МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

МАТЕМАТИЧЕСКАЯ ОБРАБОТКА И АНАЛИЗ РЕЗУЛЬТАТОВ ПСИХОЛОГИЧЕСКОГО ТЕСТИРОВАНИЯ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 248 группы
направления 09.04.03 — Прикладная информатика

механико-математического факультета
Табакова Евгения Олеговича

Научный руководитель
доцент, к. ф.-м. н., доцент _____ Е. В. Гудошникова

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2020

Введение. Степень использования математических методов в практической и исследовательской деятельности является показателем зрелости различных наук. Психология не является исключением, психологами используются методы различных математических дисциплин, в том числе математической статистики, теории множеств, теории вероятностей, математической логики.

В связи с развитием теорий в гуманитарных дисциплинах, выделением новых отраслей и направлений возрастает роль математических методов для описания и анализа изучаемых явлений, наблюдается стремление выразить открываемые законы в математической форме.

Проникновение математических методов в психологию в первую очередь связано с развитием экспериментальных и прикладных исследований. С одной стороны, применение данных методов вносит новые возможности в изучение свойств и явлений. В то же время это задает более высокие требования к постановке исследовательских задач и их решению.

Накопление экспериментального материала с использование тестов, вопросников, других измерительных инструментов предполагает необходимость обобщения полученных результатов в соответствии с поставленными задачами. Получаемые результаты представленные в виде чисел, необходимо обрабатывать, что предполагает решение вопросов об организации и сборе информации, чтобы они были доступны обработке в соответствии с поставленными задачами, выбором метода обработки. В связи с этим применение математики становится необходимым этапом решения поставленных задач.

Использование математических методов позволяет исследователю измерять, описывать и сравнивать качественные и количественные характеристики психологических феноменов, моделировать отдельные психические процессы, свойства и состояния явлений, делать выводы и прогнозы.

Целью магистерской работы является создание базы данных в Microsoft Access, позволяющей обрабатывать поступающие заявки на прохождение тестирования, создавая единый список тестируемых, разработка SQL запросов для обработки данных в таблицах базы данных, создание процедур на языке VBA, позволяющих обрабатывать условия добавления и обновления результатов в таблицах базы данных, проведение корреляционного, факторного и

дисперсионного анализа, с использованием методов статистики языка R.

Исходя из поставленной в магистерской работе цели необходимо решить следующие задачи:

1. Раскрыть основные понятия корреляционного, факторного и дисперсионного анализа;
2. Рассмотреть методы корреляционного, факторного и дисперсионного анализа;
3. Ознакомиться с языком программирования R;
4. Провести корреляционный, факторный и дисперсионный анализ с использованием методов статистики языка R;
5. Ознакомиться с языками программирования VBA и SQL;
6. Создать базу данных и запросы к ней с использование языка SQL;
7. Создать процедуры на языке VBA, позволяющие обрабатывать условия добавления и обновления результатов в таблицах базы данных.

Магистерская работа состоит из введения, пяти разделов, заключения, списка используемых источников и приложения. В первом разделе описываются методы и алгоритмы создания списка тестируемых. Второй раздел содержит описание методов и алгоритмов процедуры внесения результатов тестирования. В третьем разделе описывается корреляционный анализ и его применение. В четвертом разделе рассмотрены основные положения факторного анализа и его применение. Пятый раздел посвящен дисперсионному анализу и его применению.

Объектом исследования являются результаты тестирования.

Предметом исследования является математическая обработка результатов тестирования с помощью средств языка R, процедур созданных на языке VBA и SQL запросов к базе данных.

Работа прошла апробацию на различных конференциях, в частности, на ежегодной студенческой конференции "Актуальные проблемы математики и механики", которую проводил механико-математический факультет СГУ в апреле 2019 г., в секции "Анализ данных", в VIII Международной молодежной научно-практической конференции «Математическое и компьютерное моделирование в экономике, страховании и управлении рисками», ноябрь 2019 г.

Основное содержание работы. Задание по формированию списка:

перед началом формирования списка указываются возможные даты тестирования и максимально возможное количество тестируемых на каждую дату. При заполнении списка должны предлагаться свободные даты тестирования (с учетом записавшихся ранее).

Для выполнения этой задачи была использована СУБД Microsoft Access. Microsoft Access — реляционная система управления базами данных корпорации Microsoft. СУБД работает в 2-х режимах: проектировочном и пользовательском. Первый применяется при создании или изменении структуры базы данных и создании ее объектов. Второй режим используется при непосредственной работе с ранее подготовленными объектами для наполнения БД или получения данных из нее.

Основными особенностями Microsoft Access являются:

— создание форм, т.е. формирование программных интерфейсов для занесения, изменения или просмотра данных;
— возможность использования SQL-запросов к БД;
— построение отчетов с последующим выводом на печать;
— возможность устанавливать связь с внешними таблицами и БД;
— использование встроенного языка программирования Visual Basic for Applications (VBA) для построения бизнес-логики приложений.

SQL (structured query language — «язык структурированных запросов») — декларативный язык программирования, применяемый для создания, модификации и управления данными в реляционной базе данных, управляемой соответствующей системой управления базами данных.

Visual Basic for Applications (VBA) – это объектно-ориентированный язык программирования, который является общим инструментом для всех приложений Microsoft Office и позволяет решать задачи автоматизации.

Код VBA используется для проверки вводимых пользователем значений, для работы с элементами управления на форме, для переключения между формами, отчетами, другими элементами управления, для обращения к внешним объектным моделям и т.п. Основными компонентами программы на VBA являются процедуры и функции. С помощью VBA можно разрабатывать программы, которые включают компоненты нескольких приложений Microsoft Office и способствуют тем самым интеграции и совместному исполь-

зованию данных.

В Microsoft Access была создана база данных, которая содержит таблицы: Общая_информация, Дата_Количество, Опросник, Субтест1, Субтест2, Субтест3, Субтест4, Субтест5, Субтест6, Субтест7, Субтест8, Субтест9.

В таблицу "Дата_Количество" через созданную форму вносятся даты тестирования и количество тестируемых.

Чтобы добавить новую запись в таблицу "Дата_Количество" нужно заполнить поля формы и нажать на форме для заполнения таблицы "Дата_Количество" кнопку "Добавить новую запись".

В таблицу "Общая_информация" через форму вносятся новые записи о тестируемых.

Для того чтобы добавить новую запись в таблицу "Общая_информация" нужно заполнить поля формы и нажать на форме для заполнения таблицы "Общая_информация" кнопку "Добавить новую запись в список".

Причём при заполнении на форме поля "Дата прохождения" будет выполняться созданная на языке VBA процедура проверки введённой даты прохождения. Если введённая дата прохождения отсутствует в таблице "Дата_Количество", то будет выведено уведомление, что дата отсутствует в списке и будет предложено выбрать другую дату из списка.

Если же на введённую дату прохождения тестирования число записавшихся максимально, то будет выведено уведомление "На данную дату не осталось мест, выберите другую дату из следующих:". При этом поле "Дата прохождения" будет оставаться пустым до тех пор, пока не будет введена дата прохождения, удовлетворяющая условиям проверки.

Результаты тестирования, которые необходимо импортировать в БД передаются в виде таблиц Excel:

1. 9 листов, в каждом из которых содержится №, ФИО, Дата прохождения и результаты заданий в формате «0» или «1».
2. 1 лист, в котором содержится №, ФИО, Дата прохождения и результаты тестирования в формате «да», «нет», «может быть», «не знаю». Результаты должны быть оцифрованы в соответствии с шаблоном.

Для того чтобы добавить новые и обновить имеющиеся результаты тестирования в базе данных была создана форма "Обновление и добавление

записей в таблицах". На этой форме расположены кнопки: "Заполнить и обновить таблицы субтестов", "Заполнить и обновить таблицу опросник", "Заполнить и обновить таблицу общая информация", "Обработать таблицу опросник", "Получить список с результатами".

Для добавления и обновления данных в таблицах субтестов, "Опросник", "Общая информация" были разработаны процедуры на языке VBA с использованием языка структурированных запросов SQL, выполняющиеся после нажатия соответствующих кнопок на форме "Обновление и добавление записей в таблицах".

При этом в этих процедурах добавления и обновления результатов используются вспомогательные функции filep() и ifTableExists(). Функция filep() используется для вызова формы диалогового окна выбора файла и считывания пути до файла в переменную. Функция ifTableExists() используется для проверки существования временной таблицы в базе данных, благодаря чему в последствии эта временная таблица с результатами очищается в процедурах.

Обработка результатов таблицы "Опросник" производится с использованием созданной на языке VBA процедуры, которая содержит SQL запрос, обрабатывающий эту таблицу с использованием шаблона. Для того чтобы обработать результаты в таблице "Опросник" нужно нажать на форме кнопку "Обработать таблицу опросник", которая представлена на форме "Обновление и добавление записей в таблицах".

При оценке данных опросника используются четыре фактора. Они включают: Любознательность (Л) , Воображение (В), Сложность (С) и Склонность к риску (Р). Отверстия в шаблоне показывают ответы, соответствующие оценке два (2) балла, также на шаблоне отмечены коды для четырех факторов, оцениваемых в тесте. Все ответы, находящиеся на клетках, не попадающих в отверстия, получают один (1) балл, кроме последней колонки «Не знаю». Ответы в этой колонке получают минус один (— 1) балл в сырых баллах и вычитаются из общей оценки.

Для того, чтобы получить список с результатами тестирования была разработана процедура на языке VBA, которая содержит SQL запрос на объединение таблиц "Общая информация", "Опросник" и таблиц, содержащих результаты субтестов (причём путём преобразований в список вносятся сумма

баллов по каждому субтесту, а также сумма всех субтестов). Для выполнения этой процедуры, нужно нажать на форме "Обновление и добавление записей в таблицах" кнопку "Получить список с результатами".

Корреляционный анализ был проведен с использованием коэффициентов корреляции Пирсона и Спирмена. Также были рассчитаны коэффициенты взаимной сопряженности Пирсона и Чупрова.

Расчеты в рамках корреляционного анализа были произведены с использованием статистического пакета R. R — язык программирования для статистической обработки данных и работы с графикой, а также свободная программная среда вычислений с открытым исходным кодом. Главные особенности языка R:

- эффективная обработка данных и простые средства для сохранения результатов;
- набор операторов для обработки массивов, матриц, и других сложных конструкций;
- большая, последовательная, интегрированная коллекция инструментальных средств для проведения статистического анализа;
- многочисленные графические средства.

Для рассчёта коэффициентов корреляции Пирсона и Спирмена были использованы данные, содержащие результаты тестирования по девяти субтестам.

В общем виде формула для подсчета коэффициента корреляции Пирсона такова:

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

где x — значения принимаемые переменной x;

y — значения принимаемые переменной y;

$\bar{x}$ — средняя по x;

$\bar{y}$ — средняя по y;

n — количество наблюдений.

Коэффициент ранговой корреляции Спирмена находится по формуле:

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{2}$$

где d_i — разность рангов для каждой i–пары из n наблюдений;
n — количество наблюдений.

Для коэффициентов корреляции Пирсона и Спирмена были построены корреляционные матрицы, а также матрицы содержащие значения p-value. На основе матриц, содержащих значения p-value для рассчитанных коэффициентов корреляции можно определить значимость коэффициентов корреляции путем сравнения значений p-value с уровнем значимости (0,05; 0,01 и т. п.), то есть если значение p-value меньше принятого уровня значимости, то коэффициент корреляции считается значимым на данном уровне значимости.

Коэффициенты взаимной сопряженности Пирсона и Чупрова используются в том случае, если по каждому из взаимосвязанных признаков выделяется число групп более двух. Факт наличия связи устанавливается с помощью критерия χ^2.

Коэффициент взаимной сопряженности Пирсона находится по формуле:

$$P = \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{\phi^2}{1+\phi^2}}, \tag{3}$$

где $\phi^2 = \frac{\chi^2}{n}$.

Коэффициент взаимной сопряженности Чупрова, вычисляется по следующей формуле:

$$C = \sqrt{\frac{\chi^2}{n \cdot \sqrt{(k_1-1)\cdot(k_2-1)}}} = \sqrt{\frac{\phi^2}{\sqrt{(k_1-1)\cdot(k_2-1)}}}. \tag{4}$$

Оба приведенных коэффициента взаимной сопряженности основаны на нормировании χ^2: погашении зависимости от числа наблюдений и размерности таблицы. Данные коэффициенты принимают все свои значения на отрез-

ке [0, 1].

Для рассчёта коэффициентов взаимной сопряженности Пирсона и Чупрова была использована матрица сопряжённости двух признаков (результатов по Субтесту7 и полу). По результатам Субтеста7 оценивается умение решать геометрические задачи, богатство пространственных представлений, конструктивные практические способности, наглядно-действенное мышление.

Признак результаты по Субтесту7 содержит значения "High_result" - высокий результат (15-20 Б), "Medium_result" - средний результат (9-14 Б), "Low_result" - низкий результат (0-8 Б), признак пол содержит значения "Male" - мужской, "Female" - женский.

С помощью коэффициентов взаимной сопряженности Пирсона и Чупрова предполагалось определить, зависят ли результаты по Субтесту7 от пола испытуемого.

В результате расчетов были получены значения коэффициентов взаимной сопряженности Пирсона и Чупрова, которые свидетельствуют о том, что зависимость между результатами по Субтесту7 и полом испытуемого очень слабая.

Расчеты факторного анализа были проведены с использованием статистического пакета R и библиотеки psych.

Факторный анализ есть углубление и развитие идеи корреляционного анализа. Факторный анализ - статистический метод, который используется при обработке больших массивов экспериментальных данных. Задачами факторного анализа являются сокращение числа переменных (редукция данных) и определение структуры взаимодействий между переменными, т.е. классификация переменных, поэтому факторный анализ используется как метод сокращения или как метод структурной классификации.

Общей моделью факторного анализа служит следующая линейная зависимость:

$$x_i = \sum_{j=1}^{m} a_{ij} F_j + b_i U_i + \varepsilon_i, \tag{5}$$

где a_{ij}— факторная нагрузка j-го общего фактора на i-ую переменную,

$i = \overline{1,k}$, $j = \overline{1,m}$;
F_j— общие факторы, $j = \overline{1,m}$;
x_i— наблюдаемые переменные, $i = \overline{1,k}$;
b_i— факторная нагрузка i-го характерного фактора на i-ую переменную, $i = \overline{1,k}$;
U_i— характерные факторы, $i = \overline{1,k}$;
ε_i — случайные ошибки, $i = \overline{1,k}$.

Для проведения факторного анализа используется методика, включающая в себя следующие этапы:

1. сбор исходных статистических данных и подготовка корреляционной (ковариационной) матрицы;
2. выделение общих скрытых факторов;
3. вращение факторной структуры;
4. содержательная интерпретация результатов факторного анализа.

Для факторного анализа методом максимального правдоподобия были использованы данные, содержащие результаты тестирования по девяти субтестам. Факторный анализ методом максимального правдоподобия проводится на основе выборочной матрицы корреляций, полученной на измеренных переменных, и гипотетической матрицы, соответствующей генеральной совокупности, из которой была взята исследуемая выборка. При этом предполагается, что наблюдаемые переменные распределены нормально, а факторы ортогональны друг другу. Величины факторных нагрузок для генеральной совокупности оцениваются путем вычисления нагрузок, максимизирующих вероятность получения наблюдаемой матрицы корреляций эмпирических данных.

В процессе факторного анализа было выделено 4 фактора и осуществлено вращение матрицы факторных нагрузок методом varimax. Вращение матрицы факторных нагрузок не привело к заметным улучшениям. Таким образом в результате факторного анализа методом максимального правдоподобия было выделено 4 фактора: ML1, ML2, ML3, ML4. Фактор ML4 объясняет 21% общей дисперсии, ML3-17%, ML1-14%, ML2-13%, вместе эти факторы объясняют 65% общей дисперсии. На основании матрицы факторных нагрузок фактор ML4 включает в себя: Субтест6, Субтест2, Субтест9. Фактор ML3

включает в себя: Субтест7, Субтест3. Фактор ML1 включает в себя: Субтест8, Субтест4. Фактор ML2 включает в себя: Субтест1, Субтест5.

Двухфакторный дисперсионный анализ был проведён с использованием статистического пакета R.

Уравнение двухфакторного ДА имеет вид:

$$Y_{IJK} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, \tag{6}$$

где μ— генеральное среднее;
α_i— i-й дифференциальный эффект фактора A;
β_j— j-й дифференциальный эффект фактора B;
$(\alpha\beta)_{ij}$— эффект взаимодействия i-го уровня фактора A с j-м уровнем фактора B;
e_{ijk}— ошибка.

Дисперсионный анализ - это анализ изменчивости признака под влиянием каких-либо контролируемых переменных факторов.

Другими словами, дисперсионный анализ — система статистических методов исследования влияния переменных факторов на изучаемую переменную по дисперсии. Автором метода является Р.А. Фишер.

Сущность дисперсионного анализа состоит в том, чтобы представить общую дисперсию в виде суммы дисперсий, обусловленных влиянием контролируемых (независимых) переменных и, оценивая дисперсионное отношение, определить меру влияния факторов на средние значения изучаемой (зависимой) переменной.

В качестве данных для двухфакторного дисперсионного анализа без повторений (не связанных выборок) была использована таблица, содержащая информацию о результатах тестирования по Субтесту9 (Запоминание), на основании которого оценивается уровень развития кратковременной памяти. При проведении двухфакторного дисперсионного анализа, в качестве зависимой переменной была использована переменная Subtest9 (результаты Субтеста9), а в качестве независимых переменных (факторов) переменные age_group (возрастная группа) и gender (пол). Фактор А (возрастная группа), фактор В (пол). Было сформировано 3 комплекса гипотез.

По результатам двухфакторного дисперсионного анализа, был сделан вывод, что нулевая гипотеза H_0 принимается в комплексах гипотез 2 и 3, то есть влияние фактора B (пол) и влияние взаимодействия между факторами A (возрастная группа) и B (пол) статистически незначимо. Для фактора A (возрастная группа) в таблице дисперсионного анализа было получено значение p-value=0,0144, что меньше чем 0,05, следовательно для первого комплекса гипотез была принята гипотеза H_1 и был сделан вывод, что влияние фактора A (возрастная группа) на результаты Субтеста9 статистически значимо при уровне значимости $\alpha = 0,05$.

Заключение. В рамках магистерской работы в Microsoft Access была создана база данных, позволяющая обрабатывать поступающие заявки на тестирование, путем создания списка тестируемых в таблице базы данных. Было проведено ознакомление с языками программирования VBA, SQL и R.

На языке VBA были разработаны процедуры, позволяющие обрабатывать условия добавления и обновления результатов в таблицах базы данных, через созданные формы. Результаты тестирования импортируемые в базу данных из файлов могут обрабатываться с помощью созданных на языке SQL запросов.

С использованием методов статистики языка R были проведены корреляционный, факторный и дисперсионный анализы на основании реальных данных, а также рассмотрены теоретические аспекты данных видов анализа. Таким образом можно сделать вывод, что цель поставленная в магистерской работе была достигнута.