ВЗВЕШИВАНИЕ

Гектор Малетта[1]
ред. от 12 марта 2007

1. Задача взвешивания

В работе рассмотрены вопросы взвешивания данных, роль взвешивания в статистическом анализе и процедуры взвешивания в SPSS[2]. Вот типичный пример ситуации, в которой требуется взвешивание.

Где-то в Соединённых Штатах было проведено исследование для оценки электоральных предпочтений населения перед грядущими национальными выборами. Некоторым случайным образом в выборку отобрали 1000 будущих избирателей, 500 из которых были афроамериканцами (AA), а 500 – не были афроамериканцами (NAA). Опрос показал, что 60% респондентов, попавших в выборку, собираются голосовать за демократов. Политологи сделали несколько поспешный вывод, что демократы одержат на этих выборах убедительную победу. Оказалось, это был неверный прогноз: 57% голосов было отдано за республиканцев, и только 43% – за демократов.

Подвело политологов то, что они не учли избыточной представленности чёрного населения в выборке, в то время, как оно было исключительно продемократически настроено (около 80% голосов отдавалось демократам). Белое и прочее нечёрное население оказалось недостаточно представленным, тогда как их электоральные предпочтения более равномерно распределялись между двумя партиями. Прогноз относительно 60% голосов за демократов сложился из 40% голосов за демократов среди NAA и 80% голосов за демократов среди AA в выборке. Поскольку из обеих групп избирателей было опрошено равное количество респондентов, итоговый прогноз составил 60%. На самом же деле процент чёрного населения в данном регионе США и близко не подходит к 50. Их доля в этом регионе – около 18% всего населения. Если бы каждой группе был сопоставлен вес на основе демографических данных (82% нечёрным и 18% чёрным), прогноз был бы совсем иным:

$$(0.82 \times 40) + (0.18 \times 80) = 32.8 + 14.4 = 47.2\% \text{ за демократов.}$$

И это уже не так далеко от реального результата (43%). Если бы сотрудники центра, проводившего исследование, знали о взвешивании, они могли бы сработать лучше.

Ошибка затронула лишь общую цифру прогноза. Данные по отдельным расам (говорящие о 80% демократически настроенных избирателей среди AA и о 40% – среди NAA) не были искажены. Иного не стоило и ожидать, так как проценты были получены на основе случайной выборки из лиц, принадлежащих к каждой из групп. От чрезмерной представленности афроамериканцев в выборке пострадал лишь общий результат, построенный на основе объединения данных по разным группам. Вообще, вопрос взвешивания возникает тогда, когда разные объекты (в данном случае – люди) формируют

---

[1] Hector Maletta, университет Сальвадора (Universidad del Salvador), Буэнос-Айрес, Аргентина. E-mail: hmaletta@fibertel.com.ar, или hector.maletta@mail.salvador.edu.ar

[2] Я признателен К. Дугласу (King Douglas) и Р. Левек (Raynald Levesque) за комментарии к черновику этой статьи; Р. Левек также за публикацию работы на www.spsstools.net; и Антону Балабанову, который в ходе перевода статьи на русский язык обнаружил в тексте несоответствия и неточности, требующие исправления. Я также признателен некоторым участникам форума SPSSX-L за их вопросы относительно взвешивания в SPSS, которые побудили меня подробно разобрать здесь эту тему.

некую **общую статистику**, например, процент. Назначение отдельным частям выборки большего веса, чем тот, который им присущ, приводит к ошибкам оценивания.

2. Взвешивание наблюдений и взвешивание переменных

Набор данных обычно формируется из **наблюдений** (случаев), для каждого из которых известны значения нескольких **переменных**. Таким образом, набор данных – прямоугольный массив из *n* наблюдений и *k* переменных. Статистический анализ обычно проводится путём обобщения наблюдений и (или) переменных некоторым осмысленным образом. Так, например, значения наблюдений можно обобщить по одной переменной, чтобы получить сумму или среднее значение[3]. Аналогично, значения нескольких переменных могут быть обобщены по каждому из случаев в отдельности в некоторый итоговый показатель, например, средний балл успеваемости студента на основе усреднения оценок по разным предметам.

Обычной является ситуация, когда в данных есть неравные по объёму группы **наблюдений**. Например, когда вы оцениваете средний доход по общей совокупности на основе усреднения средних доходов по группам или географическим регионам, следует обеспечить соответствующий вес каждой группы в общем среднем доходе. Скажем, у вас имеются данные по среднедушевому ВВП в Канаде, США и Мексике и вы хотите оценить среднедушевой ВВП в целом в Северной Америке. Простое усреднение трёх средних в данном случае даст некорректный результат, поскольку население США значительно превосходит по численности население двух других стран. Нужную оценку даст умножение каждого национального среднедушевого ВВП на численность населения страны, сложение произведений и деление суммы на общую численность населения в Северной Америке. Эквивалентный результат даст умножение каждого национального среднедушевого ВВП на долю населения страны в общей численности населения на континенте и дальнейшее сложение произведений.

Похожее происходит, когда взвешиваются не наблюдения, а **переменные**. Допустим, вы желаете построить шкалу на основе нескольких переменных, но некоторые из них вы считаете более важными, чем остальные. Например, некоторые оценки студентов могут быть получены за короткие недельные спецкурсы, а некоторые – за полноценные семестровые курсы, требующие заметно больших усилий (и, соответственно, большего числа аудиторных часов). Умножение каждой оценки на некоторую меру длительности и важности курса (как, например, на число кредитов в американской системе), даст более адекватную шкалу успеваемости студентов, чем простое усреднение оценок по всем предметам.

Данная статья посвящена, главным образом, взвешиванию **наблюдений**, а не переменных. А ещё точнее – взвешиванию наблюдений в тех случаях, когда этого требуют **выборочные процедуры** (в отличие от других ситуаций, как, например, из-за разных размеров трёх североамериканских государств, упоминавшихся выше). Однако же общие принципы, изложенные здесь, применимы ко всем случаям взвешивания наблюдений.

Взвешивание переменных – совсем иное дело. Когда создаётся шкала из нескольких переменных, каким-то из них можно приписать больший вес по самым разным соображениям. Обобщение значений по нескольким переменным с одинаковым весом, как, например, усреднение оценок для получения общего балла успеваемости, не всегда будет удачным, поскольку (например) оценки за более объёмные и сложные курсы должны получать больший вес. Веса могут назначаться извне (например, с учётом системы кредитов), а могут возникать из самих данных. Например, в случае с факторным анализом, когда стараются перейти от исходных переменных к малому количеству латентно воздействующих на них факторов, каждая из наблюдаемых переменных имеет

[3] В данной работе «среднее» всегда будет означать среднее арифметическое.

свой вес в каждом из извлечённых факторов. Как сказано выше, вопросы взвешивания переменных здесь не исследуются и далее затронуты не будут.

3. Выборочная процедура

Выборки, как известно, бывают случайными и неслучайными. Большинство видов статистического анализа базируется на предположении, что исследователь использует **случайную** выборку, и большая часть этой статьи посвящается взвешиванию именно случайных выборок. Взвешивание, разумеется, не может чудесным образом обратить неслучайную выборку в случайную (хотя в чём-то и может улучшить выводы, получаемые на её основе). Осуществление случайной выборки из данной генеральной совокупности, как правило, основывается на использовании одного или нескольких следующих инструментов.

Простая случайная выборка. Наблюдения отбираются из перечня наблюдений, составляющих целевую генеральную совокупность. Пример: генеральная совокупность состоит из действующих членов Американской медицинской ассоциации, списком которых Ассоциация располагает, и выборка осуществляется путём отбора одного члена Ассоциации из каждой сотни в списке с использованием некоторой случайной процедуры. Другой пример: случайный отбор 20 сенаторов США из полного списка Сената. Иногда список сам по себе не существует, но используется какой-то другой эквивалентный механизм. Так, можно отбирать каждый 10-й дом на улице, просто идя вдоль по ней и отсчитывая дома. Начальный дом для отсчёта также можно выбрать случайно. Допустим, с помощью таблицы случайных чисел можно выбрать один из первых 10 домов с северного (южного) конца улицы.

Стратифицированная выборка. Генеральная совокупность целиком разделяется на 2 или более страт, после чего в каждой страте без исключения осуществляется простая случайная выборка. Как правило, это способствует более равномерной представленности всех страт в выборке, уменьшая вероятность непропорциональной концентрации выборки в какой-то определённой части генеральной совокупности. Кроме того, если принцип стратификации как-то связан с переменными, интересующими исследователя, дисперсия этих переменных **в каждой из страт** будет меньше, чем их дисперсия по всей совокупности. Это **снижает общую ошибку выборки**. Допустим, совокупностью исследования является всё население (все домохозяйства) Нью-Йорка, а исследуемая переменная – доход домохозяйства. Если рассматривать в качестве страт районы Нью-Йорка, то, скорее всего, средний доход будет различаться между стратами, а дисперсия дохода внутри каждого района (относительно районного среднего) будет меньше, чем для всего города в целом (относительно общегородского среднего), поскольку богатое и бедное население обычно концентрируется в определённых районах (то есть, доходы будут более однородны, если рассматривать районы Квинс или Манхэттен отдельно, чем доходы для всего города без разделения на страты).

Кластеризованная (гнездовая) выборка. Генеральная совокупность или каждая страта подразделяется на некоторое количество кластеров или подгрупп по какому-то принципу; затем некоторые из этих кластеров отбираются для исследования и **внутри каждого выбранного кластера** осуществляется простой случайный отбор наблюдений в выборку. Допустим, каждый район Нью-Йорка мы разделяем на микрорайоны по принципу одинаковых почтовых индексов (точнее – ZIP-кодов – прим. перев.), затем какое-то количество микрорайонов случайным образом отбирается, и внутри каждого отобранного микрорайона осуществляется простой случайный отбор. В отличие от стратификации, гнездовая выборка получается двухступенчатой: первый раз отбираются кластеры, затем отбираются конкретные наблюдения внутри отобранных кластеров. Опять же, в отличие от стратификации, гнездовой отбор обычно **увеличивает общую ошибку выборки**. Такая опасность более высока, если совокупность разделена на относительно небольшое число мало похожих друг на друга кластеров, из которых в

исследование попадают единицы. Допустим, для построения национальной выборки США разделяются на 3 макрорегиона (3 страты) в соответствии с временными зонами (восточной, центральной и западной), а затем внутри каждого макрорегиона выбираются 2 штата (кластера). Результаты исследования будут существенно различаться, если, скажем, вместо штатов Нью-Йорк и Мэн будут отобраны штаты Джорджия и Алабама. Если же имеется множество небольших кластеров (как, например, в случае с ZIP-кодами), и количество отобранных кластеров также высоко, вероятность существенных искажений заметно снижается.

Стратификация снижает ошибку выборки для выборок заданного размера, либо позволяет минимизировать размер выборки для заданных границ ожидаемой ошибки, что делает стратификацию почти обязательной процедурой в реальных исследованиях неоднородных совокупностей. Гнездовой отбор, напротив, увеличивает ошибку выборки, что делает этот способ не всегда приемлемым, если только не учитывать его существенные стоимостные преимущества. При концентрации объектов исследования на небольших географических пространствах он позволяет сэкономить значительные средства на полевой работе. А, кроме того, получение полного перечня объектов внутри кластера, например, перечня домохозяйств внутри одного микрорайона с одинаковым ZIP-кодом – более простая задача, чем получение такого перечня для всей совокупности.

Во многих реальных исследованиях указанные три инструмента часто комбинируются. В частности, крупные социальные исследования обычно сочетают все три вида отбора. Так, например, страна может быть подразделена на региональные страты, затем внутри страт выделяются и отбираются кластеры (районы, города, поселения и т.д.), затем, возможно, выделяются и отбираются более мелкие кластеры (микрорайоны, кварталы и проч.), после чего уже конкретные индивиды или домохозяйства отбираются внутри небольшого кластера.

Выборки с возвращением и без. Обычная выборка из генеральной совокупности осуществляется без возвращения (WOR[4]). Единожды отобранный объект не может быть отобран повторно, и дальнейший выбор делается из оставшейся совокупности. Иногда используется иной метод отбора, называемый выборкой с возвращением (WR). В этом варианте отобранный объект возвращается в совокупность и, таким образом, имеет шанс быть отобранным повторно. В таком случае все объекты отбираются из одной и той же совокупности, не исключая объекты, которые уже были отобраны. В случае WR один и тот же объект имеет шансы быть отобранным несколько раз и, таким образом, быть представленным в результирующей выборке в нескольких экземплярах. Это может рассматриваться как разновидность взвешивания: объект, отобранный 3 раза, имеет вес, равный трём, а объект без повторов будет иметь вес, равный единице. К этой идее мы вернёмся позже, но далее будем полагать, что выборка осуществлялась способом без возвращения.

Пропорциональный и непропорциональный отбор. Выборка – часть соответствующей генеральной совокупности. Эта часть характеризуется **выборочным отношением (долей)** – объемом выборки, делённым на объём генеральной совокупности (n/N). Это отношение – вероятность того, что некоторый объект из генеральной совокупности попадает в выборку. Если мы отбираем каждое десятое домохозяйство, отношение составляет 1/10 = 0.10, то есть, каждое домохозяйство имеет вероятность 10% оказаться в выборке. Рассуждая более широко, если n_k – размер выборки из группы генеральной совокупности, размером N_k, то выборочное отношение $f_k = N_k/n_k$.

В стратифицированных и гнездовых выборках выборочное отношение по разным причинам может варьироваться от одной группы генеральной совокупности к другой. В зависимости от того, происходит это или нет, выборки могут быть подразделены на выборки с переменным или постоянным выборочным отношением. В выборках с

[4] WithOut Replacement – без возвращения – примеч. перев.

переменным отношением, например, из-за непропорционального отбора, для некоторой пары индивидов вероятности попадания каждого из них в выборку могут различаться.

Например, если выборка составляется из списков членов нескольких научных обществ, таких как Американская ассоциация экономистов (AEA) и Американская ассоциация психологов (APA), возможны ситуации, что шансы экономистов попасть в выборку составляют 1:10, а шансы психологов – 1:200, если выборочные отношения (а значит и вероятности для членов ассоциаций попасть в выборку) составляют, соответственно, 0.100 и 0.005.

Планируемое и фактическое выборочное отношение. Порой реальное число наблюдений в выборке отличается от запланированного по каким-то причинам (например, некоторые респонденты отказались от интервью, либо отсутствовали дома при повторных визитах, либо анкеты были заполнены, но потом отбракованы из-за низкого качества и т.д.). Тогда как планируемое отношение обычно воплощается в продуманных процедурах отбора, таких, как использование случайных чисел или случайном отборе из перечня, отказы от интервью или отбракованные анкеты могут подчиняться неслучайным закономерностям. Допустим, высокодоходные домохозяйства чаще могут отказываться отвечать на вопросы, а низкодоходные домохозяйства чаще могут возвращать заполненные вопросники низкого качества (особенно если те даются респондентам для самостоятельного заполнения).

В некоторых исследованиях неудачные интервью замещаются другими в предположении, что замена равноценна. Такие предположения чаще обоснованы, если замещается домохозяйство, представителей которого никак не удавалось застать дома, но менее правдоподобны, если замещается домохозяйство, отказавшееся от интервью, поскольку оно может принципиально отличаться от своего «заместителя», согласившегося отвечать на вопросы.

Наличие замен, равно как и незапланированных изменений размеров выборки, может быть следствием предположений и волевых решений, в некотором смысле произвольных, которые и будут определять, какой размер, планируемый или фактический, будет использован при расчёте выборочного отношения (и, таким образом, весов, как будет показано ниже). Подобные вопросы специфичны для конкретного исследования и, в интересах достижения целей данной работы, не рассматриваются. Посему n_k является **просто** размером выборки для k-й группы генеральной совокупности, без дальнейших объяснений. Обычно это **фактический** размер выборки, но в некоторых случаях он может включать поправки на отказы или замены.

4. Взвешивание выборки

Выборка, по определению, меньше, чем соответствующая генеральная совокупность. А кроме этого, она может отличаться от совокупности своими пропорциями, если некоторые группы представлены в выборке избыточно или недостаточно. То есть, выборка отличается от генеральной совокупности масштабом, пропорциями, либо и тем, и другим. Приписывание весов выборкам имеет общую цель сделать выборку более похожей на представляемую генеральную совокупность. Две основные (не исключающие друг друга) цели взвешивания таковы:

а) **исправление масштаба**;

б) **исправление пропорций**.

Исправление масштаба. В некоторых случаях важно оценить по выборке абсолютные значения в генеральной совокупности, для чего в таблицах отражается реальный размер генеральной совокупности, а не размер выборки. Например, справка о том, что трансляцию финала Кубка мира посмотрели 2.5 миллиарда телезрителей, звучит интереснее, чем факт, что его посмотрели 15 365 опрошенных телезрителей в разных странах (откуда, возможно, и появилась оценка в 2.5 миллиарда).

Корректировка масштаба от выборки к генеральной совокупности обычно осуществляется с помощью весов общего вида $w=N/n$, когда выборочные оценки домножаются на обратное выборочное отношение. Общий, или единый **масштабирующий коэффициент** будем обозначать через v и определим его как

$$v = \frac{N}{n} \quad \textbf{(1)}$$

Допустим, имеется простая случайная выборка врачей из Американской медицинской ассоциации. Просто отобрали n врачей из списка N врачей. Вы хотите установить, сколько врачей-членов АМА курят. Вместо цифр относительно курящих в вашей выборке вы приводите результат, умноженный на $v=N/n$, который теперь близок к общему числу курящих врачей в США. Тут используется единый масштабирующий коэффициент для всех наблюдений, так как выборка являлась простой случайной.

Пропорциональное взвешивание. В сложных выборочных планах, например, в стратифицированной или гнездовой выборках, выборочные отношения могут меняться. Как результат, пропорции в целом в выборке могут не совпадать с пропорциями в генеральной совокупности. Корректировка проводится с помощью пропорциональных весов, которые определяются отдельно для каждой выборочной подгруппы, имеющей единое выборочное отношение (например, для каждой страты) и обозначаются как π_k с общей формулой π_k = *% страты в ГС / % страты в выборке*:

$$\pi_k = \frac{N_k/N}{n_k/n} \quad \textbf{(2)}$$

При таких весах страта k получает пропорциональный вес $\pi = 1$, если данная группа представлена в выборке в той же пропорции, что и в генеральной совокупности. Но $\pi_k < 1$, если группа была избыточно представлена в выборке (то есть, её доля n_k/n больше, чем доля N_k/N). И $\pi_k > 1$, если группа была недостаточно представлена (её доля в выборке была меньше доли в генеральной совокупности). Пропорциональные веса «утяжеляют» недостаточно представленные наблюдения и «облегчают» избыточно представленные.

Смешанные или интегрированные веса. Пропорциональные веса вида (2) не расширяют результаты выборки до цифр генеральной совокупности. С их помощью приводятся в соответствие лишь доли. С другой стороны, мы видели, что простое увеличение выборочных показателей до масштаба генеральной совокупности с помощью единого масштабирующего коэффициента N/n не исправляет доли в выборке. Смешанный вес, обладающий и тем, и другим свойством, получается перемножением двух рассмотренных ранее весов:

$$w_k = v\,\pi_k \quad \textbf{(3)}$$

Если множитель, заданный уравнением (1), умножается на множитель (2), для отдельно взятой страты результат упрощается до обратного выборочного отношения:

$$w_k = \frac{N}{n} \times \frac{N_k/N}{n_k/n} = \frac{N_k}{n_k} \quad \textbf{(4)}$$

Веса, определяемые по формуле (4) осуществляют разом две функции: приводят цифры выборки в соответствие с масштабом генеральной совокупности и исправляют дисбаланс в выборочных отношениях от страты к страте. В этом смысле они могут называться смешанными или интегрированными весами.

Взвешивание при наличии субстрат и кластеров. Когда план выборки включает иерархическую схему разложения крупных страт на субстраты, либо некоторую форму кластеризации совокупности, веса, в принципе, подсчитываются по тем же законам. Итоговое выборочное отношение есть результат последовательных перемножений выборочных отношений разных уровней; то же самое справедливо и для обратных показателей. Допустим, например, что стратификация первого уровня разбивает совокупность на k страт (например, штаты США), затем каждую страту подразделяют h субстрат (например, округа). Затем некоторая доля кластеров p_{kh} (например, районов с

ZIP-кодами) выбирается внутри субстраты, и, наконец, из каждого кластера (т.е. из каждого выбранного района с определённым ZIP-кодом) выбирается некоторая доля конечных единиц выборки q (например, доля домохозяйств). Для простоты положим, что у нас отсутствует информация о числе домохозяйств, приходящихся на один район с ZIP-кодом, и это число предстоит оценить на основе подсчёта домохозяйств в выбранных районах. Каждое домохозяйство из m-го ZIP-района k-го округа, относящегося к i-му штату, получает вес, равный

$$w_{mki} = \frac{H_{mki}}{h_{mki}} \frac{Z_{ki}}{z_{ki}} \quad \textbf{(5)}$$

Проще говоря, вес равен обратному вероятности того, что домохозяйство будет выбрано в своём ZIP-районе, умноженному на обратное вероятности того, что ZIP-район будет выбран из множества ZIP-районов в пределах данного округа. Поскольку в выборку включаются все округа и все штаты (это страты), вероятность попадания каждого из них в выборку равна 1, а потому эти вероятности опущены в выражении (5) для данного примера. Вообще говоря, вероятность того, что некоторый объект будет включён в выборку, равна произведениям вероятностей отбора на всех стадиях от верхнего уровня до нижнего. Если некоторый начальный элемент выборки отбирается с вероятностью p_i, следующие элементы в рамках первого отбираются с вероятностью $p_{(i)j}$, т.е. с вероятностью p_j в рамках i-й отобранной единицы, и так далее, пока элементарная единица не будет отобрана с вероятностью $p_{(ijkm\ldots)z}$ из предпоследней единицы отбора. Тогда итоговая вероятность выбора элементарной единицы будет равна

$$p_{ijkm\ldots z} = p_i p_{(i)j} p_{(ij)k} p_{(ijk)m} \cdots p_{(ijkm\ldots)z} \quad \textbf{(6)}$$

Итоговый смешанный вес (осуществляющий как масштабирующее, так и пропорциональное взвешивание) является просто обратным от этой вероятности:

$$w_{ijkm\ldots z} = \frac{1}{p_{ijkm\ldots z}} = \frac{1}{p_i p_{(i)j} p_{(ij)k} p_{(ijkm)} \cdots p_{(ijkm\ldots)z}} = \frac{1}{p_i} \frac{1}{p_{(i)j}} \frac{1}{p_{(ij)k}} \frac{1}{p_{(ijkm)}} \cdots \frac{1}{p_{(ijkm\ldots)z}} \quad \textbf{(7)}$$

Умножив веса $w_{ijkm\ldots z}$ на n/N (т.е. на простое общее выборочное отношение), мы преобразуем их к пропорциональным весам без масштабирующего эффекта, ранее обозначавшимся как π_k (в данном случае их надо было бы обозначить как $\pi_{ijkm\ldots z}$ из-за многоступенчатой процедуры отбора).

5. Выборка и генеральная совокупность: закон больших чисел

От выборочных результатов мы ожидаем отражения законов генеральной совокупности, из которой сделана выборка. Эти ожидания формализуются в математической теории статистического оценивания, краеугольным камнем которой является так называемый *закон больших чисел*. В упрощенном понимании этот закон гласит, что если повторно осуществлять случайные выборки одного и того же объёма из одной и той же генеральной совокупности и вычислять каждый раз некоторую статистику по выборке (например, выборочное среднее), являющуюся оценкой генерального параметра (например, генерального среднего), то выборочные оценки могут отличаться друг от друга, но распределение оценок будет подчинено нормальному закону со своим средним, равным значению генерального параметра (оцениваемого среднего).

Допустим, были осуществлены несколько выборок по n взрослых мужчин из соответствующей ГС, и в каждой выборке был измерен рост каждого мужчины в сантиметрах. Все выборки – одного и того же объёма (например, по 100 человек). Для каждой выборки подсчитан средний рост. Для k-й выборки средний рост обозначим как H_k. Выборочные средние, скорее всего, будут различаться между выборками. Если осуществляется много выборок, распределение выборочных средних всё больше будет напоминать нормальное с центром в «средней из средних» (т.е. средней из выборочных средних), и это среднее из средних будет всё больше похоже на средний рост мужчин в генеральной совокупности. Мы говорим «будет напоминать» и «будет похоже», так как

всё это будет функцией от а) числа сделанных выборок и б) размера выборок. Чем больше будут выборки и чем больше мы их осуществим, тем сильнее распределение средних будет напоминать колоколообразное распределение, и тем лучше их общее среднее будет приближать генеральное среднее.

Распределение переменной и выборочное распределение. Переменная в генеральной совокупности (или выборке) может иметь любое частотное распределение. Индивидуальные величины роста в выборке мужчин (или их ГС) включают великое множество разных значений; некоторые из этих мужчин выше среднего, некоторые – ниже. В случае с ростом распределение индивидуальных показателей роста внутри выборки или внутри ГС само по себе, скорее всего, будет похоже на нормальное, поскольку большинство биологических показателей имеют этот тип распределения, тогда как другие переменные (например, доход), могут быть скошенными или иметь иные отклонения от нормального. Некоторые переменные имеют равномерное распределение, некоторые – U-образное распределение, когда наблюдения концентрируются на полюсах отрезка значений с меньшим их количеством в середине, какие-то имеют убывающую частоту, когда присутствует много небольших значений и всё меньше и меньше всё более крупных, а вообще – форма распределения может быть любой, какую только можно представить. Другими словами, распределение индивидуальных значений переменной в выборке или в генеральной совокупности заранее неизвестно и почти не относится к нашему разговору.

Каким бы ни было распределение переменной в выборке, оно всегда будет иметь среднее. Если осуществляется множество выборок, мы будем иметь дело с множеством **выборочных средних**. **Выборочное распределение** переменной – это не распределение её индивидуальных значений в выборке вокруг выборочного среднего, а распределение **выборочных средних** вокруг **«среднего средних»** генеральной совокупности. Первое понятие существует «в пределах выборки», второе понятие существует «среди выборок».

Правила статистического вывода требуют, чтобы выборки были случайными, так как существует доказанная математическая теорема (подтверждаемая и эмпирическими результатами), что выборочное распределение выборочных средних, полученных из множества случайных выборок, сделанных из одной и той же генеральной совокупности, имеет тенденцию к нормальному распределению, средняя которого имеет тенденцию к совпадению с генеральным средним[5]. Ничего подобного не предполагается и не требуется относительно распределения индивидуальных значений внутри выборки вокруг выборочного среднего. Особо подчеркнём этот важный факт, поскольку многие люди ошибочно понимают нормальность выборочного распределения как требование к переменным иметь нормальное распределение относительно своих выборочных средних.

Статистическая значимость выборочных результатов. Когда в исследовании имеется лишь одна выборка, как это чаще всего случается в реальной жизни, аналитик знает, что выборочное среднее, скорее всего, не ушло слишком далеко от генерального, поскольку средние множества выборок имеют нормальное распределение и, кроме того, среднее этого нормального распределения (среднее средних) очень близко к среднему значению для всех индивидов в генеральной совокупности. На самом деле имеющаяся выборка может иметь среднее очень далёкое от генерального, но вероятность этого относительно мала, если выборка случайная, а размер её достаточно большой. Так, например, есть вероятность около 95%, что выборочный средний рост окажется в пределах двух стандартных отклонений от генерального среднего роста. Никто, впрочем, не знает, не попала ли наша конкретная выборка в оставшиеся 5% «неудачных» выборок

[5] Выборочное среднее здесь относится к непрерывным или интервальным (не путать с переменными, чьи значения объединены в интервалы – примеч. перев.) переменным, таким как возраст или доход, либо к среднему бинарных переменных (например, процент женщин, который эквивалентен среднему из 1 (для женщин) и 0 (для мужчин)). Таким образом, средние и проценты рассматриваются как эквивалентные категории.

со средними, заметно меньшими или большими истинного (возможно, выборка почти полностью составилась из карликов или, напротив, игроков NBA, несмотря на то, что была результатом совершенно корректно выполненного случайного отбора).

Но аналитики готовы к такому риску. Будучи реалистами, они ничего не утверждают точно и не стремятся к полной уверенности. Они довольствуются определённым уровнем вероятности. Их вполне устроила бы, например, такая формулировка: «По выборке, объёмом n случаев, средний рост был оценён в 175 см, и с вероятностью 95% эта оценка не отклоняется от истинного среднего более, чем на 3 см в любую сторону». Доверительный интервал ±3 см на уровне доверия 95% – это тот интервал, куда, как ожидается, попадут 95% всех возможных выборок. Всегда существует вероятность совершения ошибки в случае с выборкой, которая попадёт в остальные 5%, чьи выборочные средние находятся вне пределов доверительного интервала вокруг генерального среднего, но эта вероятность достаточно низка и считается допустимым риском при работе с выборками[6].

Как было сказано, выборочный интервал определяется некоторым числом **стандартных отклонений выборочного распределения**. Для уровня доверия 95% интервал определяется 1.96 стандартного отклонения (СО) с каждой стороны выборочного среднего. Разумеется, указание на то, что выборочное среднее находится в пределах 1.96 СО от генерального эквивалентно утверждению, что генеральное среднее находится в пределах 1.96 СО от выборочного. Стандартное отклонение, о котором идёт речь – это **не** стандартное отклонение переменной в выборке (изменчивость роста между индивидами), а стандартное отклонение **выборочного распределения** (изменчивость среднего роста между выборками из одной и той же ГС).

Стандартное отклонение выборочного распределения для переменной X называется также **стандартной ошибкой** выборочной оценки для X. Она определяется по следующей формуле:

$$SE_X = \frac{\sigma_X}{\sqrt{n}} \qquad \textbf{(8)}$$

В этой формуле σ_X – стандартное отклонение (стандарт) переменной в генеральной совокупности, а n – размер выборки. Доверительный интервал на уровне 95% – это ±1.96 SE от выборочной оценки. Обратите внимание, что SE всегда меньше (обычно – заметно меньше), чем стандартное отклонение переменной в ГС, поскольку для получения SE стандартное отклонение переменной делится на квадратный корень размера выборки. То есть, например, для выборки в 100 единиц SE будет в 10 раз меньше, чем стандартное отклонение, поскольку квадратный корень из 100 равен 10.

Величина σ_X, числитель выражения (8), **генеральное СО** переменной X, разумеется, неизвестно, так же как и генеральное среднее. Известны лишь выборочные среднее и стандарт. Обычно предполагается, что приемлемую оценку генерального стандарта, σ_X, даёт s_X – стандартное отклонение **в выборке**, одной единственной выборке, которая была фактически сделана. То есть, на самом деле, стандартная ошибка оценивается как

$$SE_X = \frac{s_X}{\sqrt{n}} \qquad \textbf{(9)}$$

Разумеется, осуществлённая выборка – лишь одна из многих возможных выборок. Так же, как и выборочные средние варьируются в разных выборках вокруг генерального среднего, выборочные стандартные отклонения s_X варьируются в разных выборках вокруг генерального стандарта σ_X. Нет гарантий, что наблюдаемый выборочный стандарт s_X совпадает с генеральным стандартом σ_X (или даже находится поблизости от него), но

[6] Изменение уровня вероятности даст нам более узкий или широкий интервал. Другие часто встречающиеся доверительные интервалы – это 90% (более узкий) и 99% (более широкий). В общем, чем больший уровень доверия вы хотите иметь для своей оценки, тем более широким должен быть доверительный интервал.

обычно другой возможности оценить *SE* просто нет. Утешением служит тот факт, что изменчивость выборочных СО обычно ниже, чем изменчивость выборочных средних, так что наблюдаемое выборочное СО, скорее, является достаточно хорошей оценкой генерального (при условии, что выборка достаточно велика и осуществлена методом случайного отбора).

Теперь – как вычисляется s_X? В простой случайной выборке несмещённую оценку генерального стандартного отклонения даёт выборочное стандартное отклонение, но со знаменателем n-1 вместо обычного n:

$$s_x = \sqrt{\left(\frac{\sum(x_i - \bar{x})^2}{n}\right)\left(\frac{n}{n-1}\right)} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \qquad \textbf{(10)}$$

В **стратифицированной** выборке лучшая оценка генерального стандарта даётся не этой формулой, которая соответствовала бы стандартному отклонению в отдельной страте. В этом случае требуется **взвешенное среднее стандартных отклонений в каждой страте**, подсчитанных **относительно среднего в каждой страте** вместо использования общего среднего. Полагая, что каждая страта репрезентирует долю f_k генеральной совокупности и составляет долю g_k в выборке, оценка генерального стандартного отклонения есть среднее стандартных отклонений, вычисленных по всем стратам, взвешенное по доли страт в генеральной совокупности (f_k).

$$\hat{\sigma}_X = \sum_k f_k \sqrt{\frac{\sum_i (x_{ik} - \bar{x}_k)^2}{n_k - 1}} \qquad \textbf{(11)}$$

Полученная в результате взвешивания результатов выборки оценка генерального стандартного отклонения и должна быть использована затем в выражении (9) для оценки стандартной ошибки среднего по стратифицированной выборке. Выборочное среднее, в свою очередь, также рассчитывается как взвешенное среднее по выборочным средним из разных страт. Таким образом, в случае стратифицированной выборки оценка стандартной ошибки и доверительных интервалов зависит от доли страт в генеральной совокупности, а также от размера выборки (n). Если средние в разных стратах различаются, *SE* по стратифицированной выборке объёма n будет меньше, чем по простой случайной выборке того же объёма.

При оценке параметров по **кластеризованным** выборкам сложность оценивания связана с тем, что в каждой страте отобраны лишь несколько кластеров, и это может исказить оценку изменчивости переменной во всей страте. Исследуя одни кластеры и игнорируя остальные, мы как бы делаем предположение, что отобранные кластеры репрезентируют страту в целом, что не всегда справедливо (представьте случайную выборку в нескольких отобранных штатах США; отобранные штаты могут оказаться, главным образом, западными, либо с северо-западного побережья, либо преимущественно южными, с вероятно, очень разными характеристиками). Легко представить, что изменчивость переменной внутри кластеров меньше, чем во всей страте, поскольку объекты из одного кластера географически близки, какие-то из их характеристик могут полностью совпадать, а остальные – быть очень схожими. Например, вариация денежных доходов в пределах одного квартала будет меньше, чем в районе или во всём городе. Наблюдая лишь некоторые кластеры, можно вообразить изменчивость более низкой, чем она есть на самом деле. Более того, когда имеется небольшое число крупных кластеров, лишь некоторые из которых попадают в выборку, существует реальная опасность того, что выбранные кластеры нерепрезентативны для среднего или стандартного отклонения в генеральной совокупности. Так, средний уровень некоторой политической оценки в выборке из преимущественно южных штатов может не совпадать со средней оценкой в целом по стране, а **изменчивость** этой переменной в выборке может быть значительно меньшей, чем изменчивость, существующая по стране в целом.

Другими словами, использование кластеризованных выборок увеличивает выборочную ошибку, но увеличивает её на неизвестную величину (если только у нас нет какого-то внешнего источника информации об изменчивости, характерной для совокупности или страты из которой отбираются кластеры). Аналитики часто делают удобное предположение, что подобных неоднородностей не существует, то есть вариация между членами отобранных кластеров равна вариации, которую можно было бы наблюдать во всей генеральной совокупности или страте, но такое предположение разумно лишь в том случае, если общее число кластеров относительно велико, равно как относительно велико и число отобранных кластеров. Таким образом, случайный отбор трёх округов в штате с девятью округами более опасный вариант, чем случайный отбор 300 ZIP-зон на территории, состоящей из 900 ZIP-зон. В первом случае неоднородность между округами может быть слишком велика и (например) отбор округа, где расположена столица штата приведёт к иным результатам, чем в случае, если бы этот округ не был отобран. С 300 из 900 ZIP-зон риск распределяется более равномерно и поводов для беспокойства заметно меньше.

Короче, если только у вас нет оценок относительно масштабов возможного увеличения выборочной ошибки из-за кластеризации, применение приведённых выше формул для выборок, в которых на каком-то этапе использовалась кластеризация, может привести к недооценке выборочной ошибки. Если оценка генерального стандартного отклонения производится по формулам (10-11), она может оказаться заниженной, и далее заниженной окажется и стандартная ошибка оценивания. Напротив, если кластеризация не использовалась, формулы (10-11) для случайных стратифицированных выборок дадут несмещённую оценку стандартной ошибки.

Взвешенные и невзвешенные данные в проверках значимости. Для выборок, построенных по принципам, отличным от принципа простого случайного отбора, подсчёт статистик значимости на основе невзвешенных данных даст смещённую оценку значимости, как и при любом другом статистическом оценивании (средние, стандартные отклонения, проценты и всё, что угодно).

Выборочные отношения и, следовательно, веса, должны приниматься в расчёт при вычислении достоверных статистик критериев значимости по выборочным данным. Однако здесь критическим является **пропорциональный** аспект взвешивания (приписывание каждому наблюдению соответствующего относительного веса), а не эффект **масштаба** (вычислительное расширение размера выборки до размера генеральной совокупности).

6. Взвешивание в SPSS

По умолчанию SPSS не применяет веса к наблюдениям в матрице данных. Если матрица включает *n* наблюдений, итоговое значение в частотной таблице будет равно *n*. Можем рассматривать это как ситуацию, когда каждое наблюдение имеет вес, равный 1. Но пользователь может использовать индивидуальные значения любой переменной в качестве весов для наблюдений. Например, команда WEIGHT BY X включит взвешивание наблюдений по значениям переменной X.

Веса могут быть постоянными (все наблюдения имеют один и тот же вес) или переменными (каждое наблюдение может иметь свой, иногда – уникальный, вес). Если задан постоянный вес, например, 100, тогда каждое наблюдение реплицируется 100 раз, а итоговое значение в таблице частот будет в 100 раз больше размера выборки. Использование таких постоянных весов влияет на **масштаб** частот без изменения **пропорций** между ними. Если выборка включает, к примеру, 50 мужчин и 40 женщин, взвешенные частоты будут равны 5 000 и 4 000, соответственно. Взвешенная частота – это, фактически, сумма всех весов (в данном случае: 50 x 100 = 5 000 и 40 x 100 = 4 000).

Пользователь может также приписать особый вес каждому случаю в отдельности, что является более распространённой ситуацией. Допустим, что выборка была

стратифицированная и в трёх стратах наблюдения были отобраны случайным образом с выборочными отношениями 0.05, 0.005 и 0.0005. Веса (по формуле N_k/n_k) будут обратными этим долям, то есть, 20, 200 и 2000, соответственно. Пользователь может приписать эти веса путём создания переменной (назовём её X), которая будет равна 20 для каждого наблюдения из первой страты, 200 для каждого наблюдения из второй страты и 2000 для наблюдений из третьей страты. Затем используется команда WEIGHT BY для указания переменной X в качестве весовой. С того момента, как будет выполнена команда WEIGHT BY, SPSS будет умножать каждое наблюдение на его вес всякий раз, когда потребуется проделать какую-либо процедуру статистического анализа. Рассмотрим пример файла данных с 7 наблюдениями, относящимися к трём стратам. В начале файл не взвешивается.

Средний возраст рассмотренных субъектов будет найден путём умножения каждого возраста на соответствующий вес (1, по умолчанию), сложения этих произведений и деления полученной суммы на **сумму весов**, которая в данном случае равна числу наблюдений (7). Результат: 44.57 лет для семи наблюдений, рассматриваемых в примере.

Номер набл.	**Вес**	**Страта**	**X**	**Возраст**	**Возраст x Вес**
1	1	1	20	18	18
2	1	1	20	21	21
3	1	2	200	42	42
4	1	2	200	38	38
5	1	2	200	37	37
6	1	3	2 000	72	72
7	1	3	2 000	84	84
Итоги					
n=7	**Sum=7**			**312**	**312**

Среднее = 312 / 7 = 44.57

Строго говоря, это взвешенное среднее, но, поскольку все веса равны 1, ситуация тривиальна. Поэтому мы называем это невзвешенным средним, но не потому, что взвешивание не производится, а потому, что все веса равны 1.

После выполнения команды WEIGHT BY X значения переменной X используются как веса[7]. Среднее будет рассчитываться по той же формуле: умножаем значения возраста на веса, складываем и делим на сумму весов, как показано в следующей таблице.

Вес	**Страта**	**X**	**Номер набл.**	**Возраст**	**Возраст x Вес**
20	1	20	1	18	360
20	1	20	2	21	420
200	2	200	3	42	8 400
200	2	200	4	38	7 600
200	2	200	5	37	7 400
2 000	3	2 000	6	72	144 000
2 000	3	2 000	7	84	168 000
Итоги					
Sum=4 640			**n=7**	**312**	**336 180**

Среднее = 336 180 / 4 640 = 72.45

[7] При включении взвешивания пользователь указывает в качестве веса саму переменную, а не её конкретные *значения* на текущий момент времени. Если значения X впоследствии изменятся, например, после команды преобразования или импорта значений из внешнего источника, *новые* значения переменной X автоматически будут использованы как веса без повторного указания X в качестве весовой переменной.

Сумма весов теперь равна 4 640 и средний возраст, как легко проверить, 72.45 лет. Средний возраст теперь выше, чем в предыдущем случае, поскольку пожилым индивидам (наблюдения 6 и 7) был приписан заметно больший вес, чем остальной части выборки. Строка «Итоги» показывает не только сумму возрастов (312) и число наблюдений в выборке (7), но и сумму весов, а также сумму произведений возраста на вес. Отношение двух последних показателей 336 180 / 4 640 = 72.45 и есть **взвешенный средний возраст** для семи субъектов в выборке. Это есть соответствующим образом взвешенная оценка среднего возраста в генеральной совокупности, которая представлена данной выборкой. Ещё раз обратим внимание, что в обоих случаях среднее было получено делением суммы произведений «Возраст x Вес» на сумму весов, но в первом случае все веса были равны единице.

Веса, использовавшиеся во втором примере – **смешанные** или **интегрированные** веса. Они увеличивают масштаб выборки и в то же время корректируют выборочные пропорции. По этой причине общее «число наблюдений» (равное сумме весов) оказывается равным 4 640 вместо 7. Такие веса, вообще говоря, являются обратными выборочным отношениям: N_k/n_k.

Все частотные таблицы, строящиеся по взвешенному файлу, покажут общий итог, равный 4 640, т.е. сумме весов, поскольку 7 наблюдений в файле «репрезентируют» 4 640 субъектов генеральной совокупности.

В некоторых случаях нам нежелательно иметь такой эффект масштаба, увеличивающий итоговые частоты. Смешанные веса мы можем легко заменить чисто пропорциональными. Для этого создадим новую переменную Y как произведение X на n/N = 7 / 4 640 = 0.00091623.

```
COMPUTE Y = X * 7 / 4 640.
FORMAT Y (F12.8).
```

[Команда FORMAT заставляет SPSS отображать максимум 12 символов для значений переменной Y, включая десятичную точку и 8 знаков после неё. Это не влияет на внутреннюю точность хранения переменной (SPSS всегда использует точность примерно до 15 знаков после точки), только на отображение на экране.] В результате веса получились следующими:

Номер набл.	X	Y = X*n/N
1	20	0.03017241
2	20	0.03017241
3	200	0.30172414
4	200	0.30172414
5	200	0.30172414
6	2 000	3.01724138
7	2 000	3.01724138
Итого	**4 640**	**7.00000000**

Поскольку Y равна старой переменной X, делённой на N и умноженной на n, сумма значений Y равна 7, т.е. размеру выборки (с учётом погрешностей округления). Теперь взвешенное число наблюдений (т.е. сумма весов) вновь равно 7, без изменения масштаба. Любая частотная таблица, построенная по данным, взвешенным по переменной Y, даст общий итог в 7 наблюдений. Однако же теперь веса не все равны 1. Какие-то из них меньше, чем 1, какие-то больше. Первые два наблюдения считаются программой как приблизительно 0.03 наблюдения каждое, следующие три наблюдения – как 0.3 наблюдения каждое, последние два наблюдения – примерно как 3 наблюдения каждое.

Оценка среднего возраста для семи наблюдений производится по взвешенным данным, то есть, с умножением возраста на соответствующий вес. Среднее, которое не зависит от масштаба выборки, неизменно равно 72.45 (как было с использованием смешанных, масштабирующих весов по переменной X). Следующая таблица показывает, как в данном случае был вычислен средний возраст по вымышленной выборке в 7 наблюдений. Обратите внимание, что сумма весов (первая колонка) равна числу наблюдений в выборке.

Вес	Страта	X	Y	Номер набл.	Возраст	Возраст x Вес
0.03017241	1	20	0.03017241	1	18	0.5431034
0.03017241	1	20	0.03017241	2	21	0.6336207
0.30172414	2	200	0.30172414	3	42	12.6724138
0.30172414	2	200	0.30172414	4	38	11.4655172
0.30172414	2	200	0.30172414	5	37	11.1637931
3.01724138	3	2 000	3.01724138	6	72	217.2413793
3.01724138	3	2 000	3.01724138	7	84	253.4482759
Итого						
Sum=7				**n=7**		**507.1681034**

Среднее = 507.1681034 / 7 = 72.45

Подытожим. SPSS по умолчанию полагает, что всем наблюдениям приписан единичный вес, но пользователь может указать любую переменную в качестве весовой. Обычно веса являются переменными из файла данных, как X и Y в рассмотренных выше примерах. Любая [числовая] переменная может быть задана в качестве весовой посредством команды синтаксиса WEIGHT BY … Альтернативный способ взвешивания – через графический интерфейс пользователя посредством меню Data / Weight data – Weight by … Когда файл взвешен по какой-либо переменной, он остаётся таковым пока это порядок не будет изменён командой отключения веса WEIGHT OFF или не будет указана другая переменная взвешивания. Если файл с включенным взвешиванием был сохранён, он останется таким и после открытия в следующий раз. Обратим внимание, что взвешивание использует **текущие** значения весовой переменной: если переменная была пересчитана, файл автоматически перевзвешивается с использованием новых значений весов. В SPSS нет необходимости назначать эту переменную весовой повторно только из-за того, что изменились её значения.

Если сумма весов не равна числу наблюдений, т.е. если веса являются **мастшабирующими**, итоговые значения в таблицах будут изменены до масштаба, заданного суммой весов, обычно – до масштаба генеральной совокупности. Это повлияет на все частоты и итоговые характеристики, включая, например, число наблюдений в таблице или суммарный доход всех домохозяйств в таблице, и так далее.

Веса могут изменять **масштабы** выборки, а также **пропорциональную** значимость индивидуальных наблюдений по отношению к остальным наблюдениям. Но даже при наличии масштабирующих весов **относительные** характеристики, такие как средние или проценты, меняются лишь за счёт **пропорциональной** составляющей весов. Если заданы индивидуальные (различные) веса для каждого наблюдения, то даже если сумма весов совпадает с числом наблюдений в файле, взвешенные средние, скорее всего, будут отличаться от простых, невзвешенных средних.

Если же ко всем наблюдениям будет применён одинаковый, постоянный вес (например, «размножение» каждого наблюдения в N/n раз), будут изменены итоги, зависящие от масштаба, но такие статистики, как средние и проценты не будут изменены, поскольку пропорциональные соотношения между наблюдениями не меняются.

Веса принимаются в расчёт только если они положительны. Любое неположительное значение (будь то отрицательное, нулевое или пропущенное) в весовой

переменной не используется. Если некоторое наблюдение имеет отрицательный, нулевой вес или пропущенное значение в переменной веса, оно не включается в анализ, а SPSS при этом выдаёт предупреждение.

7. Проверки значимости и взвешивание в SPSS

Как сказано выше, SPSS всегда рассматривает данные как взвешенные, даже если по умолчанию все веса постоянны и равны 1, а потому могут не приниматься в расчёт. Когда статистическая процедура требует подсчёта общего числа наблюдений, SPSS неизменно использует в качестве этого значения сумму весов, а не число реальных наблюдений в файле, даже если эти значения совпадают (что чаще всего и происходит).

Это означает, например, что при подсчёте стандартных ошибок, где используется квадратный корень из объёма выборки (см. (8) и (9)), стандартное отклонение делится на квадратный корень из суммы весов. Если сумма весов отличается от фактического числа наблюдений в файле, как в случае с масштабирующими весами, SPSS может разделить стандартное отклонение на величину, значительно большую квадратного корня из фактического объёма выборки. Если это так, оценка стандартной ошибки будет обманчиво мала.

Для многих выборочных обследований исследователям доступны веса смешанного (интегрированного) типа, объединяющие в себе функции масштабирования и корректировки пропорций выборки. Для исследователя это палка о двух концах:

- Если проводить проверки значимости по взвешенным данным, используя веса с функцией масштаба, пропорции между стратами будут соответствовать ГС, но стандартные ошибки будут недооценены, поскольку масштабирующие веса «заставят» SPSS «думать», что размер выборки столь же велик, как и размер генеральной совокупности. Если, скажем, выборка в 1 000 человек взята из ГС, размером в 20 миллионов, множество мелких различий, которые не будут значимы при таком размере выборки, после взвешивания станут значимыми, поскольку SPSS сочтёт, что эти различия проявляются на «выборке», размером в 20 миллионов.
- Если оставить данные невзвешенными, проверки значимости также могут давать неверные результаты, если выборочные отношения варьируются между группами наблюдений. Масштаб выборки будет верный, но пропорции между стратами могут быть искажены. Выборочные наблюдения из недостаточно репрезентированной страты получат тот же вес, что из избыточно репрезентированной страты, вызывая ошибки в оценках как средних, так и стандартных ошибок (средний возраст в примере выше будет оценён как 44 вместо 72 лет, а также может быть неверной и стандартная ошибка).

Идея в том, что проверки значимости корректно применять лишь с пропорциональными весами, не создающими эффекта масштаба. Это достижимо путём домножения существующих (смешанных) весов на n/N, после чего сумма весов станет равна n, а правильные пропорции между стратами сохранятся.

Такой подход будет давать несмещённые оценки и корректные проверки значимости то тех пор, **пока в расчёт не надо будет принять кластеризацию выборки**. Фактически, ограничением здесь является внутрикластерная корреляция переменных. Краткая характеристика такого ограничения такова: «Степень недооценки [стандартных ошибок] зависит от размера внутрикластерного коэффициента корреляции для анализируемых переменных. Чем выше внутрикластерная корреляция, тем в большей степени можно недооценить изменчивость переменных»[8].

[8] Donna Brogan, “Pitfalls of using standard statistical software packages for sample survey data” из Rollins School of Public Health, Emory University, Atlanta, 1997, опубликовано как статья в «**Энциклопедии биологической**

На практике полагают, что если кластеры многочисленны (то есть страта разделена на **большое** число относительно **небольших** кластеров, таких как ZIP-зоны в штате, а затем **значительное** число из этого множества кластеров участвовало в выборке), вклад кластеризации в ошибку оценки не будет большим. С другой стороны, в случае «редких» кластеров (то есть страта разделяется на **небольшое** число относительно **больших** кластеров, таких как округа в штате, а затем лишь **малая** часть из этих кластеров попадает в выборку) опасность недооценки стандартной ошибки будет больше. Это практическое правило зиждется на том, что в случае большого числа небольших кластеров внутрикластерные корреляции, скорее, будут относительно невелики. Когда кластеризация не используется или все внутрикластерные коэффициенты корреляции весьма малы, где-то в районе нуля, тогда использование чисто пропорциональных весов будет давать корректные оценки изменчивости данных и стандартных ошибок и, таким образом, обоснованные оценки значимости.

Для целей работы со сложными выборочными планами вообще, и особенно – для выборок с гнездовым отбором – SPSS предлагает специальный модуль, **Complex Samples** («сложные выборки» – примеч. перев.), который считает стандартные ошибки для большого числа выборочных планов[9]. Подобные инструменты существуют и в других статистических программных пакетах, например в SUDAAN. По возможности, проверки значимости на сложных выборках с кластеризацией должны проводиться с применением такого программного обеспечения. Пропорциональные веса при наличии крупных кластеров и ненулевых внутрикластерных корреляций, скорее всего, приведут к недооценке ошибок и переоценке значимости.

Дисперсия выборочных распределений статистик (средних, процентов, регрессионных коэфициентов и т.д.) может быть также оценена так называемыми бутстреп-методами (bootstrapping). По существу эта вычислительная процедура связана с генерацией большого числа подвыборок одного и того же объёма из существующей выборки. В каждой из таких подвыборок вычисляется интересующая исследователя статистика (например, среднее), после чего становится возможным оценить параметры выборочного распределения средних на основе сопоставления результатов по всем подвыборкам[10].

статистики» (Encyclopedia of Biostatistics, edited by P. Armitage and T. Colton, Wiley, 1998), а также во втором её издании 2005 года.

[9] Ранее Complex Samples был независимым компонентом, производившимся другой компанией, WesVar, и распространявшимся SPSS. Позднее, начиная с 13 версии, корпорация SPSS включает Complex Samples как интегрированный модуль в пакет SPSS.

[10] У нас имеется лишь одна выборка размера *n* и мы не знаем выборочное распределение статистики для всех возможных выборок этого объёма. Бутстреп-метод оценивает выборочное распределение, беря большое число подвыборок **из единственной выборки**, которая у нас имеется. Выборка объявляется генеральной совокупностью размера *n* из которой производится множество вторичных выборок объёма *n* **с возвращением**. То есть, некоторое наблюдение А, будучи раз отобранным во вторичную выборку, имеет шансы быть отобранным в эту же вторичную выборку ещё и ещё раз. Таким образом, требуемый размер вторичной выборки *n* может быть достигнут ещё до того, как некоторое другое наблюдение (например, В) хоть раз поучаствует в ней. Иными словами, возможно, что новая выборка из *n* наблюдений будет включать две или более копий наблюдения А и ни одной копии наблюдения В, таким образом, давая особую оценку среднего интересующей нас переменной. Ситуация эквивалентна тому, когда имеется множество копий одной и той же исходной выборки, но всякий раз в этой выборке разным наблюдениям приписываются разные веса (по числу повторений наблюдений в бутстреп-выборке). В примере выше, если наблюдение А попало в подвыборку 3 раза, а наблюдение В – ни разу, результатом будет выборка, где наблюдению А приписан вес 3, а наблюдению В – вес 0. После повторения этой процедуры много раз (возможно, много тысяч раз), распределение средних по подвыборкам будет иметь свою дисперсию, которая может являться оценкой генеральной дисперсии выборочного распределения. Такая процедура не может включать в анализ наблюдения из генеральной совокупности, которые не попали в начальную, базовую выборку, но, как показано, оценки, полученные таким способом, как правило, достаточно хороши. См. краткое описание метода в C.Z. Mooney and R.D. Duval, **Bootstrapping: A Nonparametric Approach to Statistical Inference** (Sage Publications, Quantitative Applications in the Social Sciences Series No. 95, Newbury Park, Ca, 1993).

Подводя итоги раздела, скажем, что проведение проверок значимости в SPSS без применения модуля Complex Samples при наличии в выборке неравных выборочных отношений требует взвешивания данных; более того, требуется, чтобы веса были чисто пропорциональными (то есть не изменяли масштаба выборки, а сохраняли сумму весов, равную размеру выборки). Если выборочный план включает кластеризацию и особенно – «редкую» кластеризацию в смысле, описанном выше (несколько крупных кластеров вместо множества мелких), и/или много или все внутрикластерные коэффициенты корреляции нельзя назвать мелкими, тогда следует использовать модуль SPSS Complex Sampling (или SUDAAN).

8. Прочие вопросы взвешивания в SPSS

8.1. «Ловушки» дробных весов

Когда в выборку попадает каждый объект из 20, то есть, выборочное отношение равно 0.05, масштабирующий вес будет равен 20. Это означает, что для моделирования генеральной совокупности каждое наблюдение из выборки будет повторено 20 раз. Однако не всегда обратное выборочное отношение составляет аккуратное целое число. Иногда вес дробный. Например, если размер страты в генеральной совокупности составляет 125 467, из которых в выборку попало 1 283 наблюдения, выборочное отношение составит 1 283 / 125 467 = 0.0102258 (округлено до 6 знаков после запятой). Обратное этому числу: 97.791894 и, таким образом, для моделирования совокупности каждое наблюдение в выборке воспроизводится более, чем 97 раз, но менее, чем 98 раз.

В целом, SPSS принимает дробные веса, хотя это и привносит некоторые сложности в работу, о которых полезно упомянуть. Одна проблема состоит в том, что существует статистическая процедура, которая **не** принимает дробные веса, а именно – регрессия Кокса (Cox Regression – основной инструмент анализа выживаемости, который оценивает влияние независимых переменных на шансы испытуемого сохранить исходное состояние («выжить») в течение более или менее длительного периода времени перед переходом в другое состояние). Если исследователь использует дробные веса, они должны быть усечены до целых, иначе процедура не будет работать.

Другой вопрос с дробными весами – **округление**. SPSS автоматически округляет взвешенные **частоты** до ближайшего целого[11]. Округление происходит на уровне итоговых частот, а не индивидуальных весов. Из-за этого могут возникать небольшие несоответствия в результирующих таблицах. В следующем гипотетическом примере в таблице сопрягаются две переменные: утверждение в должности в случаях, когда кандидат знакомился с некоторой информацией и в случаях, когда кандидат с информацией не знакомился. Шесть кандидатов не прошли утверждение, 3 из них были знакомы и 3 – не были знакомы с информацией. Они отображаются как 3 без ознакомления и 3 с ознакомлением до взвешивания, как 293 без ознакомления и 293 с ознакомлением после взвешивания, однако общее число не прошедших утверждение после взвешивания не равно 293+293=586, а равно 587.

Невзвешенные

	Без ознакомления	С ознакомлением	Итого
Прошли	…	…	…
Не прошли	3	3	6 = 3 + 3

Взвешенные

	Без ознакомления	С ознакомлением	Итого
Прошли	…	…	…
Не прошли	293	293	587 ≠ 293 + 293

[11] Поздние версии SPSS дают возможность пользователю управлять округлением в процедурах типа CROSSTABS.

Небольшое расхождение в значениях – почти обязательное следствие округления при использовании дробных весов. Когда веса «большие», округление до целых не меняет их сути слишком сильно. Между 98 и 97.79, например, относительная разница лишь 0.2% – вероятно, нет поводов для беспокойства. Но в случае «небольших» весов, близких к 1, отбрасывание дробной части числа может привести к значительным скачкам в значениях весов. Это часто встречается в случае «чисто пропорциональных весов» общего вида $(N_k/N)/(n_k/n)$. Обычно – это числа чуть меньшие или чуть большие единицы, например, 0.8 или 1.45. Оба значения будут округлены до 1, что стирает все следы взвешивания. Допустим, для некоторого наблюдения вес равен 0.3. Если в некоторой ячейке таблицы находится одно наблюдение с весом 0.3, взвешенный (округлённый) итог для этой ячейки будет равен 0. Если другая ячейка содержит 2 наблюдения из этой же страты, суммарная частота будет составлять 0.6, или 1 в округлённом варианте. Если третья и четвёртая ячейки имеют, соответственно, по 3 и 4 наблюдения из этой же страты, суммы весов будут составлять 0.90 и 1.20, и, опять же, округлятся до одного наблюдения. Это означает, что одно наблюдение в выборке обращается оценкой в виде пустой категории, тогда как 2, 3 и 4 наблюдения дадут оценку частоты категории, равную 1. Таким образом, переход от пустой ячейки к ячейке с одним наблюдением (одна ячейка вовсе без наблюдений, а другая – с частотой 0.3, что «соответствует» нулевой же частоте) не производит какого-либо видимого эффекта. Удвоение числа наблюдений от 1 к 2 даёт ячейку с частотой 1, но последующее удвоение с 2 до 4 вновь не даёт каких-либо видимых отличий.

Всё это выглядит не вполне корректно, разумеется. Когда рассматриваются большие выборки, добавление или удаление пары наблюдений не вызывает большой тревоги, но в детализированных таблицах со множеством ячеек многие ячейки могут «пострадать» от обозначенной выше проблемы, например, возможно появление «пустых» ячеек, которые на самом деле «населены» наблюдениями, либо появление ячеек с одинаковыми частотами, которые на самом деле скрывают разное количество наблюдений. Кроме статистической стороны вопроса, читатели могут быть сбиты с толку такими «визуальными эффектами».

Аналогичная, но особая проблема, действительно дающая поводы для беспокойства, такова, что SPSS округляет лишь взвешенные **частоты**, но не **статистики для интервальных переменных** в таблицах. Допустим, некоторая ячейка показывает частоту, например, количество людей, а соседняя с ней ячейка – их суммарный доход. Допустим, что в первую ячейку попал лишь один человек с весом 0.3, который имеет доход 10 000 долл. Значение первой ячейки SPSS округлит с результатом в ноль наблюдений (1 набл. x 0.30 = 0.30), но значение второй ячейки для общего дохода будет равно 10 000 x 0.30 = 3 000 долл. Создастся впечатление, что это «ничей» доход, поскольку количество людей указано нулевое.

Чтобы избежать подобных проблем, в поздних версиях SPSS в таблицах не отображаются средние или любые итоги, вычисленные при суммах весов, меньших 1, но это, в свою очередь, может приводить к несоответствиям в суммах по строкам или столбцам. Но даже если сумма весов превосходит 1, SPSS будет отображать неокруглённые суммы и средние, что может вызвать некоторые несоответствия. При использовании дробных весов, такие веса, как 14.05 и 14.49, округлятся как 14, тогда как средние доходы, например, округляться не будут. Допустим, в двух ячейках стоит одинаковый взвешенный суммарный доход 4 214.53 долл., суммарные веса ячеек: 14.02 и 14.49. Частоты в обеих ячейках отобразятся как 14, итоговый доход – как 4 214.53 долл., но средние доходы будут разными: 300.6 для первой ячейки и 290.85 – для второй. Для больших частот, скажем, сотен или тысяч, несостыковка в знаках после запятой вряд ли будет проблемой, но для ячеек, содержащих одно- или двухзначные частоты, подобные несоответствия могут сбивать с толку. Желательно в таких случаях сопровождать таблицы пояснениями о неточностях, связанных с округлениями.

Чтобы избежать подобных проблем в целом, лучше использовать целые веса для вычислений, связанных с частотами, но, как показано ранее, для вычислений, связанных со стандартными ошибками и проверками на значимость, предпочтительнее иметь чисто пропорциональные веса, сохраняющие размер выборки. Такие веса, как правило, небольшие и содержат значимую дробную часть. Таким образом, пользователям можно иметь два набора весов для одних и тех же данных. Один набор – **целые** веса с эффектом масштаба, другой набор – чисто пропорциональные веса с возможной дробной частью. Если пользователю требуется получить частотную таблицу **и** связанную с ней проверку значимости (например, таблицу сопряжённости и соответствующую статистику хи-квадрат), лучше, если сама таблица будет построена по целым весам (возможно, с эффектом масштаба) для соответствия всех частот, а проверка значимости проведена по чисто пропорциональным весам. Так, например, процедура CROSSTABS SPSS должна быть выполнена дважды, с разными весами. Табличные результаты пользователь берёт из результатов первого запуска, а статистику хи-квадрат – из результатов второго. Следующий синтаксис показывает этот пример.

```
VAR LABEL X 'Смешанные или интегрированные веса' Y 'Чисто пропорциональные веса'.
WEIGHT BY X.
CROSSTABS OPINION BY SEX.
WEIGHT BY Y.
CROSSTABS OPINION BY SEX / STATISTICS = CHISQ.
WEIGHT OFF.
```

В этой последовательности команд сначала наблюдения взвешиваются по переменной X, представляющей набор смешанных или интегрированных целых весов, корректирующих как масштаб, так и пропорции. Затем запрашивается таблица сопряжённости, демонстрирующая распределение некоторых мнений по полу. Результирующая таблица (итоги которой соответствуют значениям генеральной совокупности, поскольку веса масштабируют выборку) будет использована для представления данных в табличной форме. Затем данные перевзвешиваются по переменной Y, содержащей набор чисто пропорциональных весов, сохраняющих размер выборки. Во втором вызове CROSSTABS вместе с частотной таблицей запрашивается статистика хи-квадрат. Таблицу сопряжённости мы не рассматриваем, но статистику хи-квадрат будем использовать для проверки нулевой гипотезы о независимости мнений и пола. Для процедур, не строящих частот, например, REGRESSION, рекомендуются пропорциональные веса как не влияющие на размер выборки и, таким образом, дающие корректные результаты проверок значимости.

Но в выборках, включающих кластеризацию, даже тут тесты могут давать смещённые оценки значимости; тогда желательно использовать модуль Complex Samples.

8.2. Пустые категории в таблицах и графиках

Рассмотрим ещё один, иногда очень удобный вариант использования взвешивания в SPSS. До 12-й версии SPSS категории без наблюдений не отображались в таблицах и графиках. Например, если в вопросе о том, какую страну посещали последний раз, ни один из респондентов не указал «Швейцария», Швейцария не будет отображена в таблице или графике к этому вопросу. Однако же, часто пользователь желает, чтобы отображались все категории переменной, даже «пустые». В SPSS не существует специальной установки, которая бы позволяла это делать, однако можно пойти на хитрость со взвешиванием. Хитрость заключается в добавлении **нового фиктивного наблюдения**, для которого последняя страна посещения – Швейцария, и приписывании ему очень малого веса, например, 0.000001. Такой вес будет округлён до нуля и это наблюдение не будет видно в частотных таблицах, но категория «Швейцария» будет показана.

Проиллюстрируем этот трюк. Предположим, что вы уже добавили вручную нужное наблюдение в конец набора данных. Допустим, вам нужно, чтобы Швейцария отображалась в таблице посещённых стран по полу путешественников. Валидные значения последнего наблюдения необходимы лишь для 3-х переменных: ID, LASTTRIP и SEX[12]. Допустим, что для фиктивного наблюдения ID=99999 и значение переменной LASTTRIP содержит код, соответствующий Швейцарии. Пол воображаемого путешественника может быть мужским или женским – без разницы. Допустим, пол женский (SEX=2 в соответствии с принятой в исследовании классификацией). Допустим, что ранее файл был взвешен по переменной OLDWGT. Далее можно использовать следующие команды:

```
COMPUTE NEWGT=OLDWGT.
IF (ID=99999) NEWGT=0.000001.
WEIGT BY NEWGT.
CROSSTABS LASTTRIP BY SEX.
```

В построенной таблице страны посещения по полу будет отображена строка, соответствующая Швейцарии, однако вся строка будет пустая, так как единственное наблюдение в этой группе (женщина) получило практически нулевой вес. Чтобы быть более точным, ячейка для женщин будет содержать ноль, а ячейка для мужчин будет вовсе пустой. Чтобы иметь нули в обеих ячейках, вы можете создать **два** фиктивных наблюдения – одно «мужское» и одно «женское». Данная процедура не повлияет на взвешенную численность женщин или общее число наблюдений. Ещё раз заострим внимание, что фиктивные наблюдения должны иметь значения не только для страны посещения, но и для пола, так как если значение пола будет пропущено, процедура CROSSTABS будет рассматривать это наблюдение как пропущенное и в любом случае не включит его в таблицу. Разумеется, как только фиктивное наблюдение выполнит свою задачу, его лучше удалить из файла чтобы предотвратить ошибки или путаницу в будущем. Кстати говоря, когда пользователь заполняет фиктивное наблюдение значениями других переменных, он может подставить такие значения, чтобы «пустые» категории отображались и для других переменных как в примере со страной посещения выше.

Однако есть одно предостережение. До нуля округляется частота фиктивного наблюдения, но не значение интервальной переменной, его характеризующей. Допустим, в таблице анализируется взаимосвязь страны последнего посещения и годового дохода. Опять же, чтобы Швейцария появилась в этой таблице вы должны добавить фиктивное наблюдения с валидными значениями как для страны посещения, так и для дохода. Как сказано выше, в последних версиях SPSS суммы и средние не отображаются для ячеек с суммой весов, меньших 1, но неизвестно, так ли это в ранних версиях. Если цифра указанного дохода достаточно велика, а версия программы не слишком новая, в соответствующей ячейке может отобразиться какое-то ненулевое значение. Например, если воображаемый годовой доход для фиктивного наблюдения составляет 100 000 долл, а вес наблюдения равен 0.0001, ячейка с частотой для Швейцарии будет содержать нулевое значение, но суммарный доход в соседней ячейке отобразится как 10 долл., что будет выглядеть некорректно. Чтобы этого избежать, в любом случае, убедитесь, что взвешенный доход достаточно мал (меньше, чем 0.50 долл.), чтобы гарантировать его округление до нуля. Например, если фиктивному наблюдению приписан доход 100 долл., его взвешенное значение составит 0.01 долл. Равновозможное решение – указывать достаточно малый вес (например, 0.000000000001 вместо 0.0001). Это гарантирует, что, если фиктивное наблюдение в дальнейшем останется в файле, оно не повлияет на статистику дохода для реальных наблюдений в других процедурах.

[12] Идентификационный номер, страна посещения и пол, соответственно – Примеч. перев.

Впрочем, использование весов для отображения пустых категорий постепенно отходит в прошлое. В SPSS версий 12 и выше процедура CTABLES позволяет пользователю указать, следует ли включать пустые категории подкомандой /EMPTY = INCLUDE. Аналогично, в SPSS версии 14 подобное можно сделать и в графиках.

8.3. Анализ готовых таблиц

Другой специальный случай использования взвешивания в SPSS связан с анализом таблиц, взятых из учебников, статей или любых других источников. Иногда в ходе исследования мы обнаруживаем какую-то таблицу в какой-то публикации и хотим проверить, скажем, значимость различий в долях или степень взаимосвязи между интересующими нас переменными. Чтобы сделать это пользователю следует создать небольшой файл данных, воспроизводящий **ячейки** этой таблицы. В таблице сопряжённости каждая ячейка соответствует некоторой комбинации значений переменных, а частоты в ячейках соответствуют числу случаев, в которых наблюдалась каждая комбинация. Допустим, есть таблица следующего содержания. Она показывает характер занятости по регионам проживания.

	Рабочий	Служащий	Специалист	Самозанятый
Восток	332	418	125	62
Запад	465	328	211	87
Юг	225	248	152	112

Исследователю интересно знать, есть ли какая-то взаимосвязь между регионом и характером занятости, но источник, в котором была найдена таблица, подобной информации не даёт. Чтобы узнать это самостоятельно, исследователь должен назначить коды всем категориям обеих переменных и подготовить в SPSS файл следующей структуры, где каждая ячейка представлена отдельным «наблюдением»:

region	**occup**	**freq**
1	1	332
1	2	418
1	3	125
1	4	62
2	1	465
2	2	328
2	3	211
2	4	87
3	1	225
3	2	248
3	3	152
3	4	112

Значение 1 переменной region, разумеется, означает Восток, 2 – Запад, 3 – Юг, аналогично – с характером занятости. Если в SPSS создать файл такой структуры и использовать переменную freq для взвешивания, команда перекрёстной табуляции воспроизведёт исходную таблицу и приведёт искомые меры взаимосвязи. Соответствующий синтаксис может выглядеть так:

```
WEIGHT BY freq.
CROSSTABS region BY occup /STATISTICS = ALL.
```

В подкоманде STATISCICS могут быть запрошены все показатели взаимосвязи, как в этом примере, либо только те, которые хочет исследователь, например, статистика хи-квадрат (заменой ключевого слова ALL на CHISQ в приведённой выше команде CROSSTABS). Для данного примера значение критерия по статистике хи-квадрат Пирсона

оценено SPSS как 101.03436, тогда как альтернативная метрика на отношении правдоподобия даёт значение 98.03069, оба критерия значимы ($p<0.00001$).

Таким образом, команда WEIGHT в SPSS не только выполняет функцию наделения наблюдений соответствующими пропорциональными и/или абсолютными весами, но может использоваться и для других целей, как показано выше.

Перевод с английского с разрешения автора

Антон Балабанов
Нижний Новгород, 2006 г.
e-mail: a-balabanov@yandex.ru
http://www.spsstools.ru