Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

кафедра социальной информатики

СОЦИОЛОГИЧЕСКИЙ АНАЛИЗ БАЗ ДАННЫХ В ТРАДИЦИОННОМ И ПРОГРАММНОМ РЕЖИМАХ

(автореферат бакалаврской работы)

студента 4 курса 451 группы
направления 09.03.03 - Прикладная информатика
профиль прикладная информатика в социологии
Социологического факультета
Баталина Максима Алексеевича

Научный руководитель
кандидат философских наук, доцент _____ А.И. Завгородный
подпись, дата

Зав. кафедрой
кандидат социологических наук, доцент _____ И.Г.Малинский
подпись, дата

Саратов 2017

ВВЕДЕНИЕ

Актуальность проблемы. Необходимость использования приемов анализа статистических данных в совершенно разных областях деятельности, средств статистического анализа данных в различных сферах деятельности (медицине, социологии, менеджменте, экономике и других), преимущественно в научной сфере, крайне высока. Такая потребность является причиной развития программных пакетов, предназначенных для применения широкого спектра методов обработки статистической информации. В последнее время очень активно развиваются компьютерные приложения, которые позволяют анализировать статистические (социологические) данные не только малых, но и больших объемов для того, чтобы выявлять некоторые закономерности, заниматься построением прогнозов развития событий, проводить сравнения вероятных альтернатив выбора, а также определять связи между происходящими процессами и явлениями.

Актуальность исследования обусловлена сложностью выбора наилучшего статистического пакета, позволяющего находить верные ответы на существующие вопросы. Наилучший вариант должен сочетать в себе весь требуемый функционал, отличное качество работы и относительно невысокую цену. При выборе программного обеспечения следует учитывать уровень подготовки пользователя, собирающегося использовать приложение, способность обработки больших массивов информации, возможность решения поставленных задач, системные требования, предъявляемые к компьютеру.

Степень научной разработанности данной проблемы довольно низкая, так как она имеет специфический уклон и узкую направленность. Большое внимание уделяется обзорам и сравнениям различного программного обеспечения для обработки статистической информации, которые отмечены в трудах таких ученых как А.А. Макаров (STADIA против STATGRAPHICS, или кто ваш лоцман в море статистических данных//Мир ПК, № 3, 1992.), Векслер Л.С (Статистический анализ на персональном компьютере//Мир ПК, № 2, 1992.), Сильвестров Д.С. (Программное обеспечение прикладной

статистики. - М.: Финансы и статистика, 1989.), Боровиков В.П. (Программа STATISTICA для студентов и инженеров / В.П. Боровиков – М., 2001.), Бююль А.. (SPSS: Искусство обработки информации / А. Бююль, П. Цёфель – . М., 2002).

Объектом данного исследования будет являться пакет программ SPSS Statistics.

Предметом будет расчет описательных статистик, к которым относятся меры средней тенденции и меры разброса, традиционным и программным способом.

В любом исследовании цель ориентирует его на конечный результат, теоретико-познавательный и практически-прикладной.

Целью будет выявление различий между традиционным анализом данных и с помощью программного обеспечения SPSS Statistics.

Для достижения поставленной цели необходимо выполнить следующие ***задачи***. К ним относятся определение характеристик пакета программ SPSS, описание баз данных, которые будут рассматриваться в качестве примера, проведение анализа традиционным способом, проведение анализа с помощью пакета программ SPSS Statistics. В качестве подсчитываемых показателей были выбраны меры средней тенденции и меры разброса. Также нужно охарактеризовать потенциал используемой программы.

Эмпирическую базу данной работы составляют данные двух социологических исследований. Одно из них посвящено изучению отношения студентов СГУ к здоровью. Второе связано с определением отношения людей к интернет покупкам. В качестве метода сбора был выбран опрос (анкетирование). В первом исследовании было опрошено 368 респондентов, а во втором 125.

Научная новизна обусловлена тем, что до этого рассмотрение анализа с помощью программного обеспечения и традиционным способом осуществлялось раздельно. ПО всегда считали более прогрессивным видом обработки информации. В данной работе была сделана попытка проведения

сравнительного анализа для того, чтобы узнать, какие имеются отличия при проведении подсчетов разными способами.

Структура бакалаврской работы. Работа состоит из введения, трех разделов, заключения, списка использованных источников и приложения.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении показана актуальность данного исследования, определена степень изученности по теме, объект, предмет, цель, задачи и методы исследования, эмпирическая база, показана новизна.

В ***первом разделе "Структурно-функциональные характеристики SPSS Statistics"*** обозреваются структурно-функциональные характеристики пакета программ SPSS Statistics, показываются его недостатки и преимущества, также дается интерпретация такому понятию, как традиционная обработка социологической информации.

Статистический пакет программ SPSS Statistics по признаку функциональности можно отнести к универсальному продукту или пакету общего назначения.

Пакеты такого типа не предназначены для решения специфических задач и могут использоваться для анализа данных из разных сфер деятельности. Они включают в себя достаточно крупный набор статистических методов и обладают относительно простым интерфейсом. Применять подобные программные продукты рекомендуется начинающим пользователям, которые имеют лишь некоторые представления о работе программы статистической обработки, а также тем, кто уже достаточно осведомлен в этой области на начальных этапах. Еще одним важным свойством подобного программного обеспечения является многопрофильность. С помощью нее возможно проведение тестового анализа разного рода данных, используя немалый набор статистических средств. Она позволяет провести пробный анализ разных типов данных с помощью широкого диапазона статистических методов. Большая часть представленных на рынке приложений, относящихся к универсальным, очень похожи по составу применяемых в них статистических методов.

SPSS Statistics — компьютерная программа для статистической обработки данных, один из лидеров рынка в области коммерческих статистических продуктов, предназначенных для проведения прикладных исследований в социальных науках. К достоинствам можно отнести мощный статистический инструментарий, возможность решения широкого спектра задач из абсолютно разных областей, которые требуют применения статистической обработки информации, большой набор статистических и графических процедур анализа данных, а также процедур создания отчетов, высокую скорость вычислений, интуитивно понятный интерфейс, хорошо проработанную справочную систему, которая помогает начинающему пользователю, возможность свободного скачивания пробной версии программного обеспечения на официальном сайте, многообразие поддерживаемых языков, совместимость с операционными системами Windows, Mac, Linux, а также богатый выбор литературы, позволяющей овладеть всеми процедурами.

К недостаткам относятся высокие системные требования и высокая стоимость программного продукта по сравнению со статистическими пакетами аналогичного уровня.

С помощью IBM SPSS Statistics возможно проведение всех этапов анализа. Ими являются планирование исследования, сбор данных, доступ и управление данными, всесторонний анализ, создание отчетов, хранение и распространение результатов.

Рассматриваемое программное обеспечение IBM SPSS Statistics открывает очень разнообразные возможности для обработки информации. Достаточно понятный интерфейс включается в себя все процедуры работы с данными, статистические методы и средства построения отчетов для применения анализа на каждом уровне сложности. IBM SPSS Statistics и продукты IBM SPSS Amos, Sample Power, VizDesigner, Data Collection, Collaboration and Deployment Services образуют модульный, полностью интегрированный программный комплекс.

К сферам применения рассматриваемого пакета программ относятся маркетинговые исследования и продажи, финансовый анализ, хранение и обработка информации из опросов и другие. Для социологии программный продукт предоставляет автоматизированные процедуры создания баз данных социологической информации, их хранение и анализ.

В SPSS Statistics включает в себя разного типа графические возможности. К ним относятся немалое число различных категорий и типов графиков в разных системах координат. К ним относятся научные, деловые, трехмерные и двухмерные. Также доступны специализированные статистические графики (гистограммы, матричные и категоризованные).

Многообразие статистических процедур, которые включены в программу, поражает. Приложение позволяет осуществлять практически любой вид анализа, но для этого пользователю необходимо понять основные принципы их работы. Для того чтобы в полной мере пользоваться всей мощью статистического пакета надо потратить достаточно большое количество времени, которое нужно для изучения. Поэтому не каждый способен освоить материал по данному программному обеспечению. Из-за этого его потенциал, чаще всего, не используется даже наполовину.

Во ***втором разделе "Описание используемых источников данных"*** осуществляется описание исследований, которые будут применены для проведения анализа данных.

В июле 2015 года в центре региональных социологических исследований проводилось исследование, целью которого было узнать, что думают студенты СГУ о здоровье, в частности о санатории – профилактории СГУ. С помощью исследования были выявлены основные тенденции и статистические показатели. В этой работе сбор данных осуществлялся с помощью такого метода как анкетный опрос. В опросе принимали участие 368 студентов СГУ. А сама анкета состояла из 10 вопросов. В качестве еще одного источника данных, который позволит провести анализ данных, будет выступать пилотажное социологическое исследование, проведенное студентом СНИГУ Н. И.

Чернышевского направления 09.03.03 - Прикладная информатика, профиль подготовки - Прикладная информатика в социологии социологического факультета, Королевским Александром.

Целью данного исследования было выявление отношения людей к интернет покупкам. В качестве метода опроса было использовано анкетирование. А именно его разновидность – интернет опрос. Он позволяет сэкономить большое количество ресурсов. Но в то же время получить нужные данные в короткие сроки. В опросе принимали участие 125 человек, имеющих различный социальный статус и материальное положение. А сама анкета состояла из 17 вопросов.

В ***третьем разделе "Сравнительный анализ характеристик обработки социологической информации (Традиционная и SPSS Statistics)"*** приведены результаты расчетов описательных статистик традиционным способом и с применением программного обеспечения SPSS Statistics. На базе этих результатов был выполнен сравнительный анализ. Он проводился параллельно, что позволяло делать промежуточные выводы после каждого подсчета.

Одной из важнейших характеристик при описании поведения отдельных переменных является показатель средней тенденции. Возможности использования различных мер средней тенденции для шкал разного типа различны. Возможности использования различных мер средней тенденции выглядят следующим образом:

1) Для номинальных шкал используется только показатель моды (Мо)
2) Для порядковых используются мода и медиана (Ме)
3) Для количественных используются мода, медиана, среднее арифметическое ($\bar{X}$), дисперсия (δ^2), стандартное отклонение (σ), стандартная ошибка среднего($co\bar{x}$).

Мода – это то значение в анализируемой совокупности данных, которое встречается чаще других, поэтому нужно посмотреть на частоты значений и отыскать максимальное из них. Иногда в совокупности встречается более чем

одна мода. В этом случае можно сказать, что совокупность мультимодальна. Из структурных средних величин только мода обладает таким уникальным свойством. Как правило, мультимодальность указывает на то, что набор данных не подчиняется нормальному распределению.

С помощью программы SPSS Statistics была выведена мода, которая показала полное сходство в вычислениях. Модой является ответ 2.

Центральную тенденцию данных можно рассматривать не только, как значение с нулевым суммарным отклонением (средняя арифметическая) или максимальную частоту (мода), но и как некоторую отметку (определенный уровень анализируемого показателя), делящую ранжированные данные (отсортированные по возрастанию или убыванию) на две равные части. То есть половина исходных данных по своему значению меньше этой отметки, а половина – больше. Это и есть медиана.

Для нахождения медианы была выявлена следующая частота:

- 1 вариант отметили 4 респондента
- 2 ответ отметили 20 респондентов
- 3 ответ отметили 93 респондента
- 4 ответ отметили 168 респондентов
- 5 ответ отметили 78 респондентов

Упорядочив ряд распределения по возрастанию, была найдена медиана. В связи с тем, что в ряду распределения имеется два центральных значения, необходимо подсчитать среднее арифметическое этих двух элементом. Такая ситуация возникает, когда ряд распределения четный.

Подсчеты, которые проводились в программе SPSS Statistics, показали, что результат нисколько не отличается от традиционного варианта.

Средняя арифметическая является одной из наиболее распространённых мер центральной тенденции. Она используется, когда расчет осуществляется по не сгруппированным статистическим данным, где нужно получить среднее слагаемое. Средняя арифметическая - это такое среднее значение признака, при

получении которого сохраняется неизменным общий объем признака в совокупности. Является суммой всех чисел, деленной на их количество.

Формула средней арифметической имеет следующий вид:

$$\frac{x = (x_1 + x_2 + \ldots + x_n)}{n} = \sum x_i \,/n$$

При вычислении того же показателя с помощью программного обеспечения, был найден тот же результат, что и при традиционном анализе данных. Ответом является 19944.

Дисперсия - это очень важный показатель, который активно используется в различных методах статистического анализа (проверка гипотез, анализ причинно-следственных связей и др.). Как и среднее линейное отклонение, дисперсия также отражает меру разброса данных вокруг средней величины.

Проведение подсчета дисперсии происходит по следующей формуле:

$$\delta^2 = \frac{(\bar{X} - x_1) + (\bar{X} - x_2) + .. (\bar{X} - x_n)}{N}$$

$\bar{X}$ – среднее арифметическое

x_1, x_2, x_n – значения признака

N – количество наблюдений

При подсчете дисперсии были выявлены различия в результатах. Традиционным способом получился ответ 165524864. Программная обработка дала немного отличающийся показатель 166900000.

Стандартное отклонение характеризует степень отклонения данных наблюдений или множеств от среднего значения. Небольшое стандартное отклонение указывает на то, что данные группируются вокруг среднего значения, а значительное - что начальные данные располагаются далеко от него.

Для подсчета среднеквадратичного отклонения используется формула, имеющая такой вид:

$\sigma = \sqrt{\delta^2}$

δ^2– дисперсия

√165524864 = 12865.6466608

Использовав для подсчета стандартного отклонения, программное обеспечение SPSS Statistics, было получено следующее значение:

12917,42010

Стандартная ошибка указывает на точность среднего выборки как вычисления среднего генеральной совокупности. Чем меньше стандартная ошибка, тем меньше разброс, и более вероятно, что любое среднее выборки близко к среднему генеральной совокупности.

Стандартная ошибка средней арифметической вычисляется по формуле:

$$S_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

σ – стандартное отклонение

n – количество наблюдений

$$S_{\bar{x}} = \frac{12865{,}64666}{\sqrt{125}} = \frac{12865{,}64666}{11.1803398875} = 1150.73842025$$

Подсчитав стандартную ошибку среднего с помощью машинной обработки информации, получились результаты 1155,36918.

Еще одним показателем является минимум. Рассмотрим его и определим. Минимум из чисел $a_1, a_2, \dots a_n$ есть число (числа), которое не больше всех остальных. Как правило, используется обозначение $min\{a_1, \dots, a_n\}$. Ответом является 1000.

Максимум из чисел $a_1, a_2, \dots a_n$ есть число (числа), не меньшее (не меньшие), чем все остальные. Как правило, используется обозначение $max\{a_1, \dots, a_n\}$. Ответом является 50000.

Для вычисления размаха необходимо выполнить несколько простых действий. Сначала следует записать значения совокупности данных, перечисляя все значения для определения минимального и максимального чисел. После выполнения этой несложной процедуры необходимо вычесть наименьшее число из наибольшего. Это и будет являться значением размаха.

Max – 50000

Min – 1000

R = Max – Min = 50000- 1000 = 49000

Последние три расчета позволяют говорить о полном сходстве в результатах, полученных традиционным и программным способом.

Заключение. В заключении дается краткий обзор, показывающий наиболее значимые выводы, которые были получены в ходе анализа.

Статистические программные пакеты сделали методы анализа данных более доступными и наглядными: теперь уже не требовалось вручную выполнять трудоемкие расчеты по сложным формулам, строить таблицы и графики — всю эту черновую работу взял на себя компьютер, а человеку осталась главным образом творческая работа: постановка задач, выбор методов их решения и интерпретация результатов.

При составлении программы исследования была поставлена цель:
выявление различий между традиционным анализом данных и с помощью программного обеспечения SPSS Statistics.

Согласно задачам данной работы было дано описание пакета программ SPSS, указаны основные возможности, наглядно показан его интерфейс, даны основные понятия для понимания предметной области. Также показана область применения.

Было дано достаточно полное описание эмпирической базы, к которым относятся два исследования.

Проведен анализ данных (подсчет описательных статистик) с помощью программного обеспечения, а также сравнение с традиционным анализом.
Таким образом, это позволяет сделать следующие выводы:
При подсчете описательных статистик вручную было затрачено гораздо больше времени.
Проведя различные расчеты, можно сказать, что не все показатели показывают тот же результат, что и при помощи машинной обработки информации.

Несмотря на то, что подсчитывались простейшие показатели, результаты в некоторых из них расходятся. Эта разница может быть связана с человеческим фактором, либо с разницей алгоритмов подсчета.