

Государственное образовательное бюджетное учреждение
высшего профессионального образования
«Государственный университет – Высшая школа экономики»

На правах рукописи



004611297

Мавлетова Айгуль Маратовна

**МЕТОДОЛОГИЯ ВЕБ-ИССЛЕДОВАНИЙ:
ОСНОВНЫЕ ПОДХОДЫ К ФОРМИРОВАНИЮ ВЫБОРКИ
И ВОЗМОЖНОСТИ ПОВЫШЕНИЯ КАЧЕСТВА ДАННЫХ**

Специальность 22.00.01 – теория, методология и история социологии

Автореферат
диссертации на соискание ученой степени
кандидата социологических наук

Научный руководитель
д.социол.н., профессор И.Ф. Девятко

21 ОКТ 2010

Москва 2010

Работа выполнена в государственном образовательном бюджетном учреждении высшего профессионального образования «Государственный университет – Высшая школа экономики»

Научный руководитель: Доктор социологических наук, профессор
Девятко Инна Феликсовна

Официальные оппоненты: Доктор социологических наук, профессор
Татарова Галина Гашеевна

Кандидат социологических наук, доцент
Стрельникова Анна Владимировна

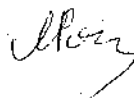
Ведущая организация: Российский Университет Дружбы Народов,
кафедра социологии.

Защита состоится «22» октября 2010 года на заседании диссертационного совета Д. 212.048.01 в Государственном университете – Высшей школе экономики по адресу: 101000, Москва, ул. Мясницкая, 20.

С диссертацией можно ознакомиться в библиотеке Государственного университета – Высшей школы экономики.

Автореферат разослан «22» сентября 2010 г.

Ученый секретарь
диссертационного совета
к.э.н.



Я.М. Рощина

Общая характеристика работы

Актуальность темы

За последние пятьдесят лет область методологии и методов социологических исследований претерпела большие изменения. В числе самых значимых изменений можно выделить развитие методов вероятностной выборки и новых подходов к многомерному статистическому анализу, распространение телефонного интервью и разработку методов панельных исследований. Однако ни одно из этих изменений не привело к таким серьезным последствиям, как применение компьютерных методов к сбору социологических данных. Это изменение является решающим и поворотным в исследовательской практике. Движение к исследованиям, основанным на компьютерном сборе данных, можно рассматривать как нечто закономерное вследствие того, что компьютер занимает центральное место во всех областях человеческой деятельности.

С распространением Интернета стремительно начала развиваться область веб-исследований. В настоящее время онлайн исследования уже могут рассматриваться не только как теоретическое нововведение, но и как реальная практика проведения социологических исследований. Вместе с тем веб-исследования пока не получили широкого признания в научной среде, и основные причины постановки под вопрос их легитимности – базовые для методологии проблемы репрезентативности и валидности данных. Важно отметить, что социальные ученые зачастую не учитывают как релевантность использования онлайн опросов в ряде исследовательских задач (например, в экспериментальных исследованиях), так и возможные методологические решения данных проблем в некоторых типах веб-исследований. Более того, не учитываются и технологические преимущества Интернета, позволяющие повысить качество онлайн данных.

Этими обстоятельствами обусловлена актуальность рассмотрения вопросов методологии веб-исследований. В работе мы рассмотрим вопросы *валидности и репрезентативности* онлайн данных сквозь призму основных *подходов к построению выборочной совокупности*, а некоторые *возможности улучшения качества данных* – сквозь призму *онлайн методологии опроса редких и «закрытых» совокупностей*, а именно, *выборки, управляемой респондентами* (respondent-driven sample – RDS).

Отметим, что, несмотря на увеличивающееся количество публикаций по веб-исследованиям, наблюдается недостаток теоретических и эмпирических работ, затрагивающих фундаментальные основы веб-исследований. Актуальность данной работы определяется и тем, что она, в том числе, предназначена восполнить пробелы в отечественной методологии социологических исследований. Обсуждение базовых методологических вопросов проведения веб-исследований, рассмотрение решений проблем валидности и репрезентативности онлайн данных, а также возможностей повышения качества данных позволит увеличить интерес к онлайн исследованиям в научной среде, и что более важно – начать движение к разработке и обоснованию новых методов сбора и анализа социологических данных.

Разработанность проблемы

Работы, представляющие ценность для нашего исследования, условно можно разделить на два основных направления: теоретико-методологические и эмпирические исследования. Первое направление рассматривает методологические вопросы веб-исследований. К примерам этих работ можно отнести исследования М.Купера¹, Д.Дилмана², Дж.Бетлеема³, Дж. Клинтона⁴ и

¹ Couper M. Web surveys: A review of issues and approaches // Public Opinion Quarterly. 2000. Vol. 64. No. 4. P. 464-494.

² Dillman D. A. Mail and Internet surveys: The tailored design method. New York: John Wiley and Sons. 2000.

Дж.Кросника⁵, а в России – работы Б.З. Докторов⁶ и И.Ф.Десятко⁷. В них был получен ряд важных результатов. В частности, М.Купер первым предложил наиболее полную типологию веб-исследований и рассмотрел методологические основания разных типов веб-исследований. Д.Дилман обосновал способы уменьшения ошибки «неответов» в веб-опросах и возможные методы повышения качества данных онлайн исследований. Дж. Клинтон и М.Деннис, фокусируя внимание на анализе онлайн панелей, показали существующую недостаточность доказательств подтверждения гипотезы эффекта «созревания» участников панелей и статистически незначимую ошибку, приносимую «осыпанием» панелей. Дж.Бетлеем предложил решение проблемы внешней валидности веб-исследований путем статистической корректировки полученных результатов; напротив, Дж.Кросник и его коллеги, сравнивая результаты онлайн и оффлайн исследований в терминах внешней валидности, пришли к выводу о принципиальной невозможности корректировки онлайн данных, полученных с помощью опроса access-панели, или так называемой панели добровольцев.

На сегодняшний день опубликовано достаточно ограниченное количество работ теоретико-методологического характера в области веб-опросов. И если имеются исследования, анализирующие тот или иной методологический аспект онлайн опросов, то практически отсутствуют работы, затрагивающие базовые вопросы методологии онлайн исследований как области в целом.

⁵ Bethlehem J. Reducing the bias of web survey based estimates. Discussion paper 07001. Voorburg/Heerlen, 2007.

⁴ Clinton J. Panel bias from attrition and conditioning: A case study of the Knowledge Networks // Paper presented at the Annual Conference of the American Association for Public Opinion Research. Montreal, Canada, 2001, P. 1-34.

⁵ Chang L., Krosnick J. National surveys via RDD telephone interviewing vs. the Internet: Comparing sample representativeness and response quality // Public Opinion Quarterly. 2009. Vol. 73, P. 641-678.

⁶ Докторов Б. Реклама и опросы общественного мнения в США: История зарождения. Судьбы творцов. М.: ЦСП, 2008.

⁷ Десятко И.Ф. Онлайн исследования и методология социальных наук: новые горизонты, новые (и не столь новые) трудности // Онлайн исследования в России 2.0 / Под ред. А.В. Шашкина, И.Ф. Десятко, С.Г. Давыдова. М: РИЦ «Северо-Восток», 2010. С. 17-30.

Эмпирические работы ведутся в нескольких направлениях. Первый пласт работ рассматривает технические процедуры проведения онлайн опросов и их администрирования. Второй пласт исследований фокусирует внимание на возможностях увеличения внутренней валидности онлайн данных. Некоторые из этих эмпирических исследований показывают данную возможность за счет технических новшеств, например, за счет включения изображений в анкету или программирования инструментария на флеш платформе⁸. Другие эмпирические работы позволяют делать выводы о том, что онлайн среда больше гарантирует анонимность процедуры опроса и тем самым увеличивает открытость респондентов в исследованиях на чувствительную тематику⁹. Третий пласт исследований показывает эффективность использования Интернета для набора и опроса определенных социальных групп, например, редких совокупностей¹⁰. Наконец, можно выделить тип работ, посвященных сравнению, обсуждению и поискам техник, помогающим повысить уровень откликов в онлайн исследованиях¹¹. Несмотря на значительное количество эмпирических исследований в данной области, вопросы улучшения качества онлайн данных посредством разработки новых инструментов и развития новых перспектив «традиционных» методов остаются недостаточно изученными.

Цель, задачи, предмет и объект исследования

В диссертационной работе мы ставим следующую *цель*: выявить некоторые возможности повышения качества данных веб-исследований и указать на потенциал использования веб-опросов для повышения валидности

⁸ Couper M., Conrad F., Tourangeau R. Visual context in web-surveys // Public Opinion Quarterly. 2007. Vol. 71. No. 4. P. 623-634.

⁹ Joinson A., Paine C., Buchanan T., Reips U-D. Measuring self-disclosure online: Blurring and non-response to sensitive items in web-based surveys // Computers in Human Behavior. 2008. Vol. 24. P. 2158-2171.

¹⁰ Bowen A., Williams M., Horvath K. Using the Internet to recruit rural MSM for HIV risk assessment: Sampling Issues // AIDS and Behavior. 2004. Vol. 8. No. 3. P. 311-319.

¹¹ Bosnjak M., Neubarth W., Couper M., Bandilla W., Kaczmirek L. Prenotification in web-based access panel surveys: The influence of mobile text messaging versus e-mail on response rates and sample composition // Social Science Computer Review. 2008. Vol. 26. No. 2. P. 213-223.

данных о труднодоступных группах населения. Соответственно, *объектом* данного диссертационного исследования являются методологические подходы, сложившиеся в области веб-исследований, *предметом* – подходы к формированию выборки и оценке качества данных в веб-исследованиях.

Сформулированная выше цель требует решения семи исследовательских *задач*.

1) Типологизировать существующие веб-исследования и рассмотреть методологические основания каждого типа.

2) Оценить качество данных в каждом выделенном типе веб-исследований в терминах внутренней и внешней валидности.

3) Рассмотреть методологические основания формирования выборки в веб-опросах, а также возможности для оценки параметров генеральной совокупности и расчета ошибок опроса.

4) Проанализировать способы статистической корректировки и ремонта эмпирических данных в веб-исследованиях.

5) Обозначить перспективы применения онлайн методологии выборки, управляемой респондентами (RDS – respondent-driven sample), для опроса труднодоступных социальных групп и институционально замкнутых совокупностей.

6) Провести онлайн RDS опрос игроков в казино и сравнить итоги с результатами оффлайн исследования с применением вероятностной выборки.

7) Определить эффективность проведения онлайн RDS опроса в терминах результативности опроса и качества данных.

Теоретические и методологические основы диссертации

Теоретико-методологической основой диссертационного исследования стали работы М. Купера и предложенная им типология веб-опросов, исследования Д. Хекаторна по разработке RDS методологии, а также работы европейских и американских социологов в области веб-исследований.

Отметим, что важным преимуществом RDS метода является возможность оценить параметры генеральной совокупности и выборочную ошибку, что приближает качество получаемых данных к результатам «стандартных» опросов с применением методов вероятностного отбора. Методология применения RDS в онлайн исследованиях находится на ранней стадии разработки, а особенности использования RDS для повышения валидности данных онлайн опросов труднодоступных совокупностей остаются малоисследованными.

Эмпирическая база диссертации основывается на методическом онлайн эксперименте, проведенным автором в мае-августе 2009 г. и нацеленным на апробацию и оценку применимости онлайн RDS метода для исследования редких совокупностей. Эксперимент был реализован на материале онлайн опроса индивидов, посещающих казино и/или играющих в онлайн казино. Эффективность и достоверность полученных результатов сравнивалась с оценками, полученными в исследовании ФОМа 2006 года с применением случайной маршрутной выборки.

С учетом поставленных задач мы можем выделить четыре *группы источников*, релевантных для нашего исследования и отражающих глубину проработки темы.

В первую группу входят методологические работы, затрагивающие основы веб-исследований. Это, прежде всего, теоретические обобщения М. Купера, а также Д. Дилмана.

Вторую группу релевантных текстов образуют эмпирические работы, обозначающие преимущества и недостатки различных типов онлайн исследований, а также оценивающих валидность и надежность получаемых данных.

Третья группа источников – это теоретические работы Д. Хекаторна по RDS методологии и эмпирические опросы, проведенные с применением этого

метода. Здесь также важно отметить работы основателя сетевого подхода Дж. Коулмана, впервые описавшего применение метода снежного кома.

И, наконец, четвертая группа релевантных текстов – базовые работы для методологии социологических исследований в целом, а именно, работы П. Лазарсфельда, институционализировавшего практику проведения эмпирического исследования в социологии, Л.Киша, разработавшего методы формирования выборки, и Д.Кэмпбелла, предложившего критерии оценки внутренней и внешней валидности экспериментальных данных.

Научная новизна исследования

Научная новизна диссертационного исследования отражена в следующих положениях.

1) Разработана типология веб-опросов, основанная на выделении концептуального объекта, или идеальной генеральной совокупности, на который можно экстраполировать полученные данные. В работе выделено пять типов веб-исследований, в которых концептуальный объект приравнивается (а) к реальному объекту; (б) к посетителям определенного Интернет-сайта; (с) к отдельным социальным группам; (d) к пользователям всемирной сети; (е) к населению, имеющему и не имеющему доступ в Интернет.

2) Выявлены угрозы внутренней валидности, вызванные случайной и систематической ошибкой измерения, а также угрозы внешней валидности, вызванные систематическими ошибками выборки, охвата и «неответов».

3) Обоснованы принципы построения выборки и оценки параметров генеральной совокупности для всех типов веб-исследований, включающие в себя: (а) явное описание или реконструкцию концептуального объекта исследования, (б) задание соответствующей концептуальному объекту основы выборки или способа её адаптивного построения.

4) Показано, что проблема репрезентативности онлайн данных решается за счет нивелирования ошибки охвата. В работе впервые систематически

проанализированы существующие возможности построения репрезентативной вероятностной выборки посетителей определенных сайтов, Интернет пользователей и всего населения.

5) Продемонстрирована относительно невысокая эффективность применения расчета корректирующих оценок склонности (“propensity score”) с целью ремонта данных в онлайн опросах с применением невероятностного отбора.

6) Обоснована эффективность применения онлайн RDS методологии для исследования труднодоступных групп (на примере игроков в казино). Основываясь на сравнении полученных результатов эмпирического методического эксперимента, проведенного автором, с результатами опроса ФОМа с применением случайной маршрутной выборки, в работе выявлены преимущества и ограничения онлайн RDS опроса.

Основные положения, выносимые на защиту

1) В работе выделено пять типов веб-исследований, исходя из характерного для каждого типа концептуального объекта, в которых: (а) концептуальный объект приравнивается к реальным участникам опроса, (b) генеральную совокупность формируют посетители конкретного Интернет-сайта, (с) генеральную совокупность формируют заданные социальные группы, (d) концептуальный объект приравнивается к пользователям Всемирной сети, (е) генеральную совокупность формируют как пользователи сети, так и люди, не имеющие доступ в Интернет. Данная типология позволяет рассмотреть вопросы внешней и внутренней валидности, а также возможности повышения качества данных.

2) Преимуществами веб-исследований являются: (а) уменьшение организационных и административных издержек на проведение исследований, (b) элиминирование временных различий и географических границ, (с) сокращение сроков проведения полевых работ, (d) технологические

новшества, позволяющие улучшить инструментарий, (е) увеличение степени заполнения и консистентности собранной информации, (ф) уменьшение вероятности получения социально-одобряемых ответов, (г) возможность опроса труднодоступных групп.

3) Наиболее важными методологическими проблемами веб-исследований являются ошибка охвата и внешняя валидность онлайн данных. В работе впервые показано, что данная проблема безупречно решается в веб-исследованиях с применением случайной выборки, репрезентирующих как веб-пользователей, так и индивидов, не имеющих доступ в Интернет. Для этого должна быть создана национальная онлайн панель, основанная на вероятностном отборе с использованием оффлайн основы выборки.

4) Интернет, с одной стороны, открывает огромные возможности для проведения опросов труднодоступных групп, с другой стороны, открывает перспективы для развития сетевых выборок. В частности, онлайн RDS методология позволяет решить проблему «доступа в поле» посредством (а) преодоления «географического пространства» и достижения места дислокации объекта исследования, (б) увеличения доверия потенциальных участников к исследованию. Это уменьшает материальные и нематериальные издержки на проведение и организацию опроса, а также значительно сокращает сроки проведения полевых работ.

5) Сравнивая результаты онлайн RDS опроса посетителей казино, проведенного автором, с результатами опроса ФОМа, в работе показана эффективность и точность полученных оценок параметров генеральной совокупности по социально-демографическим переменным. Однако методология имеет и ряд ограничений, которые выделяются и рассматриваются в исследовании.

Апробация результатов исследования

Основные положения диссертации были апробированы в трёх научных публикациях автора и представлены в докладах на следующих конференциях: «Современное Российское Общество и Социология: Семинар, посвященный памяти Юрия Левады» (Доклады на тему «Типология веб-исследований» и «Ошибка неответов в онлайн-исследованиях»; Москва, 2008-2009), «III Всероссийский социологический конгресс» (Доклад на тему «Типология веб-исследований: Новые подходы к сбору и анализу данных»; Москва 2009), «Современные проблемы формирования методного арсенала социолога (Доклад на тему «Опрос труднодостижимых групп через Интернет: выборка, управляемая респондентами»; Москва, 2010).

Материалы диссертационного исследования использовались автором при проведении семинарских занятий по курсу «Онлайн исследования в России» в Государственном Университете – Высшей Школе Экономики (2008-2010 гг.).

Теоретическая и практическая значимость исследования

Представленные в диссертации результаты могут быть использованы для дальнейших исследований в области методологии социологических веб-исследований, опроса труднодоступных социальных групп, а также в области исследований гемблинга. Приводимые в работе данные о внешней валидности различных типов выборки в веб-исследованиях могут быть использованы для научного обоснования необходимости и практического проектирования национальной онлайн панели, основанной на вероятностном отборе с использованием оффлайн основы выборки.

Представленный в работе теоретический и практический материал может использоваться при разработке лекционных курсов по методологии социологических исследований, онлайн опросам и методам построения выборки.

Логика и структура работы

Диссертационная работа состоит из введения, трех глав, заключения и библиографии. Первая глава «Типология количественных веб-исследований» нацелена на решение задачи типологизации веб-исследований. Вторая глава «Построение выборочной совокупности в веб-исследованиях» посвящена вопросам построения выборочной совокупности и возможностям оценки параметров генеральной совокупности. Третья глава «Онлайн опрос посетителей казино: опыт построения и оценки RDS выборки» выявляет научный потенциал применения онлайн RDS методологии для опроса труднодоступных групп. Объем работы составляет 169 страниц.

Основное содержание работы

Во *Введении* обосновывается актуальность и степень разработанности заявленной темы, формулируются цель и задачи диссертационного исследования.

Основная задача *Главы I «Типология количественных веб-исследований»* – каталогизация существующих подходов к проведению количественных веб-исследований. В главе приводится классификация М. Купера, выделяющего веб-опросы на основе типа выборки, и предложена авторская типология веб-опросов, исходя из концептуального объекта исследования – идеальной генеральной совокупности. Предлагаемая типология позволяет детально рассмотреть методологические основания каждого типа веб-исследования, выявить специфику формирования выборок и оценить возможности корректировки эмпирических данных.

Автором выделено и рассмотрено пять типов веб-исследований, в которых концептуальный объект приравнивается (1) к реальным участникам опроса, (2) к посетителям конкретного Интернет-сайта, (3) к определенным социальным группам, (4) к пользователям Всемирной сети, (5) к населению, имеющему и не имеющему доступ в Интернет.

В первом параграфе «Концептуальный объект = участники опроса» рассматриваются исследования, главной особенностью которых является знание, желание и готовность респондентов ответить на вопросы анкеты. Механизмы отбора респондентов на каждом этапе не определены и не контролируемы: подобные опросы подвержены ошибкам выборки, охвата, неответов, а также взвешивания и корректировки данных.

Второй параграф «Концептуальный объект = посетители веб-сайтов» посвящен веб-исследованиям, результаты которых можно обобщить на пользователей определенных сайтов. Отличительной характеристикой данного типа исследований является отбор посетителей сайта с использованием принципа систематической выборки, применяемой в опросах избирателей на выходе в день выборов. Основа выборки сужается до посетителей Интернет-сайта, и, таким образом, элиминируется проблема охвата.

В третьем параграфе «Концептуальный объект = социальные группы» анализируются веб-исследования, результаты которых можно экстраполировать на труднодоступные группы. Выявлено, что участники онлайн интервью с большей готовностью делятся своим опытом на чувствительную тему по сравнению с оффлайн опросами. Таким образом, Интернет признается социальными учеными легитимным источником рекрутирования и площадкой для опроса труднодоступных групп населения.

Четвертый параграф «Концептуальный объект = пользователи Интернета» посвящен веб-исследованиям, результаты которых можно обобщить на пользователей Всемирной сети. Один из способов построения случайной выборки пользователей сети – случайный отбор классическим способом (по базе телефонных или почтовых адресов) и опрос только тех респондентов, которые имеют доступ в Интернет. Ошибка неответов – основная проблема данного типа исследований, которая может происходить на нескольких этапах рекрутирования. Участниками опроса становятся люди, которые ответили на телефонный звонок, подтвердили, что у них есть доступ в

Интернет, предоставили свой электронный адрес, согласились стать участником панели и зарегистрировались в ней после получения приглашения по электронной почте. Подобные исследования позволяют проводить опросы, репрезентирующие Интернет пользователей в целом.

Глава завершается *пятым параграфом* «Концептуальный объект = население», в котором анализируются онлайн опросы, позволяющие строить репрезентативные выборки не только пользователей сети, но и индивидов, не имеющих доступ в Интернет. Онлайн панель рекрутируется на основе случайного телефонного или почтового опроса. Независимо от того, имеется в домохозяйстве компьютер и доступ в Интернет или нет, индивидам предлагают зарегистрироваться в панели. В случае отсутствия компьютера и выхода в Интернет потенциальным респондентам предоставляется необходимое оборудование. Таким образом, снимается основная проблема веб-опросов – ошибка охвата.

Основные проблемы, с которыми сталкиваются данные исследования: (1) ошибка неотчетов; (2) «осыпание» панели (или «истощение» панели), которое может привести к нерепрезентативности панели; (3) влияние эффекта участия в панели (или эффект «созревания» участников панели) – вероятность изменения выражаемого мнения и поведения респондентов вследствие участия в панельном исследовании.

Несмотря на возможные ошибки подобных исследований, в работе делается вывод о том, что это единственное на сегодняшний день решение методологических проблем веб-исследований и единственный способ, позволяющий конструировать случайную выборку всего населения в Интернете.

Вторая глава диссертационной работы «*Построение выборочной совокупности в веб-исследованиях*» посвящена методам построения выборок в онлайн опросах и методике корректировки полученных данных. В *первом параграфе* «Концептуальный объект = участники опроса» рассматриваются

методы построения выборочной совокупности и оценки параметров в онлайн опросах первого типа.

Каждый Интернет-пользователь k имеет неизвестную вероятность отбора p_k (где $k = 1, 2, \dots, N_I$). Ожидаемое значение вероятности p_k называют склонностью к участию индивида k в онлайн опросе. Если у каждого пользователя были бы равные шансы отбора, то в качестве оценки значения переменной по всем пользователям сети выступало простое среднее значение, однако в случае неравных шансов отбора среднее значение параметра можно оценить по формуле Горвитца-Томпсона:

$$E(\bar{y}_I) \approx \bar{Y}_I^* = \frac{1}{N_I p} \sum_{k=1}^{N_I} p_k I_k Y_k,$$

где $I_k = 1$ при наличии доступа в Интернет у k -го индивида, $N_I = \sum_{k=1}^{N_I} I_k$.

Вероятность отбора респондентов можно оценить апостериорно. Результаты данного типа опросов в случае их экстраполяции на совокупность Интернет-пользователей подвержены ошибкам охвата, выборки, ответов и измерения. Исключение составляет ситуация равных значений коэффициентов склонности к участию в опросе для всех k , когда допускается, что пропуски в данных являются полностью случайными.

Для апостериорной оценки вероятности участия в онлайн опросах первого типа, а также возможности экстраполяции полученных результатов на все население, используется метод расчета корректирующих оценок склонности. В этом методе статистического перевзвешивания в качестве контрольной группы выступает случайная выборка населения или выборка Интернет-пользователей. В работе показана невысокая эффективность применения данного метода перевзвешивания.

Второй параграф «Концептуальный объект = посетители веб-сайтов» посвящен вопросу формирования выборки и оценки параметров генеральной совокупности, когда в качестве концептуального объекта исследователь

выделяет посетителей веб-сайтов. Предположим, что исследователь должен опросить посетителей h сайтов с применением случайного систематического отбора, где h ($h=1, \dots, H$) – количество веб-сайтов. Общее число N посетителей в совокупности равно сумме посетителей на всех исследуемых сайтах $N = \sum_{h=1}^H N_h$.

Тогда среднее значение параметра по h сайтам можно оценить следующим образом:

$$\bar{Y} = \sum_{h=1}^H W_h \bar{Y}_h = \sum_{h=1}^H W_h \frac{1}{N_h} \sum_{k=1}^{N_h} a_k Y_k,$$

где $a_k = 1$, если элемент k принял участие в опросе, $a_k = 0$ – если k не принял участие, $W_h = \frac{N_h}{N}$ – доля сайта h , \bar{Y}_h – оценка среднего значения параметра Y на сайте h .

В случае учета отсутствия наблюдений по ряду индивидов ошибка среднего составит:

$$M(\bar{y}_i) = E(\bar{y}_i) - \bar{Y}_i \approx \bar{Y}_i' - \bar{Y}_i = \frac{C(p, Y)}{p},$$

где $C(p, Y)$ – коэффициент ковариации между значением переменной и коэффициентом склонности Интернет-пользователей к участию в опросе.

Третий параграф «Концептуальный объект = труднодоступные группы: выборка, управляемая респондентами» посвящен построению онлайн RDS выборки для опроса труднодоступных групп. RDS методология – это разновидность сетевой выборки, позволяющей построить выборку, независимую от изначальных субъектов, с которых начался опрос. В результате финальная выборка не зависит от того, были ли отобраны первые респонденты случайным образом или нет. Наиболее значимый результат для исследователей в данном случае – это возможность применить к полученным данным статистические методы расчетов, измерить ошибку выборки, вычислить доверительный интервал, и, таким образом, экстраполировать результаты на генеральную совокупность.

Для организации RDS опроса на первом этапе исследователь выбирает первых респондентов, которые по цепочке рекрутируют остальных участников опроса. Респондентам платят вознаграждение за участие в исследовании, а также за рекрутирование каждого нового участника. Каждому респонденту можно рекрутировать ограниченное количество участников: введенная квота позволяет увеличить количество волн, необходимых для достижения точки равновесия и получения стабильных характеристик выборки. Длительность волн, исходящих от изначально выбранных семян, позволяет получить распределение, не зависящее от характеристик первоначальных респондентов, и теоретически вероятностную выборку, когда каждый член исследуемой группы имеет шанс попасть в выборку.

Каким образом можно сделать статистические выводы о генеральной совокупности? Базовым предположением является представление цепочки опрашиваемых в виде марковской цепи. Это дает возможность предположить, что смещения, которые могли быть вызваны изначально построенной выборкой, должны быть нивелированы при последующих волнах, таким образом, что финальная выборочная совокупность не зависит от выбранных исследователям «семян». Для оценки параметров генеральной совокупности и построения доверительного интервала используется оценочная процедура бутстрэппинга, в которой симулируется процедура выборки.

При этом оценить параметр генеральной совокупности можно по следующей формуле: $\hat{p}_o^{RDS} = \frac{\hat{s}_{ba}\hat{D}_b}{\hat{s}_{ba}\hat{D}_b + \hat{s}_{ab}\hat{D}_a}$ (RDS 1), где \hat{s}_{ab} - сглаженная доля респондентов из группы А, которая рекрутировала респондентов из группы В, \hat{D}_a - средняя оценка размера социальной сети среди членов группы А.

Первое онлайн исследование с применением метода RDS было проведено в 2004 году Д. Хекаторном и С. Вейнертом. Исследователями предполагалось опросить 150 студентов. Проведение поля заняло всего 72 часа. В 2008 году С. Вейнерт проводит второе онлайн исследование, применяя данный метод

опроса и опрашивая тот же объект, что и в 2004 году, а именно, студентов. В данном исследовании было опрошено 369 респондентов за шесть недель. Онлайн исследования показали довольно высокую эффективность полученных оценок параметров генеральной совокупности.

В четвертом параграфе «Концептуальный объект = пользователи Интернета» рассматривается вопрос формирования выборки в веб-исследованиях со случайной выборкой Интернет-пользователей. Оценить среднее значение параметра генеральной совокупности можно по формуле Горвитца-Томпсона:

$$E(\bar{Y}_{HT}) = \frac{1}{N_I} \sum_{k=1}^N a_k I_k \frac{Y_k}{\pi_k}, \quad \text{где } \pi_k = E(a_k) - \text{математическое ожидание}$$

вероятности отбора элемента k . В случае простой случайной выборки $\pi_k = \frac{n}{N_I}$,

тогда:

$$\bar{Y}_I = \frac{1}{n} \sum_{k=1}^N a_k I_k Y_k$$

Если экстраполировать результаты на все население, то ошибка выборки составит:

$$B_{\bar{Y}_{HT}} = E(\bar{Y}_{HT}) - \bar{Y} = \bar{Y}_I - \bar{Y} = \frac{N_{NI}}{N} (\bar{Y}_I - \bar{Y}_{NI}).$$

Вторая глава завершается *пятым параграфом* «Концептуальный объект = население», в котором рассматриваются принципы построения выборки в веб-исследованиях со случайным отбором, репрезентирующих все население. В репрезентативных онлайн панелях рассчитываются коэффициенты результативности онлайн опросов по аналогии с коэффициентами, разработанными AAPOR (The American Association for Public Opinion Research – Американская ассоциация исследователей общественного мнения) для телефонных интервью, т.к. наибольшая угроза в данном случае – ошибка неотчетов.

Основываясь на коэффициенте рекрутирования индивида на стадии телефонного дозвона (P_{RECR}) и коэффициенте ответов на стадии заполнения профильной анкеты ($PROR$), а также учитывая уровень удержания панели ($RETR$) и процент респондентов, приглашенных участвовать в опросе и заполнивших анкету до конца ($COMR$), *кумулятивный коэффициент ответов* ($CURR$) в онлайн опросе можно рассчитать по следующей формуле:

$$CURR = P_{RECR} * PROR * RETR * COMR$$

Вычислить стандартную ошибку неответов в подобном онлайн опросе можно по следующей формуле:

$$se(\hat{y}_{ur} - y_f) = se(\hat{y}_{un}) = \sqrt{\text{var}\left[\frac{n-r}{n}(\hat{y}_{ur} - y_{un})\right]} = \left(\frac{n-r}{n}\right)\sqrt{[\text{var}(\hat{y}_{ur}) + \text{var}(\hat{y}_{un})]},$$

$$\text{где } y_f = \frac{r}{n}\hat{y}_{ur} + \frac{n-r}{n}\hat{y}_{un}$$

y_f - ответы, которые бы дали респонденты и панелисты, не откликнувшиеся на приглашение принять участие в опросе, \hat{y}_{ur} - оценка переменной на основе ответов респондентов, \hat{y}_{un} - оценка переменной индивидов, не принявших участие в опросе, n - количество респондентов и панелистов, не принявших участие в опросе (количество разосланных приглашений), r - количество респондентов.

Третья глава диссертационной работы «*Онлайн опрос посетителей казино: опыт построения и оценки RDS выборки*» посвящена проведенному автором методическому эксперименту – онлайн RDS опросу посетителей казино (май-август 2009 года). Игроки в казино являются, с одной стороны, закрытой группой с большим количеством внутригрупповых связей, с другой – труднодоступной группой для опроса. Методология RDS позволяет опросить объект исследования и количественно оценить параметры генеральной совокупности. Отметим, что опрос выпал на «переломное» время для игорных заведений: 1 июля 2009 года в России были закрыты все казино, располагающиеся вне отведенных законом четырех игровых зон. На

закрытость, сплоченность и плотность социальных связей в группе игроков указывает тот факт, что после закрытия казино многие игроки стали играть «на катранах», или, другими словами, подпольно на квартирах, куда возможно попасть только по личной рекомендации нескольких игроков.

Целью проведения методического эксперимента является апробация и оценка применимости, а также изучение особенностей использования метода онлайн RDS для исследования редких совокупностей.

Эксперимент апробирован на материале онлайн опроса индивидов, посещающих казино и/или играющих в онлайн казино. Всего в исследовании было опрошено 99 респондентов, 4 из которых выступили в качестве «первичных» участников. Четырем участникам были высланы ссылки на онлайн анкету, после заполнения которой они могли также выслать её 6 новым респондентам. Далее новым респондентам предлагалось заполнить анкету и разослать её 6 другим участникам – игрокам, посещающим казино, или индивидам, играющим в онлайн казино. Отметим, что как за участие, так и за приведение новых респондентов организаторы не выплачивали вознаграждение. Скорее всего, это повлияло как на сроки проведения полевых работ (они длились 4 месяца), так и небольшой финальный объем выборки.

География опроса: г. Москва.

В первом и втором параграфах, «Модели анализа гемблинга» и «Методы сбора данных», представлены имеющиеся паттерны анализа поведения игроков и методы сбора данных при изучении азартного поведения. Для опроса игроков, имеющих склонность к азартным играм, и, в частности, играющих в казино, используют неслучайную выборку. Это объясняется труднодоступностью объекта и отсутствием основы выборки. Наиболее часто для опроса игроков используется выборка доступных случаев, а также стихийная выборка.

В третьем параграфе «Результаты методического эксперимента» представлены результаты проведенного опроса. Используя программу RDSAT

v.6.0, разработанную Э.Вольцем для анализа RDS данных¹², автором получены следующие оценки параметров генеральной совокупности: доля мужчин среди игроков в казино в Москве составляет 0.73, доля женщин – 0.26; доля людей в возрасте до 35 лет – 0.70, в возрасте 35 лет и более – 0.30; доля людей, играющих в обычном казино – 0.47, в Интернет казино – 0.28, играющих как в обычном, так и онлайн казино – 0.25.

Как возможно измерить эффективность применения онлайн RDS в отношении посетителей казино? Следует сравнить полученные оценки с оценками параметров генеральной совокупности в исследовании с применением случайной выборки. Предположим, что изменения в численности аудитории казино в Москве за последние четыре года не повлияли на половозрастные пропорции посетителей казино, тогда есть возможность сравнить полученные оценки параметров генеральной совокупности с оценками исследования ФОМа 2006 года¹³. Имеются два важных отличия между данным исследованием и опросом ФОМа, которые следует учитывать:

- (1) ФОМ не включал в оценку индивидов, играющих в онлайн казино,
- (2) методология данного опроса - онлайн RDS, значит, в выборку могли попасть только респонденты, пользующиеся всемирной сетью.

По данным ФОМа, доля мужчин среди игроков казино в Москве – 0.73, доля женщин – 0.27. В нашем исследовании оценка доли мужчин среди игроков казино (в обычном и онлайн казино) получилась такими же. Следовательно, оценки доли параметров генеральной совокупности по переменной «пол» совпали с оценками ФОМа. Что касается возрастных оценок игроков, то в онлайн RDS оценка возрастной группы до 35 лет получилась выше – 0.70, в то время как, по оценкам ФОМа, доля данной возрастной группы составляет 0.60. Различие с оценкой ФОМа составляет 0.10 и является статистически значимым. С одной стороны, такое различие можно объяснить неточной оценкой RDS

¹² <http://www.respondentdrivensampling.org/reports/RDSAT60.htm>

¹³ Массив данных по исследованию любезно предоставлен сотрудниками ФОМа.

опроса, с другой стороны – различием между объектом исследования ФОМа и методического эксперимента.

Основываясь на опыте проведения онлайн RDS опроса, автор выделяет следующие преимущества данной методологии:

1) Онлайн RDS опрос позволяет легче установить контакт с потенциальными участниками исследования. По Интернету можно охватить большее количество игроков: как респондентов, посещающих обычные и онлайн казино, так и играющих «на катранах», или подпольно в квартирах. Вследствие того, что ссылку на заполнение анкеты присылали игроки и просили других посетителей казино принять участие в исследовании, доверие к опросу было довольно высоким, что позволило опросить в Москве 99 респондентов без предоставления вознаграждения.

2) Проведение опроса с помощью онлайн RDS методологии позволило завершить полевые работы за четыре месяца – довольно короткий срок, учитывая труднодоступность группы, низкий уровень кооперации, отсутствие вознаграждения и «переломное» время для казино. Сокращение сроков на полевые работы можно объяснить тем, что исследователю не приходилось заниматься поиском игроков в казино, договариваться об организации интервью, а также встречаться с информантами для проведения опроса. Таким образом, онлайн RDS методология (по сравнению с оффлайн) позволяет сократить как материальные, так и нематериальные издержки на планирование и проведение опроса.

3) Определяя эффективность полученных RDS оценок, отметим, что оценки параметров генеральной совокупности по переменной «пол», полученной при помощи онлайн RDS, статистически не отличались от оценок, полученных в случайном опросе, проведенном ФОМом в 2006 году.

Преимущества онлайн RDS метода очевидны, но следует также очертить и ограничения его использования. Во-первых, несмотря на то, что онлайн RDS представляет одну из наиболее удобных форм принятия участия в

исследовании, когда респондент, получая ссылку от знакомого, заполняет анкету в любое удобное для него время, уровень кооперации в группе игроков казино остается на низком уровне. В связи с отсутствием вознаграждения и низким уровнем кооперации исследователь был вынужден регулярно писать по электронной почте или звонить респондентам, принявшим участие в опросе, с просьбой выслать ссылку или письмо с напоминанием знакомым, играющим в казино, для участия в онлайн опросе.

Во-вторых, в связи с тем, что для оценки параметров генеральной совокупности базовая методология требует введения информации о том, кто именно рекрутировал тех или иных респондентов, принявших участие в исследовании, методология не обеспечивает достаточно высокий уровень конфиденциальности личной информации участников опроса. Недоверие к исследованию также может означать и то, что в опросе будут с меньшей вероятностью принимать участие респонденты, заинтересованные скрывать свою деятельность.

В-третьих, автором не было подтверждено базовое допущение RDS о случайности рекрутирования участников по переменным «пол», «возраст» и «тип казино». Отметим, что вопрос корректности использования RDS оценок в опросе, в которых не подтверждаются базовые допущения, пока остается открытым.

В-четвертых, сто опрошенных, скорее всего, является недостаточной базой для статистических выводов об игроках казино в Москве. Напомним, что в RDS опросе объем выборки должен вдвое превышать объем выборки, приемлемый в опросе с использованием случайной выборки. Недостаточный объем выборки в опросе привел к тому, что автором не совсем корректно были оценены параметры генеральной совокупности по возрасту.

В *Заключении* диссертационной работы подводятся итоги исследования, обозначается круг поставленных проблем и предложенных решений. В работе было показано, что одна из сильных сторон использования веб-исследований –

проведение опросов труднодоступных групп. Проведенное нами исследование не претендует на исчерпывающую полноту. Возможные решения проблемы валидности и надежности данных в веб-исследованиях, новые подходы к формированию выборки, а также новаторские методы онлайн исследований и анализа многообразной информации в Интернете являются для автора предметом дальнейшей научной работы.

По теме диссертационного исследования автором опубликованы следующие работы в ведущих рецензируемых научных журналах, рекомендованных ВАК Министерства образования и науки РФ:

1. Мавлетова А.М. Социологические опросы в сети Интернет: возможности построения типологии // Социология: методология, методы, математическое моделирование. 2010. №31. С. 115-134 (объемом 1 п.л.).

Другие работы, опубликованные автором по теме диссертации:

2. Мавлетова А.М. Борьба за качество и надежность данных в онлайн исследованиях: основные результаты панельной конференции CASRO 2009 // Онлайн исследования в России 2.0 / Под ред. А.В. Шашкина, И.Ф. Девятко, С.Г. Давыдова. М: РИЦ «Северо-Восток», 2010. С. 43-61 (объемом 1 п.л.).
3. Мавлетова А.М. Казино: территория эмоциональности или расчета? // Социальная реальность. 2008, №3. С. 100-110. №4. С. 96-110 (объемом 1 п.л.).

Лицензия ЛР № 020832 от 15 октября 1993 г.
Подписано в печать 20 сентября 2010 г. Формат 60x84/16
Бумага офсетная. Печать офсетная.

Усл. печ. л. 1.

Тираж 100 экз. Заказ № *269*
Типография издательства ГУ-ВШЭ, 125319
г. Москва, Кочновский пр-д, д.3.