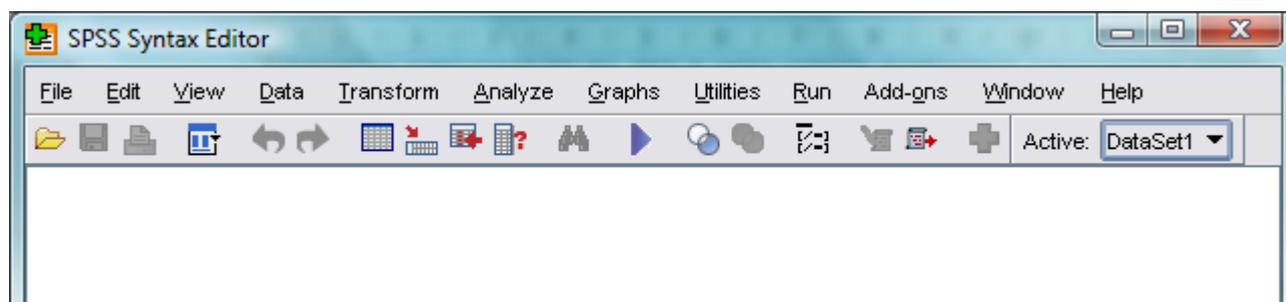
Обратите внимание на то, что в окне отчета появилась следующая запись:

```
COMPUTE trevozhnost=v_01 + v_02 + v_03 + v_04 + v_05 + v_06 + v_07
+ v_08 + v_09 + v_10 .
VARIABLE LABELS trevozhnost 'тревожность' .
EXECUTE .
```

Именно такой набор команд и был выполнен программой. Мы могли его задать и сами, создав так называемый файл синтаксиса и запустив его на выполнение.

Для этого необходимо выполнить следующие действия: File -> New -> Syntax.

В появившемся окне текстового (фактически) редактора набрать приведенный выше текст команд. Для запуска их на выполнение надо выполнить команду Run -> All.



Во многих ситуациях (например, при необходимости многократного повторения одних и тех же действий) использование файла синтаксиса предпочтительнее.

Альтернативой приведенного выше математического выражения является следующее:

```
COMPUTE trevozhnost=sum(v_01 to v_10) .
```

Здесь мы воспользовались одной из многочисленных встроенных функций программы и тем, что суммируемые переменные идут подряд в списке переменных файла данных.

Compute Variable: If Cases (Вычислить переменную: Отбор наблюдений)

Выберите в меню: Transform -> Compute variable

Затем введите имя вычисляемой переменной (тип и метку указывать необязательно, программа тип назначит самостоятельно, исходя из значений переменной).

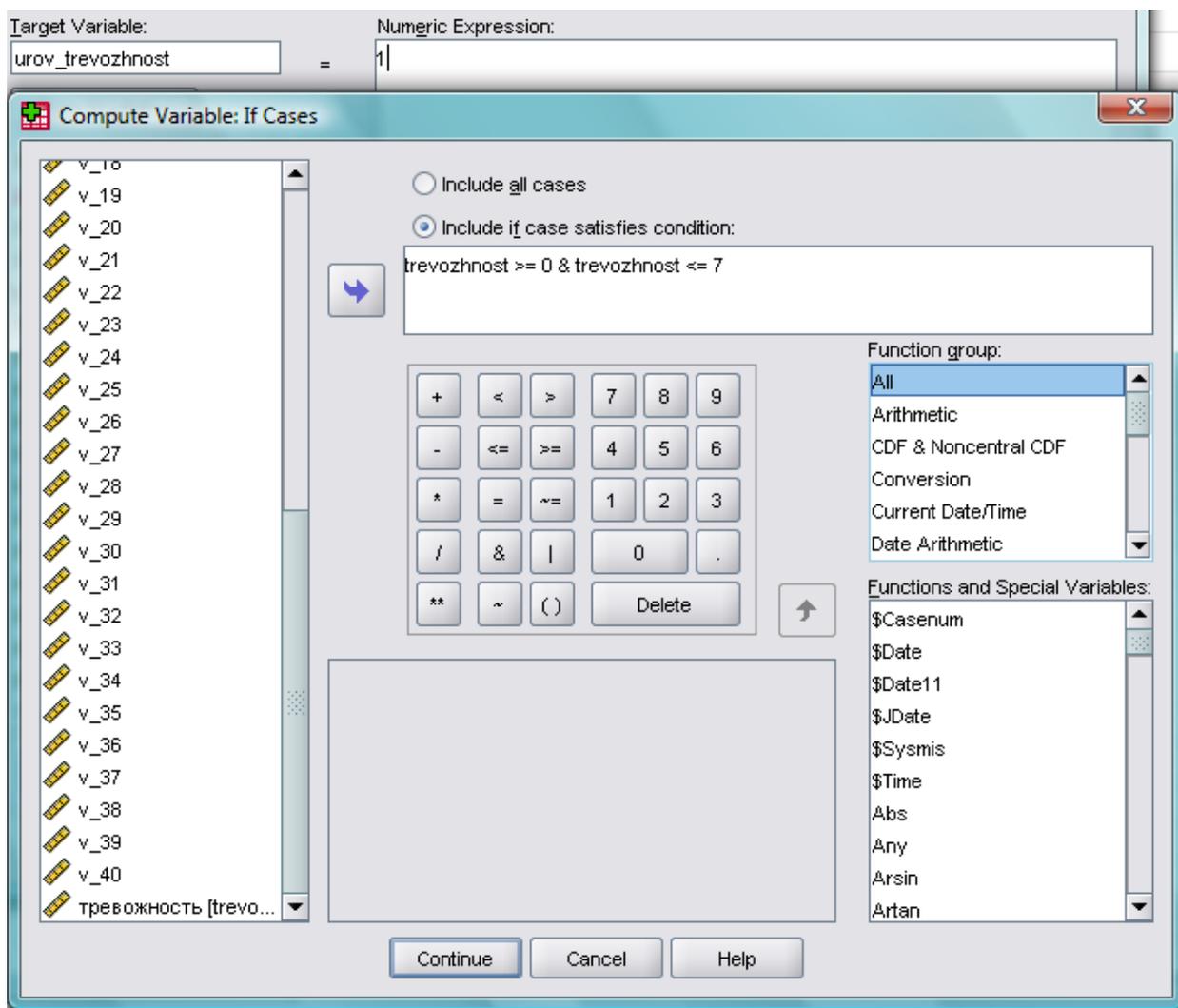
В диалоговом окне преобразования данных щелкните по кнопке If.

Диалоговое окно Compute Variable: If Cases позволяет применять преобразования к подмножествам наблюдений, используя логические выражения. Условие возвращает для каждого наблюдения значение истина, ложь или пропущенное.

- Большинство условий используют один или более из шести операторов отношения (<, >, <=, >=, =, ~=), находящихся на панели калькулятора.
- В условиях можно использовать имена переменных, константы, арифметические операторы, числовые и другие функции, логические переменные и операторы сравнения.

Пример. Вычислим показатель «Уровень тревожности» по данным теста Айзенка с использованием только что вычисленных сумм баллов по этому показателю.

Создадим новую переменную `urov_trevozhnost`, присвоим ей значение 1 и в диалоговом окне Compute Variable: If Cases введем указанное на картинке выражение (в нем используется логический оператор AND (&)).



Естественно, в файле данных появилась новая переменная со значением 1 (остальные уровни необходимо вычислить аналогично). Вот здесь нам и может пригодиться файл синтаксиса, позволяющий:

- 1) сохранять команды для последующего повторного выполнения;
- 2) упрощать набор многократно повторяющихся операций (в нашем случае предстоит выполнить 12 подобных действий – 4 переменных с 3-мя уровнями).

Перекодирование переменных

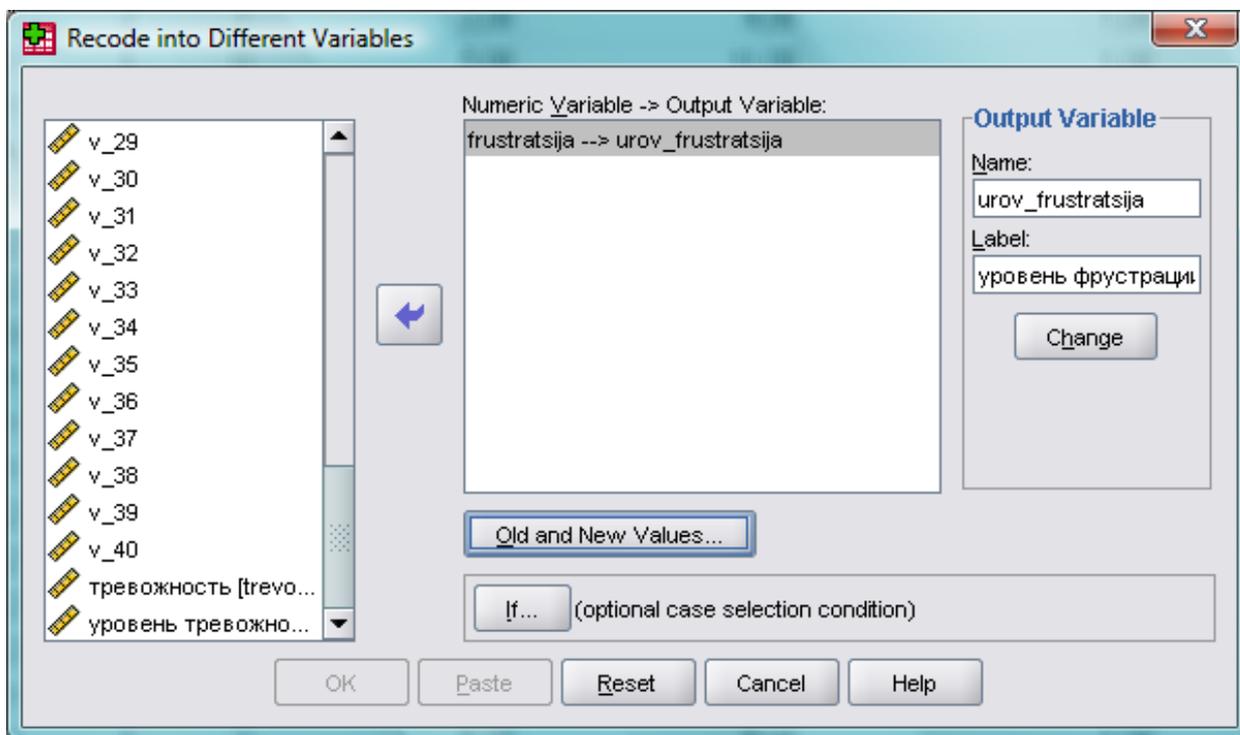
Для выполнения последнего действия (определения уровней выраженности психологической характеристики) проще воспользоваться возможностями перекодирования значений.

В программе SPSS есть две возможности перекодирования:

- по верх уже существующих;
- в новые переменные.

В нашем случае для сохранения исходных данных (сумм баллов) необходимо воспользоваться второй возможностью (различие между этими методами состоит в том, что во втором случае в диалоговых окнах нам придется указать название новой переменной).

Выполняем команду Transform -> Record into Different Variables



В появившемся окне мы должны выбрать перекодируемую переменную (frustratsija), указать имя и метку новой переменной и нажать кнопку Change.

Затем в диалоговом окне Old and New Values указать параметры перекодирования. Для нашего случая они представлены на рисунке.

Соответствующий командный код выглядит следующим образом:

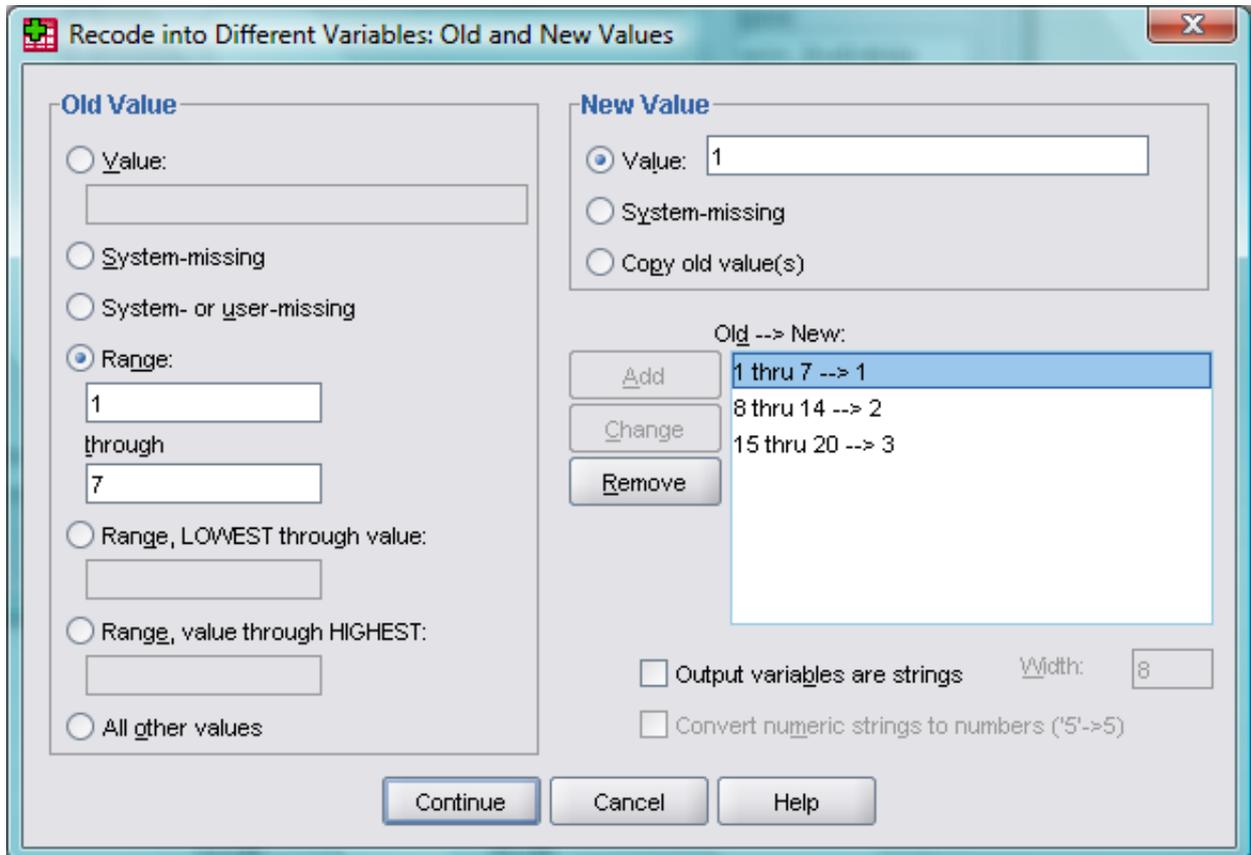
```
RECODE frustratsija (1 thru 7=1) (8 thru 14=2) (15 thru 20=3) INTO
urov_frustratsija.
```

```
VARIABLE LABELS urov_frustratsija 'уровень фрустрации'.
```

```
EXECUTE.
```

Можно перекодировать значения переменных или диапазоны значений в новые значения новой переменной. Например, можно перейти от зарплаты в числовом выражении к новой переменной зарплаты, в которой будет несколько категорий, представляющих диапазоны зарплат.

- Можно перекодировать числовые и текстовые переменные.
- Можно перекодировать числовые переменные в текстовые и наоборот.
- Если Вы перекодируете несколько переменных, то они должны быть одного типа. Не допускается одновременное перекодирование числовых и текстовых переменных.



В нашем случае для сокращения времени, затрачиваемого на перекодирование, можно воспользоваться возможностью одновременного (одинакового) перекодирования нескольких переменных.

Сортировка наблюдений

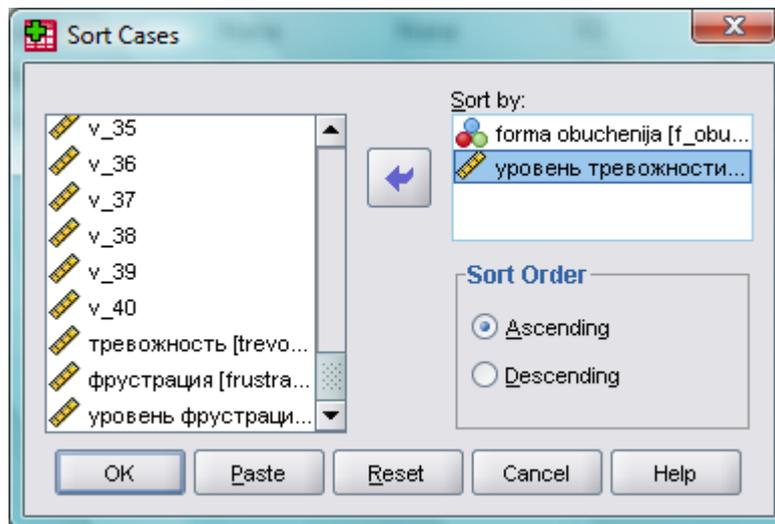
Данные в SPSS можно сортировать в соответствии со значениями одной или нескольких переменных. Рассмотрим следующий пример: Требуется упорядочить данные по форме обучения, а внутри – по уровню тревожности. Для этого поступите следующим образом:

Выберите в меню команды Data (Данные) Sort Cases... (Сортировать наблюдения)

Откроется диалоговое окно Sort Cases. Переменные файла данных будут отображены в списке исходных переменных.

Перенесем нужные переменные в список Sort by (Сортировать по). В группе Sort order (Порядок сортировки) по умолчанию выбран вариант Ascending (По возрастанию). Эта опция сортирует наблюдения в порядке возрастания значения переменной сортировки, а следующая опция, Descending — в порядке убывания.

Подтвердите настройки кнопкой ОК. В редакторе данных файл будет отсортирован в нужном порядке.



Выбранные опции соответствуют следующему командному синтаксису:

```
SORT CASES BY f_obuch(A) urov_trevozhnost(A) .
```

или, если надо сортировать по убыванию:

```
SORT CASES BY f_obuch(D) urov_trevozhnost(D) .
```

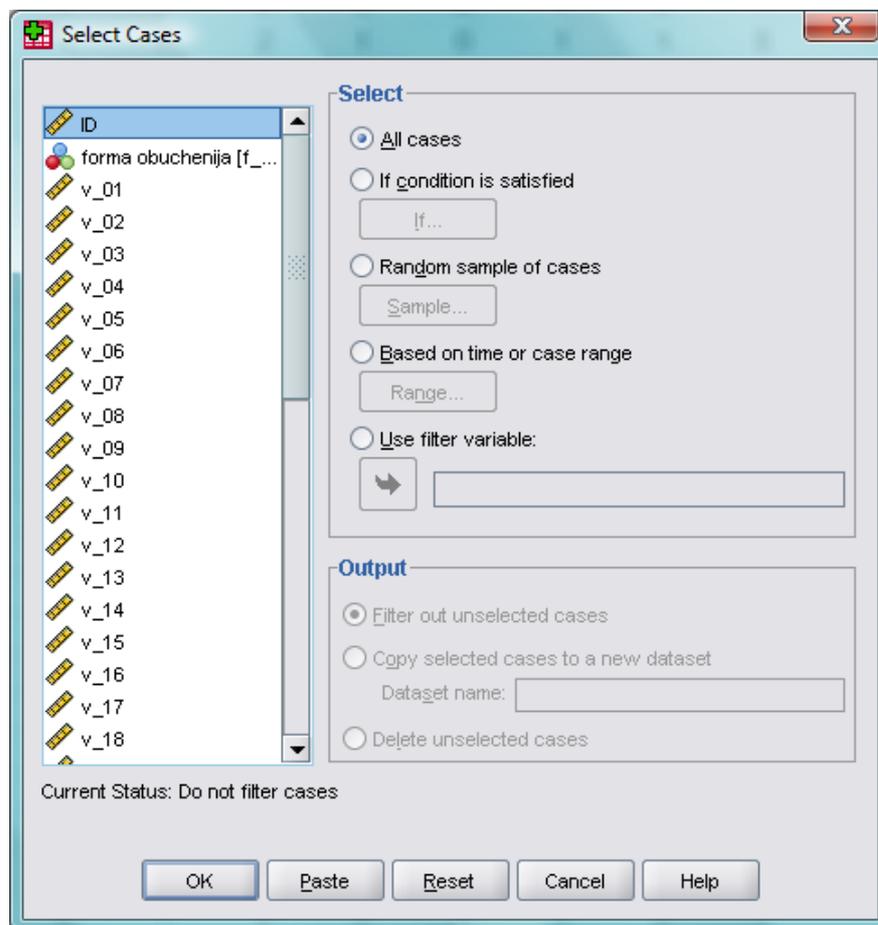
- Если выбрать несколько переменных, наблюдения сортируются по значениям каждой переменной внутри категорий, задаваемых значениями предшествующей переменной в списке сортирующих. Если, например, Вы выберете пол, а затем национальность, то наблюдения будут отсортированы сначала по значениям переменной пол, а затем внутри каждой полученной категории — по значениям переменной национальность.
- Сортировка последовательности происходит в порядке, определяемом языковым стандартом, и может не совпадать с числовым порядком кодов символов. Языковой стандарт задается в настройках операционной системы. Пользователь может изменить настройки языка на вкладке «Общие» диалогового окна «Параметры» (меню «Правка»).

Отбор наблюдений

Процедура Data -> Select Cases позволяет выбирать подмножества наблюдений несколькими способами на основе критериев, включающих переменные и сложные выражения. Можно также отобрать случайную подвыборку наблюдений.

Критерии, используемые для отбора наблюдений:

- значения и диапазоны значений переменных;
- диапазоны дат и времени;
- номера наблюдений (строк);
- арифметические выражения;
- логические выражения;
- функции.



Select (Отбор)

All cases Все наблюдения	Отключает фильтрацию наблюдений и использует все наблюдения.
If condition is satisfied Если выполнено условие	Использует условное выражение для выбора наблюдений. Отбираются наблюдения, для которых это выражение истинно. Если результат - ложь или пропущено, наблюдение не выбирается.
Random sample of cases Случайная подвыборка	Создает случайную выборку, размер которой составляет приблизительный процент общего числа наблюдений или в точности равен указанному числу наблюдений.
Based on time or case range Временной диапазон или диапазон наблюдений	Отбирает наблюдения на основании интервала номеров наблюдений или интервала дат/времени.
Use filter variable Использовать фильтрующую переменную	Использует в качестве фильтрующей выбранную числовую переменную файла данных. Отбираются наблюдения с любыми значениями переменной фильтра, кроме 0 и пропущенных.

Output (Вывод)

В этом разделе устанавливаются варианты действий с неотобранными наблюдениями. Вы можете выбрать один из следующих вариантов:

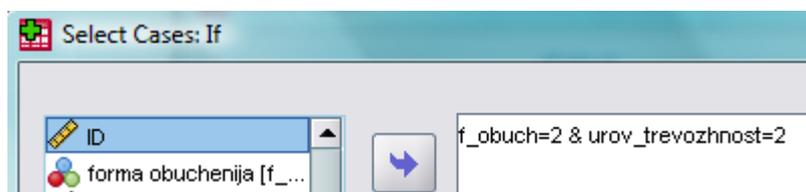
Filter out unselected cases Отфильтровать неотобранные наблюдения	Неотобранные наблюдения не включаются в анализ, но остаются в наборе данных. Вы сможете использовать неотобранные наблюдения позднее в данном сеансе, если отключите фильтрацию. Если Вы отбираете случайную выборку или отбираете наблюдения на основе условия, то генерируется переменная с именем filter_\$, имеющая значения 1 для отобранных наблюдений и значение 0 для неотобранных наблюдений.
Copy selected cases to a new dataset Копировать отобранные наблюдения в новый набор данных	Отобранные наблюдения копируются в новый набор данных, при этом исходный набор данных остается неизменным. Неотобранные наблюдения не включаются в новый набор данных и остаются в прежнем виде в исходном наборе данных.
Delete unselected cases Неотобранные наблюдения удаляются	Неотобранные наблюдения удаляются из набора данных. Удаленные наблюдения могут быть восстановлены только путем выхода без сохранения каких-либо изменений и затем повторным открытием файла. Удаленные наблюдения не восстанавливаются, если Вы сохраняете изменения в файле данных.

Если удалить неотобранные наблюдения и сохранить файл, наблюдения восстановить невозможно.

Чаще всего применяются «Если выполнено условие» и «Отфильтровать неотобранные наблюдения»

Пример. Выбрать из нашего файла записи, соответствующие студентам открытого университета со средним уровнем тревожности.

Настройки отбора представлены на рисунке.



О том, что проведен отбор данных, свидетельствуют зачеркнутые номера тех строк файла, данные в которых не соответствуют условиям отбора.

Снять фильтрацию можно выбрав в диалоговом окне All Cases.

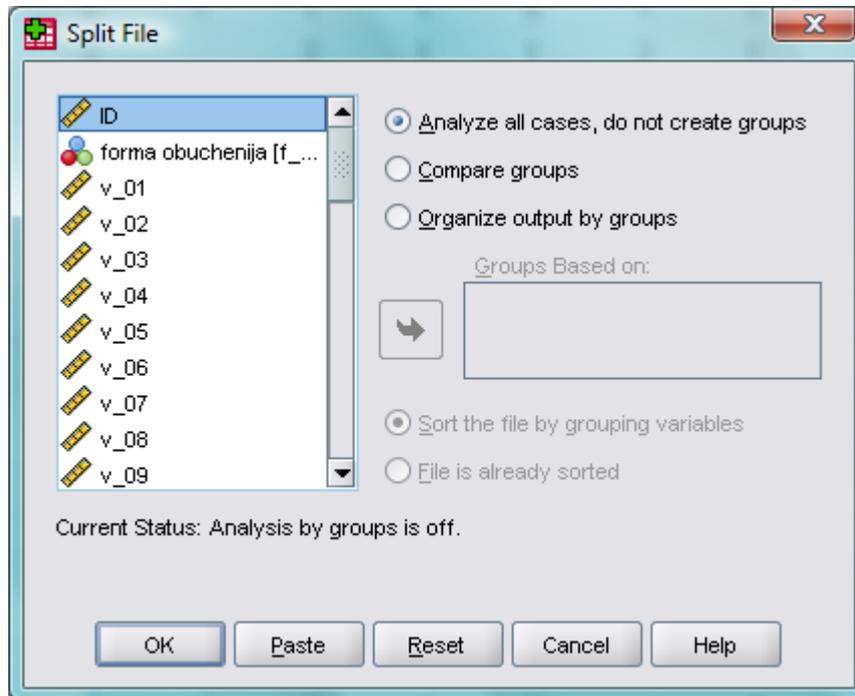
Расщепление

Процедура Data - > Split File «расщепляет» данные файла на отдельные группы для анализа, в соответствии со значениями одной или нескольких группирующих переменных. Если задано несколько группирующих переменных, то наблюдения будут группироваться по значениям каждой переменной внутри групп, образованных значениями предыдущей переменной в списке Groups Based on. Если, например, Вы выберете пол, а затем национальность, то наблюдения будут отсортированы сначала по значениям переменной пол, а затем внутри каждой полученной категории — по значениям переменной национальность.

- Можно задать до 8-ми группирующих переменных.
- Наблюдения должны быть отсортированы по значениям группирующих переменных и в том же порядке, что и переменные в списке Groups Based on. Если файл данных еще не отсортирован, выберите альтернативу Sort the file by grouping variables (эта опция

установлена по умолчанию).

О том, что файл данных находится в расщепленном состоянии, свидетельствует надпись Split File on в статусной строке программы (справа внизу окна программы).



<p>Compare groups Сравнить группы</p>	<p>Группы расщепления представляются вместе для сравнения. В случае мобильных таблиц - создается одна таблица, в которую включаются группирующие переменные, причем после построения их можно свободно перемещать между строками, столбцами и слоями таблицы. В случае диаграмм - для каждой группы в расщепленном файле создается отдельная диаграмма, и эти диаграммы выводятся вместе во Viewer (отчет SPSS).</p>
<p>Organize output by groups Организовать вывод по группам</p>	<p>Результаты каждой процедуры выводятся отдельно для каждой группы.</p>

Пример. Для получения информации по студентам разных форм обучения необходимо выполнить расщепление файла данных по соответствующей переменной.

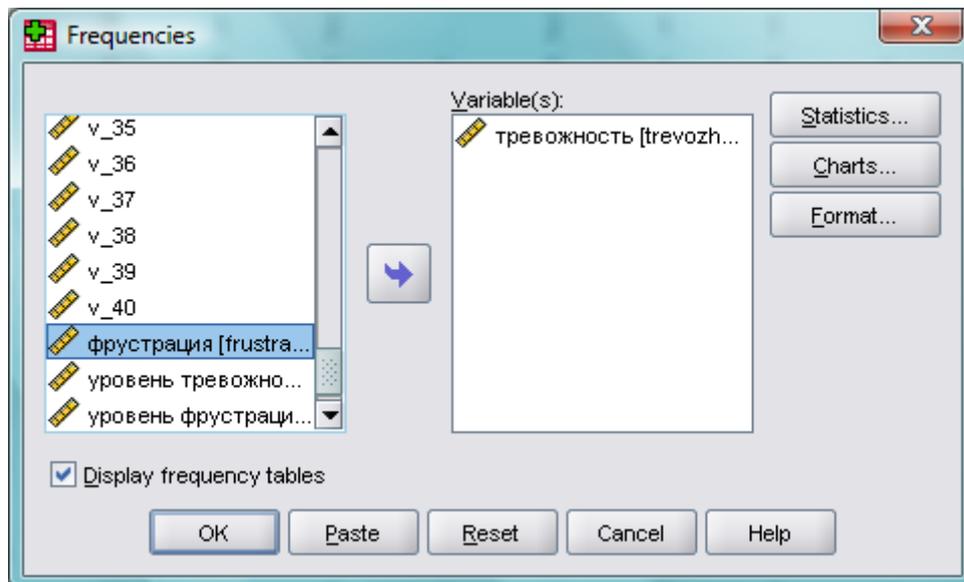
Для снятия расщепления файла необходимо выбрать опцию Analyze all cases.

Частотные таблицы и описательная статистика

Все рассматриваемые в данном курсе процедуры статистики находятся в пункте меню Analyze.

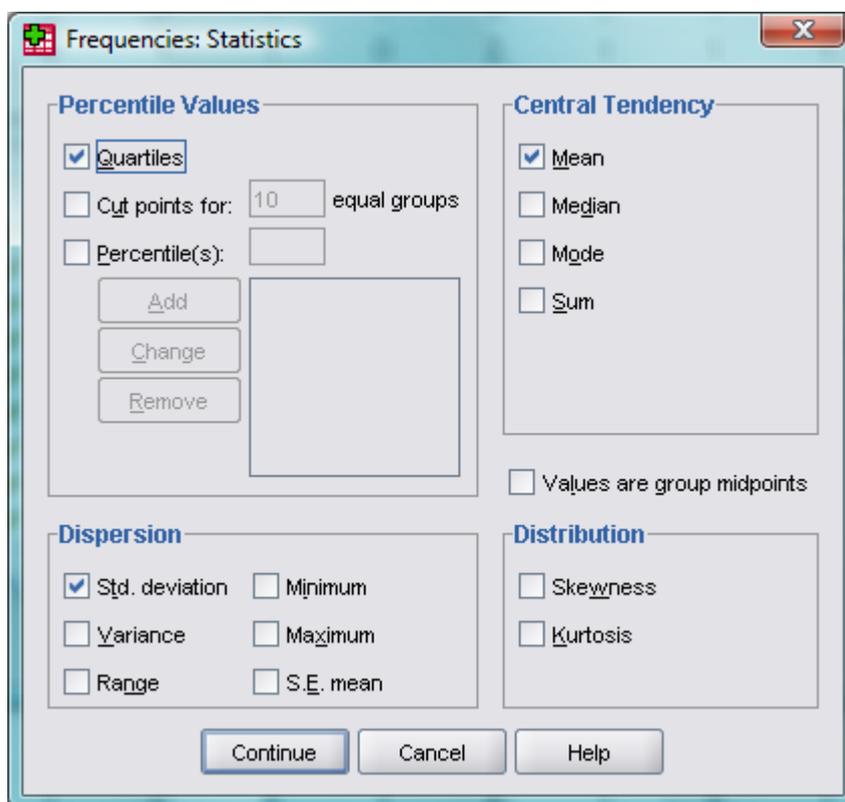
Частота вариантов (встретившихся значений переменной) – одна из базовых характеристик переменной. Обычно с частотного анализа и начинается исследование полученных данных.

Необходимо выполнить команды Analyze -> Descriptive Statistics -> Frequencies.



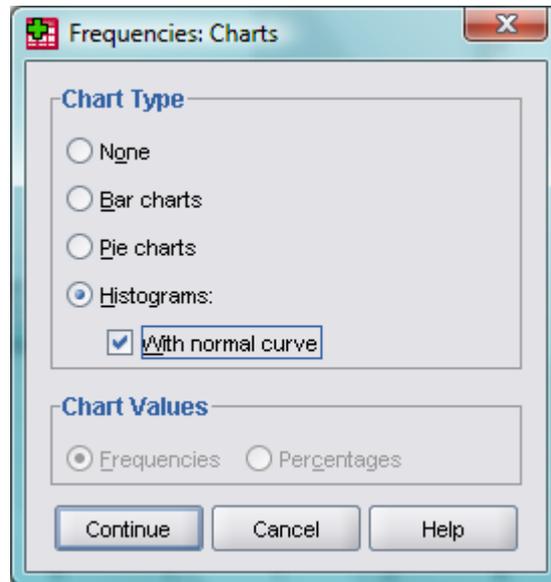
Исследуемые переменные переносим в список Variable(s).

Возможен не только вывод таблицы частот, но и расчет основных показателей описательной статистики (кнопка Statistics).

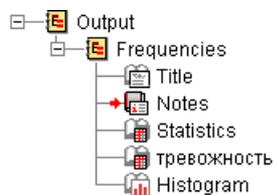


Описание соответствующих статистических характеристик было дано в предыдущем разделе курса.

Возможен также одновременный вывод графика, построенного на основе таблицы частот.



Результаты выполнения процедуры представлены на рисунках.



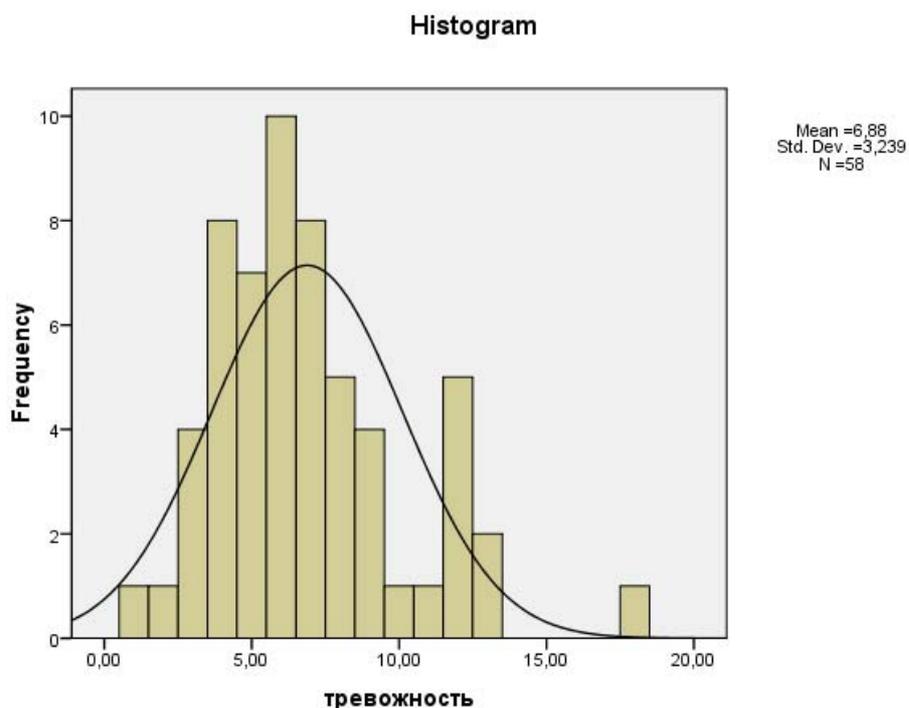
Frequencies

Statistics

тревожность		
N	Valid	58,0000
	Missing	2,0000
Mean		6,8793
Std. Deviation		3,2394
Percentiles	25	4,7500
	50	6,0000
	75	8,2500

тревожность

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1,00	1	1,7	1,7	1,7
	2,00	1	1,7	1,7	3,4
	3,00	4	6,7	6,9	10,3
	4,00	8	13,3	13,8	24,1
	5,00	7	11,7	12,1	36,2
	6,00	10	16,7	17,2	53,4
	7,00	8	13,3	13,8	67,2
	8,00	5	8,3	8,6	75,9
	9,00	4	6,7	6,9	82,8
	10,00	1	1,7	1,7	84,5
	11,00	1	1,7	1,7	86,2
	12,00	5	8,3	8,6	94,8
	13,00	2	3,3	3,4	98,3
	18,00	1	1,7	1,7	100,0
Total		58	96,7	100,0	
Missing	System	2	3,3		
Total		60	100,0		



Вычисление основных описательных статистик входит во многие инструменты анализа данных. Ниже представлены диалоговые окна другого инструмента – Analyze -> Descriptive Statistics -> Descriptives.

0	1	1	0	2	1	1	1
0	0	0	1	1	1	0	1
0	1	2	.	2	1	1	2

0	0	0	0	0	1	0	1
1	0	0	1	0	1	0	1
0	0	1	0	0	0	1	1

Имеется также возможность сохранения для выбранных переменных их стандартизованных значений. Под стандартизацией понимается пересчет величин в новые значения с равным нулю средним и единичной дисперсией).

Таблицы сопряженности

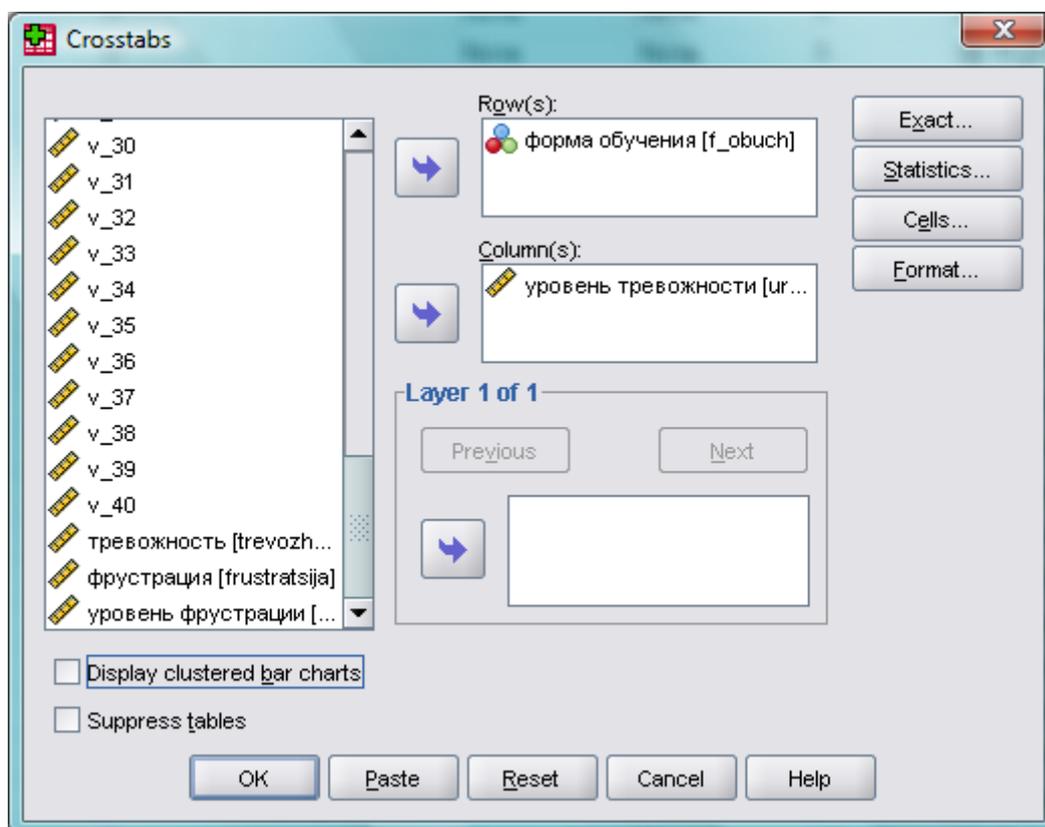
Таблицы сопряженности – один из наглядных и эффективных инструментов анализа связей между явлениями. Связь между неметрическими переменными, то есть переменными, относящимися к номинальной шкале или к порядковой шкале с не очень большим количеством категорий, лучше всего представить в форме таблиц сопряженности. Для этой цели в SPSS реализован тест χ^2 , при котором проверяется, есть ли значимое различие между наблюдаемыми и ожидаемыми частотами.

Процедура формирует двумерные и многомерные таблицы, а также вычисляет целый ряд критериев и мер силы связи для двумерных таблиц. Структура таблицы и то, упорядочены категории или нет, определяет, какие меры и критерии использовать.

Таблицы сопряженности по своей сути – аналоги сводных таблиц MS Excel.

Покажем основные шаги создания таблицы сопряженности на примере наших данных. Построим распределение по уровню тревожности студентов обеих форм обучения.

Для создания таблиц сопряженности и вычисления меры связанности на их основе, выберите в меню команды Analyze (Анализ) Descriptive Statistics (Дескриптивные статистики) Crosstabs... (Таблицы сопряженности). Откроется диалоговое окно Crosstabs.



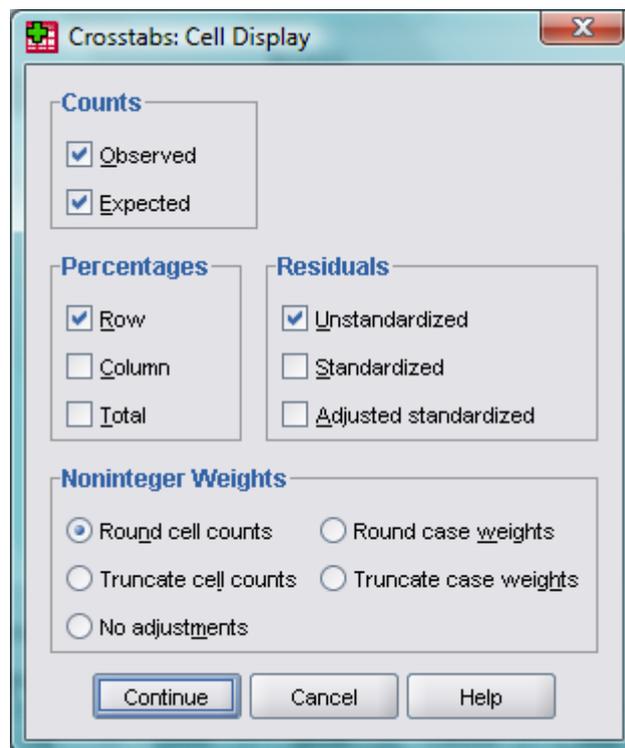
Перенесем указанные переменные в области строк и столбцов. Область слоев используется для создания многомерных таблиц сопряженности. После выполнения программой нашего задания получим следующий результат:

форма обучения * уровень тревожности Crosstabulation

Count		уровень тревожности			
		низкий	средний	высокий	Total
форма обучения	Statsionaar	15	8	0	23
	Avatud Ulikool	24	10	1	35
	Total	39	18	1	58

Повторим вызов команды построения таблицы.

Щелкнем на кнопке Cells... (Ячейки). Откроется диалоговое окно Crosstabs: Cell Display (Таблицы сопряженности: Отображение ячеек).



По умолчанию в ячейках таблицы сопряженности отображаются только наблюдаемые значения частот. В группе Counts (Частоты) можно выбрать один или более следующих вариантов отображения:

- Observed (наблюдаемые): Будут отображаться наблюдаемые частоты. Это настройка по умолчанию.
- Expected (Ожидаемые): Если установить этот флажок, будут отображаться ожидаемые частоты. Они вычисляются как произведение сумм соответствующей строки и столбца, деленное на общую сумму частот.

В группе Residuals (Остатки) можно выбрать один или более следующих вариантов отображения:

- Unstandardized (Ненормированные): Отображаются ненормированные остатки, то есть разность наблюдаемых и ожидаемых частот.
- Standardized (Нормированные): Отображаются нормированные остатки.

По умолчанию в таблицах сопряженности приводятся только абсолютные значения. Чтобы узнать, насколько эти значения важны по отношению к общему количеству, надо определить

их процентную долю для вычисления процентных значений необходимо в группе Percentages (Проценты) можно выбрать один или более из нижеследующих вариантов отображения:

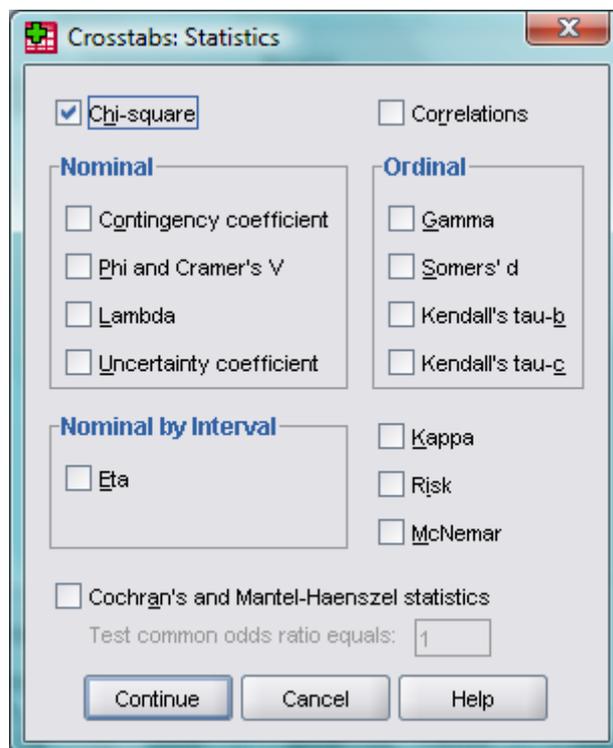
- Row (По строкам): Вычисляются процентные значения по строкам: количество наблюдений в каждой ячейке, отнесенное к сумме по строке.
- Column (По столбцам): Вычисляются процентные значения по столбцам: количество наблюдений в каждой ячейке в отношении к сумме столбца.
- Total (Полные): Вычисляются полные процентные значения: количество наблюдений в каждой ячейке, отнесенное к общей сумме наблюдений.

Результаты вычислений с указанными настройками теперь выглядят следующим образом:

форма обучения * уровень тревожности Crosstabulation

		уровень тревожности			
		низкий	средний	высокий	Total
форма обучения Stasionaar	Count	15	8	0	23
	Expected Count	15,5	7,1	,4	23,0
	% within форма обучения	65,2%	34,8%	,0%	100,0%
	Residual	-,5	,9	-,4	
Avatud Ulikool	Count	24	10	1	35
	Expected Count	23,5	10,9	,6	35,0
	% within форма обучения	68,6%	28,6%	2,9%	100,0%
	Residual	,5	-,9	,4	
Total	Count	39	18	1	58
	Expected Count	39,0	18,0	1,0	58,0
	% within форма обучения	67,2%	31,0%	1,7%	100,0%

Чтобы получить статистические критерии для таблиц сопряженности, щелкните на кнопке Statistics... (Статистика) в диалоговом окне Crosstabs. Откроется диалоговое окно Crosstabs: Statistics (Таблицы сопряженности: Статистика).



Флажки в этом диалоговом окне позволяют выбрать один или несколько критериев.

Эти критерии кратко описаны ниже. Наиболее популярным является уже знакомый нам критерий хи-квадрат.

Результаты применения этого критерия представлены в таблице (одновременно проводятся вычисления и по другим критериям):

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	,853 ^a	2	,653
Likelihood Ratio	1,204	2	,548
Linear-by-Linear Association	,001	1	,971
N of Valid Cases	58		

a. 2 cells (33,3%) have expected count less than 5. The minimum expected count is ,40.

Очень далекие от обычно используемых уровней значимости (0,05 или 0,01) значения двухсторонней значимости в последнем столбике свидетельствуют о том, что распределения студентов обеих форм обучения по уровню тревожности статистически неразличимы.

Статистики, вычисляемые для таблиц сопряженности

Хи-квадрат

Отметьте Хи-квадрат, чтобы получить значения критериев хи-квадрат Пирсона, хи-квадрат отношения правдоподобия, точного критерия Фишера и критерия хи-квадрат с поправкой Йетса (с поправкой на непрерывность) для таблиц, образованных двумя строками и двумя столбцами. Для таблиц 2×2 точный критерий Фишера вычисляется в том случае, когда

таблица, которая не является результатом наличия пропущенных строк или столбцов в таблице большего размера, имеет ожидаемое значение меньше 5 хотя бы в одной ячейке. Для всех остальных таблиц размерности 2×2 рассчитывается критерий хи-квадрат с поправкой Йетса. Для таблиц с любым числом строк и столбцов отметьте Хи-квадрат, чтобы вывести значения хи-квадрата Пирсона и хи-квадрат отношения правдоподобия. Если обе переменные в таблице являются количественными, то при пометке элемента Хи-квадрат рассчитывается критерий линейно-линейной связи.

Корреляции

Для таблиц с упорядоченными переменными по строкам и столбцам при пометке элемента Корреляции вычисляются значения коэффициента корреляции Спирмана - ρ (только для числовых данных). ρ Спирмана является мерой связи между порядковыми переменными. Если обе переменные в таблице (факторы) являются числовыми, параметр Корреляции позволяет вычислить коэффициент корреляции Пирсона r , который характеризует силу линейной связи между переменными.

Номинальные

Для номинальных данных (которые не имеют естественного порядка - например, католическое, протестантское, иудейское вероисповедание) можно выбрать одну из следующих статистик: Коэффициент сопряженности, Φ (коэффициент) и V Крамера, Лямбда (симметричное и асимметричное значения лямбда, статистика τ Гудмана и Краскала), Коэффициент неопределенности.

- Коэффициент сопряженности. Мера связи, основанная на хи-квадрат. Это значение меняется между 0 и 1, причем 0 означает отсутствие связи между переменными строки и столбца, а значение, близкое к 1, - высокую степень связи между этими переменными. Максимально возможное значение зависит от числа строк и столбцов в таблице.
- Φ и параметр V Крамера. Мера связи, вычисляется делением статистики хи-квадрат на объем выборки и взятием корня квадратного из результата. V Крамера - это мера связи, основанная на статистике хи-квадрат.
- Лямбда. Мера связи, которая отражает относительное снижение ошибки, когда значения независимой переменной используются для предсказания значений зависимой переменной. Значение 1 означает, что независимая переменная точно предсказывает значения зависимой. Значение 0 означает, что независимая переменная абсолютно бесполезна для предсказания зависимой.
- Коэффициент неопределенности. Мера связи, указывающая относительное снижение ошибки в случае, когда значения одной переменной используются для предсказания значений другой. Например, значение 0.83 указывает на то, что знание одной переменной уменьшает ошибку в предсказании значений другой на 83%. Вычисляются как симметричная, так и несимметричная версии коэффициента неопределенности.

Порядковые

Для таблиц, в которых как строки, так и столбцы содержат упорядоченные значения, пометьте Гамма (нулевого порядка для двумерных таблиц и условное для таблиц размерности от 2 до 10), τ -b Кендалла и τ -c Кендалла. Для предсказания категорий столбца по категориям строки, пометьте d Сомерса.

- Гамма. Симметричная мера связи между двумя порядковыми переменными, значения которой меняются между -1 и 1. Значения, близкие по абсолютной величине к 1, указывают на сильную связь переменных. Значения, близкие к 0, говорят о слабой связи или ее отсутствии. Для таблиц сопряженности двух переменных вычисляется гамма

нулевого порядка. Если же таблица сопряженности включает более двух переменных, для каждой подтаблицы вычисляется условная гамма.

- **d Сомерса.** Мера связи между двумя порядковыми переменными, изменяется между -1 и 1 . Значения, близкие по абсолютной величине к 1 , указывают на сильную связь между двумя переменными, а значения, близкие к 0 , — на слабую связь или ее отсутствие. Это асимметричное расширение меры гамма, отличающееся только включением числа пар, не имеющих совпадений (связей) по независимой переменной. Вычисляется также симметричная версия этой статистики.
- **Тау-b Кендалла.** Непараметрическая мера корреляции для порядковых или ранговых переменных, которая учитывает возможные совпадения значений (связи). Знак коэффициента указывает направление связи, а его модуль - силу связи, причем, чем он больше, тем связь сильнее. Значения изменяются в диапазоне между -1 и $+1$, однако -1 и $+1$ можно получить только для квадратных таблиц.
- **Тау-c Кендалла.** Непараметрическая мера связи для порядковых переменных, игнорирующая возможные совпадения значений (связи). Знак коэффициента указывает направление связи, а его модуль - силу связи, причем, чем он больше, тем связь сильнее. Значения изменяются в диапазоне между -1 и $+1$, однако -1 и $+1$ можно получить только для квадратных таблиц.

Номин./интерв

В ситуации, когда одна из переменных категориальная, а другая - количественная, выберите статистику Эта. Значения категориальной переменной должны быть закодированы числами.

Эта	Мера связи между переменными строки и столбца, значения которой изменяются от 0 (отсутствие связи) до 1 (сильная связь). Индикатор Эта подходит для зависимой переменной, измеренной в интервальной шкале (такой, как доход) и независимой переменной с ограниченным числом категорий (такой, как возраст). Вычисляются два значения эта: одно рассматривает переменную строки как интервальную переменную, а другое - переменную столбца как интервальную переменную.
Каппа	Каппа Коэна измеряет согласие мнений двух экспертов, оценивающих одни и те же объекты. Значение 1 указывает на полное согласие. Значение 0 указывает на то, что согласие - не более чем случайность. Этот индикатор определен только для таблиц, у которых обе переменные используют одинаковые значения категорий и имеют одинаковое число категорий.
Риск	Мера силы связи для таблиц 2×2 между наличием фактора и наступлением события. Если доверительный интервал для этой статистики включает 1 , предположение о том, что фактор связан с событием, будет неверным. Если наличие фактора встречается редко, то в качестве оценки относительного риска можно использовать отношение шансов.
МакНемара	Непараметрический критерий для двух связанных дихотомических переменных. Проверяет изменения в откликах с помощью распределения хи-квадрат. Полезен для выявления изменений в откликах, обусловленных экспериментальным вмешательством в планах до-и-после. Для больших квадратных таблиц выдаются результаты критерия симметричности Мак-Немара - Боукера.
Статистики Кокрена и	Статистики Кокрена и Мантеля-Хенцеля могут использоваться для проверки условной независимости дихотомической факторной

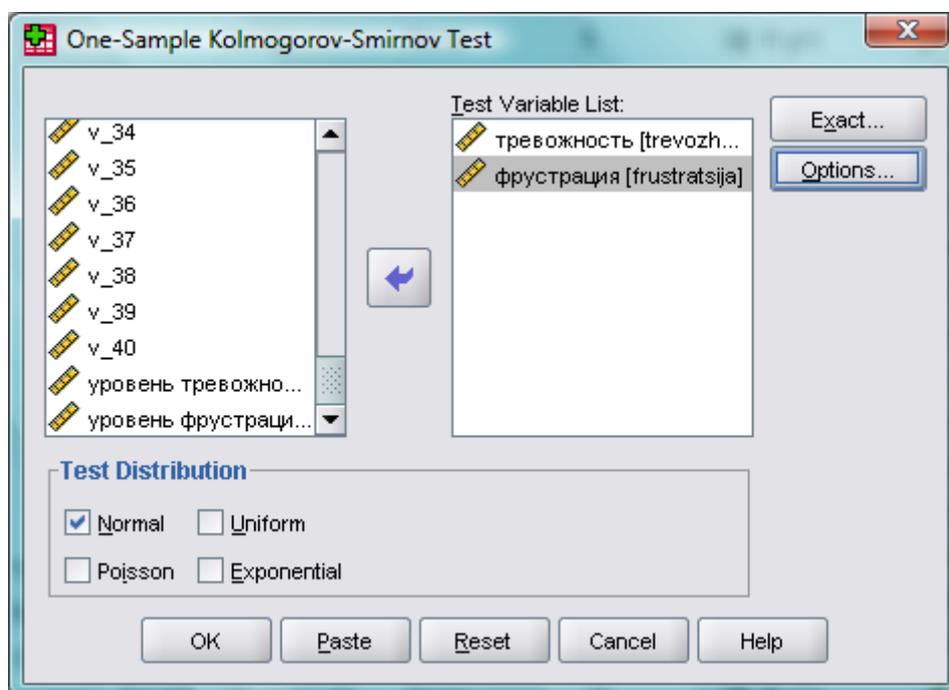
Мантеля-Хенцеля	переменной и дихотомической переменной отклика при заданных ковариационных структурах, задаваемых одной или большим числом переменных слоя (управляющих переменных). Заметим, что в то время как другие статистики вычисляются послойно, статистики Кокрена и Мантеля-Хенцеля вычисляются сразу для всех слоев.
-----------------	---

Проверка гипотезы о нормальности распределения

Для этих целей можно использовать тест Колмогорова-Смирнова для проверки формы распределения. При помощи этого теста по выбору можно проверить, соответствует ли реальное распределение переменной нормальному, равномерному, экспоненциальному распределению или распределению Пуассона. Разумеется, самым распространённым видом проверки является проверка наличия нормального распределения.

Выполним следующие команды:

Analyze -> Nonparametric Tests -> 1-Sample K-S



Перенесем интересующие нас переменные в список тестируемых. Укажем тестируемое распределение (Normal) и нажмем кнопку ОК.

		тревожность	фрустрация
N		58	60
Normal Parameters ^a	Mean	6,88	6,92
	Std. Deviation	3,239	2,948
Most Extreme Differences	Absolute	,158	,122
	Positive	,158	,122
	Negative	-,084	-,069
Kolmogorov-Smirnov Z		1,200	,945
Asymp. Sig. (2-tailed)		,112	,333

a. Test distribution is Normal.

Полученные результаты включают:

- среднее значение и стандартное отклонение
- промежуточные результаты, полученные в результате теста Колмогорова-Смирнова
- вероятность ошибки p .

Отклонение от нормального распределения считается существенным при значении $p < 0,05$; в этом случае для соответствующих переменных следует применять непараметрические тесты. В рассматриваемом примере (значение $p = 0,112$ и $0,333$) вероятность ошибки является не значимой; поэтому значения переменных достаточно хорошо подчиняются нормальному распределению.

Сравнение средних

Сравнение средних значений различных выборок относится к наиболее часто применяемым методам статистического анализа. При этом всегда должен быть выяснен вопрос, можно ли объяснить имеющееся различие средних значений статистическими колебаниями или нет. В последнем случае говорят о значимом различии.

При сравнении средних значений выборок предполагается, что обе выборки подчиняются нормальному распределению. Если это не так, то для сравнения выборок используется непараметрические тесты.

При сравнении средних значений выборок выделяют четыре различные тестовые ситуации:

- сравнение двух независимых выборок;
- сравнение двух зависимых (спаренных) выборок;
- сравнение более двух независимых выборок;
- сравнение более двух зависимых выборок.

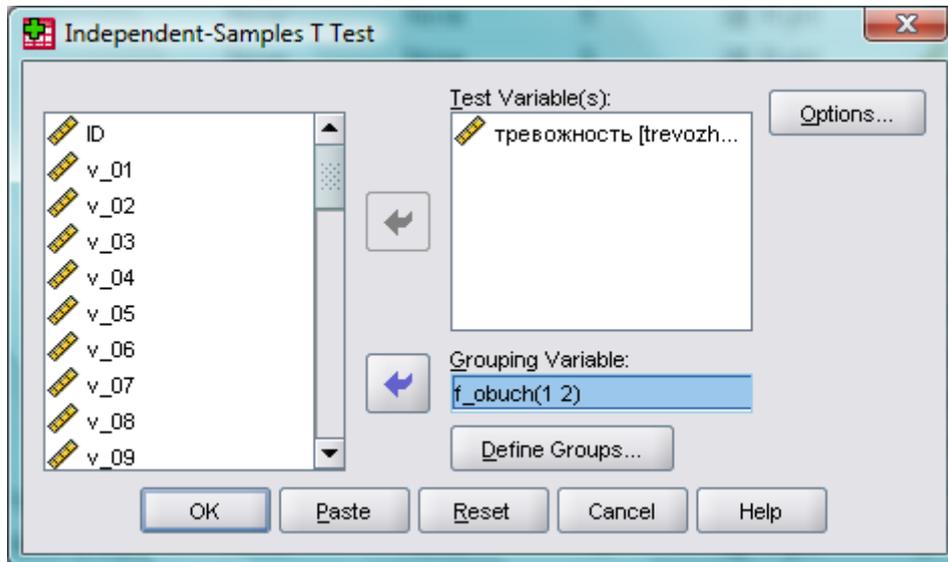
Посмотрим на примерах методику проведения этих тестов.

Сравнение двух независимых выборок

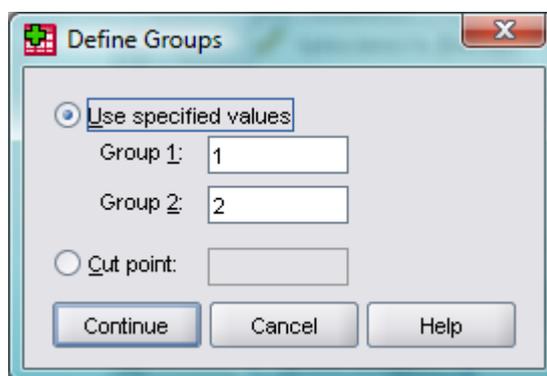
Параметрический t-тест Стьюдента

Analyze -> Compare Means -> Independent-Samples T Test

Проверим, отличаются ли полученные показатели тревожности (баллы) у студентов разных форм обучения. Для этого настроим процедуру проверки следующим образом:



Нужную переменную перевели в список тестируемых, переменную «форма обучения» назначили группирующей и определили сравниваемые группы с помощью команды Define Groups (напомним: 1- стационар, 2 – открытый университет).



Результат анализа представлен в таблице.

		Levene's Test for Equality of Variances				
		F	Sig.	t	df	Sig. (2-tailed)
тревожность	Equal variances assumed	,204	,653	-1,183	56	,242
	Equal variances not assumed			-1,223	52,195	,227

Одновременно проводится проверка равенства дисперсий сравниваемых групп. В данном случае эта гипотеза подтверждается ($0,653 \gg 0,05$) и значимость различий двух групп $p=0,242 \gg 0,05$. Следовательно, существенных различий в уровне тревожности между респондентами с разной формой обучения нет.

Для проведения этого теста файл данных не должен быть расщеплен по переменной «форма обучения».

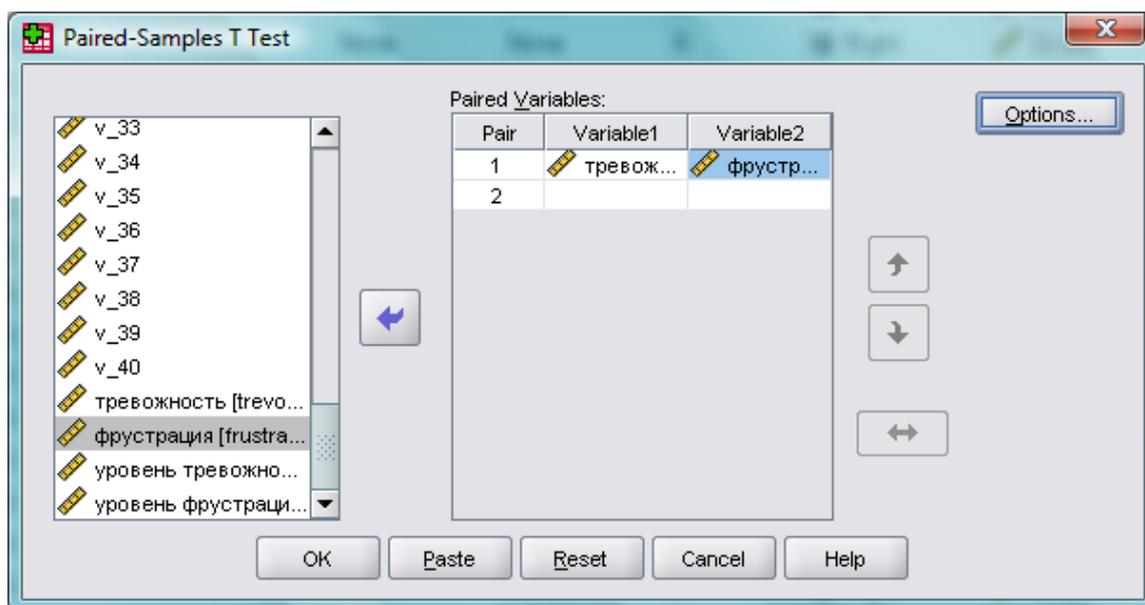
Сравнение двух зависимых (спаренных) выборок

Параметрический t-тест Стьюдента

Analyze -> Compare Means -> Paired-Samples T Test

Проверим, отличаются ли полученные показатели тревожности и фрустрации (баллы). Для

этого настроим процедуру проверки следующим образом:



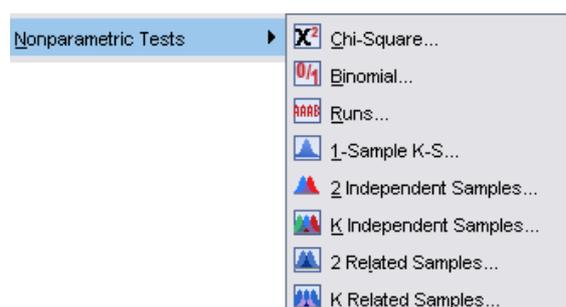
Одновременно можно тестировать несколько пар переменных.

Результат представлен в таблице.

		Pair 1
		тревожность - фрустрация
Paired Differences	Mean	,052
	Std. Deviation	3,322
	Std. Error Mean	,436
	95% Confidence Interval of Lower the Difference	Upper ,925
	t	,119
	df	57
	Sig. (2-tailed)	,906

И здесь полученный уровень двухсторонней значимости далек от критического – различий в уровне наших переменных нет.

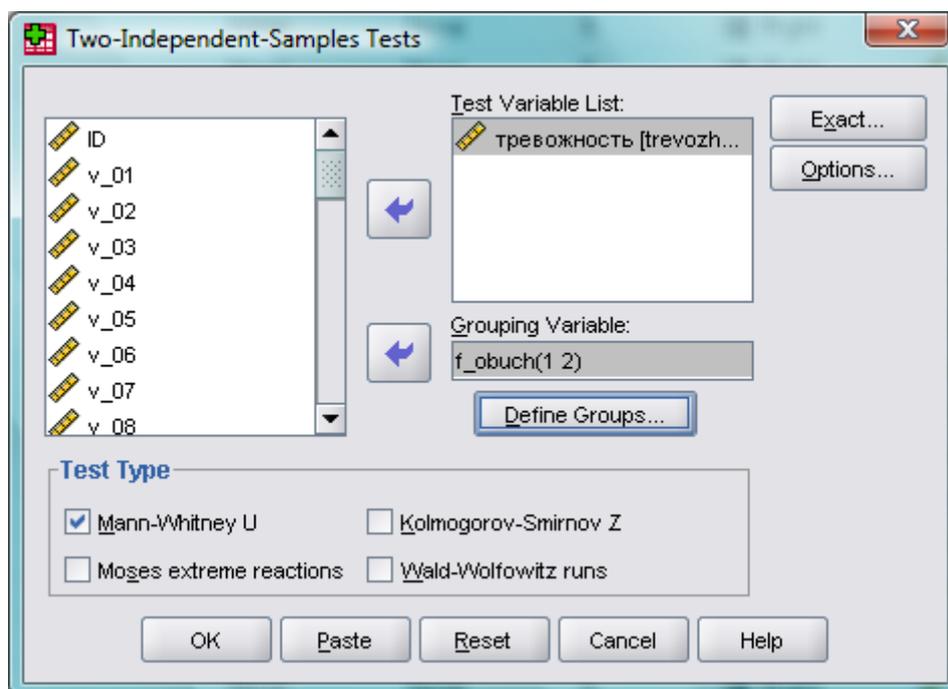
Непараметрические тесты сравнения выборок



В список входят как тесты для сравнения двух зависимых/независимых групп, так и для

выявления различий между более чем двумя группами. В последнем случае (при обнаружении таких различий) надлежит все-таки вернуться к сравнению пар выборок.

Наиболее популярным тестом сравнения двух независимых групп является U-тест Манна-Уитни.



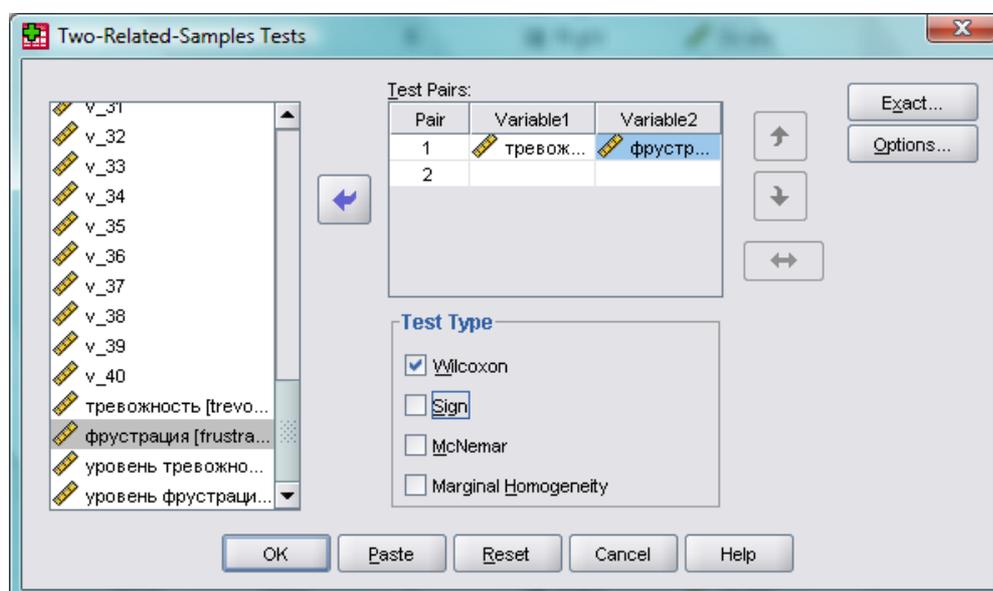
Test Statistics^a

	тревожность
Mann-Whitney U	338,000
Wilcoxon W	614,000
Z	-1,032
Asymp. Sig. (2-tailed)	,302

a. Grouping Variable: форма обучения

Результаты подтверждают вывод об отсутствии различий между двумя выборками.

Для сравнения двух зависимых групп (2-Related Samples) часто используют критерий Вилкоксона:



	фрустрация - тревожность
Z	-,038 ^a
Asymp. Sig. (2-tailed)	,970

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

И он подтверждает отсутствие статистически значимых различий.

Примечание. Отсутствие различий нас не должно огорчать – пример данных имеет чисто иллюстративное назначение.

Задание по теме «Основы работы с файлами и переменными в SPSS»

Тест «Диагностика типологий психологической защиты» состоит из 97 вопросов, на каждый из которых респондент дает ответ да/нет.

Обработка полученных значений состоит из нескольких этапов:

- подсчитать сумму баллов (количество ответов «да») по каждой из 8 шкал (см. ниже);
- «сырые» баллы перевести в проценты по приведенной шкале.

Необходимо:

1. Создать файл SPSS для ввода данных тестирования. Переменным (исходным и вычисляемым) должны быть присвоены имена, метки, метки значений и другие параметры, определяющие свойства переменных.
2. В соответствии с процедурой обработки исходных ответов (сначала в «сырые» баллы, затем – в проценты) составить файл синтаксиса SPSS, содержащий все команды такого преобразования.

Отрицание – 1, 16, 22, 28, 34, 42, 51, 61, 68, 77, 82, 90, 94

Вытеснение (подавление) – 6, 11, 19, 25, 35, 43, 49, 59, 66, 75, 85, 89

Регрессия – 2, 14, 18, 26, 33, 48, 50, 58, 69, 78, 86, 88, 93, 95

Компенсация – 3, 10, 24, 29, 37, 45, 52, 64, 65, 74

Проекция – 7, 9, 23, 27, 38, 41, 55, 63, 71, 73, 84, 92, 96

Замещение – 8, 15, 20, 31, 40, 47, 54, 60, 67, 76, 83, 91, 97

Интеллектуализация – 4, 13, 17, 30, 36, 44, 56, 62, 70, 80, 81, 87

Реактивные образования – 5, 12, 21, 32, 39, 46, 53, 57, 72, 79

Шкала перевода «сырых» оценок в процентилях

«сырые» баллы	Отриц.	Вытесн.	Регресс.	Компенс.	Проекц.	Замещ.	Интел.	Реакт. образ.
0	3	2	2	5	1	6	0	7
1	13	8	6	20	5	23	3	19
2	27	25	19	37	6	37	6	39
3	39	42	35	63	7	48	17	61
4	50	63	53	78	12	65	28	76
5	61	76	70	88	20	77	42	91
6	79	87	80	95	27	86	59	97
7	84	92	85	97	36	93	76	98
8	90	97	88	99	46	97	87	99
9	97	98	95		64	98	92	
10	98	99	97		72	99	97	
11	99		99		90		99	
12					96			
13					99			

Корреляционный анализ

В этом разделе речь пойдёт о связи (корреляции) между двумя переменными. Расчёты подобных двумерных критериев взаимосвязи основываются на формировании парных значений, которые образуются из рассматриваемых зависимых выборок.

Если коэффициент корреляции отрицательный, это означает наличие противоположной связи: чем выше значение одной переменной, тем ниже значение другой. Сила связи характеризуется также и абсолютной величиной коэффициента корреляции. Для словесного описания величины коэффициента корреляции используются следующие градации:

Значение	Интерпретация
до 0,2	Очень слабая корреляция
до 0,5	Слабая корреляция
до 0,7	Средняя корреляция
до 0,9	Высокая корреляция
свыше 0,9	Очень высокая корреляция

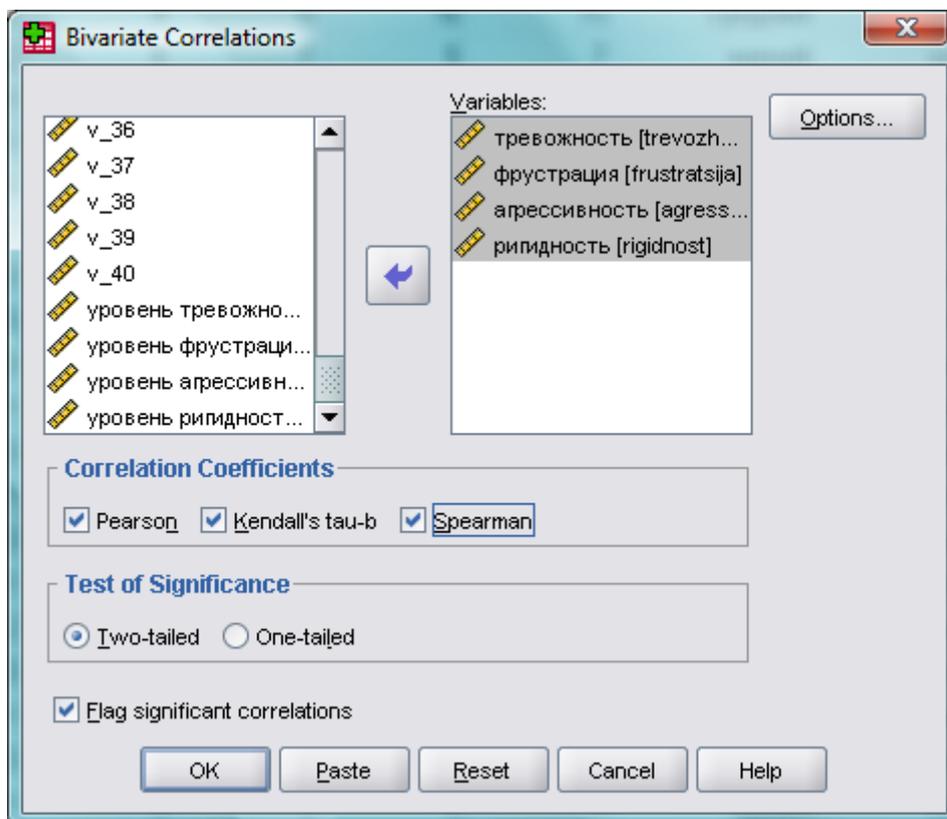
Метод вычисления коэффициента корреляции зависит от вида шкалы, которой относятся переменные.

Переменные с интервальной и с номинальной шкалой	Коэффициент корреляции Пирсона (корреляция моментов произведений).
По меньшей мере, одна из двух переменных имеет порядковую шкалу либо не является нормально распределённой	Ранговая корреляция по Спирмену или τ Кендала.
Одна из двух переменных является дихотомической: точечная двухрядная корреляция	Эта возможность в SPSS отсутствует. Вместо этого может быть применён расчёт ранговой корреляции.
Обе переменные являются дихотомическими	Четырёхполевая корреляция. Данный вид корреляции рассчитываются в SPSS на основании определения мер расстояния и мер сходства (Analyze - > Correlate - > Distances). Этот случай в данном курсе не рассматривается. Рекомендуется обратиться к системе помощи программы SPSS.

Рассмотрим процедуру вычисления корреляций по Пирсону, Спирмену и Кендалу.

Выполним команду Analyze - > Correlate - > Bivariate.

Как обычно, выберем тестируемые переменные и способ вычисления коэффициентов корреляции. Диалоговое окно позволяет выбрать все три метода.



Результаты вычислений представлены в таблицах.

		тревожность	фрустрация	агрессивность	ригидность
тревожность	Pearson Correlation	1,000	,422**	,082	,212
	Sig. (2-tailed)		,001	,539	,110
	N	58	58	58	58
фрустрация	Pearson Correlation	,422**	1,000	,045	,263*
	Sig. (2-tailed)	,001		,735	,042
	N	58	60	60	60
агрессивность	Pearson Correlation	,082	,045	1,000	,233
	Sig. (2-tailed)	,539	,735		,073
	N	58	60	60,000	60
ригидность	Pearson Correlation	,212	,263*	,233	1,000
	Sig. (2-tailed)	,110	,042	,073	
	N	58	60	60	60

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at the 0.05 level (2-tailed).

В таблице приводятся значения коэффициента корреляции Пирсона, его значимость и количество пар наблюдений, по которым проводилось вычисление (отсутствующие значения автоматически исключаются из рассмотрения).

Уровень значимых корреляций отмечен * ($p=0,05$) и ** ($p=0,01$).

Всего две пары переменных имеют значимую корреляцию (корреляционная матрица симметрична относительно диагонали). Они выделены цветом.

Аналогично выглядят и результаты вычислений двух других коэффициентов корреляции.

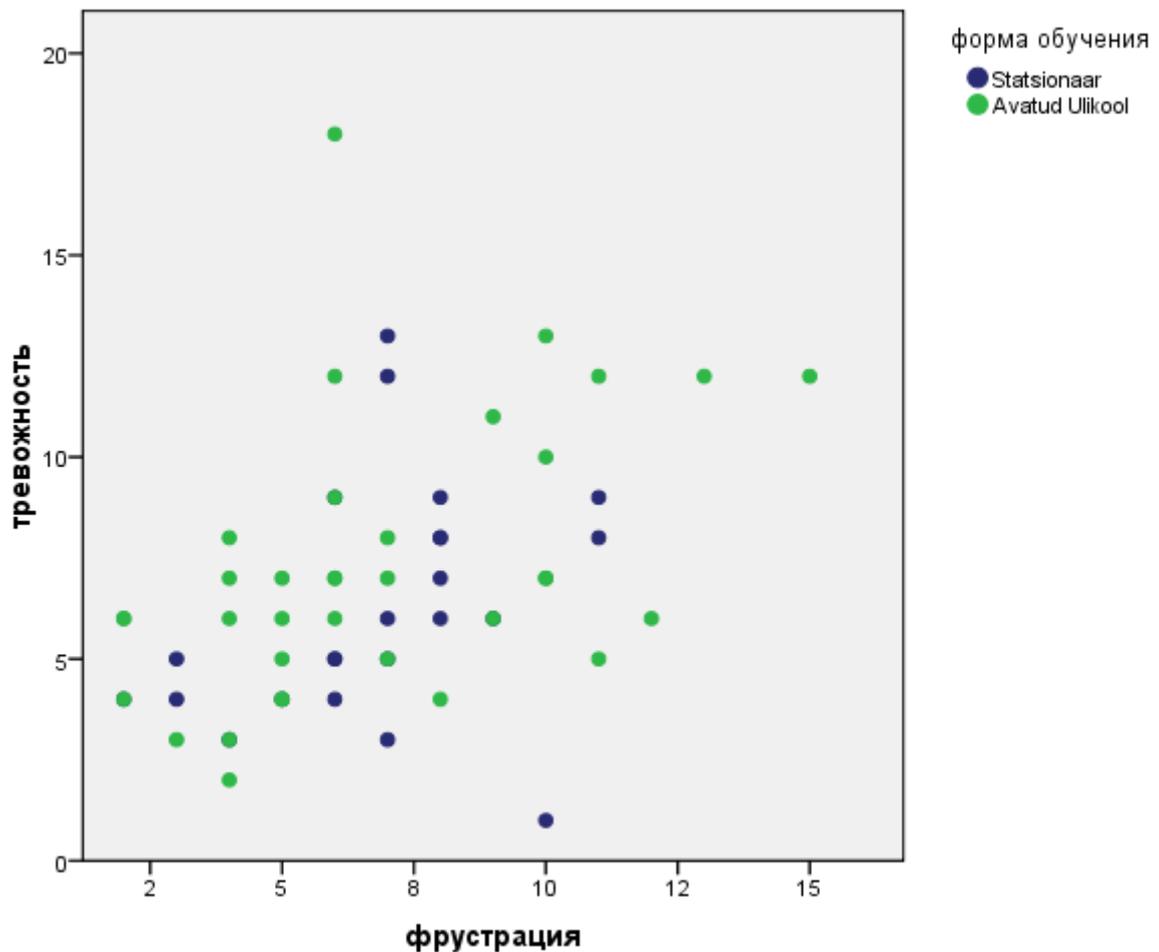
			Correlations			
			тревожность	фрустрация	агрессивность	ригидность
Kendall's tau_b	тревожность	Correlation Coefficient	1,000	,360**	,009	,163
		Sig. (2-tailed)	.	,000	,924	,097
		N	58	58	58	58
	фрустрация	Correlation Coefficient	,360**	1,000	,038	,166
		Sig. (2-tailed)	,000	.	,693	,085
		N	58	60	60	60
	агрессивность	Correlation Coefficient	,009	,038	1,000	,119
		Sig. (2-tailed)	,924	,693	.	,217
		N	58	60	60	60
	ригидность	Correlation Coefficient	,163	,166	,119	1,000
		Sig. (2-tailed)	,097	,085	,217	.
		N	58	60	60	60
Spearman's rho	тревожность	Correlation Coefficient	1,000	,477**	,005	,229
		Sig. (2-tailed)	.	,000	,968	,083
		N	58	58	58	58
	фрустрация	Correlation Coefficient	,477**	1,000	,045	,213
		Sig. (2-tailed)	,000	.	,731	,102
		N	58	60	60	60
	агрессивность	Correlation Coefficient	,005	,045	1,000	,158
		Sig. (2-tailed)	,968	,731	.	,229
		N	58	60	60	60
	ригидность	Correlation Coefficient	,229	,213	,158	1,000
		Sig. (2-tailed)	,083	,102	,229	.
		N	58	60	60	60

** . Correlation is significant at the 0.01 level (2-tailed).

Коррелированность значима только у одной пары переменных – «тревожность-фрустрация».

Для иллюстрации взаимосвязи (не)коррелируемых переменных часто используют диаграмму рассеяния. Мы уже строили ее средствами MS Excel (см. раздел 1), сейчас воспользуемся возможностями SPSS-

Для этого выполним команду Graph - > Chart Builder. В диалоговом окне выберем тип диаграммы и разместим (перетаскиванием с помощью «мышки») по осям переменные «тревожность» и «фрустрация». Имеется возможность выделить цветом данные, принадлежащие разным группам респондентов. Для этого используется номинативная (!) переменная «форма обучения».



Для редактирования элементов графика необходимо дважды на нем щелкнуть. На представленном рисунке (для лучшей «читаемости» графика) была применена заливка точек.

Задание по SPSS (вычисление переменных, анализ таблиц сопряженности, проверка гипотез, корреляционный анализ)

Исходные данные находятся в файле `ajzenk_data.xls`.

Описание методики диагностики самооценки психических состояний личности находится в файле `ajzenk_test.pdf` (Райгородский Д.Я. Практическая психодиагностика. Методики и тесты) и в приложении.

Отчет о проделанной студентом работе должен включать следующие файлы:

- 1) модифицированный файл данных;
- 2) файл синтаксиса, содержащий команды вычислений новых переменных;
- 3) файл вывода (Output) программы SPSS с результатами выполнения заданий;
- 4) в именах файлов необходимо указать свою фамилию и группу;
- 5) файлы выслать электронной почтой по адресу `it@narva.ut.ee`

Средствами SPSS выполнить следующие задания:

- 1) Импортировать в программу SPSS файл MS Excel с данными.

- 2) Подсчитать количество студентов разных форм обучения, построить гистограмму с указанием процентного соотношения численностей.
- 3) В соответствии с методикой обработки теста Айзенка вычислить 4 новые переменные, численно характеризующие уровни тревожности, фрустрации, агрессивности и ригидности респондентов. Имена переменным присвоить самостоятельно, изменить устанавливаемые по умолчанию параметры описания этих переменных на более оптимальные. В метках указать кратко смысл переменных.
Примечание. Создать соответствующий методике обработки файл синтаксиса.
- 4) На основе вычисленных в п.3 переменных провести разделение уровней тревожности, фрустрации, агрессивности и ригидности по трем группам. Границы диапазонов и словесные описания соответствующих групп указаны в описании теста (файл `ajzenk_test.pdf`).
- 5) С помощью таблиц сопряженности и диаграмм, построенных на их основе, проиллюстрировать распределение респондентов с разным уровнем рассматриваемых характеристик психического состояния среди студентов обеих форм обучения. При отображении данных в табличной и графической формах рекомендуется использовать относительные величины (%).
- 6) Проверить гипотезы об отсутствии отличий в распределении респондентов разных форм обучения по уровням исследуемых параметров. Использовать хи-квадрат тест.
- 7) С использованием t-критерия Стьюдента определить (не)значимость различий в уровнях исследуемых психологических характеристик у студентов разных форм обучения. Использовать переменные, вычисленные в п.3.
- 8) Найти коэффициенты корреляции Пирсона для переменных, вычисленных в п.3. Прodelать вычисления для всего массива данных, а также отдельно для обеих форм обучения.
- 9) Для тех пар переменных, у которых будет обнаружена статистически значимые корреляционные связи, построить диаграммы рассеяния (Scatter plots). Если же таковых не окажется, построить диаграмму рассеяния для произвольно выбранной пары переменных, вычисленных в п.3.

Кластерный анализ

Основное назначение кластерного анализа – уменьшить разнообразие объектов исследования (респондентов) путем объединения их в более-менее однородные (с близкими свойствами) группы. «Расстояние» же между образующимися группами должно быть заметно большим, чем «расстояние» между объектами, попавшими в одну группу.

Применяется кластерный анализ и для объединения многочисленных параметров (переменных).

Мы рассмотрим на примере только первое приложение кластерного анализа.

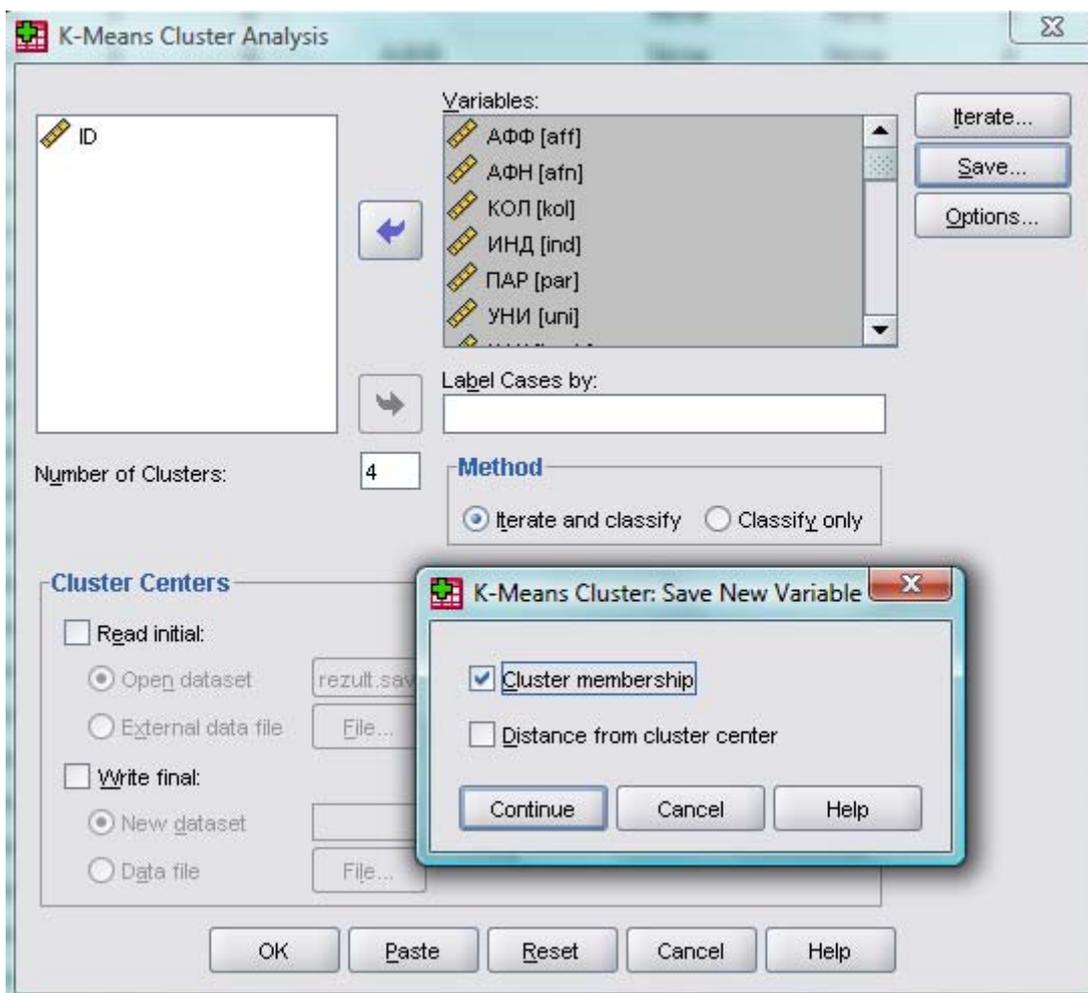
Анализируемые данные (файл `example_cluster.sav`) содержит часть информации, полученной в результате реального психолого-педагогического исследования.

Анализ проведем методом К-средних.

Выполним команду `Analyze -> Classify -> K-Means Cluster`.

В диалоговом окне перенесем все наши переменные в список переменных, на основе которых проводится кластерный анализ (кроме, естественно, переменной ID, которая содержит просто номер респондента).

Укажем предполагаемое количество кластеров (3). Укажем также на необходимость сохранить переменную, указывающую на то, к какому кластеру будет у нас отнесен тот или иной респондент. Остальные настройки оставим без изменений.



По завершении вычислений в файле данных появилась новая переменная (с названием `CL_1` – оно может отличаться номером, если проводить процедуру неоднократно), содержащая информацию о принадлежности наблюдения к кластеру.

В файле отчета, помимо информации, описывающей процедуру вычислений (количество

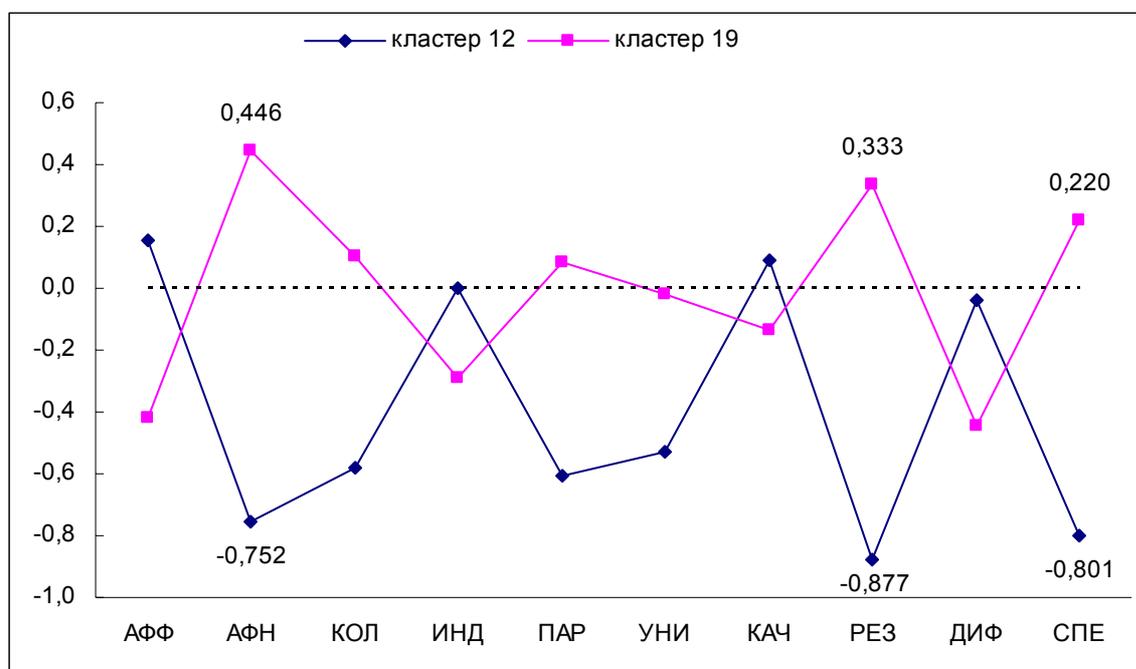
итераций), есть информация о количестве наблюдений в том или ином кластере и вычисленных координатах центров полученных кластеров

Number of Cases in each Cluster

Cluster	1	21,000
	2	13,000
	3	17,000
	4	10,000
	Valid	61,000
	Missing	4,000

На основе кластерной принадлежности объекты разбиваются на группы. Объем каждой группы получился существенным (от 10 до 21 респондента).

Полученные группы можно исследовать на предмет обнаружения значимых различий между ними. Часто рисуют так называемые профили групп, пример которых представлен на рисунке.



Здесь изображены не сами средние значения переменных, вычисленных для каждого кластера в отдельности, а средние значения стандартизованных переменных (см. описание процедуры Descriptives). Такое предварительное преобразование переменных необходимо, если они измерены в шкалах с разными диапазонами изменения переменных.

Задание по теме «Корреляционный и кластерный анализ средствами SPSS»

Все задания выполняются с данными, содержащимися в файле example_cluster.sav.

- 1) Вычислить корреляционную матрицу методом Спирмена.
- 2) Для самых сильных корреляционных связей (наибольшей положительной и наибольшей отрицательной) построить диаграммы рассеяния.

- 3) Провести кластерный анализ методом k -средних. Предположить наличие 3 кластеров. Сохранить переменную, указывающую на кластерную принадлежность данных.
- 4) Преобразовать исходные значения в Z -значения.
- 5) Сравнить попарно между собой все три кластера (Z -значения). Использовать непараметрические методы сравнения средних. Выявить статистически значимые различия ($p=0,05$).
- 6) По параметрам, значимо отличающимся, построить профили кластеров. Использовать графические возможности MS Excel.

Приложения

Описание файла данных

В учебном курсе при демонстрации методов статистической обработки данных используются результаты опроса студентов Нарвского колледжа по методике Г. Айзенка. В опросе участвовали всего 60 студентов стационарной формы обучения и открытого университета. Эти две формы обучения в файле данных (ajzenk_data.xls) закодированы единицей (1) и двойкой (2) соответственно.

Ниже приводится вопросник Айзенка и описывается методика обработки получаемых данных.

Все примеры обработки данных в данном курсе носят исключительно иллюстративный характер.

Методика диагностики самооценки психических состояний (по Г. Айзенку)

Предлагаем Вам описание различных психических состояний. Если это состояние очень Вам подходит, то за ответ ставится 2 балла; если подходит, но не очень, то 1 балл; если совсем не подходит - то 0 баллов.

Подсчитайте сумму баллов за каждую их 4-х групп вопросов:

I.	1-10 вопросы	тревожность
II.	11-20 вопросы	фрустрация
III.	21-30 вопросы	агрессивность
IV.	31-40 вопросы	ригидность

В зависимости от суммы баллов определяется уровень выраженности психического состояния

0-7 баллов	низкий уровень
8-14 баллов	средний уровень
15-20 баллов	высокий уровень

№	Психические состояния	Подходит	Подходит, но не очень	Не подходит
		2	1	0
1	Не чувствую в себе уверенности			
2	Часто из-за пустяков краснею			
3	Мой сон беспокоен			
4	Легко впадаю в уныние			
5	Беспокоюсь о только воображаемых еще неприятностях			
6	Меня пугают трудности			
7	Люблю копаться в своих недостатках			
8	Меня легко убедить			
9	Я мнительный			

10	С трудом переношу время ожидания			
11	Нередко мне кажутся безысходными положения, из которых можно найти выход			
12	Неприятности меня сильно расстраивают, я падаю духом			
13	При больших неприятностях я склонен без достаточных оснований винить себя			
14	Несчастья и неудачи ничему меня не учат			
15	Я часто отказываюсь от борьбы, считая ее бесполезной			
16	Я нередко чувствую себя незащищенным			
17	Иногда у меня бывает состояние отчаяния			
18	Чувствую растерянность перед трудностями			
19	В трудные минуты жизни иногда веду себя по-детски, хочу чтобы меня пожалели			
20	Считаю недостатки своего характера неисправимыми			
21	Оставляю за собой последнее слово			
22	Нередко в разговоре перебиваю собеседника			
23	Меня легко рассердить			
24	Люблю делать замечания другим			
25	Хочу быть авторитетом для окружающих			
26	Не довольствуюсь малым, хочу наибольшего			
27	Когда разгневаюсь, плохо себя сдерживаю			
28	Предпочитаю лучше руководить, чем подчиняться			
29	У меня резкая, грубоватая жестикация			
30	Я мстителен			
31	Мне трудно менять привычки			
32	Нелегко переключаю внимание			
33	Очень настороженно отношусь ко всему новому			
34	Меня трудно переубедить			
35	Нередко у меня не выходят из головы мысли, от которых следовало бы освободиться			
36	Нелегко сближаюсь с людьми			
37	Меня расстраивают даже незначительные нарушения плана			
38	Нередко я проявляю упрямство			
39	Неохотно иду на риск			
40	Резко переживаю отклонения от принятого мною режима			

Критические точки распределения Стьюдента

Число степеней свободы k	Уровень значимости α (двусторонняя критическая область)					
	0,10	0,05		0,01		0,001
1	6,31	12,7		63,7		637,0
2	2,92	4,30		9,92		31,6
3	2,35	3,18		5,84		12,9
4	2,13	2,78		4,00		8,61
5	2,01	2,57		4,03		6,86
6	1,94	2,45		3,71		5,96
7	1,89	2,36		3,50		5,40
8	1,86	2,31		3,36		5,04
9	1,83	2,26		3,25		4,70
10	1,81	2,23		3,17		4,59
11	1,80	2,28		3,11		4,44
12	1,78	2,18		3,05		4,32
13	1,77	2,16		3,01		4,22
14	1,76	2,14		2,98		4,14
15	1,75	2,13		2,95		4,07
16	1,75	2,12		2,92		4,01
17	1,74	2,11		2,90		3,96
18	1,73	2,10		2,88		3,92
19	1,73	2,09		2,86		3,88
20	1,73	2,09		2,85		3,85
21	1,72	2,08		2,83		3,82
22	1,72	2,07		2,82		3,79
23	1,71	2,07		2,81		3,77
24	1,71	2,06		2,80		3,74
25	1,71	2,06		2,79		3,72
26	1,71	2,06		2,78		3,71
27	1,71	2,05		2,77		3,69
28	1,70	2,05		2,76		3,66
29	1,70	2,05		2,76		3,66
30	1,70	2,04		2,75		3,65
40	1,68	2,02		2,70		3,55

60	1,07	2,00		2,66		3,46
120	1,66	1,98		2,62		3,37
Число степеней свободы k	0,05	0,025		0,005		0,0005
	Уровень значимости α (односторонняя критическая область)					

Критические значения линейного коэффициента корреляции Пирсона

n	P		v	P	
	0,95	0,99		0,95	0,99
5	0,75	0,87	27	0,37	0,47
6	0,71	0,83	28	0,36	0,46
7	0,67	0,80	29	0,36	0,46
8	0,63	0,77	30	0,35	0,45
9	0,60	0,74	35	0,33	0,42
10	0,58	0,71	40	0,30	0,39
11	0,55	0,68	45	0,29	0,37
12	0,53	0,66	50	0,27	0,35
13	0,51	0,64	60	0,25	0,33
14	0,50	0,62	70	0,23	0,30
15	0,48	0,61	80	0,22	0,28
16	0,47	0,59	90	0,21	0,27
17	0,46	0,58	100	0,20	0,25
18	0,44	0,56	125	0,17	0,23
19	0,43	0,55	150	0,16	0,21
20	0,42	0,54	200	0,14	0,18
21	0,41	0,53	300	0,11	0,15
22	0,40	0,52	400	0,10	0,13
23	0,40	0,51	500	0,09	0,12
24	0,39	0,50	700	0,07	0,10
25	0,38	0,49	900	0,06	0,09
26	0,37	0,48	1000	0,06	0,09

Критические значения выборочного коэффициента корреляции рангов (по Спирмену)

Связь достоверна, если $r_{s \text{ эмп}} \geq r_{s 0,05}$, и тем более достоверна, если $r_{s \text{ эмп}} \geq r_{s 0,01}$.

n	α		n	α		n	α	
	0,05	0,01		0,05	0,01		0,05	0,01
5	0,94		17	0,48	0,62	29	0,37	0,48
6	0,85	—	18	0,47	0,60	30	0,36	0,47
7	0,78	0,94	19	0,46	0,58	31	0,36	0,46
8	0,72	0,88	20	0,45	0,57	32	0,36	0,45
9	0,68	0,83	21	0,44	0,56	33	0,34	0,45
10	0,64	0,79	22	0,43	0,54	34	0,34	0,44
11	0,61	0,76	23	0,42	0,53	35	0,33	0,43
12	0,58	0,73	24	0,41	0,52	36	0,33	0,43
13	0,56	0,70	25	0,49	0,51	37	0,33	0,43
14	0,54	0,68	26	0,39	0,51	38	0,32	0,41
15	0,52	0,66	27	0,38	0,49	39	0,32	0,41
16	0,50	0,64	28	0,38	0,48	40	0,31	0,40

Критические значения критерия χ^2 для уровней статистической значимости $\alpha < 0,05$ и $\alpha < 0,01$ при разном числе степеней свободы

Различия между двумя распределениями могут считаться достоверными, если $\chi^2_{\text{эмп}}$ достигает или превышает $\chi^2_{0,05}$, и тем более достоверными, если $\chi^2_{\text{эмп}}$ достигает или превышает $\chi^2_{0,01}$,

k	α=0,05	α=0,01	k	α=0,05	α=0,01	k	α=0,05	α=0,01
1	3,841	6,635	35	49,802	57,342	69	89,391	99,227
2	5,991	9,210	36	50,998	58,619	70	90,631	100,425
3	7,815	11,345	37	52,192	59,892	71	91,670	101,621
4	9,448	13,227	38	53,384	61,162	72	92,808	102,816
5	11,070	15,086	39	54,572	62,428	73	93,945	104,010
6	12,592	16,812	40	55,758	63,691	74	95,081	105,202
7	14,067	18,475	41	56,942	64,950	75	96,217	106,393
8	15,507	20,090	42	58,124	66,206	76	97,351	107,582

9	16,919	21,666	43	59,304	67,459	77	98,484	108,771
10	18,307	23,209	44	60,481	68,709	78	99,617	109,958
11	19,675	24,725	45	61,656	69,957	79	100,749	111,144
12	21,026	26,217	46	62,830	71,201	80	101,879	112,329
13	22,362	27,688	47	64,001	72,443	81	103,010	113,512
14	23,685	29,141	48	65,171	73,683	82	104,139	114,695
15	24,996	30,578	49	66,339	74,919	83	105,267	115,876
16	26,296	32,000	50	67,505	76,154	84	106,395	117,057
17	27,587	33,409	51	68,669	77,386	85	107,522	118,236
18	28,869	34,805	52	69,832	78,616	86	108,648	119,414
19	30,144	36,191	53	70,993	79,843	87	109,773	120,591
20	31,410	37,566	54	72,153	81,069	88	110,898	121,767
21	32,671	38,932	55	73,311	82,292	89	112,022	122,942
22	33,924	40,289	56	74,468	83,513	90	113,145	124,116
23	35,172	41,638	57	75,624	84,733	91	114,268	125,289
24	36,415	42,980	58	76,778	85,950	92	115,390	126,462
25	37,652	44,314	59	77,931	87,166	93	116,511	127,633
26	38,885	45,642	60	79,082	88,379	94	117,632	128,803
27	40,113	46,963	61	80,232	89,591	95	118,752	129,973
28	41,337	48,278	62	81,381	90,802	96	119,871	131,141
29	42,557	49,588	63	82,529	92,010	97	120,990	132,309
30	43,773	50,892	64	83,675	93,217	98	122,108	133,476
31	44,985	52,191	65	84,821	94,422	99	123,225	134,642
32	46,194	53,486	66	85,965	95,626	100	124,342	135,807
33	47,400	54,776	67	87,108	96,828			
34	48,602	56,061	68	88,250	98,028			