Межгосударственный статистический комитет
Содружества Независимых Государств

Хрестоматия
практико-ориентированного комплекса
учебно-методических материалов
по курсу
«Организация выборочных обследований»

Хрестоматия

практико-ориентированного комплекса учебно-методических материалов

по курсу

«Организация выборочных обследований»

Обзор системы социально-демографических обследований домашних хозяйств и населения России: состояние и перспективы развития

При планировании и осуществлении деятельности по развитию системы государственной статистики России Росстат исходит из цели обеспечения реализации прав граждан и организаций на доступ к статистической информации и руководствуется постановлением Правительства Российской Федерации от 12 февраля 2003 года № 98 «Об обеспечении доступа к информации о деятельности Правительства Российской Федерации и федеральных органов исполнительной власти».

Федеральная служба государственной статистики обеспечивает государство своевременной и исчерпывающей информацией о социальном, экономическом, демографическом и экологическом положении страны. Одним из важнейших источников статистической информации являются выборочные обследования домашних хозяйств и населения (ОДХ).

В настоящее время Росстат на регулярной основе (с месячной или квартальной периодичностью) проводит четыре таких обследования: обследование бюджетов домашних хозяйств и населения (ОБДХ), обследование населения по вопросам экономической активности, занятости и безработицы (ОЭАН), обследования личных подсобных хозяйств населения (ОЛПХ), обследования потребительских ожиданий населения (ОПОН).

Особую актуальность развитию системы сбора, анализа и распространения итогов выборочных обследований домашних хозяйств и населения придает тот факт, что Концепцией социально-демографического развития Российской Федерации предусмотрено создание целостной системы социально-демографических обследований населения, включающей в себя дополнительно к существующим еще 11 выборочных статистических обследований населения с различной периодичностью, а именно:

Условия жизни населения,

Репродуктивные планы населения,

Использование суточного фонда времени населением,

Поведенческие факторы, влияющие на состояние здоровья населения,

Рацион питания населения,

Доходы населения и участие в социальных программах,

Качество и доступность услуг в сферах образования, здравоохранения и социального обслуживания, содействия занятости населения.

Использование труда мигрантов,

Участие населения в непрерывном образовании,

Трудоустройство выпускников учреждений профессионального образования,

Микроперепись населения.

Ответственным за создание системы социально-демографических обследований населения определен Росстат.

Требование создания целостной системы социально-демографических обследований населения предполагает необходимость разработки концептуально-методологических основ ее построения. При этом должно быть обеспечено единство организационно-методологических подходов к проведению всех обследований, а также принципов автоматизации работ по их подготовке, проведению, обработке результатов обследований и распространению итогов. Принцип системности должен соблюдаться и в процессе формирования выборок для каждого обследования, и в обеспечении возможности «сквозного» анализа (причем в любой разрезности) характеристик целевых групп населения, получаемых по различным обследованиям, и в создании единой системы метаданных (включая справочники, классификаторы, модели описания статистических показателей, форматов обмена данными). Все разрабатываемые проектные решения в части автоматизации выборочных обследований не должны противоречить требованиям технического проекта на создание «Информационно-вычислительной системы Росстата».

Существующая система сбора, анализа и распространения итогов выборочных обследований домашних хозяйств и населения в целом обеспечивает реализацию основных функций, а именно:

- ввод и контроль анкет с ответами респондентов;
- редактирование и импутацию микроданных ОБДХ на основе программного продукта CANCEIS;
- получение регламентных таблиц ОДХ (в том числе на основе первичных данных, прошедших процедуры редактирования и импутации с использованием CANCEIS);
- создание на основе отредактированных массивов данных информационного фонда, содержащего индивидуальные характеристики домохозяйств и членов домохозяйств по ограниченному перечню показателей, определенных к размещению в системе открытого доступа на данном этапе;
- предоставление ключевых метаданных в части описания методологии организации и проведения ОДХ.

В то же время существующая система сбора, анализа и распространения итогов выборочных обследований домашних хозяйств и населения уже не устраивает в полной мере ни пользователей статистической информации, ни сам Росстат. В настоящее время Росстатом поставлена цель: создание максимально автоматизированной «фабрики» обследований.

Так, «массовому» пользователю нужно обеспечить удобство поиска нужной информации (нужного показателя) и ее понимания, а также повысить удобство (наглядность) представления и анализа информации.

«Продвинутые» пользователи требуют, во-первых, подробного описания статистической информации (в том числе в целях понимания ее сопоставимости) и метаданных, а во-вторых, удобного доступа ко всем информационным ресурсам Росстата, включая микроданные. При этом необходимо обеспечить безусловную конфиденциальность первичных данных.

Со стороны федеральных органов исполнительной власти – получателей статистической информации регулярно поступают просьбы срочно внести те или иные

изменения в программу и методологию обследования в целях расширения перечня наблюдаемых показателей или повышения их достоверности.

Жесткие бюджетные и кадровые ограничения в системе Росстата требуют существенного сокращения трудоемкости и стоимости на всех этапах работ, прежде всего при подготовке обследования (планирования выборки, проектировании бланков обследования, формирования экономического описания постановки задачи, описания метаданных), ведения нормативно-справочной информации, при разработке программного обеспечения, вводе данных, их копировании, контроле, анализе и распространении.

Наименее формализованным этапом в настоящее время является этап планирования обследования, и особенно процедура формирования выборки. Особую методологическую сложность эти работы приобретают при переходе от планирования отдельного обследования к системе различных обследований, причем различных и по тематике, и по срокам и периодичности, и по размерам выборки, и по типам обследований (лонгидюдные или параллельные). Таким образом, решение методологических проблем планирования системы социально-демографических обследований является для Росстата исключительно актуальной задачей.

Еще одним направлением работы Росстата является развитие информационной системы открытого доступа к итогам обследования домашних хозяйств на основе публикации баз микроданных с пользовательским интерфейсом, позволяющим аналитикам производить дополнительный анализ полученных итогов ОДХ собственными силами.

В настоящее время реализован первый этап этой системы, который обеспечивает:

пользовательский доступ к метаданным, микро и макроданным нескольких ОБДХ через иерархический многоуровневый рубрикатор.

управление правами доступа пользователей;

функции администрирования.

Система открытого доступа ОБДХ в настоящее время предоставляет внешним пользователям возможность просматривать метаинформацию, микро- и макроданные и загружать их на свой компьютер для осуществления анализа данных и формирования дополнительных агрегированных итогов. На Интернет-сайте Росстата в системе открытого доступа ОБДХ представлен ограниченный перечень показателей (квартальные и годовые): микроданные - в форматах dBASE и SPSS, макроданные – в формате Excel.

Файлы информационного фонда по ОБДХ содержат индивидуальные характеристики домашних хозяйств и членов домохозяйств, полученные в результате обследования (в системе открытого доступа предоставлена информация в соответствии с перечнем показателей, определенных к размещению).

Для дальнейшего развития системы с учетом описанных выше проблем в настоящее время разрабатывается частное техническое задание на развитие системы сбора, анализа и распространения итогов выборочных обследований домашних хозяйств и населения. Система будет создаваться в ближайшие 3-4 года.

1. Методы выборочных обследований.

1.1. Введение в теорию и практику выборочных обследований.

В современных условиях существует значительная потребность в социально-экономической информации о качественно определенных массовых явлениях и процессах общественной жизни, характеризующей их количественно. Такого рода информация может использоваться для эффективного менеджмента и маркетинга в бизнесе, государственного управления и планирования в социальной и других сферах. Количественная информация об определенных социально-экономических множествах элементов (или единиц), каждое из которых составляет *генеральную совокупность* - объект исследования, называется ***статистическими данными*** и является результатом проведения обследований и опросов.

Чтобы удовлетворить потребности общества в социально-экономической информации, специализированные государственные и негосударственные организации регулярно собирают необходимые сведения – *первичные данные*, характеризующие единицы целевых генеральных совокупностей. На основе собранных первичных данных рассчитываются параметры исследуемых совокупностей, такие как средние и суммарные величины, их отношения, а также характеристики структуры и распределения единиц по значениям варьирующих признаков.

Например, могут собираться сведения о размере и структуре бюджетов домашних хозяйств регионов страны, о видах и объемах производимой продукции предприятиями районов или населенных пунктов, о занятости и безработице экономически активной части населения страны, о среднем количестве отрабатываемых за плату часов в неделю наемными работниками в зависимости от рода их занятий, о спросе и предложении на различные виды товаров и услуг.

Таким образом, *обследование* как одна из важнейших стадий статистического исследования заключается в планомерном, научно организованном и, как правило, систематическом сборе данных о явлениях и процессах общественной жизни путем регистрации заранее намеченных существенных признаков с целью получения в дальнейшем обобщающих характеристик этих явлений и процессов.

Базовым принципом получения исходной информации в статистике является метод массовых наблюдений, однако часто несплошных. Закономерен вопрос: на чем основана уверенность в надежности получаемых результатов? Обоснованием этого являются фундаментальные законы теории вероятностей и математической статистики, известные как Закон больших чисел и Центральная предельная теорема.

Смысл закона больших чисел заключается в том, что при осреднении большого числа (n) случайных слагаемых все менее ощущается характерный для случайных величин неконтролируемый разброс в их значениях. Так что в пределе при $n \to \infty$ этот разброс исчезает вовсе или, как принято говорить, *случайная величина вырождается в неслучайную.* Однако при любом конечном числе слагаемых n случайный разброс у среднего арифметического этих слагаемых остается. Поэтому возникает вопрос о характере этого разброса при $n \to \infty$.

Ответ дает «центральная предельная теорема», которая заключается в том, что для широкого класса независимых случайных величин предельный (при $n \to \infty$) закон распределения их нормированной суммы вне *зависимости от типа распределения* слагаемых стремится к нормальному закону распределения.

Заметим, что существует несколько вариантов точных формулировок центральной предельной теоремы, отличающихся друг от друга степенью общности и видом постулируемых ограничений, в том числе имеется формулировка этой теоремы и для конечной генеральной совокупности.

Статистическое исследование представляет собой процесс изучения с целью получения количественной характеристики социально-экономических явлений на основе математико-статистических методов и систем статистических показателей. Последние разработаны для описания конкретных исследуемых совокупностей и представляют собой стандартные наборы параметров (характеристик), отвечающих цели решаемых в реальных условиях практических задач исследования.

Когда речь идет об обследовании, и неспециалист и профессионал скорее всего подразумевают выборку, которую по всей видимости следует рассматривать в качестве базовой составляющей процесса проведения обследования. В идеальной ситуации конечно можно получить представляющие интерес первичные сведения от каждого респондента, относящегося к исследуемой генеральной совокупности. И когда изучаются малочисленные группы, мы в состоянии опросить каждого представителя, как возможно и следует поступать. Но поскольку в большинстве случаев проводящие обследования заинтересованы в изучении многочисленных совокупностей объектов, организаций и/или индивидов, выборка становится необходимой как технология, сберегающая время и другие ресурсы.

Проведение обследований в целях исследования генеральных совокупностей является одновременно и искусством и наукой. Все что связано с выборкой – это в большинстве своем научная составляющая, т.е. имеется обусловленное множество методов, техник и уравнений, которые гарантируют определенный (вероятностный) результат, если им следовать. Приняв это, следует создавать собственную выборку «по книге», на сколько это возможно. Поступая так получаем большую уверенность в результатах собственных исследований.

Начнем знакомство с выборочными методами с определения некоторых базовых терминов.

Выборка – это любое подмножество элементов нечто большего. Выборка может быть отобрана с помощью вероятностного метода, а может и нет.

Выборка отбирается из *генеральной совокупности*, определяемой как множество всех элементов (единиц), которые мы хотим изучить. Наиболее часто элементы генеральной совокупности – домашние хозяйства, индивиды или организации (юридические лица), а также могут быть многим другим, что возможно строго определить. Итак, генеральная совокупность должна быть строго и ясно определена (ограничена рамками исследования) прежде, чем может быть осуществлен отбор выборки. Так генеральной совокупностью могут быть:

- все организации, постоянные клиенты некоторой компании, воспользовавшиеся ее услугами в прошлом полугодии;

- взрослые, достигшие 18 лет и старше, проживающие в домашних хозяйствах на территории Российской Федерации. В этом случае исключаются дети и часть взрослых в институциональных заведениях (например, в тюрьмах и психиатрических лечебницах);

- наемные работники организаций розничной торговли в некотором субъекте РФ;

- все пациенты больницы, прошедшие определенный курс лечения;

- сельхозугодия (гектары) некоторого региона, отведенные под посевы яровой пшеницы.

Обычно определить рамки генеральной совокупности исходя из цели исследования не составляет большой сложности. Однако это важный первый шаг. Без аккуратного его выполнения не возможно будет точно узнать к какому массиву элементов применимы полученные результаты исследования.

Основа выборки - это список элементов или единиц, относящихся к целевой генеральной совокупности, который фактически имеется у организатора. Этот перечень объектов генеральной совокупности непосредственно используется для отбора выборки. таким образом, генеральная совокупность является абстрактным объектом исследования, в то время как основа выборки – это, например, компьютерный файл или перечень элементов на бумаге, относящихся к генеральной совокупности, с сопутствующей базовой информацией по каждому элементу. Например, базовой информацией могут быть контактные сведения (адрес, телефон и т.д.).

Также, как необходимо ясно и точно определять границы генеральной совокупности, нужно понимать как основа выборки соотносится с реально существующей генеральной совокупностью: исчерпывает ли основа выборки все элементы генеральной совокупности. А может быть имеются недостающие или, наоборот, присутствуют лишние элементы?

Вероятностная (случайная) выборка. В определении выборки, которое было дано выше, не использовалось понятие вероятности. Причина этого заключается в том, что способов формирования выборки очень много, и только в некоторых из них присутствует элемент случайности. *Вероятностная выборка* – это такой способ формирования выборки, при котором каждый элемент генеральной совокупности имеет известный неравный нулю шанс оказаться включенным в выборку. Существует несколько типов вероятностных выборок, наиболее простой из которых – *простая (или собственно) случайная выборка*. Важность вероятностных выборок состоит в том, что они позволяют рассчитать точность или, другими словами, ошибку выборки для статистик (функций), вычисленных на основе данных выборки. Также они позволяют чувствовать себя уверенными, делая статистически значимые выводы о генеральной совокупности на основе выборочных результатов. Для всех же остальных выборок какими бы аккуратными ни представлялись выборочные результаты научно обоснованно их нельзя применять ни к какой-либо бо́льшей группе объектов.

Неслучайная выборка. *Неслучайная выборка* - это выборка, для которой невозможно рассчитать вероятности включения в выборку для всех элементов генеральной совокупности. Статистическую теорию обоснованно невозможно применить для анализа данных не вероятностной выборки, то есть статистические критерии, рассчитываемые для проверки гипотез по данным не вероятностной выборки, не имеют никакого смысла. Другими словами, неслучайная выборка дает информацию только об элементах, включенных в выборку, но ни о какой бо́льшей (генеральной) совокупности, из которой выборка была отобрана.

Имеется много разновидностей неслучайных выборок, в том числе выборка методом «снежного кома», удобная, целевая и квотная выборки, которые кратко обсудим ниже. Конечно же, существуют ситуации, когда использование не вероятностных выборок оправдано, например, при проведении фокус-групп практически всегда используются не вероятностные выборки. Формирование фокус-групп ориентировано на то, чтобы детально изучить мнения различных представителей генеральной совокупности. Но поскольку размер фокус-групп обычно невелик, использование вероятностных выборок для отбора участников может не позволить отобрать индивидов, представляющих все многообразие жизненного опыта. Поэтому, чтобы гарантировать разнообразие мнений, выбирать участников лучше целенаправленно, основываясь на их характеристиках.

Тем не менее, использование не вероятностных выборок ограничивается специальными ситуациями. Обычно их недостатки перевешивают преимущества, такие как удобство и экономичность.

Виды неслучайной выборки. Кратко рассмотрим не вероятностные выборки различных видов, наиболее часто применяемых при проведении обследований. Если есть возможность выбора, то всегда лучше остановиться на одной из вероятностных выборок (вероятностные выборки рассматриваются далее). Однако бывают ситуации, когда набор возможных вариантов ограничен и/или уровень точности, требующейся от результатов выборки, меньше обычного. В таких случаях можно рассматривать возможность использования не вероятностной выборки. К этим случаям относятся предварительное тестирование, разведочные исследования, а также ситуации, в которых невозможно создать основу выборки.

Ниже описаны четыре наиболее распространенных типа неслучайной выборки.

Удобные выборки. Эти выборки формируются исходя из удобства исследователя. Интервью, взятые репортерами у прохожих на улице, основываются на удобных выборках, также как и интервью с учениками одного класса, и опрос первых 20 человек, посетивших концерт. Нет никакой гарантии, что удобная выборка представляет целевую генеральную совокупность. Удобные выборки могут использоваться для проведения пилотных обследований, для предварительного тестирования вопросников и для проведения фокус групп.

Квотные выборки. При квотном плане выборки генеральная совокупность разбивается на важные для целей исследования подгруппы, а затем для каждой подгруппы устанавливается квота на количество интервью. Квоты обычно устанавливаются в соответствии с данными генеральной совокупности или в соответствии с нужными характеристиками (например, равное количество мужчин и женщин). Квотные выборки были введены в практику для преодоления общей проблемы, состоящей в том, что не вероятностные выборки обычно не соответствуют генеральной совокупности по ключевым характеристикам. Для обеспечения установленной квоты на последнем этапе отбора интервьюерам предоставляется свобода выбирать респондентов по собственному усмотрению. Например, если интервьюеры уже опросили достаточно мужчин, чтобы обеспечить требуемую квоту, то опрос мужчин прекращается.

Существует мнение, что применение квотных выборок оправдано, так как они гарантируют, что труднодостижимые респонденты (например, мужчины моложе 25 лет) будут включены в выборку в пропорции, равной их пропорции в генеральной совокупности. Однако здесь могут возникнуть несколько проблем. Во-первых, поскольку интервьюеры могут сами выбирать респондентов, возможно возникновение смещения при

отборе, то есть характеристики респондентов могут влиять на их шансы быть отобранными. В отсутствии строгих инструкций (применение которых, в любом случае, было бы весьма проблематичным) интервьюеры могут, например, выбирать людей, проживающих в более привлекательных домах, более дружелюбных или находящихся дома, когда интервьюер звонит по телефону первый раз. Во-вторых, поскольку на последнем этапе формирования выборки не используется вероятностный отбор, то расчет стандартных ошибок статистик оказывается практически невозможным. В-третьих, оказывается замаскированной проблема неответов, поскольку при использовании квотной выборки респонденты, которые не хотят участвовать в обследовании или недоступны, заменяются другими респондентами, но эти замены не находят отражения в окончательной выборке.

Интересно, что из-за высоких расходов на формирование по-настоящему вероятностных выборок многие крупные негосударственные обследования проводились (а некоторые проводятся и до сих пор) с использованием квотных выборок. Эти проблемы не относится к телефонным опросам, поскольку формировать (сложные) вероятностные выборки для телефонных опросов достаточно легко.

Целевые/по суждению выборки. При использовании выборки этого типа исследователь выбирает каждый элемент выборки по своему усмотрению, основываясь на различных критериях важных в этом исследовании. Например, если изучаются городские организации, может понадобиться включить в выборку наиболее значимые для города (градообразующие) организации, независимо от того, были бы они включены в вероятностную выборку или нет. Чаще всего целевые выборки используются для небольших предварительных тестов, в которых важно протестировать вопросник на различных респондентах, представляющих различные мнения и жизненный опыт.

Выборки по принципу снежного кома/по рекомендации. Для труднодостижимых генеральных совокупностей может использоваться такая техника отбора, при которой сначала ведется поиск нескольких представителей генеральной совокупности, а затем у найденных представителей выясняется информация о других представителях генеральной совокупности. У последних, в свою очередь, также запрашивают сведения о других представителях генеральной совокупности и так далее. Понятно, что такая выборка не является вероятностной, но если требуется исследовать преступников, наркоманов, алкоголиков, богатых людей или любую другую генеральную совокупность, списки которой сложно или невозможно создать, такой подход может оказаться единственным шансом. Совершенно очевидно, что представители такой генеральной совокупности должны быть связаны друг с другом (должны знать друг друга). Выборки по принципу снежного кома наиболее эффективно применяются в пилотных исследованиях и при изучении небольших генеральных совокупностей.

Всегда следует помнить, что при отборе любой неслучайной выборки неизбежно возникает неустранимый фактор субъективности, являющийся основным недостатком таких выборок наряду с невозможностью получения статистически обоснованных выводов о генеральной совокупности.

Вероятностные (случайные) выборки.

Перейдем к рассмотрению наиболее часто используемых в обследованиях типов вероятностных выборок. Хотя специалисты в области выборок разработали множество

планов сложных выборок, почти все они основаны на планах выборок, рассматриваемых далее. Приступая к рассмотрению планов выборок, вспомним, что вероятностная выборка вовсе не обязательно подразумевает, что все элементы в основе выборки имеют равные шансы попасть в выборку. Вместо этого, вероятностная выборка подразумевает, что для каждого элемента мы знаем или можем вычислить вероятность попадания в выборку.

Характеристики эффективной выборки.

Что же требуется для того, чтобы сформированная выборка была бы эффективной, то есть такой, которая позволит с уверенностью делать выводы об исследуемой генеральной совокупности? Возможно, приведенные ниже характеристики выборки являются наиболее критичными с точки зрения эффективности.

1) Вероятностная выборка предпочтительнее. Вероятностная выборка позволяет корректно проверять статистические гипотезы, рассчитывая критерии, строить доверительные интервалы для статистик и делать выводы о генеральной совокупности в целом. Возникающие при этом проблемы связаны с размером ошибок выборки.

2) Выборка должна быть достаточно большого объема (многочисленной). Выборка слишком маленького объема не обеспечит подходящей статистической мощности, необходимой для обнаружения истинных различий между группами или выявления влияния одной переменной на другую. Хотя объем выборки может быть чрезмерно большим, что влечет неоправданные затраты ресурсов и времени, более распространенная проблема - выборки слишком маленького объема, не позволяющие обоснованно анализировать данные.

3) Выборка должна быть представительной. Для формирования выборки может быть использован один из вероятностных методов и ее объем может быть достаточным для обоснованных выводов, но если выборка непредставительная для исследуемой генеральной совокупности, то в ней заложены ошибки смещения. Следовательно такая выборка не будет эффективной. Смещение в итогах обследования происходит из-за того, что некоторых единиц слишком мало или слишком много в выборке. Это может происходить по различным причинам. Например, из-за допущенных ошибок при формировании выборки (то есть именно из-за того, как формировалась выборка) или из-за случаев недостижимости и неответов респондентов, когда отказавшиеся принять участие в обследовании отличаются от опрошенных. Указанные проблемы связаны с представительностью выборки в смысле ошибок охвата генеральной совокупности.

Таким образом, чтобы быть эффективными выборки должны обладать указанными тремя характеристиками.

Далее рассматриваются вопросы, связанные с формированием выборки, а проблема недостижимости и неответов респондентов здесь не рассматривается.

Процесс формирования выборки.

Процесс формирования выборки для проведения обследования состоит из нескольких этапов, включая:

- определение генеральной совокупности;
- создание основы выборки;
- выбор между вероятностными и не вероятностными методами отбора;

- определение плана выборки;
- определение объема выборки;
- непосредственное формирование выборки согласно плану.

Некоторые из указанных этапов представлены на схеме (см. рис.1) вместе с важными сопутствующими проблемами. Основа выборки – это фактически имеющийся у организатора обследования список элементов генеральной совокупности. Формируемая выборка, независимо от метода отбора, - это подмножество элементов основы выборки. В то время как окончательная выборка респондентов, ответивших на вопросы обследования, это, в свою очередь, - подмножество сформированной выборки.

Если процесс формирования выборки был выполнен корректно и не возникло существенных проблем с опросом респондентов, то собранные данные должны быть представительной выборкой для генеральной совокупности. Так за счет аккуратного применения этой методологии выборка объемом 2500 индивидов может достоверно представлять мнение всех граждан страны.

В процессе формирования выборки может возникнуть несколько проблем. Первая связана с ошибками охвата, то есть если основа выборки плохо представляет генеральную совокупность. Обычно эту проблему удается решить или, по крайней мере, учесть в итогах обследования, однако для этого необходимо знать о смещении. Например, если организатор собирается проводить опрос по электронной почте, то, очевидно, что основой выборки будет составленный список адресов электронной почты респондентов, в котором не представлены индивиды, у кого нет своего адреса электронной почты.

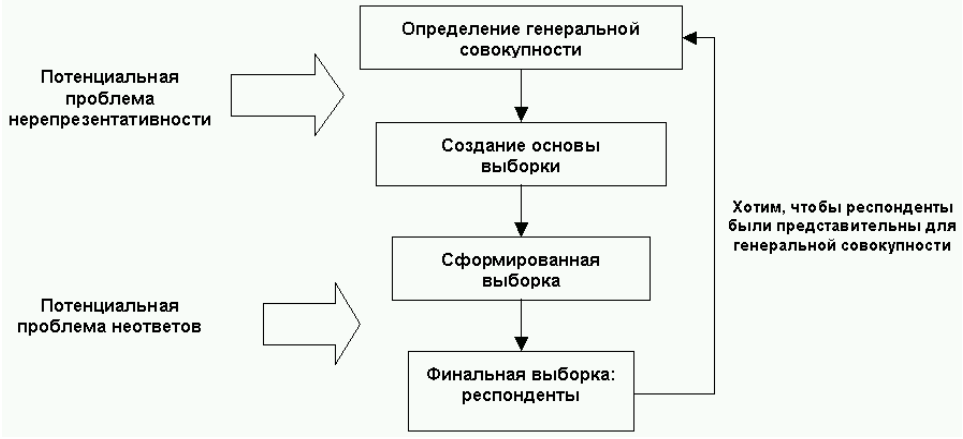


Рисунок 1. Процесс формирования выборки

Вторая более серьезная и распространенная проблема возникает из-за неответов респондентов. Даже если исходная выборка сформирована при помощи техник вероятностного отбора, неответы респондентов (которые бывают практически в любом обследовании) могут привести к смещениям в результатах обследования. Неответившие респонденты могут систематически отличаться по одной или нескольким характеристикам от ответивших. Следовательно их специфическая информация не будет представлена в итоговой выборке, а значит, и в основе выборки. Поэтому рассчитанные результаты обследования для генеральной совокупности в целом будут смещенными. Это

крайне серьезная проблема, которая подлежит контролю со стороны организатора обследования.

Если выборка формируется при помощи вероятностных методов, то характеристики исходной выборки должны соответствовать характеристикам основы выборки. Однако, если выборка формируется при помощи не вероятностных методов, исходная выборка может и не соответствовать основе выборки.

Теперь приступим к детальному рассмотрению этапов процесса формирования выборки и начнем с основы выборки.

Основа выборки. Независимо от того, какое обследование проводится, у его организаторов должен быть актуальный список элементов генеральной совокупности, а в случае его отсутствия - метод составления такого списка посредством выполнения определенного набора последовательных действий. Список элементов генеральной совокупности или набор действий, обеспечивающих его создание, и есть то, что называется основой выборки. В идеальной основе должны присутствовать все элементы генеральной совокупности, причем каждый элемент присутствует только один раз, не должно быть ошибок и посторонних элементов. Список всех больниц - это основа выборки для генеральной совокупности больниц страны.

В ходе формирования основы выборки необходимо дать четкое определение *единицы отбора* и *единицы* анализа в обследовании. Первые - это элементы, которые будут отобраны в выборку и опрошены; вторые те, которые собственно будут изучаться. Обычно все достаточно просто, единицы отбора и анализа совпадают. Они могут быть индивидами, организациями или другими объектами. Но в некоторых случаях определение единицы анализа требует тщательного размышления. В системе здравоохранения единицами отбора могут быть непосредственно больницы. Однако изучаться могут как деятельность больниц, так и отделения больниц, персонал отделений или пациенты. Также в фокусе исследования может оказаться конкретный метод лечения или процедура.

Большинство организаторов обследований сталкивается с неполными основами выборок, в которых содержатся не все элементы генеральной совокупности. К примеру, пусть требуется изучить совокупность ресторанов в большом городе. Ясно, что объехать весь город и составить список всех ресторанов проблематично, поскольку это заняло бы слишком много времени. В этом случае для создания основы выборки, во-первых, можно воспользоваться административными источниками данных (ресторанная деятельность не возможна без регистрации в муниципальных органах). Одним из недостатков такого подхода является обычно невысокая актуальность административных данных. Поэтому, для составления списка местных ресторанов имеет смысл обратиться ко всем доступным рекламным изданиям и справочникам, включая печатные и Интернет-версии, а также к путеводителям по ресторанам. Составленный список и станет основой выборки.

Однако при таком подходе высока вероятность неполного охвата генеральной совокупности. Ясно, что при использовании таких источников информации, небольшие и недавно открывшиеся рестораны, скорее всего, не будут представлены в основе выборки. Конечно, лучше всего, чтобы в основу выборки были включены все элементы генеральной совокупности, либо, по крайней мере, чрезвычайно важно знать недостатки основы выборки для понимания и использования результатов обследования.

Когда основа выборки включает не все элементы генеральной совокупности, необходимо скорректировать определение генеральной совокупности обследования.

Теперь она будет состоять только из элементов, представленных в основе выборки (в приведенном примере новой генеральной совокупностью будут все рестораны, кроме очень маленьких и недавно открывшихся). Выводы, которые делаются на основе выборочных статистик, применимы только к этой группе, а не ко всем ресторанам в городе. Вы должны быть в состоянии сформулировать критерии включения/исключения в основу выборки, то есть указать, какие факторы приводят к включению или исключению в основу представителей генеральной совокупности.

Иногда основы выборок недоступны. Это бывает, когда выборка формируется на основе продолжающихся процессов, таких как поток посетителей некоторого учреждения. Когда маркетологи проводят интервью покупателей в торговых центрах или в магазинах, они используют перехватывающую выборку. При этом одновременно формируется и основа выборки, и сама выборка. Аккуратное применение методологии дает возможность формировать вероятностные выборки при помощи перехватывающей техники или техник, подобных ей.

При создании основы выборки встречаются следующие типичные проблемы.

Отсутствующие элементы. Это наиболее серьезная проблема, поскольку она означает, что некоторые элементы генеральной совокупности не включены в основу выборки и, следовательно, не могут быть отобраны в выборку. Эта проблема также известна как *ошибка охвата*. Для небольших исследований многие специалисты считают пригодным список с 80-90%-м охватом генеральной совокупности, но это также зависит от вида выполняемого анализа и степени обобщаемости результатов, которую Вы надеетесь достичь. Крупные компании, проводящие большие обследования при помощи телефонных интервью широких слоев населения, заведомо исключают тех, у кого нет телефона. Как уже отмечалось, типичным решением проблемы отсутствующих элементов является корректировка определения обследуемой генеральной совокупности для исключения из нее отсутствующих элементов. Наилучшее же решение состоит в том, чтобы найти дополнительные источники сведений, которые позволят включать отсутствующие элементы в основу выборки.

Ошибочные/неподходящие и пропущенные элементы. Если контактная информация ошибочна, элемент основы выборки будет невозможно найти. В результате элемент становится пропущенным. Если создаются списки индивидов, эта проблема возникает из-за смертей, переездов и из-за некоторых ошибок (неверных адресов и так далее).

Также в основу могут быть включены элементы, которые на самом деле не относятся к генеральной совокупности (в приведенном выше примере, бары вместо ресторанов). Если основа выборки содержит много таких ошибок, то приходится отбирать исходную выборку намного большего объема, поэтому уменьшение количества ошибочных элементов крайне важно для снижения стоимости обследования. Также, если индивиды, контактная информация которых неверна, чем-то отличаются от индивидов, контактная информация которых верна, выборка будет смещенной.

Дублирующиеся элементы: Если элемент встречается в основе выборки более одного раза, он имеет больший шанс быть включенным в выборку. Если список невелик (или введен в компьютер), дублирующиеся элементы можно удалить. Эта проблема часто возникает в тех случаях, когда для создания основы выборки используется много списков, как в примере с ресторанами. Другой возможный способ решения проблемы

дублирующихся элементов - это взвешивание данных во время проведения анализа, но это может оказаться технически сложным.

Кластеризованные элементы. Кластеризованные элементы это элементы в списке основы выборки, не являющиеся независимыми. В качестве примеров можно привести членов домашнего хозяйства, жильцов дома и учащихся одной школы. Указанные группы элементов гораздо более схожи между собой, чем с другими случайно отобранными элементами соответствующих генеральных совокупностей. В случае наличия кластеризованных элементов, при выборе элементов необходимо принять специальные меры. Например, в обследовании членов домашних хозяйств на уровне домохозяйства можно использовать метод отбора, называемый *таблицей Киша.* Этот метод гарантирует, что кластеризованные члены домохозяйства будут отобраны случайно (то есть интервьюер не будет опрашивать первого человека, который откроет дверь).

Основы выборок должны оцениваться исходя из стоимости и реальности их создания, а также охвата генеральной совокупности. Эти факторы должны быть сбалансированы, как, впрочем, и все аспекты процесса проведения обследования. Крупная сумма, израсходованная на то, чтобы включить в основу выборки последние 3%, скорее всего, будет бесполезной тратой денег. И эти средства будет лучше израсходовать на получение более высокого уровня отклика.

1.2. Вероятностная выборка. Общие положения

В контексте проведения опросов и обследований *генеральная совокупность* – это обычно многочисленное, но конечное множество реально существующих элементов, обладающих рядом представляющих интерес характеристик, которое полностью охватывает изучаемое социально-экономическое явление. Элементами генеральной совокупности наиболее часто являются индивиды, домохозяйства, предприятия, а также могут быть территориальные единицы и др., что может быть строго определено.

Выборка - любое подмножество элементов изучаемой генеральной совокупности, отобранных для наблюдения.

Выборка непосредственно отбирается из основы выборки, т.е. из составленного организатором обследования списка относящихся к генеральной совокупности элементов с базовой информацией. Под базовой информацией понимается набор характеристик, известных до проведения обследования для каждого элемента основы выборки. Такими характеристиками могут быть, например, наименование организации (юридического лица), адрес места нахождения, контактный телефон, вид осуществляемой экономической деятельности, численность персонала и пр.

В определении выборки, которое было дано выше, не использовалось понятие вероятности. Причина этого состоит в том, что известных и используемых на практике способов формирования выборки достаточно много. Причем только в некоторых из них присутствует элемент случайного отбора.

Вероятностная или случайная выборка предполагает такую процедуру отбора, при которой каждый элемент генеральной совокупности имеет известный неравный нулю шанс оказаться включенным в выборку. Детально разработано и обычно применяется для проведения многомерных многоцелевых обследований ограниченное число вариантов

вероятностной выборки, что не ограничивает широты и надежности применения выборочного метода. Базовый вариант - простая (собственно) случайная выборка, которая также часто применяется для непосредственного отбора элементов на конечной стадии формирования более сложной выборки (см., например, выборка расслоенная случайная).

Важность вероятностной выборки состоит в том, что она позволяет научно обоснованно рассчитать ошибки выборки и доверительные интервалы для статистик, вычисленных по данным самой выборки. Также она позволяет проверять критерии и делать статистически значимые выводы о генеральной совокупности на основе выборочных результатов.

Для неслучайных выборок выборочные результаты (какими бы аккуратными они не представлялись) научно обоснованно можно применять только к совокупности элементов самой выборки, но ни к какой-либо большей группе объектов.

В теории выборки рассматриваются вероятностные выборки, отбираемые из конечной генеральной совокупности, включающей некоторое число N различимых и опознаваемых между собой элементов или единиц. Общее число выборок объема n, которые могут быть извлечены из генеральной совокупности объема N, равно числу различных сочетаний элементов совокупности по n единиц (C_N^n).

План (дизайн) выборки – это вероятностная схема формирования списка выборочной совокупности. Формально план выборки p(s) можно определить как закон распределения вероятностей отбора всех непустых подмножеств элементов генеральной совокупности $\{U\}$, такой что

$$\forall s \in \{s\} \;\; \mathrm{p(s)} \geq 0 \qquad \text{и} \qquad \sum_{s \subset \{U\}} p(s) = 1.$$

где s – выборка.

Для формирования списка элементов выборки определяются приемлемый объем выборки, зависящий от имеющихся ресурсов, и схема случайного отбора, которая приведет к формированию выборки с наилучшими свойствами в смысле обеспечения наименьшего уровня ошибок оценок, связанных с выборкой. Этот процесс называется планированием выборки.

Хотя выборка используется для многих целей, наиболее часто интерес представляют такие показатели, как среднее или суммарное значение наблюдаемого признака, отношение суммарных или средних значений, а также доля единиц в совокупности, отвечающих некоторому критерию. Оценивание параметров генеральной совокупности по данным выборки основывается на следующей базовой формуле оценки суммарного показателя, называемой π-оценкой:

$$\hat{Y}_\pi = \sum_{k=1}^{n} \frac{y_k}{\pi_k} = \sum_{k=1}^{n} w_k y_k$$

Здесь y_k - значение признака элемента k;

π_k - вероятность включения элемента k в выборку;

$w_k = 1/\pi_k$ - выборочный вес элемента k.

$Y = \sum_{k=1}^{N} y_k$ - суммарный показатель признака *y* по генеральной совокупности объемом *N*;

$\hat{Y}_{\pi}$ - оценка по данным выборки объемом *n* значения суммарного показателя признака *y*;

Соответственно оценка среднего значения признака *y* равна отношению оценки суммарного показателя ($\hat{Y}$) и объема генеральной совокупности (*N*), если он точно известен. В противном случае объем генеральной совокупности можно оценить по базовой формуле π-оценки как суммарный показатель признака тождественно равного «1» для всех элементов генеральной совокупности.

Дисперсию базовой π-оценки суммарного показателя можно оценить по данным выборки с помощью следующей общей формулы:

$$Var\left(\hat{Y}_{\pi}\right) = \sum_{k=1}^{n} \frac{y_k^2}{\pi_k^2}(1-\pi_k) + \sum_{k=1}^{n}\sum_{\substack{l=1 \\ l\neq k}}^{n} \frac{y_k y_l}{\pi_k \pi_l} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}}.$$

где π_{kl} - совместная вероятность включения пары элементов *k* в *l* в выборку.

Важно подчеркнуть, что приведенные выражения оценки суммарного показателя и оценки дисперсии этой оценки являются универсальными, т.е. применимы к данным любой вероятностной выборки. Для их использования помимо данных выборки необходимо знать вероятности включения элементов в выборку (π_k), а также парные вероятности включения (π_{kl}), что не очень удобно. Поэтому для каждого конкретного плана выборки, применяемого на практике, в теории получены более простые и удобные индивидуальные выражения для вычисления дисперсии оценок.

Достоверность результатов, получаемых по выборке, основывается на известном распределении значений оценки изучаемого параметра по множеству всевозможных выборок. В случае конечной генеральной совокупности применима Центральная придельная теорема. А именно, при достаточно большом объеме выборки оценки параметров, рассчитанные по выборке, близки к истинным значениям, причем ошибка выборки, т.е. отклонение оценки от истинного значения, распределена приблизительно по нормальному закону распределения. Поэтому по данным выборки может быть рассчитан доверительный интервал, в пределах которого с заданной вероятностью (обычно 0,95 или 0,9) находится истинное для генеральной совокупности значение оцениваемого параметра. Половина длины доверительного интервала называется *предельной ошибкой выборки*. Это величина, которую с заданной доверительной вероятностью не превышает отклонение рассчитанной по выборке оценки параметра от его истинного значения.

Кроме этого, имеется возможность определения необходимого объема выборки для обеспечения требующейся точности результатов, т.е. чтобы значение предельной ошибки выборки не превосходило заданной величины.

Например, в случае простой случайной выборки достоверность результатов (*L* – предельная ошибка выборки) оценки доли (*P*) признака в совокупности и необходимый для этого объем выборки (*n*) связаны следующим приближенным соотношением:

$$n \cong \frac{4p(1-p)}{L^2}$$

Квадратичная функция $y = p($ достигает своего абсолютного максимума при значении доли $p = 0{,}5$. С учетом этого на основе приведенного соотношения можно вычислить, что при фиксированной величине предельной ошибки выборки: L = 0,01 и любой истинной доле, в том числе при P = 0,5, нужный объем выборки составляет не более чем n = 10000 элементов для наблюдения. Также простая случайная выборка объемом n =2500 единиц при истинном значении оцениваемой доли $p = 0{,}5$ может обеспечить только 2%-ую точность оценки по данным выборки (L = 0,02).

Перейдем к рассмотрению наиболее часто используемых в обследованиях типов вероятностных выборок. Хотя специалисты в области выборок разработали множество планов сложных выборок, почти все они основаны на планах выборок, рассматриваемых далее. Приступая к рассмотрению планов выборок подчеркнем, что вероятностная выборка вовсе не обязательно подразумевает, что все элементы в основе выборки имеют равные шансы попасть в выборку. Вместо этого, вероятностная выборка подразумевает, что для каждого элемента мы знаем или можем вычислить вероятность попадания в выборку.

1.3. Простая случайная выборка и другие сходные методы отбора.

Большинство специалистов, думая о выборке, имеют в виду простую случайную выборку. Это отчасти потому, что выборки такого типа рассматриваются в учебниках по теории вероятностей и математической статистики, а также по причине их простоты.

План простой случайной выборки состоит в том, что любая выборка фиксированного объема имеет *равную* вероятность быть отобранной. Простая случайная выборка обеспечивает всем элементам основы выборки равную вероятность попасть в выборку. Однако простая случайная выборка не единственный план отбора, обеспечивающий это условие. Если в университете учатся 15 000 студентов и администрация планирует обследовать 500 человек для изучения мнений студентов, то простая случайная выборка гарантирует, что каждый студент имеет 1 шанс из 30 (500/15000) оказаться в выборке. Хотя более сложные планы выборок имеют свои преимущества, простым случайным выборкам часто отдают предпочтение по причине их простоты как теоретической, так и практической.

Для формирования простой случайной выборки нужен только список элементов генеральной совокупности – основа выборки. Больше никакой информации об элементах не требуется. Для выборок других типов требуется значительно больше информации.

Систематическая случайная выборка.

Процесс отбора простой случайной выборки может оказаться довольно продолжительным и утомительным, если основа выборки не находится в компьютерном файле. Рассмотрим исследование членов профессиональной ассоциации, в котором основа выборки представляет собой распечатанный список 110 000 имен членов ассоциации, причем электронная версия этого списка недоступна. Процедура присвоения членам ассоциации идентификационных номеров слишком трудоемка и поэтому неэффективна.

Или представьте основу выборки, которая состоит из тысяч папок с документами в шкафах в государственном учреждении. К счастью, существует лучший способ отобрать выборку из основ такого типа.

Этот способ состоит в отборе из основы выборки *систематической случайной выборки*. Сущность систематического отбора заключается в отборе из основы каждого k-го элемента, начиная с первого элемента, который отбирается случайно. Величина шага отбора K выбирается таким образом, чтобы количество отобранных элементов было равно, требуемому объему выборки. Хотя свойства статистик при систематическом отборе несколько сложнее, в большинстве случаев (если элементы в основе выборки располагались в случайной последовательности, например, индивиды по алфавиту) можно предполагать с достаточной степенью уверенности, что систематическая случайная выборка эквивалентна простой случайной выборке того же объема. Это значит, что для оценки дисперсий и средних значений можно использовать те же самые формулы, и также можно уверенно делать выводы о генеральной совокупности.

Приведем пример формирования систематической выборки. Пусть хотим отобрать выборку объемом 500 элементов из генеральной совокупности 15000 сотрудников организации. Это означает, что *доля выборки* составляет 500/15000 или 1/30, поэтому шаг отбора k равен 30. Далее, случайным образом выбираем целое число в промежутке от 1 до 30. Обозначим его S_n. В нашем примере S_n может оказаться равным 14. В этом случае выбираем 14-го сотрудника в списке и включаем его в выборку в качестве первого элемента выборки. После чего прибавляем k к S_n. В результате получаем 14 + 30 = 44. Соответственно выбираем 44-го сотрудника в списке и включаем его в выборку в качестве второго элемента. Затем прибавляем k к 44, получаем 74, и выбираем 74-го сотрудника. Далее продолжаем действовать аналогично, пока не достигнем конца списка.

В дополнение к списку всех элементов мы должны знать количество элементов в основе выборки. Это необходимо для возможности расчета длины интервала отбора. Если используется очень большой не пронумерованный список, объем генеральной совокупности необходимо оценить. В этом случае в сформированной выборке может оказаться слишком мало или слишком много элементов.

Часто бывает, что результатом деления объема генеральной совокупности на объем выборки оказывается нецелое число. Для таких случаев есть технические приемы, корректирующие описанный выше процесс для указанной ситуации.

У систематической случайной выборки есть одна специфическая проблема: упорядоченные списки. Если имеется список жилых помещений (квартиры) в домах квартала, а рассчитанный интервал отбора таков, что в выборку оказались включенными значительное число квартир, расположенных на первых этажах, то рассчитанные итоги по выборке окажутся смещенными, так как квартиры на первых этажах обычно стоят дешевле, чем остальные в тех же самых домах. В теории это очень серьезная проблема, но на практике такая ситуация встречается нечасто. Например, списки индивидов в алфавитном порядке фамилий обычно достаточно беспорядочны в плане любых ключевых характеристик, поэтому их можно использовать без дополнительной перетасовки. Если список упорядочен, то лучшим решением является его сортировка в случайном порядке и перенумерация. Еще одно решение проблемы - выбор другого начального элемента отбора.

Последовательная выборка. Похожий тип выборки - последовательная случайная выборка. При ее формировании, начиная со случайно выбранного начального элемента

отбора, выбираются следующие подряд элементы. Выборка такого типа может оказаться неприемлемой, если характеристики единиц отбора связаны с их местом в списке (например, клиенты, упорядоченные по дате первой покупки; пациенты, упорядоченные по больницам). На практике используются более сложные алгоритмы, гарантирующие, что отобранные элементы разбросаны по всей основе выборки.

Отбор с возвращением и без возвращения. Отбор может быть произведен *с возвращением* (СВ). Это означает, что любой элемент генеральной совокупности может быть включен в выборку более одного раза. Отбор *без возвращения* (БВ) означает, что любой элемент генеральной совокупности может быть отобран не более одного раза. В зависимости от того, как осуществлялся отбор, применяются несколько отличающиеся статистические модели.

Ошибка выборки, связанная с отбором. Даже если удастся избежать всех остальных ошибок, характерных для обследований, например, некачественных формулировок вопросов, плохой работы интервьюеров и так далее, неизбежно придется столкнутся с ошибкой, причины которой кроются в разбросе выборок. Какую бы выборку из генеральной совокупности мы не взяли, оценки значений любых показателей для генеральной совокупности, сделанные на основе выборки, не будут точно совпадать со значениями в генеральной совокупности. Значение показателя для генеральной совокупности называется *параметром*, а его эквивалент для выборки – *статистикой*. Другими словами, мы используем статистику для оценки параметра.

Ошибка, связанная с выборкой, не является непреодолимым препятствием для проведения анализа. Иначе не было бы обследований! Теория выборок позволяет вычислить ожидаемую величину ошибки, если задан объем выборки, метод отбора, а также выбрана интересующая нас статистика. Вообще говоря, ошибка выборки для любой статистики непосредственно связана со стандартной ошибкой, которая определяется как:

$$\text{Стандартная ошибка} = \sqrt{\frac{\text{Дисперсия}}{\text{Объем выборки}}}$$

Так что с увеличением объема выборки стандартная ошибка статистики уменьшается обратно пропорционально корню из объема выборки. соответственном, чем больше дисперсия, тем больше и стандартная ошибка.

Доверительные интервалы. Важность стандартной ошибки состоит в том, что она используется при проверке статистических гипотез и при вычислении доверительного интервала для статистики. Доверительный интервал позволяет нам утверждать, например, что процент избирателей, которые планируют проголосовать на выборах за определенного кандидата, составляет 45 ±5%. Стандартная ошибка преобразуется в доверительный интервал по следующей формуле:

Доверительный интервал = статистика $\pm$ **(**Z**)(стандартная ошибка)**

Стандартная ошибка умножается на масштабирующий фактор, обозначаемый как Z и являющийся z-значением для желаемого доверительного уровня. Для 95%-ого доверительного уровня *z*-значение равно 1,96, так что в этом случае для получения доверительного интервала стандартная ошибка умножается на 1,96. Для 90%-ого

доверительного уровня *z*-значение равно 1,64, так что в этом случае для получения доверительного интервала стандартная ошибка умножается на 1,64.

Связь дисперсии с типом выборки. Те, кто изучал статистику, сталкивались с формулой для дисперсии нескольких статистик. Наиболее распространенным примером является дисперсия выборочного среднего значения, рассчитываемая как:

$$Дисперсия = \frac{s^2}{n},$$

где n – объем выборки, а $s^2 = \frac{\sum\left(X_i - \overline{X}\right)^2}{n-1}$ - выборочная дисперсия.

Эта формула выведена при условии выполнения двух предположений, смысл которых кажется вполне очевидным в рамках теоретических курсов.

Предполагается, что:

1. Выборки формируются с использованием простого случайного отбора;
2. Отбор производится из идеальной генеральной совокупности бесконечного объема.

На практике, при проведении обследований, первое предположение иногда оказывается выполненным, но второе не выполняется никогда (хотя для всех практических задач, часто можно считать, что второе предположение выполняется). Рассмотрим каждое из этих предположений и начнем с первого, касающегося плана выборки.

Большинство программ, предназначенных для проведения статистического анализа, по умолчанию используют приведенную выше формулу для вычисления дисперсии оценки среднего значения. Это значит, что эти программы предполагают, что данные собирались по простой случайной выборке. Поэтому, если выборка для обследования действительно отобрана простым случайным образом, то можно уверенно использовать вычисленные для нее в программе дисперсии, а, значит, стандартные ошибки и границы доверительных интервалов. Но это не верно, даже если выборка была отобрана систематическим методом, при котором реальная дисперсия оценок может быть как приблизительно равной дисперсии оценки при простом случайном отборе выборки того же объема, так и во много раз больше или меньше.

Связь дисперсии с объемом генеральной **совокупности.**

Существует весьма распространенное заблуждение о том, что объем выборки должен зависеть от объема генеральной совокупности. Это заблуждение является одной из причин, по которой люди с трудом верят, что выборка объемом в 1500 человек может позволить сделать выводы о населении России с точностью примерно ±3%. Простые рассуждения о выборках приводят к мысли о том, что объем выборки не зависит от объема генеральной совокупности, и в большинстве случаев это верно.

Сформулированное выше второе предположение, которое должно выполняться при вычислении дисперсии, - это предположение о бесконечном объеме генеральной совокупности. Само собой разумеется, что это предположение никогда не выполняется при проведении обследований. Что же означает этот факт для формулы дисперсии и для утверждения о том, что объем выборки не зависит от объема генеральной совокупности? Это означает, что предположение о бесконечном объеме генеральной совокупности

технически неверно, и что, действительно, объем выборки зависит от объема генеральной совокупности. В конце концов, здравый смысл подсказывает, что выборка объемом в 100 единиц из генеральной совокупности объемом 200 единиц должна иметь меньшую ошибку, чем выборка объемом в 100 единиц из генеральной совокупности объемом в 1 000 000 единиц. Каким именно образом объем выборки связан с объемом генеральной совокупности иллюстрирует формула для дисперсии статистики при использовании простой случайной выборки из генеральной совокупности конечного объема:

$$Дисперсия_{скоррект} = (1 - \frac{n}{N})\frac{s^2}{n}$$

где n объем выборки, а N объем генеральной совокупности. Величина (1-n/N) называется *поправкой на конечность генеральной совокупности*. На практике, когда объем выборки превышает 5%-10% от объема генеральной совокупности, использование поправки на конечность генеральной совокупности имеет смысл. Если внимательно изучить приведенную формулу, станет ясно, что скорректированная дисперсия статистики меньше, чем нескорректированная. Однако, в основном процедуры в статистических программах при расчетах предполагают бесконечный объем генеральной совокупности и использование простой случайной выборки. В этом случае, при прочих равных условиях, дисперсии и стандартные ошибки будут слишком большими (приводя к более «пессимистичным» результатам расчета статистических критериев и толкая к увеличению объема выборки, в чем, на самом деле, нет никакой необходимости).

Таблица 1

Основные расчетные формулы при простом случайном отборе

Статистические показатели		Истинное значение	Оценка
А	№	1	2
Суммарное значение признака	1	$Y = \sum_{i=1}^{N} y_i$	$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} y_i$
Среднее значение признака	2	$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$	$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
Доля элементов	3	$P = A/N$	$p = a/n$
Дисперсия оценки доли элементов	4	$V(p) = (1-n/N)\frac{N}{N-1}\frac{P(1-P)}{n}$	$v(p) = (1-n/N)\frac{p(1-p)}{n-1}$
Коэффициент вариации оценки доли	5	$CV = \frac{\sqrt{V(p)}}{P}100\%$	$cv = \frac{\sqrt{v(p)}}{p}100\%$
Дисперсия признака	6	$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$
Среднее квадратическое отклонение признака	7	$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2}$	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$
Ковариация средних значений признаков X и Y	8	$COV(\bar{X},\bar{Y}) = \frac{N-n}{N\cdot n}\cdot\frac{1}{N-1}\cdot\sum_{i=1}^{N}(x_i-\bar{X})(y_i-\bar{Y})$	$\mathrm{cov}(\bar{x},\bar{y}) = \frac{N-n}{N\cdot n}\cdot\frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$
Парный коэффициент корреляции признаков X и Y	9	$R = \frac{COV(\bar{X},\bar{Y})}{S_Y\cdot S_X}$	$r = \frac{\mathrm{cov}(\bar{x},\bar{y})}{s_Y\cdot s_X}$
Коэффициент вариации признака	10	$CV_{np} = \frac{S}{\bar{Y}}100\%$	$cv_{np} = \frac{s}{\bar{y}}100\%$
Дисперсия оценки суммарного значения признака	11	$V(\hat{Y}) = \frac{N^2S^2}{n}\left(\frac{N-n}{N}\right)$	$v(\hat{Y}) = \frac{N^2s^2}{n}\left(\frac{N-n}{N}\right)$
Дисперсия оценки среднего значения признака	12	$V(\bar{y}) = \frac{S^2}{n}\left(\frac{N-n}{N}\right)$	$v(\bar{y}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right)$
Стандартная ошибка оценки суммарного значения признака	13	$S(\hat{Y}) = \frac{NS}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N}\right)}$	$s(\hat{Y}) = \frac{Ns}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N}\right)}$
Стандартная ошибка оценки среднего	14	$S(\bar{y}) = \frac{S}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N}\right)}$	$s(\bar{y}) = \frac{s}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N}\right)}$

значения признака			
Коэффициент вариации оценки	15	$CV = \frac{S(\bar{Y})}{\bar{Y}}100\% = \frac{S(\bar{y})}{\bar{Y}}100\%$	$cv = \frac{s(\bar{Y})}{\bar{Y}}100\% = \frac{s(\bar{y})}{\bar{y}}100\%$

В табл.1 использованы следующие обозначения:

N - количество элементов совокупности;

n - количество элементов выборочной совокупности;

A - количество элементов совокупности, удовлетворяющих некоторому условию;

a - количество элементов выборки, удовлетворяющих некоторому условию;

i - номер элемента;

y_i - значение признака y ***i***-го элемента; ***i = 1,2… N*** или ***i = 1,2… n***.

x_i - значение признака x ***i***-го элемента; ***i = 1,2… N*** или ***i = 1,2… n***.

1.4. Расслоенная (стратифицированная) случайная выборка.

Существенно более эффективная процедура формирования выборки, которую может выполнить даже новичок в организации обследований, добавляет в процесс формирования выборки один шаг. При *расслоенной выборке* основа выборки сначала группируется в слои, элементы которых идентичны по некоторым характеристикам. Затем внутри каждого слоя выполняется простой или систематический случайный отбор, а полученные в результате выборки объединяются в одну.

В качестве иллюстрации в примере с ресторанами можно создать слои по типу кухни: европейская, восточная, смешанная – одна из возможных группировок. Далее случайно отбираются рестораны отдельно в каждой из этих групп.

Организаторы обследований применяют расслоенный отбор по трем причинам. Это может:

- уменьшить вариацию (разброс переменных) в выборке;
- обеспечить пропорциональную представленность наблюдений в каждом слое;
- увеличить в итоговой выборке число элементов из подсовокупностей и, таким образом, повысить надежность статистического анализа.

Две последние причины побуждают непрофессиональных организаторов обследований использовать расслоенные случайные выборки. Чтобы воспользоваться преимуществом уменьшенного разброса при расслоенной выборке, необходимо знать о связях между переменными расслоения и переменными, представляющими интерес для изучения. Кроме того, для корректного расчета стандартных ошибок необходимы либо формулы расчета ошибки выборки для расслоенной выборки, либо специализированное программное обеспечение, такое как пакет SPSS (компонет Complex Samples).

Тем не менее, если расслоение всегда полезно для обеспечения пропорционального представительства наблюдений из каждого слоя, почему бы ни использовать этот тип выборки чаще? Проблема заключается в том, что для формирования расслоенной выборки требуется не только список элементов генеральной совокупности. Также нужна базовая информация, по крайней мере, об одной характеристике для каждого элемента основы выборки, которая будет использована для создания слоев. Такой характеристикой может быть пол, статус клиента или место жительства. Однако, иногда такая информация недоступна. Но, если она доступна, расслоение следует всегда иметь в виду, особенно, если основа выборки имеется в электронном виде, что делает работу с основой выборки достаточно простой.

Формирование расслоенной выборки включает только один дополнительный шаг по сравнению с простой случайной выборкой. Сначала генеральная совокупность должна быть разделена на подгруппы. После чего могут использоваться техники случайного или систематического отбора для отбора конкретных элементов внутри каждого слоя. При этом каждый слой рассматривается как отдельная основа выборки. Расслаивать можно по нескольким характеристикам.

Типы расслоенной выборки.

Расслоенные выборки бывают двух типов (проиллюстрировано на рисунке 2).

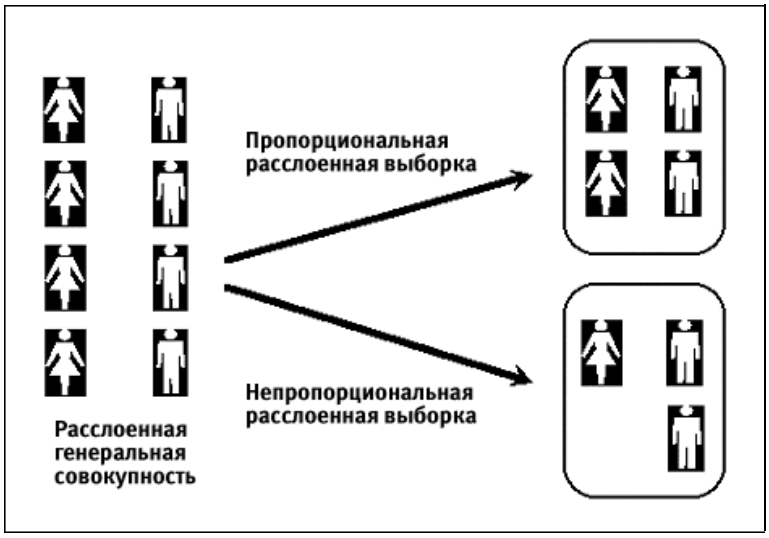


Рисунок 2. Два типа расслоенных выборок

При *пропорциональной расслоенной выборке* объем выборки в каждом слое пропорционален объему этого слоя в генеральной совокупности. То есть, если некоторая группа в генеральной совокупности содержит 10% элементов всей генеральной совокупности (основы выборки), то и в итоговой выборке 10% элементов будут относиться к этой группе. Так, проводя обследование сотрудников крупной компании, можно сделать расслоение по категории занятости сотрудников, и затем такой отбор, при котором доля сотрудников в каждой категории занятости в выборке будет соответствовать доли сотрудников в этой категории занятости в компании в целом.

Важно отметить, что при прочих равных условиях, хорошо сформированная простая случайная выборка будет близка к пропорциональной выборке. То есть, если женщины составляют 55% генеральной совокупности, то в выборке должно оказаться около 55% женщин. Но это не совсем справедливо. Если нужно, чтобы проценты в выборке точно соответствовали процентам в генеральной совокупности, то необходимо сформировать расслоенную выборку.

При *непропорциональной расслоенной выборке* в каждом слое отбирается такое количество элементов, что объем выборки для слоя *непропорционален* объему генеральной совокупности для этого слоя. Эта техника обычно используется организаторами обследований, чтобы повысить статистическую мощность при выявлении различий между группами (слоями). Так в примере с ресторанами можно увеличить долю в выборке ресторанов со смешанной кухней (если таких ресторанов не так много, как

ресторанов других типов), чтобы получить в выборке по 100 ресторанов европейской, восточной и смешанной кухни. Равное число наблюдений в каждой подгруппе максимизирует возможность обнаружения различий между группами при прочих равных условиях.

Профессиональные организаторы обследований применяют непропорциональное расслоение по двум причинам. Если стоимость отбора существенно различается в зависимости от слоя, можно отобрать выборку оптимального объема с точки зрения минимизации затрат. Если дисперсии в слоях существенно различаются, можно отобрать больше наблюдений в слоях с большей дисперсией и, тем самым, повысить точность статистических оценок.

Для лучшего понимания описанных выше причин использования непропорциональной расслоенной выборки, поясним их в терминах *эффекта плана* конкретной выборки. Эффект плана - это отношение дисперсии статистики (скажем среднего значения количественной переменной), вычисленной для текущей выборки, к дисперсии той же самой статистики, вычисленной в предположении, что выборка - простая случайная того же объема. Эффект плана больше 1 показывает, что план выборки менее эффективен (дает б*о*льшую ошибку для выборки того же объема), чем простая случайная выборка. Эффект плана меньше 1 свидетельствует об обратном.

Допустим, что сформирована простая случайная выборка клиентов банка объемом 1000 и рассчитали среднее значение и дисперсию остатков на их счетах (в качестве иллюстрации предположим, что получены значения 934.45 доллара и 447.45 долларов соответственно). Затем сформировали пропорциональную расслоенную выборку того же объема (1000), слои в которой были основаны на типах клиентов (различное количество счетов и так далее). Для расслоенной выборки рассчитали те же две статистики (используя соответствующую формулу для дисперсии среднего значения для расслоенной выборки) и обнаружили, что среднее значение остатков на счетах клиентов равно 934.45 долларов (такое же, что обычно не наблюдается), но дисперсия среднего значения стала меньше (402.98 доллара). Эффект плана – это отношение 402.98 к 447.45 равное 0.90. Используя расслоенную выборку, мы эффективно уменьшаем дисперсию статистик и, соответственно, стандартные ошибки. Это увеличивает статистическую мощность.

Расслоенные выборки помогают уменьшить ошибку выборки, обычно обеспечивая лучшие результаты, по крайней мере, не худшие, по сравнению с простыми случайными выборками. Другими словами, эффект плана расслоенных выборок обычно меньше 1. Это особенно справедливо, если слои относительно однородны по представляющим интерес переменным.

Взвешивание непропорциональных расслоенных выборок.

При использовании непропорциональных расслоенных выборок элементы генеральной совокупности имеют неравные вероятности попадания в выборку. Без соответствующей корректировки статистические оценки, рассчитанные для параметров генеральной совокупности, будут смещенными. В примере с ресторанами, если доля ресторанов со смешанной кухней в выборке больше, чем в генеральной совокупности, то средний оборот ресторанов, вычисленный для всей выборки, не является представительным для всех ресторанов города, поскольку в выборке рестораны со смешанной кухней будут влиять на средний оборот больше, чем в генеральной

совокупности. Эту ситуацию можно исправить при помощи *взвешивания*, которое снижает вес, или влияние, ресторанов со смешенной кухней. В данном случае выборку нужно взвесить таким образом, чтобы она отражала истинные доли трех типов ресторанов в городе.

Таблица 2.

Основные расчетные формулы при расслоенном случайном отборе

Статистические показатели		Истинное значение	Оценка
А	№	1	2
Суммарное значение признака	1	$Y = \sum_{h=1}^{L} Y_h$	$\hat{Y}_{st} = \sum_{h=1}^{L} N_h \bar{y}_h$
Среднее значение признака	2	$\bar{Y} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{Y}_h$	$\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h$
Дисперсия оценки суммарного значения признака	3	$V(\hat{Y}_{st}) = \sum_{h=1}^{L} N_h (N_h - n_h)\frac{S_h^2}{n_h}$	$v(\hat{Y}_{st}) = \sum_{h=1}^{L} N_h (N_h - n_h)\frac{s_h^2}{n_h}$
Дисперсия оценки среднего значения признака	4	$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h (N_h - n_h)\frac{S_h^2}{n_h}$	$v(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h (N_h - n_h)\frac{s_h^2}{n_h}$
Стандартная ошибка оценки суммарного значения признака	5	$S(\hat{Y}_{st}) = NS(\bar{y}_{st}) = \sqrt{\sum_{h=1}^{L} N_h (N_h - n_h)\frac{S_h^2}{n_h}}$	$s(\hat{Y}_{st}) = Ns(\bar{y}_{st}) = \sqrt{\sum_{h=1}^{L} N_h (N_h - n_h)\frac{s_h^2}{n_h}}$
Стандартная ошибка оценки среднего значения признака	6	$S(\bar{y}_{st}) = \sqrt{\frac{1}{N^2}\sum_{h=1}^{L} N_h (N_h - n_h)\frac{S_h^2}{n_h}}$	$s(\bar{y}_{st}) = \sqrt{\frac{1}{N^2}\sum_{h=1}^{L} N_h (N_h - n_h)\frac{s_h^2}{n_h}}$
Коэффициент вариации оценки	7	$CV = \frac{S(\hat{Y}_{st})}{Y}100\% = \frac{S(\bar{y}_{st})}{\bar{Y}}100\%$	$cv = \frac{s(\hat{Y}_{st})}{\hat{Y}_{st}}100\% = \frac{s(\bar{y}_{st})}{\bar{y}_{st}}100\%$
Доля элементов совокупности	8	$P_{st} = \sum \frac{N_h P_h}{N}$	$p_{st} = \sum \frac{N_h p_h}{N}$
Дисперсия оценки доли элементов	9	$V(p_{st}) = \frac{1}{N^2}\sum \frac{N_h^2 (N_h - n_h)}{N_h - 1}\frac{P_h(1-P_h)}{n_h}$	$v(p_{st}) = \frac{1}{N^2}\sum N_h^2 (N_h - n_h)\frac{p_h(1-p_h)}{n_h - 1}$
Коэффициент вариации оценки доли	10	$CV = \frac{\sqrt{V(p_{st})}}{P_{st}}100\%$	$cv = \frac{\sqrt{v(p_{st})}}{p_{st}}100\%$

В табл.2 использованы следующие обозначения:

L - Число слоев;

H - Номер слоя;

Y_h - Суммарное значение признака ***y*** в ***h***-м слое генеральной совокупности;

N_h - Объем ***h***-го слоя генеральной совокупности;

$\bar{y}_h$ - Среднее значение признака ***y*** в ***h***-м слое выборки;

N - Объем генеральной совокупности;

$\bar{Y}_h$ - Среднее значение признака ***y*** в ***h***-м слое генеральной совокупности;

N_h - Объем ***h***-го слоя выборки;

S_h^2 - Истинное значение дисперсии для ***h***-го слоя: $S_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(y_{hi} - \bar{Y}_h)^2$

I - Номер элемента внутри слоя;

y_{hi} - Значение признака ***y*** ***i***-го элемента слоя ***h***;

s_h^2 - Несмещенная оценка дисперсии для ***h***-го слоя: $s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2$

P_h Доля элементов в слое *h*: $P_h = A_h / N_h$, где A_h - количество элементов слоя, удовлетворяющих некоторому условию;

p_h Оценка доли элементов в слое *h*: $p_h = a_h / N_h$, где a_h – количество элементов слоя *h* выборки, удовлетворяющих данному условию;

1.5. метод отбора элементов с вероятностями пропорциональными их величине.

Если размер всех элементов совокупности, задан значениями некоторой вспомогательной переменной, известными до начала проведения обследования, тогда отбор элементов в выборку можно осуществить с неравными вероятностями пропорциональными размеру (ВПР-метод отбора).

ВПР метод отбора позволяет повысить точность оценивания, если используемая для определения вероятностей вспомогательная переменная размера приблизительно пропорциональна изучаемым признакам. ВПР выборку выгодно использовать, когда имеются отдельные единицы с большими значениями признака. Фактически это вероятностный аналог метода наблюдения «основного массива единиц» (см. курс «Теория статистики»).

ВПР выборка часто применяется в обследованиях населения при отборе территориальных единиц. при этом вероятности включения элементов в выборку пропорциональны численности проживающего населения в территориальных единицах отбора. В качестве примера использования ВПР выборки можно привести случай проведения выборочного обследования бюджетов домашних хозяйств и обследования занятости населения.

1.6. Кластерная (сериальная) выборка.

Кластерная выборка подразумевает отбор групп элементов или кластеров. Например, кластерами могут быть школы, больницы или (географические) территории, а единицами отбора - учащиеся, пациенты или жители. Кластерные выборки могут снизить стоимость обследования за счет концентрации единиц наблюдения в физически или географически мелких областях при проведении учета и при проведении интервью.

Кластерный отбор может оказаться необходимым в тех случаях, когда основа выборки не является списком единиц наблюдения. Например, если проводится обследование учащихся средних школ страны, нам вряд ли удастся получить общий список всех учащихся всех школ, из которого затем будет отбираться выборка. В этом случае, лучше всего использовать многоэтапную выборку, когда сначала отбираются районы, затем школы в отобранных районах, а затем классы или учащиеся в отобранных школах. В таком многоэтапном процессе учащиеся отбираются без использования основы выборки, в которой содержится список всех учащихся средних школ.

По этим причинам кластеризация часто используется в многоэтапных планах (смотрите ниже) и при формировании территориальных (географических) выборок.

Основной недостаток кластерных выборок - меньшая точность оценок, чем у простых случайных выборок такого же объема. Этот недостаток возникает из-за того, что единицы выборки внутри кластеров (например, учащиеся одной школы), как правило, более однородны (гомогенны), чем простая случайная выборка из генеральной совокупности, так что в кластерной выборке, как правило, содержится меньше информации.

Таблица 3

Основные расчетные формулы при кластерном отборе

Статистические показатели	Истинное значение	Оценка
Суммарное Значение признака	$Y = \sum_{i=1}^{M} T_i$; $\bar{T} = \frac{1}{M}\sum_{i=1}^{M} T_i$	$\hat{Y} = \frac{M}{m}\sum_{i=1}^{m} T_i$
Среднее значение признака	$\bar{Y} = \frac{Y}{N}$	$\hat{\bar{Y}} = \frac{\hat{Y}}{\hat{N}}$
Дисперсия оценки суммарного значения признака	$V(\hat{Y}) = M^2(1-\frac{m}{M})\frac{1}{m}\frac{1}{M-1}\sum_{i=1}^{M}(T_i-\bar{T})$	$v(\hat{Y}) = M^2(1-\frac{m}{M})\frac{1}{m}\frac{1}{m-1}\sum_{i=1}^{m}(T_i-\hat{\bar{T}})^2$
Дисперсия оценки среднего значения признака	$V(\hat{\bar{Y}}) = \frac{1}{N^2}\left[M^2(1-\frac{m}{M})\frac{1}{m}\frac{1}{M-1}\sum_{i=1}^{M}(T_i-\bar{T})^2\right]$	$v(\hat{\bar{Y}}) = \frac{1}{N^2}\left[M^2(1-\frac{m}{M})\frac{1}{m}\frac{1}{m-1}\sum_{i=1}^{m}(T_i-\hat{\bar{T}})^2\right]$
Стандартная ошибка оценки суммарного значения признака	$S(\hat{Y}) = \sqrt{M^2(1-\frac{m}{M})\frac{1}{m}\frac{1}{M-1}\sum_{i=1}^{M}(T_i-\bar{T})^2}$	$s(\hat{Y}) = \sqrt{M^2(1-\frac{m}{M})\frac{1}{m}\frac{1}{m-1}\sum_{i=1}^{m}(T_i-\hat{\bar{T}})^2}$
Стандартная ошибка оценки среднего значения признака	$S(\hat{\bar{Y}}) = \frac{1}{N}\sqrt{M^2(1-\frac{m}{M})\frac{1}{m}\frac{1}{M-1}\sum_{i=1}^{M}(T_i-\bar{T})^2}$	$s(\hat{\bar{Y}}) = \frac{1}{N}\sqrt{M^2(1-\frac{m}{M})\frac{1}{m}\frac{1}{m-1}\sum_{i=1}^{m}(T_i-\hat{\bar{T}})^2}$
Коэффициент вариации оценки	$CV = \frac{S(\hat{Y})}{Y}100\% = \frac{S(\hat{\bar{Y}})}{\bar{Y}}100\%$	$cv = \frac{s(\hat{Y})}{\hat{Y}}100\% = \frac{s(\hat{\bar{Y}})}{\hat{\bar{Y}}}100\%$

В таблице 3 использованы следующие обозначения:

i - Номер гнезда (кластера);

M - Количество гнезд (кластеров);

T_i - Суммарное значение признака в ***i***-м гнезде (кластере).

$\bar{T}$ - Среднее значение признака ***Y*** по гнездам (кластерам);

m - Количество выбранных гнезд (кластеров);

N - Объем генеральной совокупности;

$\hat{N}$ - Оценка количества элементов генеральной совокупности;

$\hat{\bar{T}}$ - Оценка среднего значения признака ***Y*** по гнездам (кластерам):

$$\hat{\bar{T}} = \frac{1}{m}\sum_{i=1}^{m} T_i$$

1.7. Многоступенчатая (многоэтапная) выборка.

Иногда оказывается невозможным получить список элементов генеральной совокупности, которую нужно обследовать. Когда организаторы опросов общественного

мнения хотят получить представительную выборку всего взрослого населения страны, у них нет возможности получить основу выборки в виде списка, так как списка всего взрослого населения страны с адресами и телефонами просто не существует. Есть два наиболее часто используемых подхода к проведению таких обследований. Можно случайно отобрать телефонные номера (существующие номера телефонов известны). Вместо этого можно случайно отбирать географические территории, постепенно переходя к более мелким территориям, затем к домохозяйствам и, наконец, к индивидам в домохозяйствах. И в том, и в другом случае мы начинаем отбирать не те, элементы, которые собираемся обследовать, то есть телефонные номера, населенные пункты и так далее. Другими словами, мы формируем *многоэтапные выборки*.

В рассмотренных ранее типах выборок, элементы, составляющие основу выборки, соотносились с элементами генеральной совокупности как один к одному. Другими словами, если планировалось исследовать больницы, то основа выборки состояла из списка больниц. Если планировалось изучение клиентов, то основа выборки представляла собой список клиентов или метод непосредственного, за один шаг, их нахождения (группировка элементов списка по слоям не противоречит этому, так как слои, по-прежнему, состоят из представителей генеральной совокупности.)

Если требуется обследовать учащихся средних школ по всей территории страны, то ясно, что списка учащихся средних школ не существует. Поэтому возникает проблема: как создать вероятностную выборку, по данным которой можно делать выводы о генеральной совокупности всех учащихся средних школ? Решением этой проблемы будет создание плана многоэтапной выборки, который может включать следующие этапы:

- отбор случайной выборки первичных единиц отбора, которые состояли бы из различных городских местностей и сельских административно-территориальных единиц;
- отбор случайной выборки населенных пунктов, относящихся к отобранным первичным единицам;
- отбор случайной выборки районов в отобранных населенных пунктах;
- отбор случайной выборки средних школ в отобранных районах;
- отбор случайной выборки классов в отобранных школах;
- отбор случайной выборки учащихся в отобранных классах.

В приведенном примере хорошо видно, насколько отличается такой подход от других рассмотренных планов выборки. Вплоть до последнего шага, основа выборки (которая в действительности представляет собой несколько разных основ выборок) - это не список учащихся, то есть не генеральная совокупность, которую мы планируем обследовать, а список элементов более высокого уровня агрегации. Вот почему такие выборки называются многоэтапными.

Именно многоэтапные планы выборки используются в обследовании домашних хозяйств, а также в любых других обследованиях населения страны. Рисунок 3 иллюстрирует процесс формирования выборки в этих обследованиях. Как и в случае гипотетической выборки учащихся средних школ, отбирается случайная выборка городских и сельских территорий. В отобранных территориях отбираются меньшие территориальные единицы (например, населенные пункты, избирательные участки и так далее). В меньших территориальных единицах отбираются квартиры и дома. На последнем шаге происходит отбор потенциальных респондентов внутри домохозяйств.

Рисунок 3. Иллюстрация многоэтапной выборки

На самом деле существует два наиболее распространенных типа многоэтапных выборок. Если в примере с выборкой учащихся средних школ обследованы все ученики в отобранных классах, то формировалась *кластерная выборка* (на последнем этапе отобраны все элементы). Если, вместо этого, в каждом отобранном кластере будет отобрана только выборка элементов, то такая выборка будет настоящей многоэтапной. Иногда эти термины используются в качестве синонимов. И в том, и в другом случае на каждом этапе для отбора элементов обычно применяется простой случайный отбор, хотя также может применяться и расслоенный отбор.

Чтобы многоэтапная выборка была эффективной, единицы самого высокого уровня должны охватывать всю генеральную совокупность. Если при проведении обследования авиапассажиров сначала отбираются авиакомпании и в выборку оказываются включенными только крупнейшие авиакомпании, такая выборка не будет представительной для пассажиров всех авиакомпаний. Но с проблемой элементов, отсутствующих в основе выборки мы сталкивались и ранее, и обычное решение этой проблемы состоит в переопределении целевой генеральной совокупности, если вся генеральная совокупность недоступна.

Для многоэтапных выборок, как и для расслоенных выборок, статистические программы, включая SPSS, по умолчанию, не рассчитывают корректные стандартные ошибки статистик. Почти во всех случаях эффект плана многоэтапной выборки больше 1. Это значит, что многоэтапные выборки увеличивают дисперсию по сравнению с простыми случайными выборками тех же объемов.

Все это вовсе не означает, что планы выборки такого типа не следует использовать. Основная цель формирования выборок заключается в формировании вероятностной выборки, которая аккуратно представляет изучаемую генеральную совокупность. А

иногда единственный способ изучить определенную группу - это использовать многоэтапную выборку. Конечно, не следует забывать о необходимости вычисления корректных стандартных ошибок, но это не должно мешать Вам в выборе подходящего плана выборки.

Принятие решения об использовании многоэтапной выборки зависит от нескольких факторов, среди которых:

- Отсутствие списка элементов генеральной совокупности.

- Существование хорошо очерченных кластеров (районы в городе, типы военных частей при изучении военнослужащих и так далее).

- Обоснованная оценка количества элементов в кластерах. Такая же информация требуется для простой случайной выборки.

На завершающем этапе формирования выборки кластеры должны быть достаточно малы. Это дает возможность снизить стоимость обследования. Если в обследовании населения России в качестве (последних) кластеров использовать районы, то нам вряд ли удастся снизить стоимость обследования.

Чаще в целом общий объем выборки должен быть весьма значительным, чтобы оправдать применение многоэтапной выборки, например, больше 1000.

Таблица 4

Основные расчетные формулы при двухэтапном кластерном отборе

Статистические показатели	Истинное значение	Оценка
Суммарное значение признака	$Y = \sum_{i=1}^{N} Y_i$; $Y_i = \sum_{j=1}^{M_i} Y_{ij}$	$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i} y_i$; $\hat{Y}_i = \frac{M_i}{m_i} y_i$
Количество элементов в генеральной совокупности	$M = \sum_{i=1}^{N} M_i$	$\hat{M} = \frac{N}{n}\sum_{i=1}^{n} M_i$
Среднее значение признака	$\bar{Y} = \frac{Y}{M}$; $\bar{Y}_c = \frac{1}{N}\sum_{i=1}^{N} Y_i$; $\bar{Y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} Y_{ij}$	$\bar{y} = \frac{\hat{Y}}{\hat{M}}$; $\bar{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$
Количество элементов в среднем на группу	$\bar{M} = \frac{1}{N}\sum_{i=1}^{N} M_i = \frac{M}{N}$	$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$
Дисперсия оценки суммарного значения признака	$V(\hat{Y}) = K_1 S_b^2 + K_2 S_{wi}^2$, где $K_1 = \frac{N^2(N-n)}{Nn}$; $K_2 = \frac{N}{n}\sum_{i=1}^{N}\frac{M_i^2(M_i - m_i)}{M_i m_i}$	$v(\hat{Y}) = K_1 s_b^2 + K_2 s_{wi}^2$, где $K_1 = \frac{N^2(N-n)}{Nn}$; $K_2 = \frac{N}{n}\sum_{i=1}^{n}\frac{M_i^2(M_i - m_i)}{M_i m_i}$
Дисперсия оценки среднего значения признака	$V(\bar{y}) = \frac{1}{M^2}\left[K_1 S_b^2 + K_2 S_{wi}^2\right]$	$v(\bar{y}) = \frac{1}{M^2}\left[K_1 s_b^2 + K_2 s_{wi}^2\right]$
Стандартная ошибка оценки суммарного значения признака	$S(\hat{Y}) = \sqrt{K_1 S_b^2 + K_2 S_{wi}^2}$	$s(\hat{Y}) = \sqrt{K_1 s_b^2 + K_2 s_{wi}^2}$
Стандартная ошибка оценки среднего значения признака	$S(\bar{y}) = \frac{1}{M}\sqrt{K_1 S_b^2 + K_2 S_{wi}^2}$	$s(\bar{y}) = \frac{1}{M}\sqrt{K_1 s_b^2 + K_2 s_{wi}^2}$
Коэффициент вариации оценки	$CV = \frac{S(\hat{Y})}{\hat{Y}}100\% = \frac{S(\bar{y})}{\bar{Y}}100\%$	$cv = \frac{s(\hat{Y})}{\hat{Y}}100\% = \frac{s(\bar{y})}{\bar{y}}100\%$

В таблице 4 использованы следующие обозначения:

N - Количество групп в генеральной совокупности;

i - Номер группы;

Y_i - Сумма признака в ***i***-й группе совокупности;

M_i - Объем в ***i***-й группе, $i=1, \ldots, N$;

j - Номер элемента в группе;

Y_{ij} - Величина признака для ***j***-го элемента в ***i***-й группе совокупности;

n - Количество выбранных групп;

m_i - Количество выбранных элементов в ***i***-й выбранной группе,

$i=1, \ldots, n$;

$y_i = \sum_{j=1}^{m_i} y_{ij}$ - Суммарное значение признака по выборке в ***i***-й выбранной группе;

y_{ij} - Величина признака для ***j***-го выбранного элемента в ***i***-й группе;

$\overline{Y}_c$ - Среднее значение признака по совокупности;

$\overline{Y}_i$ - Средняя величина характеристики признака в ***i***-й группе совокупности;

$\hat{Y}_i$ - Несмещенная оценка для Y_i;

$\overline{y}_i$ - Среднее значение характеристики выбранного элемента в ***i***-й выбранной группе;

S_b^2 - Дисперсия совокупности между группами:

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y}_i)^2;$$

S_{wi}^2 - Дисперсия внутри ***i***-й выбранной группы:

$$S_{wi}^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(Y_{ij} - \overline{Y}_c)^2;$$

s_b^2 - Несмещенная оценка дисперсии совокупности между группами:

$$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(M_i y_i - \frac{1}{n}\sum_{i=1}^{n} M_i \overline{y}_i)^2;$$

s_{wi}^2 - Несмещенная оценка дисперсии внутри ***i***-й выбранной группы:

$$s_{wi}^2 = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_i)^2 \cdot$$

1.8. Оценка по отношению (с помощью коэффициента) и постстратификация

Предположим, что имеется дополнительная информация о некотором вспомогательном признаке, измеренном для элементов выборки, который связан с изучаемыми характеристиками, представляющими основной интерес в обследовании. Дополнительной информацией может быть известный суммарный или средний показатель вспомогательного признака по всей генеральной совокупности.

Рассчитанная по выборке оценка суммарного показателя, как правило, не совпадает с известным истинным значением. Поэтому возникают два вопроса:

1) как учесть имеющуюся вспомогательную информацию так, чтобы дать более точную оценку представляющего основной интерес суммарного показателя?

2) Каким образом объяснить расхождение итоговых данных относительно вспомогательного показателя?

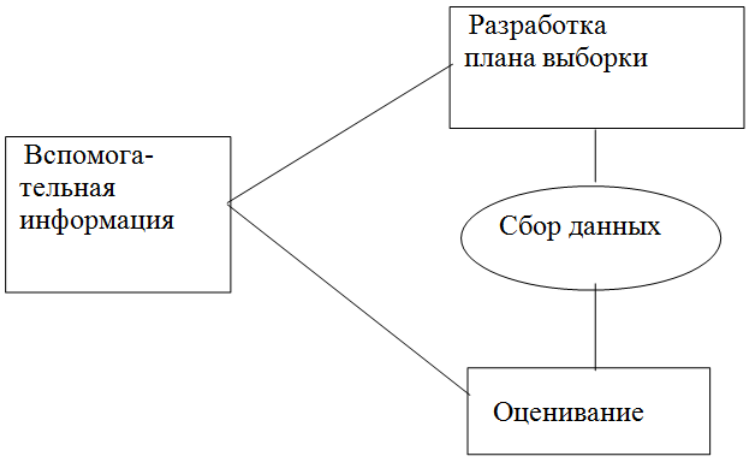


Рис. 4. Схема использования вспомогательной информации.

В данном случае вспомогательная информация будет использоваться не на стадии формирования выборки (как при расслоенном отборе), а на стадии оценивания, т.е. после проведения наблюдения. Кроме того, в основе выборки нет индивидуальных данных каждого элемента по вспомогательной переменной.

В рассматриваемом случае для возможности использования вспомогательной информации на этапе проведения наблюдения собирают два типа информации по каждому элементу, включенному в выборки: а) по представляющей основной интерес переменной; и б) по вспомогательной переменной, суммарный показатель которой по генеральной совокупности известна заранее. После этого на основе данных выборки рассчитывают оценки суммарных значений целевой и вспомогательной переменных.

Обычно оценка суммарного показателя для вспомогательной переменной отличается от известного истинного значения. Поэтому с помощью информации о наблюдаемом расхождении (вычисляется коэффициент различия) корректируется значение оценки целевого показателя. Такой подход обоснован, если выполняется

предположение, что целевая переменная хотя бы приблизительно пропорциональна вспомогательной переменной.

Таким образом, на основе выборки получают π-оценки суммарных значений переменных *y* и *x*, а затем рассчитывают новую оценку с помощью корректирующего коэффициента:

$$Y_\pi = \sum_{k \in s} \frac{y_k}{\pi_k}; \quad X_\pi = \sum_{k \in s} \frac{x_k}{\pi_k}; \quad Y_\pi = KX.$$

Постстратификация. Для некоторых удобных в целях расслоения характеристик слой, к которому принадлежит тот или иной элемент, нельзя указать до того, как будут собраны данные. Типичными примерами могут служить такие характеристики как пол, возраст или уровень образования респондента. Объемы этих слоев (страт) могут быть известными с достаточной точностью, например, по итогам последней переписи населения, но распределить элементы случайной по слоям можно только после проведения обследования. Поэтому можно отобрать, например, простую случайную выборку некоторого объема, а затем распределить обследованные элементы по слоям. И далее, например, вместо выборочного среднего использовать другую, постстратификационную, оценку генеральной средней.

Таким образом, суть пострасслоения состоит в том, что после обследования в выборке определяют группы единиц (слои), называемые **постстратами**. При этом предполагается, что распределение генеральной совокупности по этим постслоям известно. Нет никаких оснований предполагать, что такое распределение совокупности точно совпадает с распределением в выборке. При оценивании учитываются известные пропорции выделенных групп элементов в генеральной совокупности.

Основное отличие от расслоенной выборки состоит в том, что распределение выборки по постслоям не контролируется: объем постслоев в выборке становится известным только после обследования. Это случайные величины, зависящие от отбираемой выборки**.**

1.9. Оценивание по регрессии

Пусть целью выборочного обследования является оценка суммарного показателя изучаемой переменной (y) для генеральной совокупности $\{U = 1,2,...,k,...,N\}$, где N – объем генеральной совокупности $\{U\}$:

$$Y = \sum_{k \in U} y_k$$

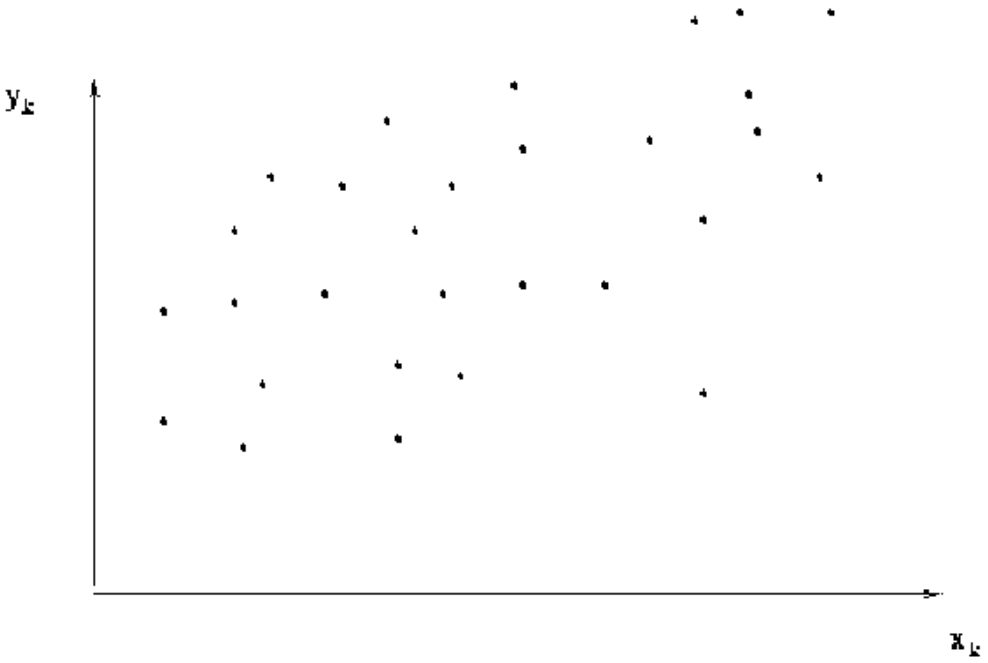
Пусть для единиц генеральной совокупности $\{U\}$ имеется вспомогательная переменная (x) со значениями:

$x_1,...,x_k,...,x_N$

Соответственно суммарным показатель по совокупности для вспомогательной переменной будет

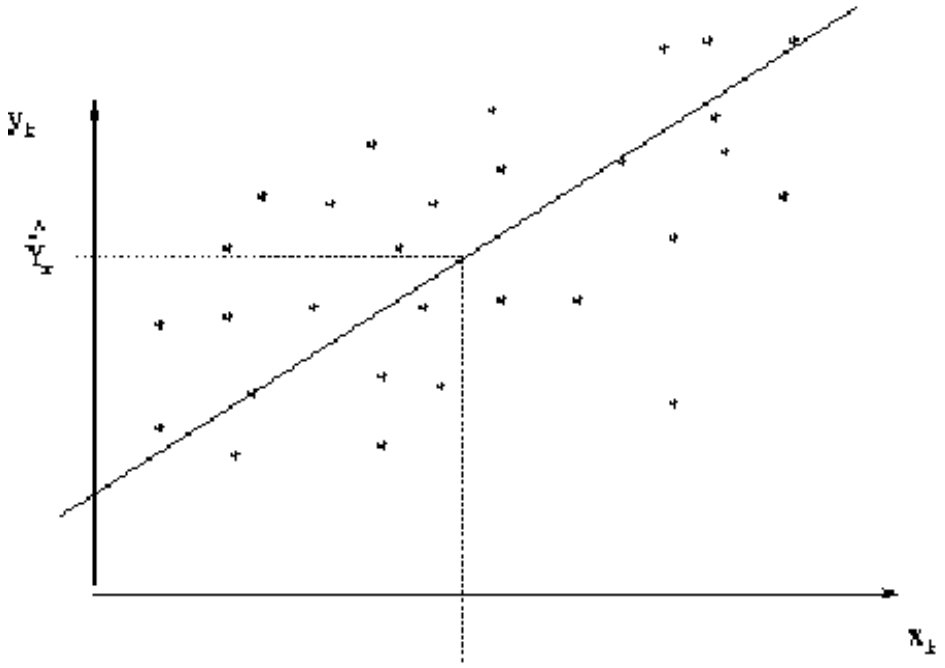
$$X = \sum_{k \in U} x_k$$

Пусть в выборке имеется более или менее линейная связь между значениями x_k и y_k



(см. следующую диаграмму).

Графически возможные оценки среднего значения располагаются приблизительно в центре скопления точек.



Если истинное среднее значение признака (*x*) известно, то можно вычислить оценку по регрессии:

$$\hat{Y}_{reg} = \hat{Y}_{\pi} + (\bar{X} - \hat{X}_{\pi})\hat{b},$$

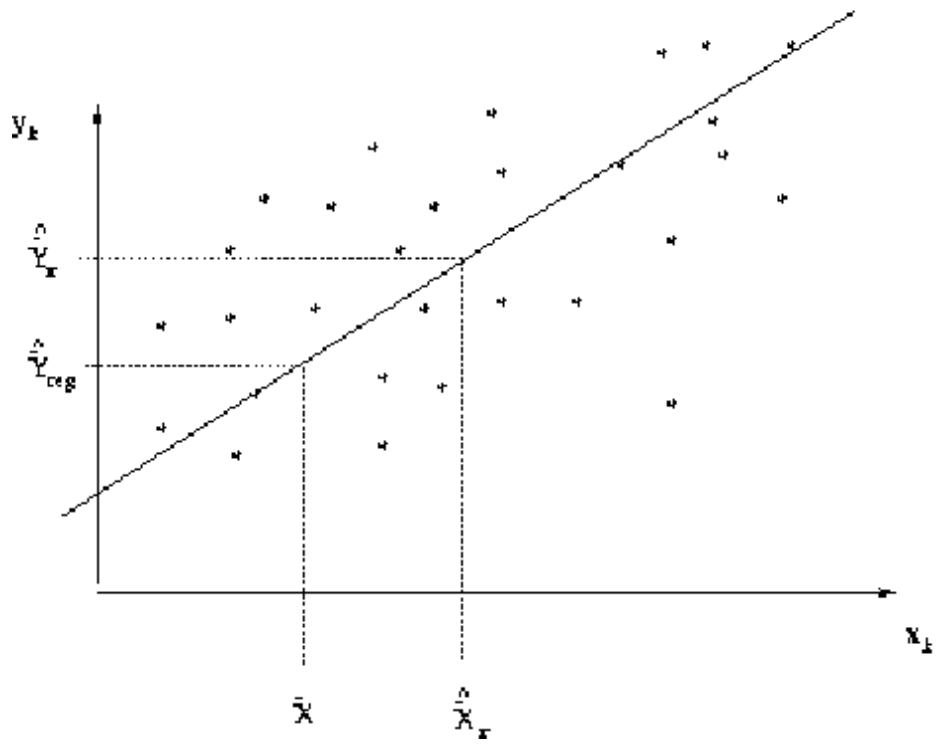
где $\hat{Y}_{reg}$ - оценка по регрессии;

$\hat{Y}_{\pi}$ - базовая оценка среднего значения признака (*x*), соответствующая плану выборки ;

$\hat{X}_{\pi}$ - базовая оценка среднего значения признака (*y*), соответствующая плану выборки ; $\hat{b}$ - оценка коэффициента регрессии b , который также должен быть оценен по данным выборки (коэффициент b характеризует наклон линии регрессии).

Соответственно оценка по регрессии суммарного значения определяется следующим образом:

$$\hat{Y}_{reg} = \hat{Y}_{\pi} + (X - \hat{X}_{\pi})\hat{b}$$

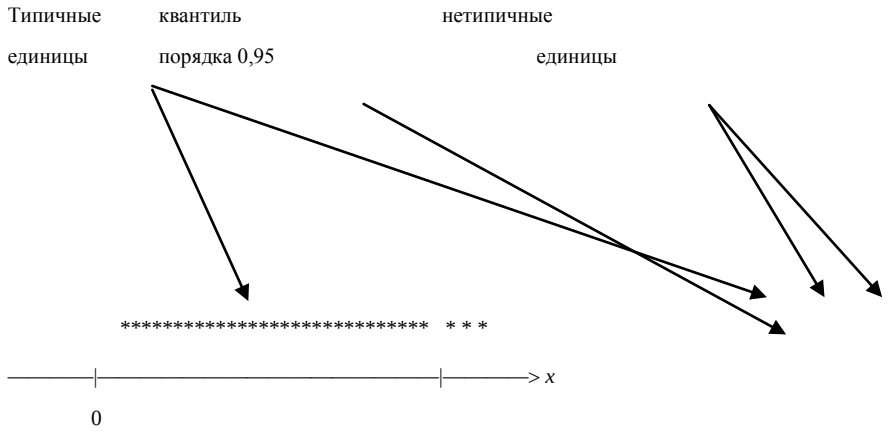

y_k
$\hat{Y}_x$
$\hat{Y}_{reg}$
x_k
$\bar{x}$
$\hat{X}_x$

1.10. Введение в проблему нетипичных единиц и импутацию неполных данных

До проведения статистического наблюдения нетипичные единицы можно идентифицировать по данным основы выборки. В силу своей исключительности такие единицы не могут «представлять» собой другие, не включенные в выборку, согласно выборочному плану. Это привело бы к значительному смещению агрегированных результатов обследования.

Например, нетипичными единицами являются единицы совокупности, имеющие подавляющий вклад в некоторый показатель наблюдения, либо исключительные в каком-либо другом качественном смысле.

Практически в выборочных обследованиях нетипичные единицы определяются в группах классификации как единицы со значением базового признака (например, численность населения населенного пункта) выше установленной верхней границы. В качестве такой границы обычно выступает квантиль порядка 0.95.



Нетипичные единицы, выявленные до начала проведения обследования, должны быть помещены в отдельный массив (слой) и специально обрабатываться.

Для обработки значений признаков нетипичных единиц, выявленных после получения отчетов (свыше установленной квантили распределения исследуемого показателя, например, 0,95), веса этих единиц уменьшаются, а веса других единиц в данной группе (слое) увеличиваются.

При расчете статистических показателей по регламентным разрезам единицы, идентифицированные как нетипичные, обрабатываются специальным образом. Им обычно присваивается единичный выборочный вес, с соответствующим перевзвешиванием других единиц в слоях. Также нетипичные единицы не учитываются при определении средних значений признаков в слоях.

При проведении любого статистического обследования на этапе обработки собранных данных приходится сталкиваться с проблемой пропусков в собранной пообъектной информации. Пропуски в данных связаны: 1) с наличием ошибок, выявленных на этапе ввода и проведения логического и арифметического контроля,

которые не удается устранить из-за невозможности связаться с респондентом по различным объективным причинам; 2) из-за недостижимости респондентов или отказа предоставить запрашиваемую информацию – неответы респондентов.

В обследованиях различают два вида неответов респондентов – полные, когда информация об объекте наблюдения отсутствует полностью, и частичные, когда отсутствуют ответы на отдельные вопросы программы обследования. Как уже отмечалось причиной полных неответов может являться недостижимость респондента или отказ от участия в обследовании. В обоих случае крайне важно установить по крайней мере сам факт осуществления респондентом деятельности. Причинами же неответов на отдельные вопросы обследования может являться следующее:

- респондент отказывается или затрудняется ответить на отдельные вопросы анкеты;

интервьюер не задает некоторые вопросы обследования или некорректно регистрирует ответы;

- ответ респондента вступает в конфликт, логическое противоречие, с ответами на другие вопросы или не соответствует предусмотренным категориям ответов (ошибки в данных);
- оператор допускает ошибки и пропуски при вводе данных.

Для исправления ошибок заполнения в данных выборки и обработки случаев неответов респондентов обычно применяется два подхода – метод перевзвешивания (заключается в корректировке выборочных весов) и метод импутации (вменение конкретных величин пропущенным значениям). В проводимых обследованиях обычно комбинируются оба метода.

Так нецелесообразно применение методов импутации для генерации отсутствующей информации, что вызвано фактами полных неответов респондентов. Основным методом их учета при вычислении оценок показателей обследования обычно является корректировка выборочных весов собранных наблюдений. При этом для расчета новых весов в обследовании индивидуальных предпринимателей лучше всего использовать внешнюю информацию о количестве действующих предпринимателей на территории субъекта РФ из базы данных МНС РФ.

Далее детально рассматривается задача корректировки результатов обследования в связи с неответами респондентов на отдельные вопросы обследования.

Частичные пропуски в данных – неответы на отдельные вопросы обследования - являются потенциальным источником следующих проблем:

трудности при получении статистических оценок показателей по неполным наборам данных, в особенности, если оценка основывается на нескольких учетных признаках;

несогласованность оценок показателей обследования по причине различий в базе для их получения;

смещение оценок, вызванное тем, что структуры пропусков в данных не являются случайными.

Процесс заполнения пропущенных значений называют импутацией. Различают три основных подхода к проведению импутации: детерминистский метод, называемый также редактированием; методы, основанные на использовании данных донора; и методы

вычисления оценок отсутствующих значений. Классификация методов импутации приведена на следующей диаграмме (см. рис.5).

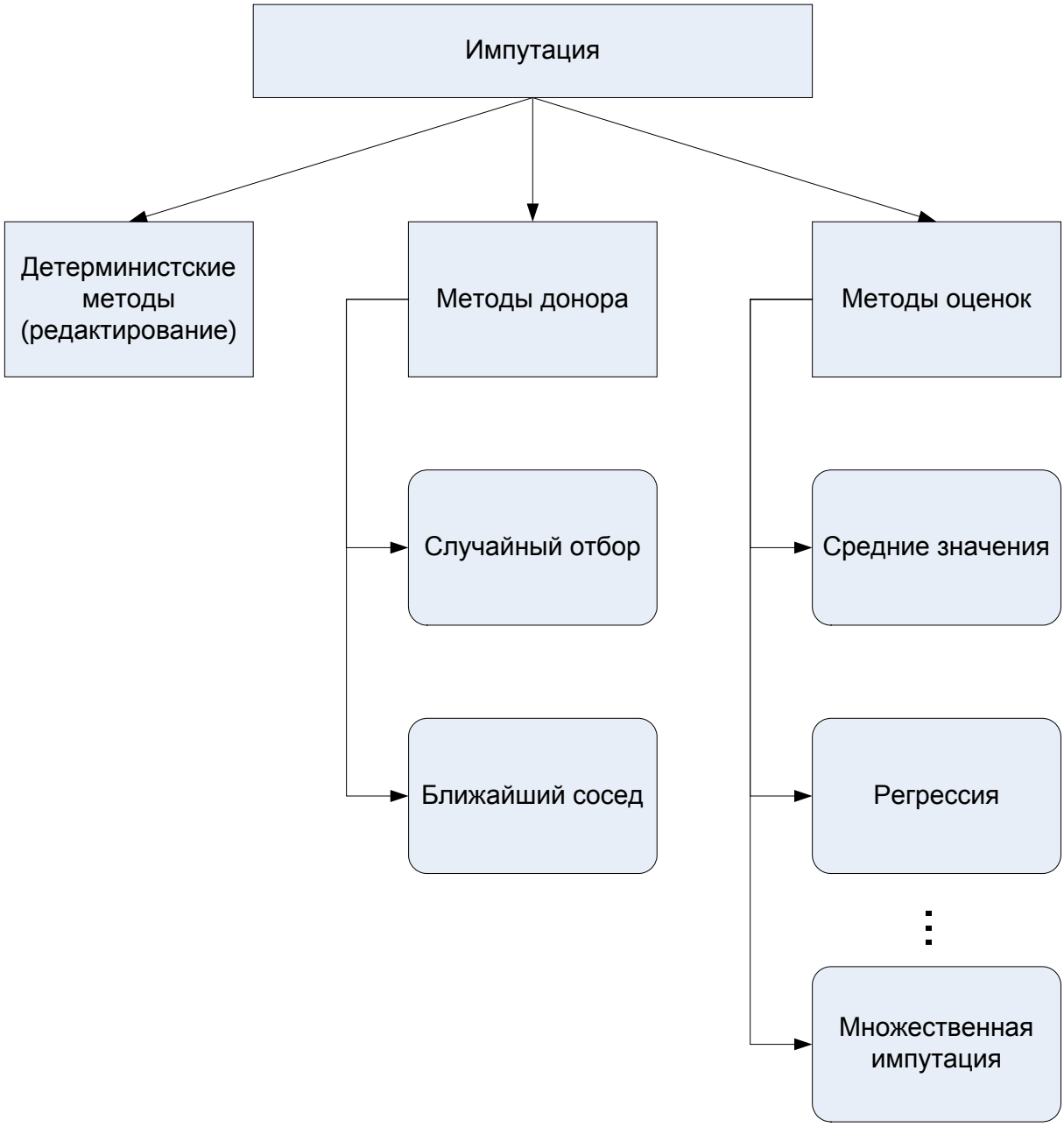


Рисунок. 5. Классификация методов импутации

Детерминистская импутация – редактирование.

Детерминистский метод импутации называют также редактированием данных и заключается он в определении импутируемого значения из логических и арифметических соотношений между переменными.

Методы донора. В методах донора импутируемое значение для записи-реципиента берется из другой записи-донора, выбираемой по определенным правилам. В зависимости от способа выбора записи-донора различают несколько разновидностей метода донора, основными из которых являются случайный выбор в классах (группах) и выбор ближайшего по некоторой метрике соседа. К записям, используемым в качестве донорских, предъявляются некоторые требования. Обычно эти требования заключаются в том, чтобы донорская запись удовлетворяла всем правилам редактирования и содержала только ответы респондента, то есть не содержала импутированных значений.

Случайный выбор донора. Случайный выбор записей, используемых в качестве доноров, обычно осуществляется в некоторых специально определенных классах. Классы представляют собой достаточно однородные по заданным параметрам группы наблюдений.

В качестве критериев отнесения записей с неответами респондентов к классам импутации обычно применяются некоторые вспомогательные данные, имеющиеся по каждому объекту наблюдения. Это могут быть отрасли экономики, виды деятельности, географические районы и другие данные.

Если в записи-реципиенте подлежат импутации несколько полей, то они все берутся из одной отобранной записи-донора. Это обеспечивает сохранение соотношений между переменными и гарантирует выполнение правил редактирования.

Метод ближайшего соседа. Импутация методом ближайшего соседа предполагает выбор такой записи из числа донорских, которая имеет наименьшее расстояние от импутируемой записи. В качестве метрики для определения расстояния могут выбираться различные функции, возможно, учитывающие некоторые априорные или исторические сведения о взаимосвязанности некоторых параметров.

В методе ближайшего соседа необходимо определить не только метрику, но и переменные по которым будет вычисляться расстояние. Для каждой требующей импутации записи набор переменных, по которым производится подбор ближайшего соседа, может быть различным. Кроме того, может оказаться, что все переменные, присутствующие в метрике, требуют импутации в данной записи. В этом случае донор может быть подобран по другому критерию, например, методом случайного выбора.

Не рассматривая различные примеры определения расстояний между записями, отметим только, что при реализации данного метода часто применяются различные преобразования. Целью таких преобразований является устранение возможного преобладания в выборе одного из параметров вследствие различных единиц измерения. В качестве преобразования может применяться нормирование, ранжирование и другие.

Методы оценок. Существует множество методов оценок, применяемых для импутации пропущенных или несостоятельных значений в данных. Общая идея этих методов заключается в подстановке вместо отсутствующего значения некоторой оценки, которая может основываться на данных текущего обследования или на исторических данных предыдущих обследований. Следует отметить, что в случае применения любого из методов оценок всегда подразумевается в явном или неявном виде применение некоторой модели данных. Поэтому выбор той или иной оценки должен основываться на некоторых априорных или экспериментальных данных относительно модели. В случае применения

методов оценок не гарантируется, что после импутации методом оценок запись будет удовлетворять всем правилам редактирования. Объясняется это тем, что оценки в основном вычисляются независимо для каждой импутируемой переменной. Поэтому после импутации методом оценок может потребоваться повторная проверка выполнения правил редактирования и, возможно, повторная импутация.

2. Методы выборочных обследований: численные примеры

2.1. Процедуры отбора простой случайной и систематической выборки номеров с помощью таблицы случайных чисел.

Для получения случайной последовательности целых чисел в интервале от 1 до заданного значения, например, между 1 и 273, можно выполнить следующие действия:

Выясняем количество цифр в максимальном числе заданного интервала. Так в числе «273» имеется три значащие цифры.

В таблице случайных чисел произвольно выбираем столько же колонок цифр сколько в максимальном числе промежутка.

В примере 3 колонки. При этом каждая строка в наборе выбранных колонок задает случайное число.

Начиная с верхней (или любой другой) строки выбираем целые числа, номера единиц, которые принадлежат заданному интервалу. Числа большие максимального значения в интервале пропускаются.

В примере выбираем любые встреченные числа от «001» до «273», а остальные пропускаем, например, «575».
Примечание. В практических целях число «0» интерпретируется как «10», «00» - как «100», «000» - как «1000» и т.д.

Движение по выбранным колонкам вниз (или вверх), причем с любой строки, продолжается до тех пор, пока не будут выбраны случайные числа в необходимом количестве.

Если встречается уже отобранное случайное число, то оно пропускается (в случае отбора без возвращения).

Пример 1.

С помощью таблицы случайных чисел:

А) нужно отобрать 3 единицы из 10.

Так как число «10» обозначается как «0», то в таблице случайных чисел нужно произвольно выбрать одну колонку. Например, №5 (см. следующую табл.). Отбор начнем с первой строки.

Фрагмент таблицы случайных чисел:

№ колонки	
1234	5678
____	____

1089 8719 …

9385 7902 …

6934 8660 …

0052 1007 …

5736 9249 …

1901 5988 …

5372 6212 …

… …

«8» – единица №8 включается в выборку;

«7» – единица №7 включается в выборку;

«8» – единица №8 уже отобрана;

«1» – единица №1 включается в выборку.

В результате в выборку включены единицы: №№ «1», «7» и «8».

Б) нужно отобрать 5 единиц из 80.

Так как число «80» записывается 2 цифрами, то в таблице нужно произвольно выбрать две колонки, например, №№1 и 2. Начинаем отбор с первой строки:

№ колонки

1234 5678

____ ____

1089 8719 …

9385 7902 …

6934 8660 …

0052 1007 …

5736 9249 …

1901 5988 …

5372 6212 …

… …

«10» – единица №10 включается в выборку;

«93» – единица 93 > 80, поэтому пропускаем;

«69» – единица №69 включается в выборку;

«00» – это число 100 > 80, поэтому пропускаем;

«57» – единица №57 включается в выборку;

«19» – единица №19 включается в выборку;

«53» – единицу №53 включается в выборку.

В результате в выборку включены единицы: №№10, 19, 53. 57, и 69.

Пример 2. Процедура формирования систематической выборки.

Пусть генеральная совокупность состоит из 285 (N) элементов. Требуется сформировать систематическую выборку объема 12 единиц (n). Предполагается, что список единиц совокупности упорядочен случайно.

Вычисляем длину интервала отбора: $q = N / n = 285/12 = 23{,}75$ (в значении q 4 значащих цифры).

В таблице случайных чисел выбираем 4 колонки и определяем случайное число. Пусть это 1979. Данное число задает точку начала отбора: $r_1 = 19{,}79$.

Осуществляем отбор каждой q–ой единицы списка, начиная с элемента, отобранного первым.

№ п/п	Уровни кумулятивного ряда	№ отобранной единицы: (целая часть числа) + 1
1	19,79	20
2	43,54 = 19,79 + 23,75	44
3	67,29 = 43,54+23,75	68
4	91,04 = 67,29+23,75	92
5	114,79	115
6	138,54	139
7	164,29	165
8	186,04	187
9	209,79	210
10	233,54	234
11	257,29	258
12	281,04	282
13	304,79	305 > 285

Замечание.

«Отбрасывать» дробную часть в значении интервала отбора q нельзя. Если в примере значение q округлить до целых (24) и если начало отбора также окажется равным 24, тогда получим следующую систематическую выборку:

24 48 72 96 120 144 168 192 216 240 264 288
(номер 12-ой единицы выборки окажется больше N = 285).

2.2 Сравнение простой случайной и расслоенной выборки.

Пример 3. В следующей таблице представлены данные генеральной совокупности состоящей из 5 элементов.

Расслоение	Слой 1			Слой 2	
Совокупность {U}	(1)	(2)	(3)	(4)	(5)
Признак (y)	13	15	17	25	30

Простая случайная выборка.

Характеристики совокупности:

$$N = 5 \qquad \overline{Y} = 20 \qquad S_y^2 = \frac{1}{N-1}\sum_{k \in U}\left(y_k - \overline{Y}\right)^2 . = 52$$

Если извлекать простую случайную выборку объема 2 из всей совокупности {U}, то возможно получить одну из 10 выборок:

y_1, y_2	13, 15	13, 17	13,25	13,30	15,17	15, 25	15,30	17,25	17,30	25,30
$\overline{y}$	14	15	19	21. 5	16	20	22. 5	21	23.5	27. 5

Арифметическое среднее 10-ти выборочных средних равно 20 (свойство несмещенности выборочного среднего)

Дисперсия оценки: $V(\overline{y}) = (1 - \frac{n}{N})\frac{S^2}{n} = \left(1 - \frac{2}{5}\right)\frac{52}{2} = 15.6$,

Следовательно коэффициент вариации оценки будет равен

$$CV(\overline{y}) = \frac{\sqrt{V(\overline{y})}}{\overline{Y}}100\% = \frac{\sqrt{15.6}}{20}100\% = 19.7\%$$

Расслоенная случайная выборка.

Характеристики слоев:

$$N_1 = 3 \qquad \overline{Y_1} = 15 \qquad S_1^2 = 4$$

$$N_2 = 2 \qquad \overline{Y_2} = 27.5 \quad S_2^2 = 12.5$$

Из каждого слоя отбираем по одной единице.

Всего возможно извлечь 6 выборок:

$y_1 \;\; y_2$ $\qquad \overline{y} = \frac{1}{N}\sum_h w_h y_h$

$$\overline{y}_1 = \frac{1}{5}\left(3 \cdot 13 + 5 \cdot 25\right) = 17.8$$

13 25

$$\overline{y}_2 = 19.8$$

13 30

15 25 $\bar{y}_3 = 19$

15 30 $\bar{y}_4 = 21$

17 25 $\bar{y}_5 = 20.2$

17 30 $\bar{y}_6 = 22.2$

Арифметическое среднее 6-ти выборочных средних:

$\bar{y}$ = 20 (оценка также несмещенная)

Дисперсия оценки $\bar{y}$:

$$V(\bar{y}) = \sum_h \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}$$

$$= \left(\frac{3}{5}\right)^2 \left(1 - \frac{1}{3}\right) \frac{4}{1} + \left[\frac{2}{5}\right]^2 \left(1 - \frac{1}{2}\right) \frac{12.5}{1} = 1.96$$

Соответственно коэффициент вариации оценки $\bar{y}$:

$$CV(\bar{y}) = \frac{\sqrt{V(\bar{y})}}{\bar{Y}} 100\% = \frac{\sqrt{1.96}}{20} 100\% = 7\%$$

Пример 4.

Свободные места на авиалайнерах являются причиной потери дохода авиакомпаний. Предположим авиакомпании требуется оценить среднее количество свободных пассажирских мест в расчете на рейс за последний год. Для этого из файла базы данных, содержащего 4500 (*N*) записей о рейсах, были отобраны случайным образом 225 (*n*) записей. Вычисленное среднее значение и среднее квадратическое отклонение составили:

$\bar{y} = 11.6$ мест и $s = 4.1$ места.

Тогда с помощью 90%-го доверительного интервала среднее количество свободных мест на рейс ($\bar{Y}$) оценивается как

$$\bar{y} \pm z_{0.9} \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N}} = 11.6 \pm 1.64 \frac{4.1}{\sqrt{225}} \sqrt{\frac{4500-225}{4500}}$$

$$= 11.6 \pm 0.44 \Leftrightarrow (11.16; 12.04)$$

Если пренебречь поправкой на конечность совокупности, что считается допустимым при $n \le 0{,}05N$, то соответствующий доверительный интервал составит

$$11{,}6 \pm 0.45 \Leftrightarrow (11.15; 12.05)$$

Оценим общее количество свободных мест на рейсах за прошедший год (Y) с помощью 90%-го доверительного интервала.

$$\hat{Y} \pm z_{0.9}\frac{Ns}{\sqrt{n}}\sqrt{\frac{N-n}{N}} = 11.6\cdot 4500 \pm 1.64\frac{\cdot 4500\cdot 4.1}{\sqrt{225}}\sqrt{\frac{4500-225}{4500}} =$$

$$= 52200 \pm 1972 \Leftrightarrow [50227;54172]$$

Если пренебречь поправкой на конечность совокупности, то соответствующий доверительный интервал будет

$$52200 \pm 2025 \Leftrightarrow [50175;54225].$$

Определим необходимый объем выборки, для которого с 90-% доверительной вероятностью истинное среднее значение числа свободных мест на авиарейсах не отличалось бы от оценки более чем на 0,1.

Для определения объема выборки используется следующая формула:

$$n = \frac{\frac{z^2 s_y^2}{L^2}}{1+\frac{z^2 s_y^2}{NL^2}} = \frac{n_0}{1+\frac{n_0}{N}}, \text{ при } n_0 = \frac{z^2 s_y^2}{L^2}$$

Тогда

$$n_0 = \frac{1.64^2 4.1^2}{0.1^2} = 4521.2176$$

$$n = \frac{4521.2176}{1+\frac{4521.2176}{4500}} = 2255.3$$

Следовательно, для решения поставленной задачи объем выборки (n) должен составить не менее 2256 записей в файле данных.

Пример 5.

Предположим, что цель состоит в оценке доли работников, страдающих некоторым профессиональным заболеванием, относительно всех работающих на предприятии с численностью 1500 человек. Известно, что обычно на предприятиях такого типа этой болезнью страдают три работника из десяти. Какого объема должна быть выборка, чтобы общая длина 0,95-го доверительного интервала не превышала бы 0,02 (предполагается простой отбор с возвращением и без возвращения)?

Решение.

1. Отбор с возвращением.

Длина доверительного интервала для среднего значения вычисляется как

$$\Delta(0,95) = \left[\hat{Y} - z_{1-\alpha/2}\sqrt{\frac{\tilde{s}_y^2}{m}}, \hat{Y} + z_{1-\alpha/2}\sqrt{\frac{\tilde{s}_y^2}{m}}\right].$$

Если обозначить через $\hat{p}$ оценку доли при отборе с возвращением, то можно записать

$$\Delta(0{,}95) = \left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{m-1}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{m-1}} \right].$$

Так как в этом случае оценкой дисперсии будет

$$Var(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{m-1}.$$

Для того, чтобы общая длина доверительного интервала не превышала бы 0,02 необходимо, чтобы

$$2z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{m-1}} \le 0{,}02.$$

После деления на 2 и возведения в квадрат, получим

$$z^2_{1-\alpha/2}\frac{\hat{p}(1-\hat{p})}{m-1} \le 0{,}0001,$$

что дает

$$m-1 \ge z^2_{1-\alpha/2} \times \frac{\hat{p}(1-\hat{p})}{0{,}0001}, \qquad m = 1 + 1{,}96^2 \times \frac{0{,}3 \times 0{,}7}{0{,}0001} = 8068{,}36.$$

Рассчитанный объем выборки m = 8069 оказался больше, чем объем всей совокупности.

2. Отбор без возвращения.

Длину доверительного интервала для среднего значения можно выразить как

$$\Delta(0{,}95) = \left[\bar{Y} - z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{s_y^2}{n}}, \bar{Y} + z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{s_y^2}{n}} \right].$$

Если $\hat{p}$ - оценка доли, то при отборе без возвращения имеем:

$$\Delta(0{,}95) = \left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{\hat{p}(1-\hat{p})}{n-1}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{\hat{p}(1-\hat{p})}{n-1}} \right].$$

Чтобы общая длина доверительного интервала не превышала бы 0,02, необходимо

$$2z_{1-\alpha/2}\sqrt{\frac{N-n}{n}\frac{\hat{p}(1-\hat{p})}{n-1}} \le 0{,}02.$$

Разделив на 2 и возведя в квадрат, получим

$$z^2_{1-\alpha/2}\frac{N-n}{N}\frac{\hat{p}(1-\hat{p})}{n-1} \le 0{,}0001.$$

Что дает

$$n \ge \frac{0{,}0001 + 1{,}96^2 \times 0{,}30 \times 0{,}70.}{\left\{0{,}0001 + 1{,}96^2 \times \frac{1}{1500} \times 0{,}30 \times 0{,}70\right\}} = 1264{,}98$$

В данном случае объем выборки равный 1265 оказался меньше, чем объем всей совокупности. Влияние поправки на конечность совокупности может, следовательно, иметь решающее значение при выборке небольшого объема.

Вообще, если требуется оценить генеральную долю *P* с помощью выборочной доли *p* с заданной абсолютной погрешностью *L*, тогда так как

$$L = z_{1-\alpha/2}\sqrt{(1-\frac{n}{N})\frac{p(1-p)}{n-1}}$$

Следовательно

$$n = \frac{1+\frac{z^2 p(1-p)}{L^2}}{1+\frac{z^2 p(1-p)}{NL^2}} = \frac{1+n_0}{1+\frac{n_0}{N}}, \text{ при } n_0 = \frac{z^2 p(1-p)}{L^2}$$

2.3. Кластерная выборка.

Пример 6.

Предположим, что газету выписывают 40 000 подписчиков.

Требуется оценить количество подписчиков, которые захотят также подписаться и на новое приложение к газете (т.е. нужно оценить эффективность кросспродаж). Выборочный опрос подписчиков решено провести методом кластерной выборки. Для этого файл с учетными карточками подписчиков был упорядочен по переменной почтового индекса, после чего всех подписчиков сгруппировали в кластеры по 10 человек. Всего получилось $M = 4000$ кластеров.

Из совокупности кластеров простым случайным образом были отобраны 80 кластеров (m). Таким образом, по выборке были опрощены 800 подписчиков.

По выборке были получены следующие результаты:

кластер 1 - 1 положительный отклик;
кластер 2 - 2 положительных откликов;
.....
кластер 8 - 8 положительных откликов;
кластер 9 - 1 положительный отклик.
.....

Тогда для переменной положительных откликов имеем:

$$\sum_{i=1}^{80} Y_i = 360 \qquad \sum_{i=1}^{80} Y_i^2 = 2040$$

Оценим количество подписчиков, которые будут выписывать приложение к газете:

$$\hat{Y} = \frac{M}{m}\sum_{i=1}^{m} Y_i = \frac{4000}{80}\sum_{i=1}^{80} Y_i = 18000$$

95%-ый доверительный интервал вычисляется через вариацию оценки:

$$s_{Y_i}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \hat{\bar{Y}})^2 = \frac{1}{m-1}\left[\sum_{i=1}^{m} Y_i^2 - m(\hat{\bar{Y}})^2\right] = 5{,}32$$

$$V(\hat{Y}) = M^2(1-\frac{m}{M})\frac{1}{m}s^2_{Y_{cp}} = 1042720 = (1021)^2$$

95%-ый доверительный интервал для оценки суммарного значения:

18000 ± 1,96 * 1021 = 18000 ± 2001

95% доверительный интервал для оценки доли:

45 % ± 5.0 %

Замечание.

Если оценивать долю согласных подписчиков на основе формул простой случайной выборки при

n = 800 и N = 40000, тогда вариация оценки будет:

$$V(\hat{p}) = (1-\frac{n}{N})\frac{1}{n-1}\hat{p}(1-\hat{p}) = (0{,}017420)^2$$

95%-ый доверительный интервал: 45% ±3,4%

Доверительный интервал меньше!

Эффективность кластерной выборки меньше, чем простой случайной. Это связано с отличием кластеров между собой.

Для повышения точности следовало бы либо расслоить кластеры на однородные группы, либо использовать отбор с вероятностью пропорциональной размеру кластеров (если бы такая информация имелась до проведения обследования).

2.4. Алгоритм оптимального расслоения.

В случае использования расслоенной случайной выборки необходимо задать:

- группы по значениям качественной и/или количественной переменной);
- размещение объема выборки по слоям.

Пример 7.

Правило расслоения по количественной переменной Экмана (1959).

Границы слоев определяются так, чтобы

$N_h(c_h - c_{h-1}) = const$

где $c_h, h = 1,..., H$ - границы слоев.

$c_0 = x_{\min}$, а $c_H = x_{\max}$;

x – вспомогательный признак.

Нетрудно убедиться в том, что в примере 3 оптимальные границы следующие:

$c_0 = x_{\min} = 13;\ c_1 = 19.8;\ c_2 = x_{\max} = 30,$

Действительно

$N_1(c_1 - c_0) = N_2(c_2 - c_1)$

$3(19.8 - 13) = 2(30 - 19.8_1) = 20.4$

2.5. Размещение по слоям общего объема выборки.

Пример 8.

Пусть имеется совокупность, состоящая из 1060 предприятий. Требуется оценить их средний оборот по данным расслоенной выборки объема 300 (*n*) единиц. Для повышения точности оценки совокупность была расслоена на 5 слоев по вспомогательной переменной (*x*) численности персонала, известной за предыдущий год. В следующей таблице приведены основные характеристики расслоенной генеральной совокупности.

Интервалы группировки	N_h	$\bar{x}_h$	S_h^2
0-9	500	5	1.5
10-19	300	12	4
20-49	150	30	8
50-499	100	150	100
500 и более	10	600	2500
Всего	1060	29,8	7803,7

1. Применение формул простой случайной выборки для вспомогательной переменной дает следующие результаты:

$$V(\bar{x}) = 11{,}59 \qquad CV_{\bar{x}} = 11{,}42\% \qquad L_{0.95}(\bar{x}) = 22{,}38\%$$

Ясно, что прогнозируемая точность оценки при простом случайном отборе недостаточна, так как точность оценки среднего целевой переменной будет еще хуже.

2. Определим пропорциональное размещение по слоям (n_h) и оценим соответствующую точность итогов.

Интервалы группировки	$n_h = n \cdot \frac{N_h}{N}$	Дисперсия оценки (в слое)	Коэффициент вариации оценки
0-9	142	0,0076	0,0174
10-19	85	0,0338	0,0153
20-49	42	0,1351	0,0123
50-499	28	2,5333	0,0106
500 и более	3	633,333	0,0419
Итого	300	0,0378	0,0065

Тогда

$$V(\bar{y}) = 0{,}0378 \qquad CV_{\bar{y}} = 0{,}6518\% \qquad L_{0.95}(\bar{x}) = 1{,}28\%$$

Эффект плана соответственно составляет:

$$DEFF(\bar{Y}_{srs}) = \frac{V(\bar{Y}_{prop})}{V(\bar{Y}_{srs})} \cong \frac{0{,}0378}{11{,}59} = 0{,}0033$$

3. Оптимальное размещение объема выборки по слоям.

Оптимальное размещение задается следующей формулой:

$$n_h = n \cdot \frac{N_h S_h}{\sum_{k=1}^{H} N_k S_k}.$$

Тогда по данным таблицы имеем:

$$\sum_h N_h S_h = 500\sqrt{1,5} + 300\sqrt{4} + 150\sqrt{8} + 100\sqrt{100} + 10\sqrt{2500} = 3136,64$$

Применяем формулу оптимального размещения с $n = 300$:

$$\begin{cases} n_1 = 58.56 = n \times \dfrac{N_1 S_1}{\sum_h N_h S_h} = 300 \times \dfrac{500\sqrt{1,5}}{3136,64} \\ n_2 = 57.38 \\ n_3 = 40.57 \\ n_4 = 95.64 \\ n_5 = 47.82 > N_5 = 10 \Rightarrow n_5 = 10 \end{cases}$$

Включаем слой 5 целиком в выборку и повторяем расчеты с $n = 290$:

$$\begin{cases} n_1 = 67.35 = n \times \dfrac{N_1 S_1}{\sum_{h=1}^{4} N_h S_h} = 290 \times \dfrac{500\sqrt{1,5}}{2636,64} \\ n_2 = 65.99 \\ n_3 = 46.66 \\ n_4 = 109.98 > N_4 = 100 \Rightarrow n_4 = 100 \\ (n_5 = 10) \end{cases}$$

Включаем также слой 4 целиком в выборку и повторяем расчеты с $n = 190$:

$$\begin{cases} n_1 = 71.09 = n \times \dfrac{N_1 S_1}{\sum_{h=1}^{3} N_h S_h} = 190 \times \dfrac{500\sqrt{1,5}}{1636,64} \Rightarrow n_1 = 71 \\ n_2 = 69.65 \Rightarrow n_2 = 70 \\ n_3 = 49.25 \Rightarrow n_3 = 49 \\ (n_4 = 100) \\ (n_5 = 10) \end{cases}$$

В случае оптимального размещения получаем следующие характеристики точности:

$V(\bar{x}) = 0,0043$ $\qquad CV_{\bar{x}} = 0,2194\%$ $\qquad L_{0.95}(\bar{x}) = 0,43\%$

Эффект плана соответственно составляет:

$$DEFF(\bar{Y}_{srs}) = \frac{V(\bar{Y}_{opt})}{V(\bar{Y}_{srs})} \cong \frac{0,0043}{11,59} = 0,0004$$

2.6. Оценивание показателей генеральной совокупности по данным выборки с использованием вспомогательной информации.

Пример 9.

Требуется оценить среднее значение $\overline{Y}$ признака (y) на основе выборки объема 1000 единиц, сформированной методом простого случайного отбора без возвращения из совокупности в 1000000 единиц. При этом пусть известно среднее значение $\overline{X} = 15$ вспомогательного признака (x).

Пусть по выборке получены следующие результаты:

$$\overline{y} = 10,\ s_y^2 = 20,\ \overline{x} = 14,\ \ s_x^2 = 25,\ \ s_{xy} = 15,.$$

Оценим среднее значение $\overline{Y}$ признака.

При простом случайном отборе π-оценка имеет вид:

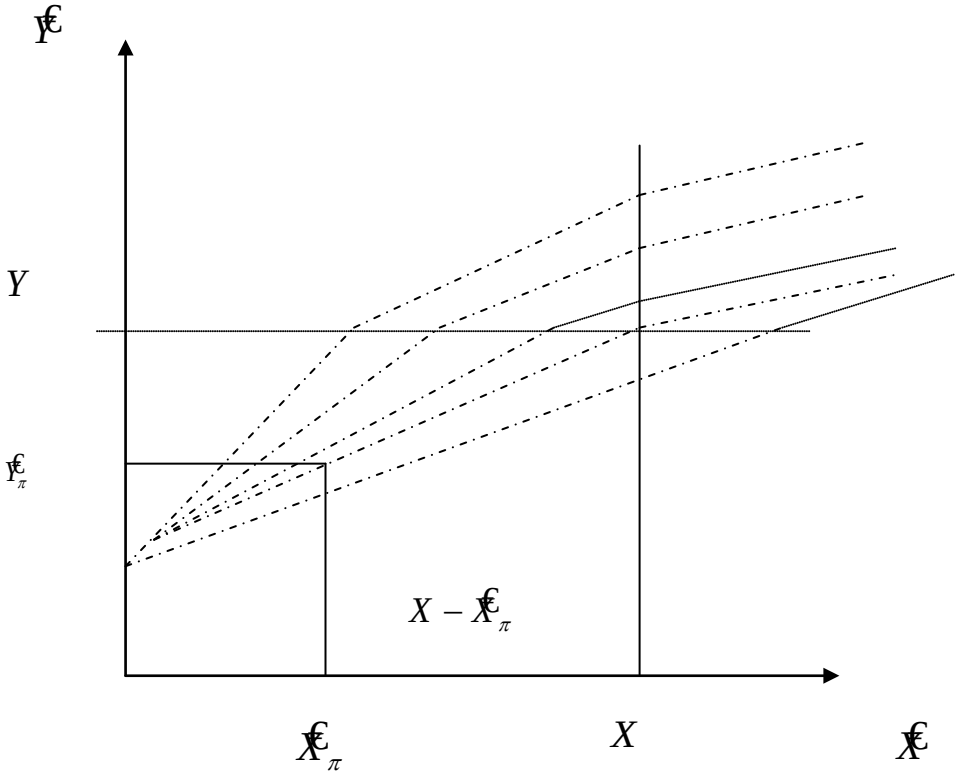
$$\hat{\overline{Y}}_\pi = \overline{y} = 10 : \hat{Y}_\pi = N \cdot \hat{\overline{Y}} = 10^6 \cdot 10 = 10^7$$

Соответствующая оценка дисперсии оценки среднего:

$$\hat{V}(\hat{\overline{Y}}_\pi) == \frac{s_y^2}{n}\left(\frac{N-n}{N}\right) = \frac{20}{10^3}\left(\frac{10^6 - 10^3}{10^6}\right) = 0{,}01998 \cdot$$

Оценивание с помощью корректирующей разности:

$$\hat{Y}_D = \hat{Y}_\pi + (X - \hat{X}_\pi)$$



Эту оценку целесообразно использовать, когда разность $Y - \hat{Y}_\pi$ в среднем приблизительно равна разности $X - \hat{X}_\pi$.

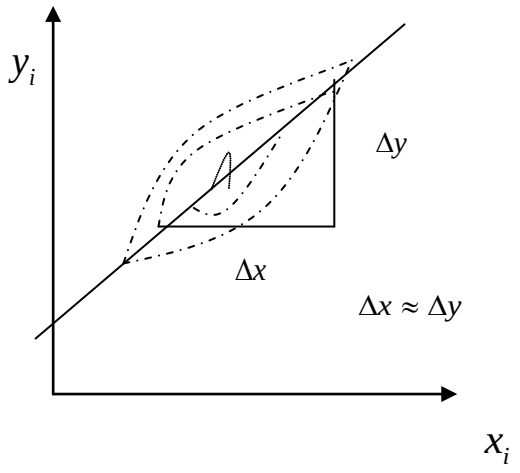
Данные необходимые для оценивания по разности.

Цель заключается в оценке суммарного значения показателя генеральной совокупности Y;

Для каждого элемента в выборке имеются значения признака *y* (данные наблюдения) и *x* (вспомогательные данные);

Известно суммарное значение *X* вспомогательного признака;

Имеется приблизительно линейная зависимость между признаками y_i и x_i, а коэффициент регрессии близок к 1.0 (или –1.0).



Пример 9 (продолжение).

Оценка с помощью разности выражается формулой:

$$\hat{Y}_D = \hat{Y}_\pi + (X - \hat{X}_\pi) = 10 + (15 - 14) = 11$$

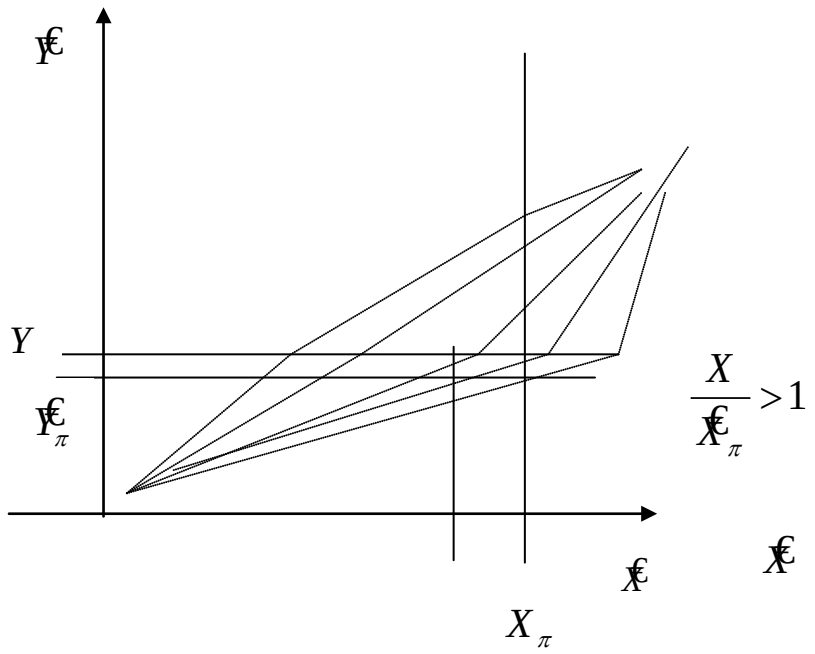
Соответствующая оценка дисперсии этой оценки:

$$\hat{V}(\hat{Y}_D) = = \left(\frac{N-n}{nN}\right)(s_y^2 - 2s_{xy} + s_x^2)$$

$$= \left(\frac{10^6 - 10^3}{10^3 \cdot 10^6}\right)(20 - 2 \cdot 15 + 25) = 0{,}014985.$$

Оценивание с помощью корректирующего отношения:

$$\hat{Y}_{rat} = \hat{Y}_\pi \cdot (X / \hat{X}_\pi)$$



Оценку по отношению целесообразно использовать, если:

разность $Y - \bar{Y}_\pi$ в среднем приблизительно пропорциональна разности $X - \bar{X}_\pi$;

и $\bar{Y}_\pi = 0$ в тех случаях, когда $\bar{X}_\pi = 0$.

Данные необходимые для оценивания по отношению.

Цель заключается в оценке суммарного значения показателя Y генеральной совокупности;

Для каждого элемента в выборке имеются значения признака y (данные наблюдения) и x (вспомогательные данные);

Известно суммарное значение X вспомогательного признака;

Имеется приблизительно пропорциональная зависимость между признаками y и x, причем линия регрессии приблизительно проходит через начало координат и коэффициент регрессии положителен.

Пример 9 (продолжение).

Оценка с помощью отношения выражается формулой:

$$\bar{Y}_{rat} = \bar{Y}_\pi \cdot (X / \bar{X}_\pi) = (10 \cdot 15)/14 = 10{,}7124$$

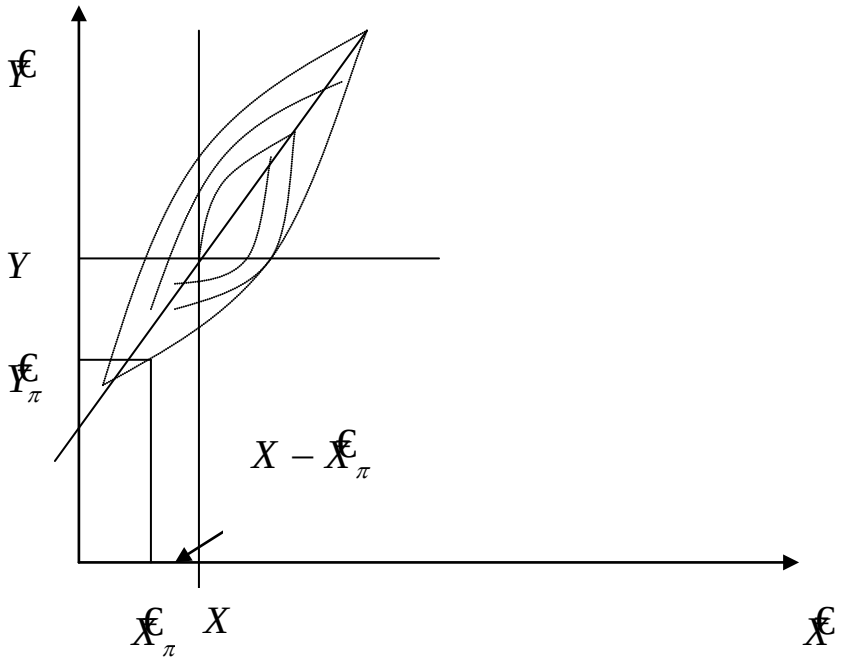
Соответствующая оценка дисперсии этой оценки:

$$V(\bar{Y}_{rat}) == \left(\frac{N-n}{nN}\right)(s_y^2 - 2\frac{\bar{Y}_\pi}{\bar{X}_\pi} s_{xy} + \frac{\bar{Y}_\pi^2}{\bar{X}_\pi^2} s_x^2)$$

$$= \left(\frac{10^6 - 10^3}{10^3 \cdot 10^6}\right)(20 - 2 \cdot \frac{10}{14} \cdot 15 + \frac{10^2}{14^2} \cdot 25) = 0{,}0113152.$$

Оценивание с помощью регрессии:

$$\bar{Y}_{reg} = \bar{Y}_\pi + B(X - \bar{X}_\pi)$$



Оценку по регрессии целесообразно использовать, если:

разность $Y - \hat{Y}_\pi$ в среднем приблизительно пропорциональна разности $X - \hat{X}_\pi$;

$\hat{Y}_\pi \neq 0$ даже когда $\hat{X}_\pi = 0$.

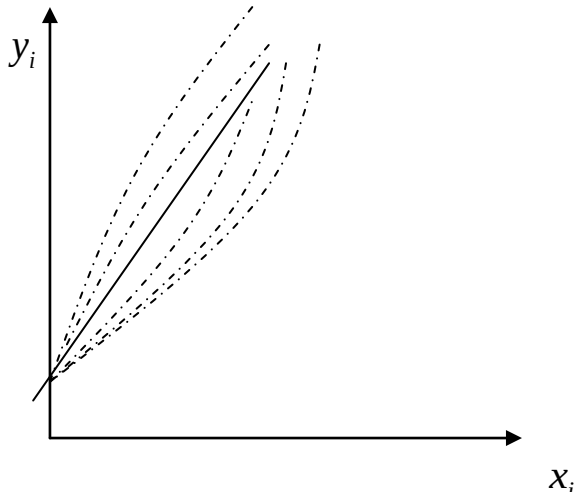
Данные необходимые для оценивания по разности.

Цель заключается в оценке суммарного значения показателя Y генеральной совокупности;

Для каждого элемента в выборке имеются значения признака y (данные наблюдения) и x (вспомогательные данные);

Известно суммарное значение X вспомогательного признака;

Имеется приблизительно пропорциональная зависимость между признаками y и x.



Пример 9 (продолжение).

Оценка с помощью регрессии выражается формулой:

$$\hat{Y}_{reg} = \hat{Y}_\pi + \frac{s_{xy}}{s_x^2}(X - \hat{X}_\pi) = 10 + \frac{15}{25}(15 - 14) = 10{,}6,$$

Соответствующая оценка дисперсии этой оценки:

$$\hat{V}(\bar{y}_{reg}) = \left(\frac{N-n}{nN}\right)\cdot s_y^2(1-r^2) = \left(\frac{N-n}{nN}\right)\cdot s_y^2\left(1-\frac{s_{xy}^2}{s_x^2\cdot s_y^2}\right)$$

$$= \left(\frac{10^6-10^3}{10^3\cdot 10^6}\right)\cdot 20\cdot\left(1-2\cdot\frac{15^2}{25^2\cdot 20^2}\right) = 0{,}010989.$$

3. Методика проведения обследований населения по проблемам занятости

3.1. Статистический инструментарий и организация сбора сведений обследования населения по проблемам занятости

Выборочные обследования населения по проблемам занятости проводятся в частных домашних хозяйствах во всех республиках, краях, областях, автономной области и округах Российской Федерации с целью получения данных о численности и составе экономически активного населения, занятых и безработных, экономически неактивного населения, уровне экономической активности, уровне занятости и уровне безработицы, продолжительности занятости и незанятости, способах поиска работы.

Обследования проводятся по состоянию на критическую (обследуемую) неделю.

Критическая (обследуемая) неделя длится с понедельника по воскресенье. С 1999 г. обследование проводилось один раз в квартал по состоянию на последнюю неделю второго месяца квартала, т.е. февраля, мая, августа, ноября. С сентября 2009 года обследование проводится ежемесячно.

Единицами наблюдения являются частные домашние хозяйства и лица в возрасте от 15 до 72 лет - члены этих домашних хозяйств. Минимальная и максимальная границы возраста обследуемого населения определены при проведении первого обследования в 1992 г. с учетом наличия экономически активного населения в этих возрастах.

Домохозяйством считается:

- один человек, проживающий в отдельном жилом помещении или части жилого помещения и обеспечивающий себя всем необходимым для жизни и не объединяющий средства для ведения общего хозяйства с кем-либо из других лиц, проживающих в данном жилом помещении;
- два человека или более, проживающие совместно в отдельном жилом помещении, части его или нескольких жилых помещениях и обеспечивающие себя всем необходимым для жизни посредством ведения общего хозяйства, полностью или частично объединяя и расходуя свои средства. Эти лица могут быть связаны отношениями родства или отношениями, вытекающими из брака, либо быть не родственниками, либо и теми, и другими.

- Лица, снимающие жилое помещение у отдельных граждан, в состав домохозяйства владельца жилого помещения не входят и считаются отдельными домохозяйствами.
- Лица, нанятые для работы по дому (домашняя прислуга) и проживающие в помещении своего работодателя, в составе домохозяйства своего работодателя не учитываются (независимо от того, что они за свой труд получают питание и проживание), а рассматриваются как отдельные домохозяйства.

В составе домашнего хозяйства могут быть учтены лица, не имеющие родственных связей с членами домашнего хозяйства, но проживающие в данном помещении и ведущие одно хозяйство (пожилые или другие лица, находящиеся на попечении хозяйства).

При проведении обследования в каждом домашнем хозяйстве учитываются лица, постоянно (обычно) проживающие по данному адресу, включая и тех лиц, которые в отчетный период и период обследования временно отсутствовали, а также постоянно проживающие в Российской Федерации иностранные граждане (т.е. лица, имеющие гражданство только зарубежного государства).

При формировании региональных итогов население учитывается по месту проживания.

Анкета выборочного обследования населения по проблемам занятости содержит вопросы по учетным признакам, которые задаются в логической последовательности. Анкета обследования и Порядок ее заполнения утверждаются Постановлением Росстата.

Кроме того, в территориальных органах Росстата ведется карточка на помещение, которая представляет собой список домохозяйств в пределах помещения с указанием количества членов домохозяйств (независимо от возраста, включая лиц моложе 15 и старше 72 лет), постоянно (обычно) проживающих в помещении, отобранном для проведения обследования.

Обследование проводится путем опроса населения и записи сведений в Карточку на помещение и Анкету.

В соответствии с целями обследования в Анкете собирается обезличенная информация по следующим основным учетным признакам:

а) сведения о респондентах: пол; возраст; семейное положение; гражданство; уровень образования; профессия или специальность, полученная по окончании учебного заведения; общее число членов домохозяйства; родственные отношения в домохозяйстве;

б) наличие оплачиваемой работы или доходного занятия: занятость в обследуемую неделю, включая наличие работы, которая временно не выполнялась;

в) признаки, характеризующие основную работу: вид экономической деятельности; занятие (профессия, должность); классификация по статусу занятых; количество работников на предприятии; количество привлекаемых наемных работников работодателями; вид основной работы и условия найма; региональное месторасположение основной работы; нормальная и фактическая продолжительность рабочей недели; причины работы меньше нормальной продолжительности рабочей недели или временного отсутствия;

г) сведения о второй работе: наличие дополнительной работы в обследуемую неделю или в предшествующий месяц; вид экономической деятельности; занятие (профессия, должность); классификация по статусу занятых; вид работы и условия найма; количество фактически отработанного времени на дополнительных работах;

д) готовность к дополнительной занятости: поиск дополнительной работы; пожелания по форме дополнительной занятости; возможная продолжительность (в часах) дополнительной занятости; способы поиска дополнительной работы; готовность приступить к дополнительной работе;

е) поиск работы: поиск работы лицами, не занятыми в обследуемую неделю; способы поиска работы; готовность приступить к работе; характер работы, к которой незанятый готов приступить; продолжительность поиска работы; регистрация в службе занятости в качестве безработного; получение пособия по безработице;

ж) прошлая деятельность лиц, не занятых в обследуемую неделю (безработных и экономически неактивных): наличие когда-либо ранее работы у незанятых; вид деятельности и занятие (профессия, должность) по последнему месту работы; продолжительность периода незанятости; причины, по которым незанятые оставили последнее место работы;

з) дополнительные сведения об экономически неактивных лицах: социальный статус; причины отказа от поиска работы; причины неготовности приступить к работе;

и) занятость производством товаров или услуг в домашнем хозяйстве: занятость в обследуемую неделю производством в домашних хозяйствах продукции сельского, лесного хозяйства, охоты, рыболовства (как для реализации, так и для собственного потребления); основной вид производимой продукции и отработанное время в обследуемую неделю на выполнении работ по производству этой продукции; занятость производством в домашнем хозяйстве промышленных товаров или услуг для получения дохода или обмена; основной вид производимых в домашнем хозяйстве промышленных товаров и услуг и отработанное время на выполнении этих работ.

Программа основного обследования может дополняться единовременным или периодическим сбором информации по блоку вопросов, характеризующих наиболее актуальные проблемы рынка рабочей силы.

Опрос населения начинается в первый понедельник после критической (обследуемой) недели и проводится интервьюерами в течение последующих двух недель путем непосредственного посещения домашних хозяйств. Участие населения в обследовании является добровольным.

При обследовании населения по проблемам занятости применяется метод ведения опроса по стандартизированному бланку Анкеты с готовым текстом вопросов. Это гарантирует, что всем респондентам задаются одни и те же вопросы в одной и той же последовательности.

Подобным образом стандартизированы способы записи ответов интервьюерами. Это гарантирует, что ответы на один и тот же вопрос от разных респондентов записаны сопоставимым методом.

При непосредственном посещении домашнего хозяйства ответы на вопросы Анкеты могут быть получены как от самого респондента, так и со слов совместно проживающих членов домашнего хозяйства в случае отсутствия респондента в момент опроса.

Подбор и инструктирование интервьюеров осуществляется в территориальных органах Росстата в соответствии с методическими рекомендациями, разработанными Росстатом России. К работе в качестве интервьюеров не привлекаются лица, которые в связи со своей профессиональной деятельностью или по другим причинам могут использовать полученные сведения в ущерб опрашиваемым гражданам.

Интервьюеры дают обязательство о сохранении конфиденциальности статистических данных, а также других сведений об опрошенных лицах, которые они получили при посещении домашних хозяйств. Обязательство остается в силе после окончания работы интервьюером.

При выполнении работы по опросу населения интервьюеры обязаны удостоверять свою личность предъявлением документа. На время работы интервьюеру выдается удостоверение установленного образца, которое подписывается руководителем администрации региона и территориального органа Росстата.

Интервьюеры (счетчики):

- в период обследуемой недели (в отчетный период) проводят предварительный обход населения по отобранным для обследования адресам с целью оповещения о проводимом обследовании и согласования даты и времени проведения непосредственно интервью;
- в карточке на помещение составляют списки домохозяйств в пределах помещения, в которых указывают количество членов домохозяйства (независимо от возраста, включая лиц моложе 15 и старше 72 лет), фактически постоянно (обычно) проживающих по включенным в выборку адресам, состав членов домохозяйства по возрастным группам в зависимости от количества полных лет, количество подлежащих опросу членов домохозяйства;
- проводят опрос лиц в возрасте 15 - 72 года по вопросам анкеты и отмечают ответы на бланке анкеты;

После завершения опроса по данному адресу в карточке на помещение по каждому домохозяйству заполняются следующие данные:

- фактически опрошено, всего;

  из них:
- длительно (1 год и более) отсутствующие, подлежащие опросу;
- находящиеся на срочной службе в Вооруженных Силах;
- отсутствующие на момент проведения опроса;
- отказались от ответа.

Далее проводится первичный логический контроль ответов на отдельные вопросы анкеты в соответствии с Инструкцией о порядке заполнения анкеты;

После завершения опроса граждан по всем отобранным для обследования адресам составляют отчет о выполненной работе интервьюера (счетчика), в котором содержатся следующие сводные данные:

- количество первоначально установленных для обследования адресов;
- число первоначально установленных для опроса респондентов;
- количество фактически обследованных помещений;
- число фактически опрошенных респондентов;
- количество необследованных помещений с указанием причин;
- количество не опрошенных респондентов с указанием причин;

По завершению указанной выше работы интервьюеру-инструктору передаются заполненный отчет о работе интервьюера (счетчика), заполненные бланки анкет, карточки на помещение.

При проведении опроса интервьюеры (счетчики), в соответствии с приведенными в анкете подсказами, последовательно зачитывают сформулированные вопросы. Изменение формулировки вопросов анкеты не допускается. При необходимости интервьюеры

(счетчики) могут дать пояснения к вопросам анкеты, руководствуясь Порядком заполнения Анкеты.

Интервьюерам (счетчикам) не разрешается опрашивать население по адресам, не представленным в выборке.

3.2. Алгоритм формирования выборки

При проведении выборочных обследований населения наиболее эффективной базовой основой при построении выборочных совокупностей являются материалы переписей населения, использование которых имеет ряд преимуществ. Информационный массив переписи населения позволяет осуществлять формирование выборки на готовых основах и приводит к исключению самой высокой статьи затрат, связанной с ее составлением при отсутствии подходящих баз данных.

При создании выборки можно достаточно детально учесть социально-экономическую и демографическую структуру населения в территориальном разрезе, содержащуюся в информационным массиве переписи населения. Кроме этого, можно реализовать двухступенчатую схему формирования выборки и определить уже на втором этапе адресную часть домохозяйств, включенных в выборку. Также наличие переписной информации на электронных носителях позволяет реализовывать при построении выборок эффективные планы отбора (заложенные в специализированных статистических пакетах программ), такие как отбор с вероятностями пропорциональными размеру, которые обеспечивают отражение в выборке территориальных и структурных особенностей как регионов страны, так и населения.

Учитывая, что обследования рынка труда по своей структуре и назначению относятся к крупномасштабным обследованиям населения, поэтому в 2007 году Росстатом была разработана методика для проведения обследований населения по проблемам занятости в 2008-2011 гг. на основе информационного массива данных Всероссийской переписи населения 2002 года. В дальнейшем при создании выборочной сети домохозяйств наиболее целесообразным и эффективным признано использовать материалы всероссийских переписей населения (2010 года и последующих).

Основным ориентиром при разработке используемых в настоящее время методологических рекомендаций по формированию выборочной сети домохозяйств и выборочных массивов по отбору первичных выборочных единиц (ПВЕ) являлось обеспечение эффективного проведения в последующие годы формирования выборочной сети домашних хозяйств для проведения обследований населения по проблемам занятости.

Особое внимание было обращено на систематизацию и подготовку к анализу информации о размещении объемов выборки первичных выборочных единиц (ПВЕ) по территории субъектов РФ при проведении обследования населения по проблемам занятости в 2008-2011 гг. Для этого потребовалась разработка методологических рекомендаций по разделению общего выборочного массива первичных выборочных единиц на годовые, квартальные и месячные независимые подвыборки для проведения обследования населения по проблемам занятости.

Базовой и теоретической основой для разработки общих методологических положений по созданию выборочных массивов единиц наблюдения является теория выборочного метода и его основные этапы, определяющие процедуру проектирования выборочных обследований и обеспечивающие вероятностный характер включения объектов и единиц наблюдения в выборочную совокупность.

При разработке общих методологических положений и их реализации в практической работе учитывался ряд требований, предъявляемых к планированию выборочных обследований, главными из которых являются:

- соблюдение международных и стандартных требований к организации крупномасштабных обследований населения;
- обеспечение достаточности выборочных массивов единиц наблюдения для построения множества подвыборок на ряд лет и реализации независимых схем ротации;
- обеспечение территориального представительства субъектов РФ при вероятностном характере формирования выборочного массива;
- обеспечение получения итогов экстраполяции выборочных данных крупномасштабных обследований населения в пределах заданной точности.

В методологическом плане для подготовки и проведения месячных обследований населения по проблемам занятости характерны три основных стадии: общее планирование и подготовка обследования (стадия 1), проведение обследования (стадия 2) и статистический анализ полученных результатов обследования (стадия 3).

При проведении месячных обследований населения по проблемам занятости, также как в других обследованиях населения, Росстатом используются многоэтапные схемы формирования выборки. Эффективной базовой основой для формирования выборки рынка труда являются материалы ВПН. Наличие информационного массива ВПН 2002 г. на электронных носителях сделало возможным применение для формирования выборки двухступенчатого отбора.

Изучаемой генеральной совокупностью, соответственно выбранной базовой основой выборки, являются все типы частных домохозяйств по месту своего постоянного (обычного) проживания, за исключением коллективных домохозяйств, т.е. части населения, состоящей из лиц, долговременно находящихся в больницах, школах-интернатах, интернатах для престарелых и инвалидов, других институциональных заведений и прочих коллективных жилых помещений.

Единицей наблюдения и единицей анализа при проведении обследования занятости населения являются домашнее хозяйство и его члены в возрасте 15-72 года.

Главной задачей при разработке плана выборки данного обследования является обеспечение, при его реализации, представительности территориальной структуры субъектов РФ и достаточности объема выборки для получения статистически надежных данных по основным признакам программы обследования, характеризующим численность и состав экономически активного населения, занятых, безработных, уровень экономической активности и безработицы.

Учитывая основное назначение плана выборки данного обследования, определение базовой основы, на которой проводится формирование выборочных массивов, и модели построения выборки, то общее решение главных вопросов, определяющих архитектуру плана выборки, а в целом и методологических положений по созданию выборочных массивов единиц наблюдения, сводится к следующему:

Создание выборочной сети домохозяйств для проведения месячных обследований рынка труда осуществляется с учетом административно-территориальной структуры Российской Федерации. Формирование выборки проводилось на четыре года, отдельно по городскому и сельскому населению.

1. При построении выборки используется двухэтапная выборка: на первом этапе формируется выборочный массив первичных выборочных единиц (ПВЕ), на втором – выборочная совокупность адресов домашних хозяйств. Отбор домашних хозяйств осуществляется в рамках ПВЕ, включенных в выборочный массив на первой ступени.

Базовой основой при создании выборочной сети домохозяйств, как отмечалось ранее, является информационный массив ВПН 2002 г., в который включается только население частных домохозяйств. Это означает, что на этапе формирования выборки проводится специальная процедура, направленная на создание первичного информационного фонда (ПИФ), включающего в основном только население частных домохозяйств.

2. В качестве первичной выборочной единицы при реализации двухэтапной выборки используется счетный участок (переписная единица третьего уровня) как по городскому, так и сельскому населению. Достоинство счетного участка, выбранного в качестве ПВЕ, заключается как в его компактности, наиболее низкой вариации его размера, достаточности объема счетного участка для формирования на втором этапе подвыборок домашних хозяйств, достаточности количественного состава данной ПВЕ для их отбора на длительный период времени, так и в возможности обеспечения получения представительного размещения этих первичных единиц по экономико-географическому положению, характеру расселения населения и типу административных районов с разной численностью населения.

Размер счетного участка по городскому населению в среднем составляет 420 человек (приблизительно в среднем 120-140 домохозяйств), а по сельскому населению – 320 человек (приблизительно 90-110 домохозяйств).

В целом по России базовый информационный массив может содержать приблизительно 371300 счетных участков (ПВЕ): 246640 счетных участков – в городской местности и 124660 – в сельской местности.

В соответствии с принятой ПВЕ основой выборки для построения выборочного массива ПВЕ на первом этапе является совокупность счетных участков, образованных при проведении ВПН 2002 г. и включенных в ПИФ при его создании.

Таким образом на первом этапе при создании выборки единицей отбора является счетный участок, реквизиты которого имеются на электронных носителях и в целом включают пять основных признаков: код населенного пункта, в котором образована данная ПВЕ, номер переписного и номер инструкторского участка, в состав которых входит данный счетный участок, номер счетного участка и номер папки, входящей в состав счетного участка.

3. Формирование на первом этапе выборочного массива ПВЕ проводится с применением модели отбора с вероятностью пропорциональной размеру (ВПР). В качестве размера используется или показатель «число домохозяйств в счетном участке», или показатель «численность постоянного населения в возрасте 15-72 года».

Входными данными для реализации по городской и сельской местности модели отбора первичных выборочных единиц с ВПР являются:

- сведения первичного информационного фонда о количестве ПВЕ, включенных в его состав, о численности постоянного населения и количестве домашних хозяйств.
    - Данная информация известна в рамках субъектов РФ по городскому и сельскому населению, по каждому административному району, в т.ч. по городу республиканского, краевого, областного, окружного подчинения, ЗАТО;
- информация о реквизитах ПВЕ (для указания кода населенного пункта используется девятизначный код ТЕРСОНА);
- информация на уровне счетного участка по трем показателям: «численность постоянного населения», «количество домохозяйств», «численность постоянного населения в возрасте 15-72 года». Последние два показателя, как отмечено выше, могут быть использованы в качестве размера. Окончательное решение о выборе показателя принимается на основе анализа результатов пробного отбора ПВЕ с ВПР. Если результаты взвешивания выборочных данных по указанным показателям отклоняются от исходной информации в допустимых пределах, то в качестве размера используется показатель «количество домохозяйств». Если же отклонения распространенных итогов от исходных итогов оказываются значительными, то предпочтение отдается показателю «численность постоянного населения в возрасте 15-72 года»;
- информация о количестве ПВЕ, подлежащих отбору на первом этапе в рамках субъекта РФ, отдельно по городскому и сельскому населению;
- информация о социально-демографической структуре населения по счетному участку по 8 признакам: пол, возраст, образование, национальность, источники средств существования, занятость, размер домохозяйства, тип жилого помещения.

Общий выборочный массив ПВЕ был сформирован на четыре года (2008-2011 гг.). Для его построения использовалась процедура систематического отбора ПВЕ с вероятностью пропорциональной размеру. Модель отбора ПВЕ с ВПР включает следующие основные этапы:

- построение кумулятивного ряда, в котором указывается нарастающим итогом значения показателя, выбранного в качестве размера ПВЕ;
- определение интервала отбора и случайного начала отбора (для этих целей используется генератор случайных чисел). Реализуется процедура отбора с ВПР на основе построенного кумулятивного ряда, включенного в основу отбора ПВЕ;
- определение совокупности ПВЕ, подлежащей включению в общий выборочный массив, и расчет вероятности включения ПВЕ в выборку (что необходимо для непосредственного оценивания показателей обследования);

- расчет для отобранной выборки ПВЕ вероятностей совместного включения ПВЕ в выборку, что необходимо для определения эффекта примененного сложного плана отбора и оценивания характеристик точности оцененных показателей обследования (не было выполнено).

4. При создании выборочной сети домохозяйств для проведения используется полностью независимая схема ротации. Это означает, что, во-первых, годовые выборочные массивы ПВЕ в рамках каждого из четырех годов, равно как и подвыборки четырех кварталов и двенадцати месячных периодов наблюдения, по каждому из четырех лет не пересекаются между собой; и, во-вторых, подвыборки домохозяйств, сформированные по непересекающимся ПВЕ, также являются непересекающимися.

5. Для разделения общего выборочного массива на четыре года также используется систематический отбор, интервал которого определяется числом лет использования общего выборочного массива (в данном конкретном случае интервал включения соответствующей ПВЕ в годовую выборку равен четырем).

6. При формировании общего выборочного массива ПВЕ используется процедура упорядочивания совокупности административных районов с учетом их географической близости. Основное назначение данной процедуры – это выделение территориальных сегментов, объединяющих ряд смежных административных районов с целью получения оптимального территориального представительства субъекта РФ и обеспечения сопоставимости результатов обследования при его проведении на основе непересекающихся выборок и подвыборок.

Для реализации этой процедуры был выполнен большой объем работ специалистами Росстата, связанный с установлением географической близости территориальных единиц субъекта РФ и анализом размещения выборки квартального обследования рынка труда (КОРТ) за 2006 г. по территории субъектов РФ.

Для установления географической близости административно-территориальных единиц в рамках субъекта РФ использовался принцип серпантинного расположения территориальных единиц с севера на юг в направлении запад-восток. Для анализа размещения выборки ПВЕ был использован картографический материал, подготовленный и представленный в Росстат территориальными органами Росстата.

7. При размещении общего выборочного массива ПВЕ по построенным территориальным сегментам в рамках субъекта РФ использовалась модель Л. Киша с коррекцией.

8. В случае исчерпывания ПВЕ при длительном проведении квартальных обследований населения по проблемам занятости, в том числе и при различного рода временных и структурных изменениях, была предусмотрена объективная процедура отбора смежной ПВЕ в рамках переписной единицы второго уровня, т.е. в рамках инструкторского участка. Для этих целей в выходной электронной таблице предусмотрена специальная графа, где были указаны все номера счетных участков, смежные с отобранным, и принадлежащие к соответствующему инструкторскому участку.

Предусмотренная процедура обеспечила проведение обследований населения по проблемам занятости с месячной периодичностью начиная с IV квартала 2009 года.

9. Формирование выборки домохозяйств проводится на второй ступени. Ее построение осуществляется в рамках каждой ПВЕ, включенной в состав общего выборочного массива ПВЕ, сформированного на первой ступени.

Для формирования выборки домохозяйств предусмотрено применение стандартной процедуры систематического отбора, при котором начало отбора определяется случайно.

Основой выборки домохозяйств на второй ступени является совокупность переписных листов, составленных на отдельное домохозяйство в пределах счетного участка, отобранного на первом этапе.

Систематический отбор домохозяйств проводится из упорядоченного списка домохозяйств с учетом их расслоения по двум признакам: размер домохозяйства и тип жилого помещения. Указанное упорядочивание списка домохозяйств в целом характеризует процедуру неявного расслоения и направлено на обеспечение представительства в выборке разнообразных типов домохозяйств по указанным признакам.

Адресная часть домашнего хозяйства, попавшего в выборку на втором этапе, определяется на основании файлов графических образов переписных листов, составляющих архив первичных носителей информации ВПН 2002 г.

10. По результатам формирования на первом и втором этапах выборочной сети домохозяйств для проведения обследования населения по проблемам занятости определяется: во-первых, общая вероятность включения в выборку домашнего хозяйства; и, во-вторых, базовый вес элементов выборки, который является основой для получения итогов экстраполяции по структурным признакам, включенным в программу месячных обследований населения по проблемам занятости.

Для расчета общей вероятности включения в выборку домохозяйства (P) используется стандартное соотношение с учетом двухэтапного характера формирования выборки:

$$P=P_1 \cdot P_2 \qquad (1)$$

P_1 – вероятность включения ПВЕ в выборку на первом этапе вычисляется как:

$$P_{1,ghi} = k_{gh} \cdot \frac{m_{ghi}}{M_{gh}} \qquad (2)$$

где k_{gh} – число ПВЕ, подлежащих отбору в g-ом субъекте РФ отдельно по городскому и сельскому населению (h);

m_{ghi} – количество домохозяйств в i-ой ПВЕ, включенной в выборку на первом этапе (в g-ом субъекте РФ отдельно по городскому и сельскому населению (h));

M_{gh} – общее количество домохозяйств по всей совокупности ПВЕ (в g-ом субъекте РФ отдельно по городскому и сельскому населению (h)).

P_2 – вероятность включения домохозяйства в выборку на втором этапе вычисляется по формуле:

$$P_2 = n_{ghi} / m_{ghi} \qquad (3)$$

где n_{ghi} – количество домохозяйств, подлежащих отбору в рамках i-ой ПВЕ (в g-ом субъекте РФ отдельно по городскому и сельскому населению (h));

m_{ghi} – количество домохозяйств в i-ой ПВЕ, включенной в выборку на первом этапе (в g-ом субъекте РФ отдельно по городскому и сельскому населению (h)).

Базовый вес домохозяйства, включенного в выборочную совокупность, определяется как обратная величина общей вероятности включения в выборку домохозяйства. Следует также отметить, что в плане выборки предусматривается применение процедур коррекции базовых весов с учетом внешней информации. В качестве внешней информации применяются данные текущей статистики населения в группировке по региональной принадлежности и по половозрастному составу. При этом на этапе распространения данных выборки на генеральную совокупность фактически используется пострасслоенный план анализа данных.

Следует отметить, что применяемая процедура для корректировки выборочных весов индивидов имеет определенные недостатки, приводящие к несогласованности итогов обследования со структурой генеральной совокупности.

Для построения выборочного массива ПВЕ использовалась процедура систематического отбора ПВЕ с вероятностью пропорциональной размеру. Модель отбора ПВЕ с ВПР должна включать кроме расчета вероятностей включения ПВЕ в выборку (что необходимо для непосредственного оценивания показателей обследования) еще и вычисление вероятностей совместного включения ПВЕ в выборку. Это необходимо для определения эффекта примененного сложного плана отбора и определения характеристик точности оцененных показателей обследования, соответствующих плану выборки.

В качестве примера вычисления вероятностей совместного включения ПВЕ в выборку рассмотрим следующую ситуацию. Пусть требуется отобрать ВПР систематическую выборку объемом 4 единицы из совокупности (территориальный сегмент), состоящей из 8 единиц, размер которых (число домашних хозяйств - признак x) известен:

№ПВЕ	1	2	3	4	5	6	7	8
Размер (x)	300	300	150	100	50	50	25	25

1. Вычислите вероятности включения элементов на основе размера предприятий.

2. Сформируйте выборку с помощью систематического отбора, используя при этом реализацию случайной величины: 0,278.

3. Вычислите вероятность включения второго порядка для единиц 3 и 5.

В условиях приведенного примера сначала нужно вычислить вероятности включения ПВЕ на основе известных размеров единиц по формуле:

$$\pi_k = n\frac{X_k}{\sum_{e\in U} X_e}$$

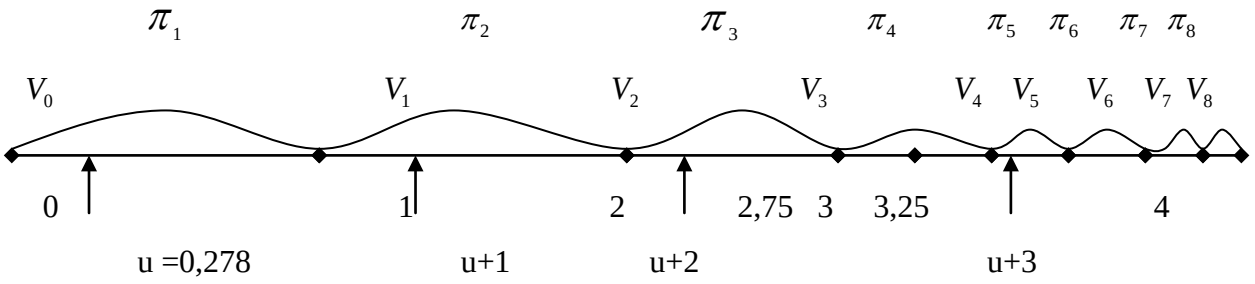
Соответственно получаем, что

$$\pi_1=1,\ \pi_2=1,\ \pi_3=0{,}75,\ \pi_4=0{,}5,\ \pi_5=0{,}25,\ \pi_6=0{,}25,\ \pi_7=0{,}125,\ \pi_8=0{,}125.$$

Накопленные вероятности включения будут следующими:

$$V_0=0,\ V_1=1,\ V_2=2,\ V_3=2{,}75,\ V_4=3{,}25,\ V_5=3{,}50,\ V_6=3{,}75,\ V_7=3{,}875,\ V_8=4.$$

Данную ситуацию ВПР систематического отбора можно представить графически следующим образом, если была сгенерирована реализация случайной величины (u) равномерно распределенной в единичном интервале: u = 0,278:



Соответственно приведенного рисунку отобраны в выборку единицы 1,2,3,5.

Заметим, что

$$\pi_{12}=1,\ \pi_{13}=0{,}75,\ \pi_{15}=0{,}25,\ \pi_{23}=0{,}75,\ \pi_{25}=0{,}25$$

Также ясно, что

$$\pi_{35}=\pi_{53}$$

Соответственно, пусть в выборку попала 5-ая единица. Это означает, что существует целое положительное i, такое что

$$u + i \in [3;3{,}25]$$
$$\Downarrow$$
$$u + (i-1) \in [2;2{,}25] \in [2;2{,}75]$$

Т.е. интервалу, соответствующему 3-ей единице. Следовательно, если в выборку попадает 5-ая единица, то тогда в выборке также будет и 3-ая единица, т.к.

$$pr(3 \in выборке / 5 \in выбоорке) = \frac{pr(3,5 \in s)}{pr(5 \in s)} = \frac{\pi_{53}}{\pi_5} = 1 \Rightarrow$$

Поэтому $\pi_{35} = \pi_{53} = \pi_5 = 0{,}25$ или графически π_{35} =-[2;2,25]∩[2;75]=0,25.

В целом же можно подчеркнуть, что реализация модели двухэтапной выборки основана на знании следующей информации.

Во-первых, информации о реквизитах ПВЕ, образованных по всем административно-территориальным единицам, количестве счетных участков в каждой территориальной единице, количестве ПВЕ, подлежащих отбору, отдельно по городскому и сельскому населению, количестве домохозяйств в каждом счетном участке первичного информационного фонда и его структурным характеристикам по выделенной системе социально-демографических признаков.

Во-вторых, информации о совокупности переписных листов, составленных на каждое домохозяйство, входящее в состав ПВЕ первичного информационного фонда, о графических образах переписных листов, включая их номера и адресную часть домохозяйств.

В-третьих, процедур по актуализации выборочных массивов ПВЕ при обнаружении в них или недействующих ПВЕ или недостижимых ПВЕ (в том числе и ликвидированных) на момент проведения обследования.

После поступления информация о недостижимости ПВЕ (получается от территориальных органов Росстата на основе анализа списка отобранных ПВЕ) выполняется автоматизировано процедура частичной актуализации из подмножества ПВЕ, где случайно был отобран недействующий счетный участок, т.е. из ряда смежных с ним ПВЕ, входящих в один и тот же инструкторский участок.

При этом используются или специальные схемы выбора ПВЕ (например, круговая схема, разработанная Л. Кишем) или же стандартные процедуры случайного подбора одной ПВЕ из ряда смежных с ней ПВЕ, например, ближайшего соседа. Для реализации указанных процедур важное значение имеют полные и точные сведения о структурном составе первичного информационного фонда.

Характерным для создания и эксплуатации выборочной сети домашних хозяйств для проведения месячных обследований населения по проблемам занятости является формирование непересекающихся выборок единиц наблюдения на первом и втором этапах выборки как на четырехлетний период (2008-2011 гг.), так и на каждый квартал и

месяц годовой выборки. Это означает, что при создании выборочных массивов для проведения обследования с целью сбора информации по ключевым признакам программы обследования применяются независимые схемы ротации.

Использование данных схем направлено, прежде всего, на повышение охвата элементов генеральной совокупности по всем субъектам РФ в годовом цикле обследований при объемах выборки, определяемых ограниченностью финансовых ресурсов, выделяемых на проведение данного крупномасштабного обследования населения с месячной периодичностью. Указанное является одной из главных особенностей независимых схем ротации в сравнении с лангитьюдным подходом, когда контролируется доля ежемесячной ротации элементов выборки и период, в течение которого наблюдается каждый элемент (т.е. создается панель индивидов). Также при этом подходе удается сократить затраты на проведение наблюдения.

Применение независимых схем ротации ограничивает возможность использования полученных информационных массивов для изучения динамики явлений и процессов, происходящих с течением времени на рынке труда (например, продолжительность безработицы). Это обусловлено отличием в рамках субъекта РФ территориальных структур годовых, квартальных и месячных выборочных массивов в каждом периоде обследования, что и является ограничением использования результатов обследований населения по проблемам занятости при изучении динамики в том числе на индивидуальном уровне.

При анализе динамики показателей, рассчитанных по данным независимых выборок нужно учитывать следующие объективные факторы.

Так если требуется оценить изменение среднего (доли) значения переменной y, произошедшее за определенное время, например, в промежутке между двумя датами: (1) и (2), т.е. разницу средних: $\overline{Y}_2 - \overline{Y}_1$, то можно поступать двумя разными методами.

При первом варианте действий можно извлечь две независимые выборки для наблюдения в моменты времени (1) и (2) (что соответствует схеме проведения обследования населения по проблемам занятости).

Тогда имеем естественную оценку по данным двух последовательных выборок: $\overline{y}_2 - \overline{y}_1$ для разности $\overline{Y}_2 - \overline{Y}_1$.

Разумеется в силу независимости обеих выборок дисперсия этой оценки может быть вычислена по формуле:

$$Var(\overline{y}_2 - \overline{y}_1) = Var(\overline{y}_2) + Var(\overline{y}_1) \qquad (4)$$

При втором варианте действий можно сформировать одну выборку в момент времени (1), которая повторно наблюдается в момент времени (2) (лонгитюдный подход). Тогда по-прежнему имеем, что $\overline{y}_2 - \overline{y}_1$ - оценка по теперь зависимым выборкам для разности $\overline{Y}_2 - \overline{Y}_1$

Дисперсию этой оценки в данном случае можно вычислить по формуле:

$$Var(\overline{y}_2 - \overline{y}_1) = Var(\overline{y}_2) + V(\overline{y}_1) - 2Cov(\overline{y}_1, \overline{y}_2) \qquad (5)$$

где $Cov(\bar{y}_1, \bar{y}_2)$ - ковариация оценок $\bar{y}_1$ и $\bar{y}_2$.

При благоприятных обстоятельствах (стабильности структуры генеральной совокупности) $Cov(\bar{y}_1, \bar{y}_2) > 0$. Следовательно с учетом (5) имеем, что

$$Var(\bar{y}_2 - \bar{y}_1) < Var(\bar{y}_2) + Var(\bar{y}_1) \quad (6)$$

Поэтому можно сделать вывод, что вариант 1 обследования постоянной группы единиц наблюдения, панели, дает более точные результаты. Однако сделанный вывод, конечно, справедлив не всегда. Например, если требуется оценить временное среднее, использование постоянной группы единиц может иметь отрицательные последствия:

$$Var\left(\frac{\bar{y}_1 + \bar{y}_2}{2}\right) = \frac{1}{4}\left[Var(\bar{y}_1) + Var(\bar{y}_2) + 2Cov(\bar{y}_1, \bar{y}_2)\right] > \frac{1}{4}\left[Var(\bar{y}_1) + Var(\bar{y}_2)\right]$$

т.е. в данном случае метод независимых выборок предпочтительнее.

Основой для достижения в определенной степени сопоставимых территориальных структур годовых массивов по субъектам РФ является, во-первых, применение специальных процедур размещения первичных выборочных единиц, направленных на максимальное отражение в выборке территориальных особенностей субъекта РФ; во-вторых, определение оптимальных объемов выборочных массивов ПВЕ и в-третьих, применение эффективных моделей размещения объемов выборки ПВЕ по территории субъекта РФ.

Важное значение для разработки и реализации указанных направлений в части построения сопоставимых территориальных структур на длительный период времени имеет анализ размещения выборочных массивов предыдущих обследований.

Резюме

Построение новой выборочной сети домашних хозяйств для проведения в 2008-2011 г.г. обследований населения по проблемам занятости сначала с квартальной, а потом с месячной периодичностью явилось одним из важнейших направлений совершенствования организации квартальных обследований рынка труда. В рамках развития данного направления особое внимание было обращено на обеспечение в выборке территориального представительства субъектов РФ и отражение в выборочной сети основных структурных особенностей населения, состава и типа домохозяйств. Указанное направлено на получение наиболее полной, надежной и сопоставимой информации о современном состоянии рынка труда и его основных факторов в течение длительного периода времени.

Разработанная структура плана выборки обследования населения по проблемам занятости в 2008-2011 г.г. включает ряд модулей, реализация которых непосредственно направлена на достижение обозначенной задачи. В частности, при построении выборочной сети домашних хозяйств впервые было реализовано территориальное сегментирование субъекта РФ, в результате которого его территория была разделена на ряд территориальных сегментов, содержащих административно-территориальные единицы (АТЕ) с учетом их географической близости (последняя определялась на основе специального серпантинного расположения АТЕ, характеризующих административно-территориальное устройство субъекта РФ).

Кроме того, впервые при построении выборки данного обследования совокупность объектов наблюдения формировалась на длительный период времени (т.е. на четыре года). Это потребовало при создании выборочной сети домохозяйств использования как моделей, наиболее полно соответствующих разработанной структуре плана выборки (в частности при определении объемов выборки ПВЕ и его размещении по множеству объектов наблюдения при формировании общего выборочного массива ПВЕ, его разделения на годовые массивы, квартальные и месячные подвыборки), так и специальных процедур систематизации единиц основы выборки на первом и втором этапах построения выборочной сети домашних хозяйств.

В случае исчерпывания ПВЕ при длительном проведении квартальных обследований населения по проблемам занятости, в том числе и при различного рода временных и структурных изменениях, была предусмотрена объективная процедура отбора смежной ПВЕ в рамках переписной единицы второго уровня, т.е. в рамках инструкторского участка. Для каждой ПВЕ, включенной в выборку, было предусмотрено создание массива номеров счетных участков, смежных с отобранной ПВЕ, принадлежащих к соответствующему инструкторскому участку. Это обеспечило переход проведения обследований населения по проблемам занятости на месячную периодичность начиная с IV квартала 2009 года.

Следует также отметить, что все процедуры, связанные с использованием моделей построения на первом и втором этапах выборочной сети домашних хозяйств, реализованы в автоматизированном режиме.

К основным недостаткам используемых в настоящее время моделей планирования выборки и распространения выборочных микроданных на генеральную совокупность следует отнести следующее.

Общий выборочный массив ПВЕ был сформирован на четыре года (2008-2011 гг.). Для его построения использовалась процедура систематического отбора ПВЕ с вероятностью пропорциональной размеру. Модель отбора ПВЕ с ВПР включает следующие основные виды работ. Определение совокупности ПВЕ, подлежащей включению в общий выборочный массив, и расчет вероятности включения ПВЕ в выборку (что необходимо для непосредственного оценивания показателей обследования). Однако для отобранной выборки ПВЕ также требуется вычисление вероятностей совместного включения ПВЕ в выборку. Это необходимо для определения эффекта примененного сложного плана отбора и определения характеристик точности оцененных показателей обследования, соответствующих плану выборки.

Кроме того, применяемая в настоящее время процедура для корректировки выборочных весов индивидов имеет существенные недостатки, приводящие к несогласованности итогов обследования со структурой генеральной совокупности.

3.3. Расчет индивидуальных весовых коэффициентов в рамках обследований, проводимых с месячной, квартальной и годовой периодичностью

По результатам проведения обследований рабочей силы по проблемам занятости для последующей компьютерной обработки собранных первичных сведений обследования создаются следующие базы данных:

на региональном уровне:

- база микроданных обследованных лиц;
- база данных домашних хозяйств, участвующих в обследовании;
- база регламентных таблиц и публикационных бюллетеней;

на федеральном уровне:

- база микроданных обследованных лиц;
- база регламентных таблиц и публикационных бюллетеней.

База микроданных обследований населения по проблемам занятости содержит фонды микроданных по результатам обследований начиная с 1992 года. Единицей хранения в базе микроданных являются данные по каждому респонденту (индивиду).

Система показателей базы микроданных включает:

а) первичные показатели, полученные в ходе опроса как ответы на вопросы Анкеты;

б) производные показатели, интегрированные показатели, полученные при обработке результатов обследования на основе:

- сочетания нескольких последовательных ответов на вопросы Анкеты;
- агрегирования значений первичного показателя в группировки более высокого уровня (группировки видов деятельности и занятий, группировки по возрасту, отработанному времени);

в) индивидуальные весовые коэффициенты: месячные, квартальные и годовые, вычисленные при обработке результатов обследования, для обеспечения получения оценок по выборке статистических показателей на основе микроданных.

Расчет квартальных и годовых индивидуальных весов для распространения данных выборки на генеральную совокупность основывается на усреднении значений месячных весовых коэффициентов по формуле средней арифметической, т.е. сумма месячных значений делится на 3 и 12 соответственно. Используемая в настоящее время процедура расчета квартальных и годовых выборочных весов индивидов потенциально может приводить к несогласованности распространенных итогов обследования. Более подробно этот вопрос будет рассмотрен после анализа процедуры вычисления месячных выборочных весов индивидов.

Методика взвешивания в целях распространения месячных выборочных данных обследования основана на присвоении соответствующего индивидуального веса каждой отдельной единице наблюдения – индивиду (персоне).

В настоящее время расчет месячных индивидуальных весов производится методом итеративного взвешивания выборки с использованием в качестве генеральной совокупности внешних данных о численности населения в возрасте 15-72 года в группировке по регионам, типу местности и половозрастному признаку. При взвешивании

месячных данных выборочной совокупности (обследованных персон) в начале года используются демографические данные на начало предыдущего года, при обработке результатов обследования в последующие месяцы - данные на начало текущего года, когда эта информация становится доступной.

Используемая процедура взвешивания заключается в определении соотношения объемов генеральной совокупности населения и выборки обследованных индивидов, распределенных по группам с учетом половозрастной, региональной принадлежности, а также по типу местности проживания.

Региональная группировка представлена делением на субъекты РФ места жительства индивидов, а по типу местности – городская или сельская. Половозрастная группировка представлена делением по половому признаку (отдельно по мужчинам и женщинам) и по возрастным группам: с 15 до 24 лет - однолетним, с 25 до 49 лет - пятилетним, с 50 до 72 лет - однолетним

При расчете индивидуальных весов взвешивание производится последовательно в пять этапов.

На первом этапе производится расчет базовых весов выборки ($w^{б}_{рег,тип,пол}$) как отношение объема генеральной совокупности населения (N) и объема выборки обследованных индивидов (n) в возрасте 15-72 года в группировке по регионам (рег), по полу (пол) и типу местности (тип). Выборочный вес каждого обследованного индивида в возрасте 15-72 года, относящегося к фиксированной группе по региональному признаку, по полу и типу местности, рассчитывается как

$$w^{б}_{рег,тип,пол} = \frac{N_{рег,тип,пол}}{n_{рег,тип,пол}} \quad (1)$$

На втором этапе с учетом вычисленных на первом базовых индивидуальных весов оценивается по выборке численность населения в возрасте 15-72 года ($\hat{N}_{рег,тип,пол,вгр}$) в группировке по регионам, полу, типу местности и по возрастным группам (*вгр*) как

$$\hat{N}_{рег,тип,пол,вгр} = n_{рег,тип,пол,вгр} \cdot w^{б}_{рег,тип,пол} \quad (2)$$

На третьем этапе оценивается по выборке численность населения в возрасте 15-72 года ($\hat{N}_{рег,пол,вгр}$) в группировке по регионам, полу и по возрастным группам (*вгр*) как сумма соответствующих оценок в отдельности по городской и сельской местностям:

$$\hat{N}_{рег,пол,вгр} = \hat{N}_{рег,тип=город,пол,вгр} + \hat{N}_{рег,тип=село,пол,вгр} \quad (3)$$

где

$$\hat{N}_{рег,тип=город,пол,вгр} = n_{рег,город,пол,вгр} \cdot w^{б}_{рег,,город,пол} \quad (3а)$$

$$\hat{N}_{рег,тип=село,пол,вгр} = n_{рег,село,пол,вгр} \cdot w^{б}_{рег,село,пол} \quad (3б)$$

На четвертом этапе вычисляется корректирующий коэффициент ($r_{рег,пол,вгр}$), равный отношению численности населения в возрасте 15-72 года ($N_{рег,пол,вгр}$) в группировке по регионам, полу и по возрастным группам и значению оценки этого показателя, рассчитанной на третьем этапе:

$$r_{рег,пол,вгр} = N_{рег,пол,вгр} \Big/ \hat{N}_{рег,пол,вгр} \quad (4)$$

На пятом этапе вычисляются значения окончательного индивидуального веса каждого элемента выборки k ($k \in рег \cap тип \cap пол \cap вгр$, где k=1, …, n – объем месячной выборки), представляющего собой произведение базового выборочного веса, рассчитанного на первом этапе, и корректирующего множителя, рассчитанного на четвертом этапе:

$$w_{рег,тип,пол,вгр;k} = w^{б}_{рег,тип,пол} \cdot r_{рег,пол,вгр} \quad (5)$$

Подставляя в (5) соотношения (1)- (4) получаем, что значения окончательного индивидуального веса каждого элемента выборки будут

$$\begin{aligned} w_{рег,тип,пол,вгр;k} &= w^{б}_{рег,тип,пол} \cdot r_{рег,пол,вгр} \\ &= \frac{N_{рег,тип,пол}}{n_{рег,тип,пол}} \times \frac{N_{рег,пол,вгр}}{\hat{N}_{рег,пол,вгр}} \\ &= \frac{N_{рег,тип,пол}}{n_{рег,тип,пол}} \\ &\times \frac{N_{рег,пол,вгр}}{n_{рег,город,пол,вгр} \cdot w^{б}_{рег,,город,пол} + n_{рег,село,пол,вгр} \cdot w^{б}_{рег,село,пол}} \end{aligned} \quad (6)$$

Сравнивая полученное выражение (6), которое фактически используется в настоящее время для расчета значений окончательного выборочного веса элементов месячной выборки, не равно естественному выражению (см. формулу (7)), обычно используемому при пострасслоенном перевзвешивании.

Так как в данном обследовании для перевзвешивания используются внешние данные о численности населения в возрасте 15-72 года в группировке (постслои) по регионам, типу местности, по полу и возрастным группам, то выборочный вес каждого обследованного индивида в возрасте 15-72 года, относящегося к фиксированной группе (по региональному признаку, типу местности, по полу и возрастной группе) может быть рассчитан как отношение объема генеральной совокупности и объема выборки в этой группе (рекомендуется консультантом):

$$w^{рек}_{рег,тип,пол,вгр} = \frac{N_{рег,тип,пол,вгр}}{n_{рег,тип,пол,вгр}} \quad (7)$$

При суммировании по списку выборки приведенных в формуле (7) выборочных весов по регламентным разрезам разработки в точности получаются значения численности населения в возрасте 15-72 года в соответствующем разрезе. Поэтому можно утверждать, что при определении окончательного выборочного веса по формуле (7) будут получаться полностью согласованные со структурой и объемом генеральной совокупности (численность населения в возрасте 15-72 года по рассматриваемым группам) распространенные месячные данные обследования рабочей силы по проблемам занятости.

В отличие от этого случая применение формулы (6) для вычисления выборочных весов по регламентным разрезам разработки приводит к рассогласованию распространенных данных выборки с генеральной совокупностью обследования. Относительно согласованности месячных и квартальных оценок статистических показателей системы изучаемых признаков при использовании существующей методики расчета индивидуальных весовых коэффициентов, в общем, можно сделать вывод об их согласованности (когда нет пропусков в месячных микроданных по используемой при взвешивании группировке).

3.4. Расчет ошибок выборки

Согласно Методологическим положениям по проведению выборочных обследований населения по проблемам занятости в 2008-2011 годах годовой объем выборочного массива по России в целом установлен в размере свыше 270 тыс. лиц в возрасте 15 - 72 года с рекомендаций его увеличения свыше 290 тыс. лиц.

По субъектам Российской Федерации применяется разная доля отбора с учетом относительной вариации по показателю "уровень безработицы". Задается фиксированная степень относительной точности по этому показателю не более:

- 1,5% стандартная относительная ошибка выборки в целом по России;
- 5% стандартная относительная ошибка выборки по ряду крупных и средних регионов;
- 8%-10% для небольших по численности населения регионов.

Устанавливается, что размер месячной выборки гарантирует получение представительных итогов в пределах заданной степени точности в целом по Российской Федерации и по разрезам разработки. По субъектам Российской Федерации представительные итоги обеспечивает совокупность трех и двенадцати месячных выборок: квартальные и годовые итоги соответственно.

При этом в качестве показателей точности оценивания статистических характеристик по результатам месячных обследований населения по проблемам занятости, т.е. количественной меры возможного отклонения оценки от действительного значения параметра, используются показатели, в основе которых лежит дисперсия оценки параметра: , где - оценка параметра .

На базе оценки дисперсии оценки, которая рассчитывается по данным выборки, вычисляются такие характеристики точности как:

- стандартная ошибка выборки равная корню квадратному из дисперсии оценки;
- коэффициент вариации оценки, т.е. стандартная относительная ошибка выборки;

- предельная ошибка выборки, равная половине длины доверительного интервала, и границы доверительного интервала, в котором заключен неизвестный параметр с заданной доверительной вероятностью (обычно в Европе и России 95%).

При подготовке публикационных материалов по результатам обследований населения по проблемам занятости должен производиться расчет характеристик точности оценивания - стандартная ошибка выборки, стандартная относительная ошибка выборки, границы доверительного интервала - для указанных ниже статистических показателей.

В целом по России:

- ✓ численность занятого населения;
- ✓ уровень занятости;
- ✓ численность безработных;
- ✓ уровень безработицы;
- ✓ численность экономически активного населения;
- ✓ уровень экономической активности;
- ✓ численность экономически неактивного населения;
- ✓ уровень экономической неактивности;
- ✓ численность экономически неактивных граждан, не выразивших желание иметь работу;
- ✓ численность экономически неактивных граждан, выразивших желание иметь работу.

По регионам России:

- ✓ численность занятого населения;
- ✓ уровень занятости;
- ✓ численность безработных;
- ✓ уровень безработицы;
- ✓ численность экономически активного населения;
- ✓ численность экономически неактивного населения в трудоспособном возрасте.

Выбор модели расчета характеристик точности базируется на принятом способе вычисления оценок. Исходя из того, что оценка статистических показателей по результатам обследования населения по проблемам занятости основывается на микроданных и индивидуальных весовых коэффициентах, при расчете точности оценивания в настоящее время применяется модель расслоенной выборки с произвольным размещением, где каждый слой представляет одну из комбинаций выделенной системы четырех признаков группировки (региональная принадлежность, тип местности, пол и возрастная группа).

Однако в данном случае использование указанной модели не вполне адекватно, так как является оптимистическим прогнозом уровня точности. Фактически в обследовании для распространения данных выборки на генеральную совокупность применяется пострасслоенная модель. Соответственно дисперсию оценки корректно рассчитывать как сумму двух компонентов. Первый, который вычисляется в настоящее время при обработке данных обследования, отражает вариацию оценивания, связанную с постслоями. Второй

же характеризует вариацию оценки, которая связана со случайностью представительства элементов выборки в постслоях.

Поэтому в данном случае для расчета дисперсии оценок рекомендуется использовать следующую формулу пострасслоенного оценивания:

$$\hat{V}ar\left(\hat{Y}_{post}\right) \cong N^2\left[\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1-\frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}+\frac{1}{n^2}\sum_{h=1}^{H}\frac{N-N_h}{N}s_h^2\right]$$

Суммирование в формуле ведется по постслоям (h=1,…, H), относящимся к фиксированному разрезу разработки данных обследования, например, по городскому населению субъекта РФ;

$\hat{Y}_{post}$ - пострасслоенная оценка суммарного показателя;

$\hat{V}ar\left(\hat{Y}_{post}\right)$ - оценка дисперсии посстрасслоенной оценки суммарного показателя;

N - объем генеральной совокупности фиксированного разреза разработки данных обследования;

n - объем выборки в фиксированном разрезе разработки данных обследования;

N_h - объем постслоя h;

n_h - объем выборки в постслое h;

s_h^2 - скорректированная дисперсия признака y в постслое h.

Также в данном случае для расчета характеристик точности оценивания можно использовать непараметрические методы автоматизированного вычисления. В частности можно применить процедуру бутстрепинга, реализованную, например, в пакете IBM SPSS (начиная с18-ой версии). Данный метод позволяет вычислить робастные оценки, в том числе средних, стандартных ошибок, доверительных интервалов и др. Бутстрепинг наиболее употребителен в качестве альтернативы для параметрических оценок в тех случаях, когда вызывает сомнение выполнение предположений этих методов оценивания.

В случае обследования рабочей силы по проблемам занятости для формирования месячной выборки применяется двухэтапный отбор с ВПР методом выборки на первом этапе и систематическим планом на втором. В целом обоснованность приведенной выше формулы расчета дисперсии оценок показателей обследования для модели с таким планом выборки и последующим пострасслоенным оцениванием не вызывает сомнений. Тем не менее, выполнение альтернативного расчета характеристик точности в данном случае оправдано.

4. Организация статистического наблюдения за трудоустройством выпускников, обеспечивающего получение представительных данных на уровне субъектов Российской Федерации

Современная отличающаяся динамизмом экономика предъявляет особые требования к рынку труда и системе профессионального образования, требует большую

гибкость при подготовке специалистов, их адаптацию к новым изменяющимся условиям на рынке труда. Перефразируя известное высказывание, лучше быть эксплуатируемым, чем непригодным к эксплуатации из-за отсутствия адекватных современной экономике профессиональных навыков.

Трудоустройство выпускников, привлечение молодых профессионалов для восполнения и развития кадрового потенциала предприятий относятся к числу тех проблем, от решения которых зависит их экономическое благополучие и конкурентоспособность, а также темпы экономического, технологического, инновационного развития страны. Вопросы привлечения квалифицированных специалистов, повышения профессионального уровня кадров находятся в центре внимания у руководителей предприятий, а вопросы трудоустройства выпускников – у руководителей образовательных учреждений и органов управления образованием.

Успешное трудоустройство выпускников является результатом сбалансированной государственной политики в сфере образования и на рынке труда, предусматривающей создание нормативных и экономических рычагов, способствующих как подготовке специалистов в необходимых объемах и с требующимися компетенциями, так и их адекватного использования.

К основным направлениям деятельности Правительства Российской Федерации на период до 2012 года, утвержденным распоряжением Правительства Российской Федерации от 17 ноября 2008 г. N 1663-р, а также к приоритетным направлениям в сфере образования относится приведение содержания и структуры профессиональной подготовки кадров в соответствие с современными потребностями рынка труда.

Одна из стратегических целей Минобрнауки России1 напрямую связана с обеспечением текущих и перспективных потребностей экономики и социальной сферы в профессиональных кадрах необходимой квалификации, созданием условий для развития непрерывного образования. Она ориентирована на достижение целей Правительства Российской Федерации по повышению уровня и качества жизни населения (в части удовлетворения потребностей граждан в образовании) и обеспечению динамичного и устойчивого экономического развития (в части обеспечения эффективной занятости населения и удовлетворения потребностей экономики в трудовых ресурсах и развития международного экономического сотрудничества). Основными индикаторами оценки достижения соответствующих задач являются: «Структура подготовки кадров в учреждениях профессионального образования по уровням (удельный вес в общей численности выпускников)» и «Удельный вес нетрудоустроенных выпускников очной формы обучения государственных и муниципальных учреждений профессионального образования в общей их численности».

Реализация задачи по приведению содержания и структуры профессионального образования в соответствие с потребностями рынка труда конкретизирована в Федеральной целевой программе развития образования на 2011–2015 годы (ФЦПРО). Одним из ее целевых показателей является «Доля выпускников дневной (очной) формы обучения по основным образовательным программам профессионального образования (включая программы высшего профессионального образования), трудоустроившихся не позднее завершения первого года после выпуска, в общей численности выпускников

[1] Доклад о результатах и основных направлениях деятельности на 2012–2014 годы Минобрнауки России.

дневной (очной) формы обучения по основным образовательным программам профессионального образования соответствующего года».

С 2009 г. Минобрнауки России совместно с Минздравсоцразвития России и Рострудом во взаимодействии с субъектами Российской Федерации и образовательными учреждениями отрабатывает механизмы трудоустройства выпускников учреждений профессионального образования, в т.ч. ежемесячного мониторинга распределения выпускников образовательных учреждений, корректировку объемов и профилей подготовки в образовательных учреждениях с учетом результатов мониторинга, организацию стажировок выпускников и их последующее трудоустройство, содействие самозанятости выпускников, включая обучение основам предпринимательской деятельности, поддержка малых инновационных предприятий, создаваемых в вузах, центров содействия трудоустройству выпускников в образовательных учреждениях, взаимодействие с работодателями. Разрабатываются модели деятельности субъектов Российской Федерации по изменению ситуации в сфере трудоустройства выпускников.

Ясно, что компетентные решения могут приниматься только при условии наличия достоверной и актуальной информации. Однако официальные статистические данные по вопросу трудоустройства выпускников не дают реальной оценки их выхода на рынок труда.

Так, форма № ВПО-1 «Сведения об образовательном учреждении, реализующем программы высшего профессионального образования» содержит данные о распределении выпускников высших учебных заведений очной формы обучения за счет средств бюджетов всех уровней. Весь выпуск распределяется на получивших направления на работу (в том числе в соответствии с заключенными договорами (контрактами), из них в рамках целевой контрактной подготовки), не получивших направления на работу (выделяется численность женщин и численность не получивших направлений на работу из-за отсутствия заявок), а также самостоятельно трудоустроившихся, продолжающих обучение на следующем уровне по очной форме и призванных в ряды Вооруженных сил. Данные разрабатываются в разрезе специальностей и направлений подготовки. Однако отсутствие показателей о распределении выпускников, обучающихся на платной основе, всех форм обучения, искажают полную достоверную картину об изучаемом явлении.

Форма № СПО-1 «Сведения об образовательном учреждении, реализующем программы среднего профессионального образования» содержит данные о распределении выпускников учреждений среднего профессионального образования очной формы обучения за счет средств бюджетов всех уровней. Показатели разрабатываются в разрезе специальностей и базовому уровню образования (лица, имеющие основное общее образование, и лица, имеющие среднее (полное) общее образование). Весь выпуск распределяется на получивших направления на работу (в том числе в соответствии с заключенными договорами (контрактами), из них в рамках целевой контрактной подготовки), не получивших направлений на работу (выделяется численность женщин и численность лиц, не получивших направление на работу из-за отсутствия заявок), а также самостоятельно трудоустроившихся, продолжающих обучение на следующем уровне по очной форме, призванных в ряды Вооруженных сил. Однако отсутствие показателей о распределении выпускников СПО, обучающихся на платной основе, всех форм обучения, не позволяют получить полную достоверную картину об изучаемом явлении.

Форма № 1 (профтех) «Сведения об образовательных учреждениях, реализующих программы начального профессионального образования» фиксирует данные об общей

численность выпускников обученных за отчетный год за счет бюджета и по договорам. Численность выпускников распределяется на направленных на работу в организации и не направленных на работу по различным причинам (призваны на военную службу, поступили в образовательные учреждения высшего и среднего профессионального образования, предоставлено свободное трудоустройство и др.). Из численности выпускников, которым предоставлено свободное трудоустройство, выделяются те, которым было предоставлено это право из-за несогласия выпускника с предложенными условиями контракта работодателя и из-за отсутствия рабочих мест.

Форма № 2-Т (трудоустройство) «Сведения о предоставлении государственных услуг в области содействия занятости населения» содержит данные о численности выпускников образовательных учреждений начального, среднего и высшего профессионального образования, состоящих на регистрационном учете в государственных учреждениях службы занятости, а также о лицах в возрасте 18–20 лет из числа выпускников учреждений начального и среднего профессионального образования, ищущих работу впервые. Однако соответствующие показатели не дают полного представления о трудоустройстве выпускников учреждений профессионального образования, поскольку фиксируют только информацию об официально зарегистрированных лицах, но, как правило, их значительно меньше, фактически.

Форма № 1-кадры «Сведения о дополнительном профессиональном образовании работников организации» фиксирует данные о приеме в организацию и выбытии выпускников очной формы обучения образовательных учреждений высшего, среднего и начального профессионального образования, независимо от того, была ли работа в данной организации не первая. В соответствующем разделе показываются: численность выпускников образовательных учреждений высшего, среднего и начального профессионального образования, окончившие эти образовательные учреждения в 2009-2010 гг. (датой окончания соответствующего образовательного учреждения является дата выдачи диплома) трудоустроившиеся в отчитывающуюся организацию в 2009г. и в 2010г.; а также численность выпускников, уволенных в этот период по всем основаниям.

На основании всех указанных источников информации сформировать некую картину положения рынке труда молодых специалистов возможно, однако она не несет в себе необходимого содержательного наполнения, особенно в части вопросов трудоустройства и закрепляемости профессиональных кадров.

Согласно международному опыту проведения исследований по проблематике трудоустройства выпускников, себя зарекомендовал подход, при котором в анкету обследования по вопросам занятости населения добавляется специализированный блок вопросов. Так, статистическими офисами разных стран в обследование рабочей силы (Labour Force Survey) регулярно включаются разделы по интересующим вопросам. Это такие модули как «Образование в течение всей жизни» и «Переход из образования в трудовую сферу». Это позволяет для получения оценок по интересующим вопросам, касающимся трудоустройства выпускников, использовать план анализа данных выборки обследования занятости населения.

Вопросы трудоустройства молодых людей и их карьерного продвижения регулярно и весьма серьезно изучаются в большинстве стран Организации экономического сотрудничества и развития. Специальные проекты непосредственно по проблематике профессионального образования и профессиональной подготовки вот уже более десяти лет реализуются Европейским фондом образования. Примером может служить

долговременный проект ЕФО «Реформа профессионального образования и обучения», главной целью которого явилось создание эффективных моделей реформирования профессионального образования и обучения, которые можно было бы использовать в различных секторах экономики и в различных странах. В рамках данного проекта проводились опросы работодателей по изучению предъявляемых требований к работникам определенных профессий. Опросы осуществлялись самими образовательными учреждениями по десятку профессий, соответствующих уровню МСКО 3, а также уровню МСКО 5В.

Таким образом, методологической основой организации статистического наблюдения для получения данных, позволяющих охарактеризовать положение на рынке труда специалистов, получивших профессиональное образование, является выборочное обследование населения по проблемам занятости (ОНПЗ), проводимое. Росстатом с 1992 г. на основе выборочного метода наблюдения с последующим распространением итогов на всю численность населения обследуемого возраста.

ОНПЗ проводится путем опроса населения по стандартизированному бланку анкеты с готовым текстом вопросов и вариантов ответов, которые расположены в логической последовательности. По его итогам для лиц, имеющих профессиональное образование, рассчитываются следующие показатели в разбивке по уровням профессионального образования, полу и возрасту: численность занятых, безработных, экономически неактивного населения; уровень занятости; уровень безработицы; наличие специальности (профессии), подтвержденной соответствующим дипломом, свидетельством, удостоверением или другим подобным документом; наличие и характер основной работы; причины поиска новой работы и время, затраченное на ее поиск и др. Однако в настоящее время выделить в их составе выпускников образовательных учреждений не представляется возможным. Поэтому включение в анкету опроса ОНПЗ соответствующих идентифицирующих вопросов позволит. Обоснованно формировать необходимую содержательную информацию о положении на рынке труда молодых специалистов.

Обследование трудоустройства выпускников учреждений профессионального образования рекомендуется проводить раз в год в составе обследования по проблемам занятости, желательно в декабре/ноябре. Данное предложение было обосновано в научно-исследовательской работе: «Разработка и практическая апробация аналитических показателей общественной и общественно-экономической эффективности мероприятий по реализации государственной политики в сфере образования в условиях новой модели финансирования - бюджетирования, ориентированного на результаты», выполненной в 2008-2010 гг. Кроме этого, это предложение подтверждается результатами исследования по проблемам взаимосвязи профессионального образования и рынка труда, которое более пяти лет проводятся по заказам Минобрнауки России в рамках Мониторинга экономики образования. Вопросы трудоустройства затрагиваются в обследованиях студентов, преподавателей и руководителей учреждений профессионального образования и в опросах домохозяйств.

Важное место при проведении обследований принадлежит применяемому методу отбора единиц наблюдения. В обследовании населения по проблемам занятости единицами наблюдения являются частные домашние хозяйства и отдельные лица в возрасте 15–72 лет – члены этих домашних хозяйств, т.е. лица, которые проживают совместно и обеспечивают себя всем необходимым для жизни, ведут общее хозяйство,

полностью или частично объединяя и расходуя свои средства. Они могут быть связаны отношениями родства или отношениями, вытекающими из брака, либо быть не родственниками, либо и теми, и другими. В данном обследовании в качестве единиц наблюдения рекомендуется учитывать всех членов домашних хозяйств, закончивших учреждение профессионального образования (начального, среднего или высшего) не более чем за пять лет до момента опроса.

Так как наблюдение рекомендуется осуществлять на выборочной основе с использованием выборки единиц наблюдения для обследований населения по проблемам занятости, поэтому формирование итогов, то есть получения распространенных данных, должно осуществляться на основе индивидуальных коэффициентов взвешивания, используемых при обработке данных выборки ОНПЗ.

Оценка характеристик точности показателей программы обследования трудоустройства выпускников учреждений профессионального **образования.**

На основе результатов пилотного статистического наблюдения за трудоустройством выпускников учреждений профессионального образования была произведена оценка точности и достоверности основных распространенных показателей, соответствующих вопросам анкеты, на предмет их возможности обоснованного использования (публикации). Следует учесть, что в обследовании важны не только количественные, но и категориальные переменные (позволяющие определить структуру явления).

Для определения характеристик точности распространенных показателей пилотного статистического наблюдения за трудоустройством выпускников учреждений профессионального образования возможно использовать несколько моделей анализа выборочных данных. Выбор модели расчета характеристик точности базируется на принятом способе вычисления оценок, доступных данных о методе планирования выборки, а также имеющейся внешней информации об исследуемой совокупности. С учетом того, что оценка статистических показателей по результатам обследования населения по проблемам занятости основывается на микроданных и индивидуальных весовых коэффициентах, при расчете точности оценивания в настоящее время Росстатом применяется модель расслоенной выборки с произвольным размещением, где каждый слой представляет одну из комбинаций выделенной системы четырех признаков группировки (региональная принадлежность, тип местности, пол и возрастная группа).

Однако в данном случае используемая Росстатом модель расчета характеристик точности не вполне адекватна, так как является оптимистичной оценкой уровня точности. Фактически в обследовании для распространения данных выборки на генеральную совокупность применяется пострасслоенная модель, базирующаяся на известной по данным текущей статистики численности населения в региональном разрезе, по типу местности и в половозрастных группах. Соответственно дисперсию оценки корректно рассчитывать как сумму двух компонентов. Первый, который вычисляется в настоящее время при обработке данных обследования, отражает вариацию оценивания, связанную с постслоями. Второй же характеризует вариацию оценки, которая связана со случайностью представительства элементов выборки в постслоях.

Поэтому в данном случае для расчета дисперсии оценок показателей обследования ($\hat{V}ar\left(\hat{Y}_{post}\right)$) точнее использовать следующее соотношение:

$$\widehat{Var}\left(\widehat{Y}_{post}\right) \cong N^2\left[\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1-\frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}+\frac{1}{n^2}\sum_{h=1}^{H}\frac{N-N_h}{N}s_h^2\right]$$

Суммирование в формуле ведется по слоям (h=1,…, H), относящимся к фиксированному разрезу разработки данных обследования, например, по городскому населению субъекта РФ;

$\widehat{Y}_{post}$ - пострасслоенная оценка суммарного показателя;

$\widehat{Var}\left(\widehat{Y}_{post}\right)$ - оценка дисперсии посстрасслоенной оценки суммарного показателя;

N - объем генеральной совокупности фиксированного разреза разработки данных обследования;

n - объем выборки в фиксированном разрезе разработки данных обследования;

N_h - объем постслоя h;

n_h - объем выборки в постслое h;

s_h^2 - скорректированная дисперсия признака y в постслое h.

Также в данном случае для расчета характеристик точности оценивания можно использовать непараметрические методы автоматизированного вычисления. В частности можно применить процедуру бутстрепинга, реализованную, например, в пакете IBM SPSS. Данный метод позволяет вычислить робастные оценки, в том числе средних, стандартных ошибок, доверительных интервалов и др. Бутстрепинг наиболее употребителен в качестве альтернативы для параметрических оценок в тех случаях, когда вызывает сомнение выполнение предположений этих методов оценивания.

Сводка результатов анализа вычисленных характеристик точности оцененных показателей трудоустройства выпускников учреждений профессионального образования по РФ в целом представлена в следующей таблице.

Таблица 1.

Сводка результатов анализа уровней характеристик точности оценок

Интервалы группировки по количеству респондентов в выборке, отметивших категорию вопроса программы обследования	Значение коэффициента вариации оценки (%)	95%-ая предельная ошибка выборки (%)
>600	< 3	< 6
500-600	3-4	6-8
300-500	4-5	8-10
150-300	5-7	10-14
100-150	7-10	14-20
50-100	10-15	20-30
20-30	15-25	30-50
< 20	> 25	> 50

Ниже представлены типичные примеры наблюдаемых уровней точности в зависимости от количества респондентов в категориях вопросов (переменных) программы наблюдения.

Таблица 2.

Форма обучения

		Оценка	95% доверительный интервал (границы)		Коэффициент вариации	Невзвешенная частота
			Нижняя	Верхняя		
Объем генеральной совокупности	Всего	2059276,090	1973463,024	2145089,156	,021	1025
	Очная	963611,120	886689,575	1040532,665	,041	496
	Очно-заочная (вечерняя)	42206,780	22920,641	61492,919	,233	20
	Заочная	183820,410	145599,586	222041,234	,106	97
	Экстернат	4593,150	-1777,734	10964,034	,707	2
	Всего	3253507,550	3213548,544	3293466,556	,006	1640

Данные таблицы 2 показывают, что достоверными оценками (значение коэффициента вариации < 0.15) являются количества выпускников очной формы обучения (963611 чел.), заочной (183820 чел) и, конечно, оценка общего числа выпускников, всего (2059276 чел.), то есть когда фактическое число респондентов превышает 50 человек.

Таблица 3.

Тип образовательного учреждения

		Оценка	95% доверительный интервал (границы)		Коэффициент вариации	Невзвешенная частота
			Нижняя	Верхняя		
Объем генеральной совокупности	Всего по форме собственности учреждения	2061826,530	1976026,743	2147626,317	,021	1026
	Государственное, муниципальное	1096462,730	1016308,592	1176616,868	,037	567
	Негосударственное	75286,950	49758,282	100815,618	,173	37
	НЕ знаю	19931,340	6937,376	32925,304	,332	10
	Всего	3253507,550	3213548,544	3293466,556	,006	1640

Аналогично предыдущей данные таблицы 3 свидетельствуют, что число респондентов в выборке меньшее 50 человек не может обеспечить достоверной оценки показателя: для типа образовательного учреждения негосударственное число респондентов в выборке составляет 37 человек, а значение соответствующего коэффициента вариации равно 0.173; для ответа не знаю число респондентов в выборке составило 10 человек, а значение соответствующего коэффициента вариации равно 0.322; в то время как для типа образовательного учреждения Государственное, муниципальное значение коэффициента вариации равно 0.037 при 567 респондентах, отнесших себя к этой категории.

Определение необходимых объемов выборки по субъектам РФ.

Для определения нужных объемов выборки выпускников в разрезе субъектов РФ, обеспечивающих приемлемый уровень точности оценок, необходимо было провести экспериментальные расчеты. В качестве информационной основы указанных расчетов были использованы данные пилотного обследования трудоустройства и закрепляемости выпускников учреждений профессионального образования (специальный модуль в Обследование населения по проблемам занятости), проведенного Росстатом в декабре 2009 г. В ходе пилотного обследования по вопросам трудоустройства и закрепляемости

выпускников учреждений профессионального образования было опрошено 1640 респондентов в 7 регионах РФ (в среднем 18% общего объема выборки ОНПЗ в эти х регионах). В ниже следующей таблице представлено их распределение по территориальному признаку. В виду ограниченного количества выпускников в разрезе субъектов РФ, охваченных пилотом, для повышения надежности результатов вычисление характеристик точности оценок показателей осуществлялось по объединенному массиву данных выборки выпускников.

Таблица 4.

Распределение опрошенных выпускников по территориальному признаку

	Количество опрошенных выпускников	Общий объем выборки	Доля (%) выпускников в выборке
Краснодарский край	294	1587	18,5
Красноярский край	255	1572	16,2
Архангельская область	139	629	22,1
Брянская область	236	1384	17,1
Гульская область	134	879	15,2
Челябинская область	316	1293	24,4
Республика Татарстан	266	1456	18,
Итого	1640	8800	18,6

По результатам анализа вычисленных характеристик точности основных показателей статистического наблюдения за трудоустройством выпускников учреждений профессионального образования можно определить необходимые для обеспечения различных уровней точности распространенных итогов объемы выборки выпускников. Для этого используется следующее соотношение в целях определения требующегося объема выборки в целом по РФ.

Если V – фиксированное значение дисперсии $V(\bar{y})$ в предположении расслоенной выборки, то формулы вычисления объема выборки принимают следующий вид в общем случае:

$$n = \frac{\sum \frac{N_h^2 S_h^2}{w_h}}{V + \sum N_h S_h^2},$$

$$\text{где } w_h = \frac{n_h}{n}$$

где n – рассчитанный объем выборки (нужный для обеспечения фиксированной точности V в заданном разрезе разработки данных обследования);

N_h - объем слоя h;

n_h - объем выборки в слое h;

s_h^2 - скорректированная дисперсия признака y в лое h.

В случае необходимости оценки характеристик категориальных показателей (для обследования трудоустройства выпускников) исходя из выше указанной формулы можно рассчитать ориентировочные объемы выборки для каждого вопроса в зависимости от

требующейся точности (95%-ой предельной ошибки выборки (*L*)) и истинной доли признака (*p*) в исследуемой генеральной совокупности, используя соотношение:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}; \; n_0 = \frac{\sum \frac{N_h}{N} p_h (1 - p_h)}{V}$$

где n – рассчитанный объем выборки (нужный для обеспечения фиксированной точности V в заданном разрезе разработки данных обследования);

N_h - объем слоя h;

p_h - доля выборки интересующей категории признака в слое h.

Рассчитанные по указанной выше формуле объемы выборки для всех вопросов обследования можно оценить сверху следующими значениями.

Таблица 4.

Зависимость объема выборки (*n*) от предельной ошибки (*L*) и истинной доли (*P*) признака в совокупности

P / L	0.05	0.10	0.20	0.30	0.40	0.50
0.005	7 600	14 400	25 600	33 600	38 400	40 000
0.01	1 900	3 600	6 400	8 400	9 600	10 000
0.02	475	900	1 600	2 100	2 400	2 500
0.03	211	400	711	933	1 066	1 111
0.04	119	225	400	525	600	625
0.05	76	144	256	336	384	400

Как видно из приведенных в таблице результатов расчетов, для обеспечения 5%-ой ошибки оценки доли (отдельной категории признака программы наблюдения) требуется порядка 400 респондентов в самом сложном для оценивания случае, когда истинное значение доли близко к 0,5. Соответственно, если имеются 5 независимых категориальных переменных программы обследования, то для обеспечения 5%-ой ошибки оценки каждой целевой доли этих переменных объем выборки должен составлять не менее 2000 респондентов.

Для отдельно рассматриваемой переменной программы обследования с 11-ю категориями, соответствующей вопросу анкеты: «Укажите причину, по которой Вы не искали работу» (табл. 5), если требуется достоверная оценка каждой категории, объем выборки (1640 единиц) должен быть увеличен в 3-4 раза. При этом нет твердой гарантии, что оценки долей, соответствующие категориям «Собственный бизнес», «По состоянию здоровья» и «Нет необходимости», окажутся достоверными.

Подводя итоги проведенных экспериментальных расчетов по определению фактических (рассчитанных по данным пилотного обследования – 1640 респондентов) характеристик точности оцененных показателей трудоустройства выпускников учреждений профессионального образования и анализа теоретически рассчитанных объемов выборки можно сделать вывод о том, что рекомендуемый объем выборки

выпускников учреждений образования должен составлять порядка 10-12 тысяч человек. С учетом 20%-ой встречаемости по данным пилота респондентов, отвечающих критерию опроса выпускников, такой объем выборки выпускников может обеспечить рассылка анкеты обследования трудоустройства и закрепляемости выпускников учреждений профессионального образования всем индивидам, включенным в месячную выборку (например, ноябрьскую) обследования населения по вопросам занятости. Только в этом случае можно ожидать достоверные оценки показателей анкеты трудоустройства выпускников.

Таблица 5.

Причины, по которым работа не искалась

		Оценка	Стд. ошибка	95% доверительный интервал (границы)		Коэфф. вариации	Невзвеш. частота
				Нижняя	Верхняя		
% от итогового значения	Всего выпускников, не осуществляющих поиск первой работы	90,7%	,8%	89,0%	92,1%	,009	1488
	1-Получили работу в соответствии с заключенным контрактом с работодателем	,5%	,2%	,2%	1,0%	,387	7
	2-Получили работу в соответствии с распределением	,4%	,1%	,2%	,8%	,414	7
	3-Собственный бизнес	,1%	,0%	,0%	,2%	,707	2
	4-Получили предложение от работодателя	,4%	,2%	,2%	,9%	,417	7
	5-Продолжили работать на том же месте, что и во время обучения	3,8%	,5%	2,9%	5,0%	,139	57
	6-Продолжили обучение в другом учебном заведении	1,0%	,3%	,6%	1,7%	,248	18
	7-По состоянию здоровья	,2%	,1%	,0%	,5%	,600	3
	8-По семейным обстоятельствам	1,0%	,3%	,6%	1,6%	,265	16
	9-Призыв в Вооруженные Силы	1,2%	,3%	,8%	1,9%	,232	22
	10-Нет необходимости работать	,1%	,1%	,0%	,5%	,736	2
	11-Другие причины	,7%	,2%	,4%	1,3%	,315	11
	Всего	100,0%	,0%	100,0%	100,0%	,000	1640