**Задание 5. Сравнение средних**

**Создание и редактирование файлов данных в PSPP**

ex01.sav

Рассмотрим Следующий пример. Он содержит информацию о 100 школьниках трех выпускных классов. О каждом из школьников имеются следующие данные:

* пол;
* класс;
* вуз (выбранный профиль вуза);
* хобби (систематическое внешкольное занятие);
* информация о пяти тестах;
* средняя годовая отметка за 10-й класс;
* средняя годовая отметка за 11-й класс.

На этапе планирования исследования структура данных может быть задана в виде предварительного списка переменных с указанием их типов и диапазонов возможных значений, например, так, как в табл.



На предварительном этапе следует выделять два типа переменных: количественные и категориальные (номинальные).

**Количественная** переменная позволяет различать объекты по уровню выраженности некоторого свойства, например: средний балл отметки, тестовый показатель и пр.

Второй тип — **категориальные** (номинальные) переменные. Обычно они используются как основания для деления объектов (испытуемых) на группы или категории: пол, класс, возрастная категория, уровень дохода и пр.

Важным свойством номинальных переменных является возможность их представления в виде набора целых чисел. Например, трем видам внешкольных увлечений (хобби) учащихся (спорт, компьютер, искусство) можно сопоставить числа 1, 2, и 3 соответственно. Числовое представление данных в компьютерных программах всегда предпочтительнее символьного, поскольку обработка чисел происходит быстрее, проще и с меньшей вероятностью ошибок. Кроме того, можно без проблем включить в анализ новые элементы. Например, если в группе окажется учащийся, увлечение которого не соответствует перечисленным, будет полезно включить в переменную хобби категорию с названием «другие» и присвоить ей число 4.

Порядок создания переменных также важен при вводе данных. Здесь следует придерживаться простого правила: наиболее важные и часто используемые переменные должны помещаться в начало файла, для остальных данных вопрос порядка следования не столь важен.

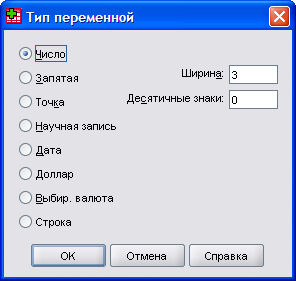
Чаще всего в начало файла следует поместить категориальные переменные, которые далее предполагается использовать для деления объектов (испытуемых) на группы, например пол, семейное положение и пр.

Вкладка **Переменные** позволяет задать структуру файла данных, то есть определить имена, метки и структуры переменных. Заголовки столбцов представляют собой параметры каждой из переменных.

**Имя переменной** должно удовлетворять следующим требованиям:

* в имени могут использоваться любые буквы, цифры, символы @, #, ., \_, $, но имя всегда должно начинаться с буквы, а символ «.» (точка) не может стоять в конце имени;
* буквы верхнего и нижнего регистров символов не различаются, то есть имена ID, id, Id и iD воспринимаются программой как идентичные;
* имена переменных не должны совпадать с каким-либо из зарезервированных слов (all, ne, eq, to, le, lt, by, or, gt, and, not, ge, with).

**Параметр Тип** – тип переменной определяется с помощью диалогового окна



В подавляющем большинстве случаев приходится иметь дело с числовыми данными. В тех редких случаях, когда значения переменных представляют собой буквы или буквосочетания (слова), необходимо установить переключатель Строка.

Как правило, строковые переменные не подлежат обработке. Поэтому их следует избегать, за исключением редких случаев, например, когда данная переменная содержит имена людей или названия городов.

Остальные 6 типов переменных, которые можно задать с помощью переключателей, присутствующих в диалоговом окне, практически не используются при обработке данных.

**Параметр Ширина** позволяет задать максимальное количество знаков, которое может иметь значение переменной, включая дробную часть.

В большинстве случаев нет необходимости менять заданную по умолчанию ширину переменной.

Зададим для всех переменных ширину 3 знака. Значения можно вводить копированием.

**Параметр Десятичные** предназначен для задания числа десятичных знаков после запятой в случае, если тип переменной допускает использование дробных чисел.

Для числовых переменных значение в ячейке Десятичные автоматически устанавливается равным 2.

В нашем примере знаки после запятой необходимы для переменных отметка1 и отметка2, причем количество этих знаков равно 2 и задано программой по умолчанию. Необходимо установить для всех остальных числовых переменных значение параметра Десятичные равным 0. Это не является необходимым, а служит для удобства отображения данных на экране или в отчете.

**Метки переменных.** Как правило, метка используется в тех случаях, когда смысл переменной недостаточно точно отражен в названии, это комментарий к имени переменной.

**Метки значений переменных**

Параметр Значения позволяет управлять наименованиями уровней (категорий)

переменной. Под уровнем, или категорией, понимается целочисленное значение переменной, имеющее определенный смысл. Например, переменная пол имеет два уровня: 1 — жен (женский) и 2 — муж (мужской). Буквосочетания или слова, поставленные в соответствие уровням переменной, например жен и муж, называются метками значений и отражают смысл разных значений переменной для исследователя.

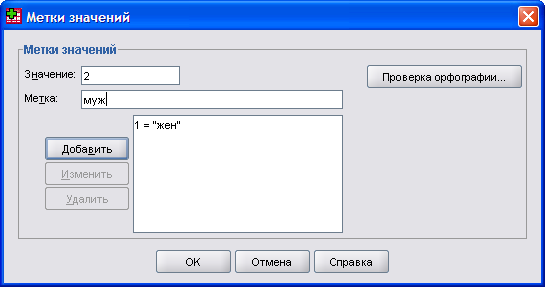
Наличие меток значений исключительно важно для удобочитаемости результатов анализа, поскольку очень трудно запомнить смысл значений всех переменных, даже если их сравнительно немного. В крупных демографических исследованиях количество переменных может превышать 100, поэтому метки значений являются практически незаменимым средством расшифровки результатов статистического анализа.

SPSS также позволяет включать метки значений в файл данных, что очень удобно при его создании. Длина метки значения ограничивается 60 символами, однако, как правило, для пользователей достаточно 3–10 символов.

Как и в случае столбца Тип, перемещение в ячейку столбца Значения приводит к появлению кнопки с многоточием. Если щелкнуть на ней, на экране появится диалоговое окно Метки значений, в котором можно задать метки для различных значений переменной.

Например для переменной пол в поле Значение введите число 1, нажмите клавишу Tab для перехода в поле Метка, введите туда слово «жен» и щелкните на кнопке Добавить.

В поле Значение введите число 2, нажмите клавишу Tab для перехода в поле Метка, введите туда слово муж и щелкните на кнопке Добавить.



После задания меток значений переменных появляется еще одна удобная возможность изменения вида окна вкладки Данные. Если в меню **Вид** установить флажок рядом с командой **Метки значений**, численные значения переменных будут заменены заданными метками. Для того чтобы вернуться к численным значениям переменных, достаточно сбросить флажок рядом с командой Метки значений.

**Параметр Пропуски** используется очень редко, поскольку программа и так позволяет учитывать пропуски в данных. Необходимость в этом параметре возникает, когда требуется различать причины пропусков значений. Например, пропуск в данных может быть обусловлен тем, что респондент еще не опрошен или, может быть, он отказался отвечать на данный вопрос.

Так, в отношении переменной вуз (предполагаемый для поступления вуз) нам необходимо различать тех учащихся, которых мы не успели опросить, и тех, которые еще не определились. Для еще не опрошенных учащихся мы будем оставлять пустую ячейку, а не определившихся учащихся будем обозначать цифрой 9.

Пропущенные значения в диалоговом окне можно задать обозначения пропусков. Если поставить переключатель *Отдельные пропущенные значения* и ввести число, например, 9, это значение не будет использоваться в дальнейшем при обработке переменной, наряду с пустыми ячейками.

**Параметр Ширина столбца** - Столбцы требуется для всех переменных. С его помощью можно управлять шириной (в символах) столбцов вкладки Данные.

Основной принцип, которым следует руководствоваться: чем меньше знаков занимает большинство значений некой переменной, тем меньшей следует задавать ширину соответствующего столбца.

По умолчанию для всех переменных значение равно 8.

В нашем примере необходимо установить следующие значения параметра Столбцы:



**Параметр** **Выравнивание** позволяет управлять расположением данных внутри ячейки.

**Шкала измерения**. Значение параметра Шкала выбирается в раскрывающемся списке из трех доступных: Количественная, Порядковая и Номинальная.

Каждое из этих значений несет дополнительную информацию о типе данных переменной.

* Значение **Количественная** указывается для количественных типов данных, допускающих арифметические операции. Значение параметра Шкала можно указать для переменных тест1, ..., тест5, отметка1, отметка2.
* Значение **Порядковая** указывается для количественных данных, отражающих измеренное качество на уровне порядка и не допускающих основных арифметических операций. Например, упорядочивание учащихся по сообразительности или кодирование ответов по степени выраженности (1 — «не нравится», 2 — «нейтрально», 3 — «нравится») — это измерения в порядковой шкале. К подобным типам можно отнести и многие дихотомические переменные (переменные, имеющие лишь два различных значения: 0 или 1, да или нет и т. п.).
* Значение **Номинальная** указывается для категориальных (неколичественных) типов данных, не отражающих количество измеряемого свойства. К таким данным относятся переменные №, пол, класс, вуз, хобби.

Иногда может оказаться затруднительным сделать выбор между значениями Количественная и Порядковая. Во всех статистических процедурах оба этих значения параметра Шкала сказываются на результатах анализа одинаковым образом.

Чаще всего нет необходимости менять значение параметра Количественная, заданное по умолчанию. Тем более что выбор значения Номинальная ограничивает набор возможных методов обработки данных.

**Ввод данных**

Для ввода данных следует перейти на вкладку Данные

Существует два метода непосредственного ввода данных в ячейки электронной таблицы: по строкам (переменным) и по столбцам (объектам). В зависимости от конкретных данных тот или другой способ оказывается предпочтительным.

**Ввод данных по переменным** предполагает последовательное заполнение всех строк одного столбца значениями одной переменной, затем другой переменной и т. д. Вы вводите значение в левую верхнюю ячейку окна, затем нажимаете клавишу ↓ или Enter. Таким образом, с визуальной точки зрения ввод данных по переменным осуществляется «по вертикали».

**Ввод данных по объектам** заключается в последовательном заполнении каждой строки значениями всех переменных. Сначала вводится значение в левую верхнюю ячейку окна, затем с помощью клавиши → или Tab осуществляется переход в соседний столбец (переменную), вводится значение переменной и т. д. до тех пор, пока вся строка не будет заполнена. В отличие от предыдущего способа ввод данных по объектам осуществляется «по горизонтали».

Некоторым пользователям привычнее вводить, хранить и редактировать данные в электронной таблице программы Excel. Перенос можно осуществить двумя способами.

Первый способ самый простой. Он предполагает предварительную подготовку специального файла Excel. Столбцы этого файла должны соответствовать переменным, строки — объектам. В первой строке должны содержаться имена переменных, удовлетворяющие требованиям SPSS. Достаточно открыть такой файл Excel из редактора данных SPSS, чтобы переместить содержащиеся в нем данные в SPSS. Для этого при помощи последовательности команд Файл>Открыть>Данные необходимо в диалоговом окне Открыть данные в списке Тип файла выбрать пункт Excel (\*.xls).

Импортируйте данные из файла ex01.xls. Сохраните готовый файл данных под именем ex01.sav. Файлы данных в SPSS имеют расширение sav.

**Окно вывода**

Это окно отражает главный результат обработки данных. Данные этого окна перед сохранением или печатью можно отредактировать.

**Задача 1**

А.Наследов. SPSS 19 Профессиональный статистический анализ данных. Стр 153

Повторите проверку средних значений отметки 2 для независимых выборок по полу.

**Задача 2.** Проведите t-тест для отметки 1 для разных классов

**Задача 3**

В лекции 4 рассмотрена процедура проверки гипотезы о том, что у различных видов температура плавления ДНК отличается.



С помощью критерия Стъюдента в SPSS сделать вывод о различии двух средних этих выборок. Сравните результаты с примером из лекции. Оформите отчет в Word документе.

**Задача 4**

Решите задачу 3 с помощью питона на основе следующего примера.

**Пример: двухвыборочный t-тест в Python**

Исследователи хотят знать, имеют ли два разных вида растений одинаковую среднюю высоту. Чтобы проверить это, они собирают простую случайную выборку из 20 растений каждого вида.

Используйте следующие шаги, чтобы провести t-тест для двух образцов, чтобы определить, имеют ли два вида растений одинаковую высоту.

**Шаг 1: Создайте данные.**

Во-первых, мы создадим два массива для хранения измерений каждой группы из 20 растений:

import numpy as np

group1 = np.array([14, 15, 15, 16, 13, 8, 14, 17, 16, 14, 19, 20, 21, 15, 15, 16, 16, 13, 14, 12])

group2 = np.array([15, 17, 14, 17, 14, 8, 12, 19, 19, 14, 17, 22, 24, 16, 13, 16, 13, 18, 15, 13])

**Шаг 2: Проведите t-критерий с двумя выборками.**

Далее мы воспользуемся [функцией ttest\_ind()](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html?ref=codecamp.ru) из библиотеки scipy.stats для проведения двухвыборочного t-теста, в котором используется следующий синтаксис:

**ttest\_ind(a, b, equal\_var=true)**

где:

**a**: массив выборочных наблюдений для группы 1

**b**: массив выборочных наблюдений для группы 2

**equal\_var**: если True, выполните стандартный независимый t-критерий с двумя выборками, который предполагает равные дисперсии генеральной совокупности.

Прежде чем мы проведем тест, нам нужно решить, будем ли мы предполагать, что две совокупности имеют одинаковую дисперсию или нет. Как правило, мы можем предположить, что совокупности имеют равные дисперсии, если отношение большей выборочной дисперсии к меньшей выборочной дисперсии составляет менее 4:1.

#find variance for each group

print(np.var(group1), np.var(group2))

7.73 12.26

Отношение большей дисперсии выборки к меньшей дисперсии выборки составляет 12,26 / 7,73 = 1,586 , что меньше 4. Это означает, что мы можем предположить, что дисперсии генеральной совокупности равны.

Таким образом, мы можем приступить к выполнению двухвыборочного t-критерия с равными дисперсиями:

import scipy.stats as stats

#perform two sample t-test with equal variances

stats.ttest\_ind(a=group1, b=group2, equal\_var=True)

(statistic=-0.6337, pvalue=0.53005)

Статистика t-теста равна -0,6337 , а соответствующее двустороннее значение p равно 0,53005 .

**Шаг 3: Интерпретируйте результаты.**

Две гипотезы для этого конкретного двухвыборочного t-критерия следующие:

H 0 : µ 1 = µ 2 (две средние совокупности равны)

H A : µ 1 ≠ µ 2 (две средние совокупности не равны)

Поскольку p-значение нашего теста (0,53005) больше, чем альфа = 0,05, мы не можем отвергнуть нулевую гипотезу теста. У нас нет достаточных данных, чтобы сказать, что средняя высота растений между двумя популяциями различна.