УДК 519.23
DOI: 10.21122/2309-4923-2022-2-47-52

В. А. ЛИВИНСКАЯ , М.А. ШАЛУХОВА

ПРИМЕНЕНИЕ МЕТОДОВ ПРИКЛАДНОЙ СТАТИСТИКИ В ИССЛЕДОВАНИИ РЫНКА ТРУДА IT-СПЕЦИАЛИСТОВ

Межгосударственное образовательное учреждение высшего образования «Белорусско-Российский университет»

Методы прикладной статистики являются надежным математическим инструментом, позволяющим обрабатывать и анализировать собранные статистические данные. По мере увеличения количества и многообразия устройств, позволяющих накапливать самые разнообразные типы данных, а также расширения сфер деятельности человека, в которых используются информационные технологии, растет необходимость применения современных инструментов статистической обработки данных.
В статье рассматриваются методы прикладной статистики, применяемые для исследования рынка труда специалистов IT-отрасли, с применением свободно распространяемых инструментов анализа данных, одним из которых является язык программирования для статистической обработки данных R.

Ключевые слова: *анализ; прикладная статистика; метод; анализ данных; рынок труда; дисперсионный анализ.*

Введение

С развитием информационных технологий появляются возможности собирать, накапливать, и обрабатывать большие массивы данных, с целью извлечения новых знаний. Единого критерия классификации не существует, однако любые массивы неоднородных данных свыше 150 Гб в сутки являются big data. такие данные требуют особенного инструментария. Данные, не являющихся big data, анализируются с помощью инструментария математической статистики. Математическая статистика играет роль фундамента для прикладной статистики, методы которой активно применяются в самых разнообразных отраслях – от технических до гуманитарных исследований. В данной статье рассматривается применение методов прикладной статистики в исследовании рынка труда ИТ-специалистов.

Объект исследования – рынок вакансий IT-специалистов на основании информации, размещенной в свободном доступе на сайте агрегаторе HeadHunter.

Предмет исследования – актуальные методы прикладной статистики, применяемые для исследования рынка труда специалистов IT-отрасли. Основной целью является анализ информации о вакансиях специалистов IT-отрасли с использованием современных, свободно распространяемых инструментов анализа данных, одним из которых является язык программирования для статистической обработки данных R. Полученные результаты могут быть полезны как для соискателей рабочих мест в этой сфере, составителей программ учебных заведений, подготавливающих IT-специалистов, так и для работодателей, оценивающих конкуренцию на рынке труда.

Основная часть

Основными этапами анализа данных являются: сбор информации, ее очистка от аномальных наблюдений, применение соответствующих инструментов анализа и интерпретация полученных результатов с их визуализацией. В качестве исходной информации для исследования были получены данные о размещаемых вакансиях на русскоязычном сайте агрегаторе. Современные информационные технологии позволяют организовать мониторинг рынка вакансий в режиме реального времени, путем сбора информации с помощью парсинга сайтов-агрегаторов.

В исследовании рассматривалась информация, содержащаяся в 49567 объявлениях о вакансиях в категории «Информационные технологии, интернет, телеком», размещенных с января по май 2021 года на сайте HeadHunter.ru. На первом этапе была написана программа-парсер, взаимодействующая с HeadHunter API. В результате работы программы был осуществлен сбор данных с сайта и преобразование в датафрейм – таблицу, содержащую конкретную информацию по каждой вакансии. Для анализа были отобраны объявления о вакансиях, размещенных в городах Беларуси, Москве и Санкт-Петербурге. Объем выборки составил 26463 вакансий. Информация о размерах предлагаемой заработной платы в этой совокупности размещалась в 42 % вакансий. Основными характеристиками вакансий были выбраны: компетенции соискателя, предметная область, опыт работы, требуемый уровень квалификации, локация предложения, уровень заработной платы. Для анализа полученного датафрейма использовались параметрические и непараметрические методы дисперсионного анализа, позволяющие сравнивать средние или медианные значения количественного признака для двух и более значений категориального признака. Основным инструментом был выбран язык анализа

данных R, в библиотеках которого реализованы все существующие методы прикладной статистики с возможностью визуализации полученных результатов.

На следующем этапе, после получения данных, для обеспечения сопоставимости из исходной совокупности были сформированы подвыборки, содержащие информацию о вакансиях по разным специализациям.

Следующий этап – обнаружение и удаление аномальных наблюдений (выбросов), так как они могли в значительной степени исказить числовые характеристики, такие как средние и дисперсия. Одним из методов, сочетающем оба способа обнаружения данных, вызывающих подозрение, является критерий межквартильного размаха (IQR), позволяющий удалять из выборки наблюдения, не попадающие в интервал.

В данном исследовании, для удаления выбросов, использовались средства языка R–функции boxplot.stats()$out. Наглядное сравнение диаграмм разброса исходного фрейма и данных после удаления выбросов представлено на рисунке (рисунок 1).

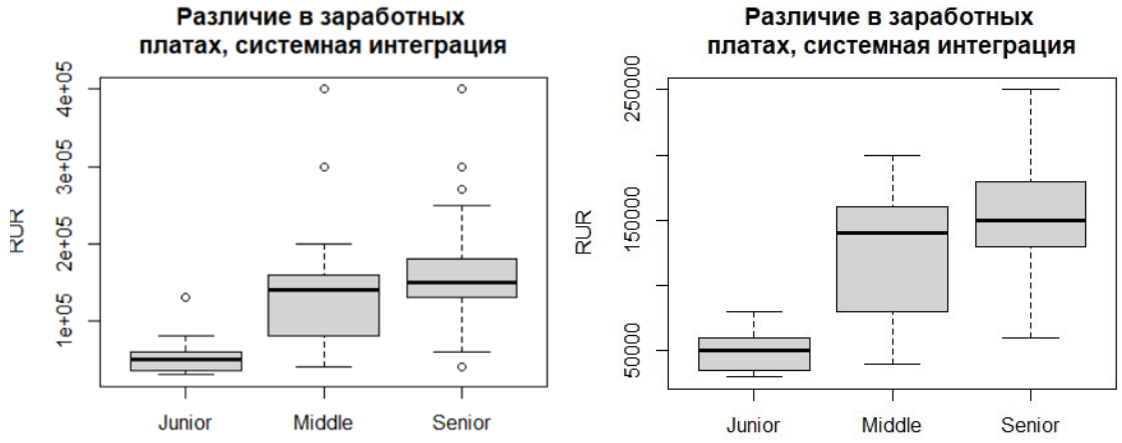

Рисунок 1- Диаграммы разброса признака а) с выбросами б) после удаления выбросов

Дисперсионный анализ (ANOVA /Analisis of Variances) применяют для изучения влияния категориальных признаков на количественную переменную. В зависимости от количества категориальных факторов, одновременно влияющих на количественный, применяется однофакторный или многофакторный, параметрический и непараметрический дисперсионный анализ.

Классические методы параметрического дисперсионного анализа основываются на следующих предпосылках, которые должны быть протестированы:

1) независимость наблюдений;

2) отсутствие выбросов в каждой группе;

3) данные в каждой группе должны быть извлечены из нормально распределенных генеральных совокупностей.

4) дисперсии экспериментальных данных приблизительно равны на различных уровнях изучаемого фактора (условие гомогенности).

В данном исследовании использовались инструменты пакета nortest, в частности тест на нормальность тест Шапиро-Уилка: shapiro.test. В случаях, когда нулевая гипотеза принималась, в дальнейшем использовался параметрический дисперсионный анализ ANOVA. Гомогенность дисперсий проверялась с помощью теста Levene.

Нарушения предпосылок параметрического дисперсионного анализа предполагает использования других подходов. Так, нарушение только условия гомогенности при подтверждении гипотезы о нормальном распределении выборок предполагает использование теста Welch one-way test is в качестве альтернативы стандартному one-way ANOVA.

В случае, когда одна из предпосылок применимости дисперсионного анализа не выполняется, используют его непараметрический аналог Краскелла-Уоллиса, реализованный в используемой библиотеке языка R (kruskal_test). Этот критерий основан на сравнении рангов значений признака в сравниваемых группах. Достоверность различий групповых медиан оценивается с помощью так называемой «статистики размера эффекта», которая показывает степень, в которой одна группа имеет данные с более высокими рангами, чем другая группа. В отличие от p-значений, на них не влияет размер выборки. В случае, когда нулевая гипотеза не может быть принята, попарное различие выясняют с помощью теста Dunn.

Согласно полученным данным, наиболее востребованными специализациями на рынках труда России, Беларуси и Украины являлись “Банковское ПО” (19,7%), “Web-инженер” (19,1%) и “Системная интеграция” (16,7%). Далее анализировалось различие в начальном уровне заработной платы у Web-инженеров, поскольку их востребованность, согласно данным представленным на сайте HeadHunter в Беларуси, наибольшая.

Востребованность Web-инженеров обеспечивается высокой концентрацией фирм, специализирующихся на разработке web и мобильных приложений, а

также общим процессом цифровизации.

В работе протестирована гипотеза о различии в заработной плате Web-инженеров в зависимости от города (Москва, Санкт-Петербург, города Республики Беларусь) и степени квалификации соискателя(grade). После удаления одного аномального наблюдения в городе Москва с заработной платой 5600 $ была протестирована гипотеза о нормальном распределении соответствующих выборок с помощью теста Шапиро-Уилка и была отвергнута ($p<0,05$). В дальнейшем применялся критерий Краскелла Уоллиса. Описательная статистика для групп представлена в таблице 1.

Таблица 1 – Числовые характеристики показателя заработная плата Web-инженера в долларах США (usd)

Числовая характеристика	Регион		
	Москва	Республика Беларусь	Санкт-Петербург
Количество наблюдений(n)	117	27	60
Минимальное значение (min)	410,959	300	219,178
Максимальное значение (max)	4680	4550	4680
Медиана (median)	1917,808	1750	1438,356
Межквартильный размах (iqr)	1506,849	2050	1643,836
Среднее значение (mean)	1972,185	2105,733	1768,742
Стандартное отклонение (sd)	1074,741	1357,899	1242,032

Гипотеза о статистически значимом различии медиан в группах не может быть отвергнута на основании теста Kruskal-Wallis (таблица 2, рисунок 2).

Таблица 2– Результаты попарного сравнения медиан с помощью теста Dunn

Группа 1	Группа 2	Количество (n1)	Количество (n2)	Statistic	Вероятность (p)
Москва	РБ	117	27	0,044689804	0,964354561
Москва	Санкт-Петербург	117	60	-1,631468846	0,102791428
Республика Беларусь	Санкт-Петербург	27	60	-1,159051883	0,246435038

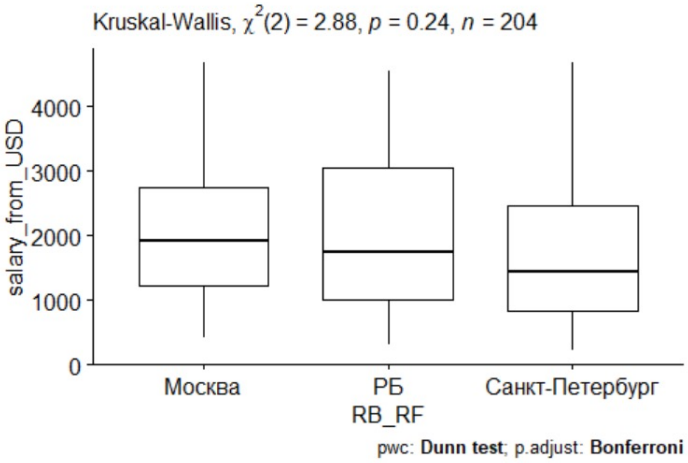

Рисунок 2. Результаты попарного сравнения медиан с помощью теста Dunn

С учетом квалификации, у соискателей с одним и тем же уровнем квалификации предлагаемая начальная заработная плата статистически не различается (рисунок 2).

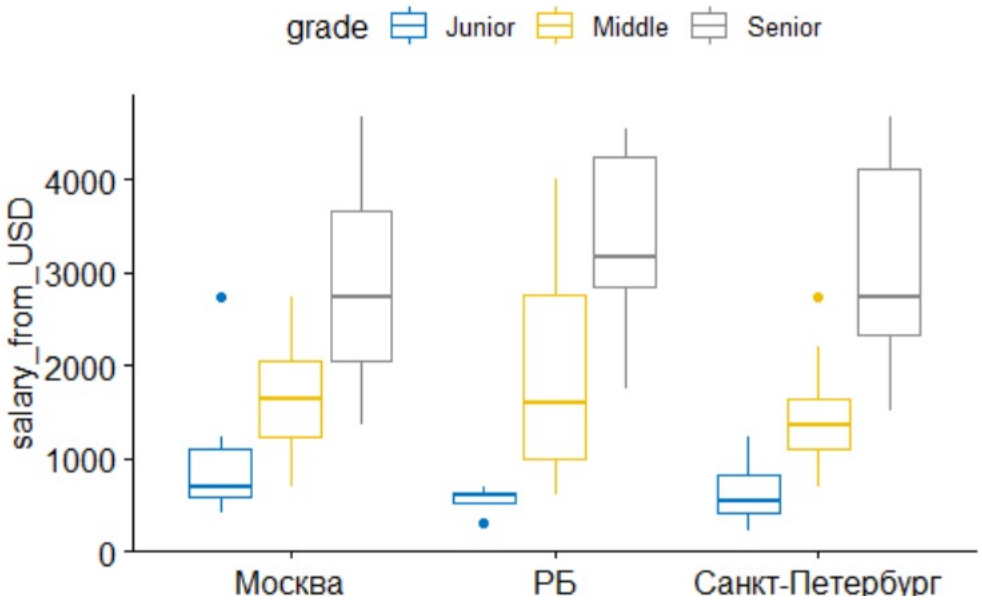

Рисунок 3. Результаты попарного сравнения медиан с помощью теста Dunn

Учитывая отсутствие различия в начальном уровне заработных плат, релокация веб-инженеров в Москву и Санкт-Петербург экономически нецелесообразна, с учетом более высокой стоимости жизни в этих городах.

Таблица 3 – Разброс в зарплате по уровню компетенции web-инженеров

Компетенция	Количество (n)	Минимальное значение (min)	Максимальное Значение (max)	Медиана (median)	Межквартильный размах (iqr)	Среднее значение (mean)	Стандартное отклонение (sd)
Junior	46	219,178	2739,726	684,932	376,712	746,456	410,94
Middle	85	600	4000	1506,849	890,411	1659,226	722,21
Senior	73	1369,863	4680	2739,726	1945,205	2991,147	994,9

Разброс в предлагаемой заработной плате имеет широкий диапазон для всех компетенций специалистов, что характерно для отрасли, в целом и связанно с особенностями найма и оплаты труда в IT компаниях.

Рассмотрим связь между категориальными признаками регион и уровень компетенции специалиста. Таблица сопряженности между категориальными признаками регион и уровень компетенции представлена в таблице 4.

Таблица 4 – Распределение вакансий по регионам и уровню компетенции

Регион	Компетенция		
	Junior	Middle	Senior
Москва	26	45	46
Республика Беларусь	4	17	6
Санкт-Петербург	16	23	21

Визуальное представление разницы структур распределения вакансий по регионам и уровню компетенции получено с помощью построения мозаичного графика (рисунок 8).

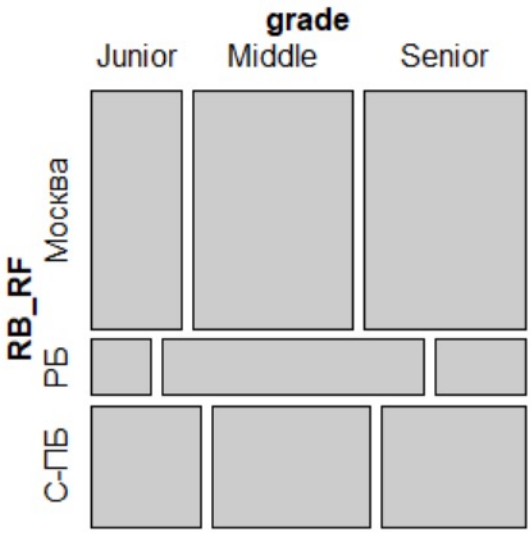

Рисунок 4. Мозаичный график распределения вакансий по регионам и уровню компетенции

Из графика видно, что во всех рассматриваемых регионах наименее востребована квалификация Junior-специалистов. В Республике Беларусь из опытных разработчиков наиболее востребованной на рынке труда квалификацией является квалификация Middle-специалистов, в Москве и Санкт-Петербурге востребованность квалификаций опытных разработчиков Middle и Senior специалистов отличается незначительно.

Для получения ответа на вопрос есть ли связь между количеством вакансий в зависимости от уровня квалификации городами, нулевая гипотеза формулируется из предположения, что связь отсутствует. Применение точного критерия Фишера для таблицы сопряженности не позволило отклонить нулевую гипотезу (p-value = 0.21). Таким образом нельзя считать, что определенная квалификация более востребована в каком – либо городе.

Заключение

Для решения задачи анализа статистических данных и визуализации его результатов оптимальным выбором является использование методов прикладной статистики. Проведенный в статье анализ рынка труда специалистов IT-отрасли был выполнен с помощью возможностей языка сценариев R. R широко используется аналитиками данных, и обладает одними из лучших возможностей визуализации графиков. В результате проведенного анализа было доказано отсутствие статистической разницы в уровне оплаты Web-инженеров рассматриваемых регионов. Заказчиками разработки программного обеспечения в Республике Беларусь преимущественно являются иностранные компании (около 80%), в Москве и Санкт-Петербурге заказчиками услуг в большинстве своем выступают российские кампании (более 80 %). Данный факт позволяет предположить, что в Российской Федерации внедрение цифровых технологий находятся на более высоком уровне. Наибольшая конкуренция наблюдается среди специалистов квалификации Junior, наиболее востребованными специализациями на рынках труда России, Беларуси и Украины в рассматриваемый период являются “Банковское ПО” (19,7%), “Web-инженер” (19,1%) и “Системная интеграция” (16,7%).

Результаты такого анализа могут использовать как потенциальные соискатели с опытом работы в соответствующей сфере, так и представители университетов (студенты, выбирающие наиболее интересное для них направление в IT, а также преподаватели, разрабатывающие контент для приобретения компетенций в этих направлениях).

ЛИТЕРАТУРА

1. **Кобзарь А. И.** Прикладная математическая статистика. Для инженеров и научных работников. – М.:ФИЗМАТЛИТ, 2006 – 816 с. –ISBN 5-9221-0707-0

2. **R Core Team (2020).** R: язык и среда для статистических вычислений. R Foundation for Statistical - Computing, Вена, Австрия. [Электронный ресурс].– Режим доступа: URL https://www.R-project.org/

REFERENCES

REFERENCES

1. **Kobzar A. I.** Applied mathematical statistics. For engineers and scientists. - Moscow:FIZMATLIT, 2006 - 816 p. -ISBN 5-9221-0707-0

2. **R Core Team (2020).** R: A language and environment for statistical computing. R Foundation for Statistical – Computing, Vienna, Austria. t [Online].– Available:URL https://www.R-project.org/ – Date: 01.11.2021.

V. A. LIVINSKAYA, M. A. SHALUCHOVA

APPLICATION OF APPLIED STATISTICAL METHODS IN THE STUDY OF THE LABOR MARKET OF IT-SPECIALISTS

Interstate Educational Institution of Higher Education Belarusian-Russian University

The methods of applied statistics are reliable mathematical tool for processing and analyzing collected statistical data. As the number and variety of devices allowing to accumulate a variety of data types increase and the spheres of human activity in which information technology is used expand, the need for modern statistical data processing tools grows.

The article describes methods of applied statistics used to study the labor market of IT-industry professionals, using freely distributed tools for data analysis, one of which is a programming language for statistical data processing R.

Keywords: *analysis; applied statistics; method; data analysis; labor market; analysis of variance; ANOVA.*

Ливинская Виктория Александровна, кандидат физико-математических наук, доцент кафедры финансов и бухгалтерского учета Белорусско-Российского университета, кафедры программного обеспечения информационных технологий, г. Могилёв.
Сфера научных интересов: прикладной статистический анализ данных, эконометрика, математическое моделирование социально-экономических процессов, использование методов машинного обучения в медицинских исследованиях.
Livinskaya V. A., PhD in Phisics and Mathematics, Associate Professor, Associate Professor of the Department of Finance and Accounting, Belarusian-Russian University, Mogilev.
Research interests: applied statistical data analysis, econometrics, mathematical modeling of socio-economic processes, the use of machine learning methods in medical research.

Email: viktorijalivinskaya@gmail.com

Шалухова Мария Александровна, магистрант Белорусско-Российского университета, г. Могилев.
Сфера научных интересов: анализ данных, системный анализ, системы управления информацией, кибер-физические системы, искусственные нейронные сети, применение нейросетевых технологий для разработки систем управления.
Shaluchova M. A., graduate student, Belarusian-Russian University, Mogilev.
Research interests: data analysis, systems analysis, information management systems, cyber-physical systems, artificial neural networks, application of neural network technology to develop control systems.

Email: shaluhova.m@gmail.com