СТАТИСТИЧЕСКИЕ МЕТОДЫ РЕСАМПЛИНГА: РАНДОМИЗАЦИЯ И БУТСТРЕП

В. К. Шитиков

Институт экологии Волжского бассейна РАН , 445003 Тольятти, ул. Комзина, 10
e-mail: stok1@list.ru , авторский сайт http://www.ievbras.ru/ecostat/Kiril

Техника скоро дойдет до такого совершенства,
что человек сможет обойтись без себя самого.
С.Е. Лец

Введение

При анализе эмпирических данных недостаточно получить точечную выборочную оценку параметра числовой случайной величины. Необходимо также изучить его статистические свойства, в первую очередь распределение полученной оценки, что является основой для построения доверительных интервалов и тестирования статистических гипотез. Поскольку точный вид распределения обрабатываемых данных, как правило, неизвестен, используют приближенные методы аппроксимации истинных свойств исследуемой статистики. Классическая теория основывается на асимптотическом методе и использует то или иное стандартное предельное (при стремлении размера выборки к бесконечности) распределение выборочных параметров. Современной альтернативой асимптотическому методу является моделирование эмпирического распределения данных с использованием методов *генерации повторных выборок*.

Понятие **повторные выборки** в общем случае отличается от обычного представления, применяемого в методах выборочного анализа. Если, например, производится анализ заболеваемости и отбирается срез данных в определенном месте и в определенный момент времени, то отобрать вторую, третью и т.д. порции информации будет уже невозможно, потому что это будут уже данные из другого места или же взятые в другой момент времени. Поэтому возникает проблема: как, имея лишь одну единственную повторность, оценить значение необходимого нам показателя и получить меру точности этой оценки.

В том случае, когда нет возможности получить истинные повторности наблюдений**,** разработаны методы, которые формируют так называемые "**псевдовыборки**", и на их основе можно получить необходимые характеристики искомого параметра: оценки математического ожидания, дисперсии, доверительного интервала. Методы "**численного ресамплинга**" {по английски – resampling, поэтому в иной транслитерации - "ресэмплинг" или "ресемплинг"} или, как их иногда называют в русскоязычной литературе, "методы генерации повторных выборок" объединяют три разных подхода, отличающихся по алгоритму, но близких по сути: рандомизация, или перестановочный тест {permutation}, бутстреп {bootstrap} и метод "складного ножа" {jackknife}.

К сожалению, русскоязычный читатель может встретить на книжных полках очень ограниченное число публикаций, посвященных этой динамично развивающейся идеологии. Настоящая статья является вольным и несколько переосмысленным переводом учебных материалов, представленных проф. Дэвидом Хауэлом (D. C. Howell) из Университета в Вермонте, автором книги «Статистические методы в психологии», выдержавшей семь изданий. Первоисточник представлен на сайте http://www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html

1. Концепции ресамплинга и общая схема его реализации

1.1. Краткий исторический экскурс

Идеи численного ресамплинга не являются принципиально новыми в статистике и относятся по крайней мере к 1935 году, но практическое применение этих методик было связано с вынужденым ожиданием, пока не появятся достаточно быстрые компьютеры. Один из первых алгоритмов, предложенный М. Кенуем в 1949 г., заключался в том, чтобы последовательно исключать из имеющейся выборки по одному наблюдению, обрабатывать всю оставшуюся информацию и предсказывать результат в исключенной точке. Совокупность расхождений, полученных таким образом по всем точкам, несет в себе информацию о выборочном смещении, которой можно воспользоваться для уточнения параметров. Дж. Тьюки активно усовершенствовал этот метод, назвав его "**jackknife**" (складной нож), и использовал для оценки дисперсии изучаемой совокупности и проверки нулевой гипотезы о том, что распределение некоторой статистики симметрично относительно заданной точки. «Понятие "складной нож" относится к универсальному методу, призванному заменить частные методики, которые не всегда пригодны, подобно бойскаутскому ножу, годящемуся на все случаи жизни» (Мостеллер, Тьюки, 1982).

Bootstrap-процедура или "бутстреп" была предложена как некоторое обобщение процедуры "складного ножа". Дело в том, что формирование подвыборок в jackknife, а тем более в других методах перепроверки, обязательно связано с уменьшением числа элементов по сравнению с исходной совокупностью. Известный американский статистик, профессор Станфордского университета Б. Эфрон (Efron, 1977) в своей ранней статье "Бутстреп-методы: новый взгляд на методы складного ножа" описал алгоритм «выбора с возвращением», в котором формально сохраняются неизменными степени свободы на каждом этапе обработки данных. Он предложил строить новые выборки, *моделируя выборки из эмпирического распределения*, или другими словами, взять конечную совокупность из n элементов исходной выборки и с помощью датчика случайных чисел сформировать из нее любое число размноженных выборок. Процедура, хотя и нереальна без ЭВМ, проста с точки зрения программирования.

По одной из версий, слово "**bootstrap**" означает кожаную полоску в виде петли, прикрепляемую к заднику походного ботинка для облегчения его натягивания на ногу. Благодаря этому термину появилась английская поговорка 30-х годов: «Lift oneself by the bootstrap», которую можно трактовать как «Пробить себе дорогу благодаря собственным усилиям» (или вытянуть себя из болота за шнурки от ботинок, что заставляет вспомнить о подвигах барона Мюнхгаузена).

Одновременно с внедрением методов планирования эксперимента начали бурно развиваться алгоритмы **рандомизации**, которые заключаются в многократном случайном перемешивании строк или столбцов таблицы наблюдений относительно уровней воздействия изучаемых факторов. При каждой итерации перестановочного теста на основе сгенерированной псевдовыборки рассчитываются имитируемые значения Q_{sim} анализируемого показателя или статистики, которые сравниваются с аналогичной величиной Q_{obs}, найденной по эмпирическим данным. Обычно в ходе перестановок не меняется ни состав исходной таблицы, ни численность групп с разными уровнями воздействия, а только происходит беспорядочный обмен элементами данных между этими группами.

Существуют мнения (Manly, 2007), что рандомизация вообще является частным случаем испытаний **Монте-Карло (**см. первые работы Бюффона в 1777 г.), в которых специфический стохастический процесс осуществляет равновероятные перестановки данных между уровнями воздействия. При обоих подходах псевдовыборки, сгенерированные в ходе имитации, используются для оценки доверительных интервалов

или проверки нулевой гипотезы. Однако различия этих методов в основных предположениях и ограничениях становятся весьма существенными в наиболее типичных для методов Монте-Карло исследованиях, когда данные наблюдений вообще не используются, чтобы смоделировать вероятностный процесс.

1.2. Параметрическая статистика и ресамплинг: принципиальные различия

Первоначально методы ресамплинга возникли просто как средство преодоления смещения параметра, обусловленного выборкой, но затем начали широко применяться для решения любых статистических задач: проверки гипотезы о законах распределения случайных величин, регрессии, дисперсионного анализа, многомерной классификации и т.д. При этом они составили эффективную конкуренцию традиционным параметрическим методам, еще больше сгустив туман неопределенности у несколько ошарашенного конечного пользователя, не знакомого с тонкой материей идеологических разногласий и вынужденного часто вслепую искать ответы на ключевые проблемы обработки данных («Что делать?», «Кто виноват?» и «Кому жить хорошо?»).

Попробуем и мы прокомментировать разницу в концептуальной основе параметрических и непараметрических тестов (хотя различие между ними не вполне ясно, и вряд ли станет полностью ясным после наших рассуждений). Чтобы сопоставить эти два подхода, рассмотрим предварительно весьма прозаичный пример, связанный с подбрасыванием монеты. Мы не знаем, действительно ли кто-то увлекался этим почтенным занятием, но анализ соотношения вероятностей “аверс-реверс” очень популярен в статистике, поскольку аналогичен многим нашим практическим ситуациям.

Итак, предположим, что у нас есть старая римская монета, у которой на передней стороне нанесено больше серебра, чем на обратной. Мы хотим оценить вероятность того, что в результате 9 подбрасываний у нас выпадет 9 “решек”, и, если она высока, то монета является "подлинной". Выполняя параметрический тест, мы формулируем нулевую гипотезу, что выпадение орла и решки являются равновероятными, и задаемся вопросом «Какова вероятность выпадания 9 решек из 9 подбрасываний, если нулевая гипотеза т.е. $p = 0.50$, является истиной?». Ответ на него могут дать некоторые простые вычисления, описанные в любом учебнике по статистике и основанные на биномиальном распределении. Еще раз обратим внимание, что мы априори установили **параметр** (p) и вычислили результат, основанный на этом параметре и некотором постулированном стандартном распределении.

При втором подходе на основе ресамплинга задаемся более лаконичным вопросом «Если эта монета подлинна, как часто мы получаем 9 решек в результате 9 бросков?». Этот вопрос не использует слово "параметр" и не нуждается априорных предположениях, что вероятность решки на любом броске в случае нулевой гипотезы равна $p = 0.50$. Также нет необходимости проводить какие-либо аналогии с биномиальным распределением. Мы просто берем 100 подлинных монет и будем их одновременно общей большой кучей 9 раз подбрасывать вверх (как это сделать – чисто техническая проблема, не связанная со статистикой). И мы можем легко подсчитать, какое количество монет из 100 легло лицевой стороной вверх все 9 раз, и принять эту вероятность как норму при оценки “подлинности или фальшивости” валюты.

Итак, параметрические тесты основываются на целом ряде априорных предположений и, если они верны, обладают несомненной надежностью и прекрасной теоретической проработанностью. Мы, однако, не станем здесь обсуждать проблему, насколько возможные отклонения от этих предположений (независимость измерений и их ошибок, однородность дисперсий, нормальность распределения и проч.) могут повлиять на обоснованность конечных выводов – на этот счет имеются различные, диаметрально противоположные мнения. Достаточно сказать, что эта проблема существует. В отличие от них, процедуры ресамплинга не требуют никакой априорной информации о виде закона распределения изучаемой случайной величины и в этом смысле могут рассматриваться

как непараметрические. Они выполняют обработку различных фрагментов исходного массива эмпирических данных, как бы поворачивая их разными гранями и сопоставляя полученные таким образом результаты. Вопрос о полной корректности такого приема остается открытым, но если признать его законным, то асимптотические достоинства ресамплинга удается доказать вполне строго. Значения параметров, построенных по размноженным подвыборкам, строго говоря, не являются независимыми, однако при увеличении *n* влияние зависимости может ослабевать и с ресамплированными значениями статистик можно обращаться как с независимыми случайными величинами.

Подробно все нюансы сравнения параметрических методов и ресамплилинга обсуждаются Д. Хауэлом (D. C. Howell) в специальной статье http://www.uvm.edu/~dhowell/StatPages/Resampling/philosophy.html.

2. Проверка статистических гипотез с использованием рандомизации

2.1. Общее описание алгоритма

Рассмотрим предварительно схему реализации рандомизационного теста, который концептуально ближе традиционным параметрическим методам, чем бутстреп-процедуры. Основная цель рандомизации состоит в том, чтобы проверить некоторую нулевую гипотезу, хотя трактовка H_0 в некотором смысле отличается от той, которую мы привыкли понимать.

Алгоритм выполнения рандомизационного теста выглядит весьма прямолинейным. В качестве примера его описания используем схему сравнения двух независимых выборок:

1. Выбираем произвольную метрику ***T***, позволяющую оценить статистическую значимость возможного фактора различий двух групп данных.
   - Для определенности будем считать, что в качестве таковой используется традиционная *t*-статистика Стьюдента, хотя возможны и иные меры, такие как разность между средними или среднее для первой выборки. Подчеркнем, что здесь мы не имеем в виду проверку гипотезы о различии между внутригрупповыми средними, а значение *t* используем просто как один из подходящих индексов, измеряющих «неодинаковость» выборок.
2. Вычисляем значение тестируемой статистической величины для исходных (эмпирических) данных, которую обозначим как t_{obs}.
3. Повторяем *N* раз следующие действия, где *N* - число, больше чем 1000:
   - объединяем данные из обеих выборок и перемешиваем их случайным образом;
   - первые n_1 наблюдений назначаем в первую группу, а остальные n_2 наблюдения отправляем во вторую;
   - вычисляем тестовую статистику t_{sim} для рандомизированных данных.
   - если $t_{sim} > t_{obs}$ увеличиваем на 1 счетчик *S* (т.е. используем односторонний тест).
4. Разделив значение *S* на *N*, получим соотношение частот, с которой метрика t_{sim} на рандомизированных данных превысила значение t_{obs} на данных, которые мы получили в эксперименте. Иными словами, вычислим оценку вероятности *p* того, что случайная величина ***T*** примет значение, большее, чем t_{obs}. По традиции, если $p > 0{,}05$, то принимается нулевая гипотеза H_0 об отсутствии значимых отличий исходных выборок от их нуль-модели по индексу ***T***, а если *p* меньше задаваемого уровня значимости, то H_0 отклоняется в пользу альтернативы.

Пусть, например, размер бицепса в группе из 4 культуристов, принимавших йогурт «Активия» от Данон, составил 22, 25, 25 и 26, тогда как группа контроля из 3 участников имел показатели 17, 21 и 23. Если йогурт не влияет на размер бицепса и нулевая гипотеза об отсутствии различий между группами верна, то любые 3 из имеющихся 7 наблюдений с одинаковой вероятностью могли бы быть приписаны к контрольной совокупности, а остальные – к группе с воздействием (т.е. наблюдалось бы явление «exchangeable» или «обмениваемости» данных). Вычислим 35 пар групповых средних для всех возможных вариантов разбиения 7 измерений на две группы. Проанализировав результаты, мы легко найдем, что есть только одна псевдо-комбинация данных, при которой среднее значение для контрольной группы было бы еще меньше (а для группы с воздействием, соответственно, еще больше), чем это получено в эксперименте. Таким образом, различие между группами, столь же большое как это зафиксировано эмпирически, произошло бы только в 2 случаях из 35, т.е. вероятность справедливости сформулированной нулевой гипотезы составляет 0.0571.

Часто для методов рандомизации употребляют термин *перестановочный* тест (permutation), имея в виду перестановку данных между отдельными группами. Этот термин не вполне удачен, поскольку в действительности мы осуществляем не перестановки, а берем различные *комбинации* данных, уникальных относительно выбранной тестовой статистики. В частности, перестановка {21, 17, 23} в контрольной

группе приведенного выше примера не является шагом рандомизации, поскольку приводит к тем же значениям групповых средних, медиан, вариаций и т.д. Фраза "рандомизационный тест", как отмечает Д. Хауэл, является хорошим компромиссом, чтобы избегнуть двусмысленности термина "перестановка".

В приведенном выше примере мы берем все возможные комбинации данных, вычисляя для каждой из них тестовую статистику. Разумеется, это часто оказывается невозможным. Например, если у нас есть три группы с 20 наблюдениями в каждой, то мы имеем 60! / (20! · 20! · 20!) или $5.78 \cdot 10^{26}$ различных комбинаций группировки наблюдений, и даже самый быстрый суперкомпьютер не будет в состоянии их перебрать. Решение состоит в том, что мы берем случайную выборку из всех возможных комбинаций, которая не будет приводить к *точному* ответу. Однако результаты уже 5000 итераций могут удовлетворить придирчивого исследователя, поскольку погрешность будет наблюдаться в 3-м десятичном разряде или менее того. В этом случае рандомизацию можно трактовать как разновидность имитационного процесса Монте Карло.

Нами был выше рассмотрен алгоритм рандомизации на примере тестирования различий между двумя группами, однако этот подход может быть распространен на решение многих других задач статистической обработки данных. Необходимо просто выбрать подходящую тестовую статистику и определить рациональный механизм генерации случайных комбинаций, т.е. ответить на два самых трудных вопроса анализа: «Что является индикатором различий?» и «Как должны перемешиваться данные?».

Уровень значимости рандомизационного теста (т.е. *p*-значение, полученное на шаге 4) целесообразно интерпретировать в контексте нуль-моделей данных, введенных для дальнейшего теоретического обобщения процедур проверки статистических гипотез. **Нуль-модель** - это образ (имитация структуры) наблюдаемых данных, сформированный из предположения, что H_0 верна, и позволяющий восстановить плотность распределения оценок вероятности анализируемого критерия. При рандомизации потенциальное влияние фактора, способствующего отклонению нулевой гипотезы, снимается путем многократно случайного перемешивания исходных измерений. Нетрудно заметить, что популярные критерии (*t*, *F* и проч.), используемые в параметрических тестах, также имеют соответствующее распределение в условиях справедливости гипотезы H_0 (т.е. соответствуют нуль-модельному распределению, но уже без каких-либо манипуляций с исходными эмпирическими данными).

Выше отмечалось, что использование априорных предположений является фундаментальным признаком, отличающим параметрические методы от ресамплинга. Однако не менее важна их концептуальная разница и в самом механизме проверки нулевой гипотезы. Пусть по нашей традиции имеется две группы наблюдений, выполненных при разных уровнях воздействия изучаемого фактора. Параметрический статистик делает предположение, что обе эти совокупности распределены по нормальному закону и у обеих одинаковая дисперсия. Принимая на веру нормальность и гомоскедастичность (или проверив эти утверждения с использованием соответствующих статистических критериев), единственный путь, который остается исследователю, чтобы оценить отличия в выборках, сводится к сравнению средних. Формулируется гипотеза H_0: $\mu_1 = \mu_2$ и с помощью *t*–критерия проверяется достаточно локальное предположение о равенстве центров распределения обеих групп.

При использовании классических непараметрических тестов (например, с использованием критерия Манна-Уитни-Вилкоксона) анализ становится менее определенным и оперирует уже не со средними, а с такими свободными и не вполне точными понятиями, как «сдвиг местоположения». В случае рандомизации также не ставится задача оценить параметры совокупности и не имеется никакого законного основания для подтверждения или отклонения H_0 о равенстве средних. Здесь ищутся законные альтернативные пути для того, чтобы проверить логическую гипотезу о наличии

эффекта воздействия на основе выявления различий в структуре данных (формально мы даже не уверены, что ее можно называть нулевой гипотезой в классическом смысле).

Как было показано выше, идея рандомизации сводится к многократно повторяемому случайному переприсваиванию (random assignment) меток групп сделанным измерениям: если наблюдения были беспорядочно перетасованы относительно уровней воздействия, и если при этом не было ощутимого изменения некоторого принятого критерия по сравнению с его эмпирически найденным значением, то можно принять гипотезу об отсутствии эффекта. Мы можем использовать любую статистику, чувствительную (опять же, по нашему предположению) к эффекту воздействия, исходя из его природы, и оценивать отличие групп по средним, медианам, дисперсиям или иным характеристикам (Edgington, 1995, p. 141). Отметим, что восстанавливая распределение любой статистики по нуль-модели, мы фактически имеем дело с неким комплексным показателем, поскольку вариабельность тех же средних значений зависит и от степени статистического разброса, и от асимметрии распределения случайной величины, коль скоро мы отказываемся от исходных допущений о нормальности и гомоскедастичности. Отказываясь от точного утверждения $\mu_1 = \mu_2$ в пользу неопределенного утверждения, что «структура данных в группах различна», мы получаем изрядную дозу гибкости, но это также делает статистическую интерпретацию теста более трудной.

2.2. Рандомизация: сравнение двух независимых выборок

Два американских психолога (Ruback and Juieng, 1997) поставили своей задачей оценить действия водителей, освобождающих место на стоянке, в зависимости от того, видят ли они, что кто-то хочет занять их место. Они измеряли секундомером время от посадки водителя в машину до того, как он благополучно покинул место на стоянке (причем с невероятно большой точностью до сотых долей секунды!). Возьмем два фрагмента их выборок для 20 водителей в каждой:

Никто не ожидает в очереди не стоянке

36.30 42.07 39.97 39.33 33.76 33.91 39.65 84.92 40.70 39.65
39.48 35.38 75.07 36.46 38.73 33.88 34.39 60.52 53.63 50.62

Имеются ожидающие освобождающегося места

49.48 43.30 85.97 46.92 49.18 79.30 47.35 46.52 59.68 42.89
49.29 68.69 41.61 46.81 43.75 46.55 42.33 71.48 78.95 42.06

Формулируется нулевая гипотеза, что вид автомобиля, ожидающего места на стоянке, не оказывает абсолютно никакого эффекта на водителя, который собирается ее покинуть. Если H_0 верна, то любая случайная комбинация этих 40 наблюдений столь же вероятна, как и тот результат, который мы наблюдали в натуре.

Прежде всего, в этом примере возможны 40! / (20! 20!) комбинаций таких разбиений, но мы не думаем, что Вы захотите вынуть карандаш и просчитать больше чем 137 миллиардов значений тестовой статистики. Поэтому будем брать случайную выборку из всех возможных комбинаций.

Другая проблема состоит в том, что мы нуждаемся в некотором способе идентифицировать то, что подразумевается под "результатом тестирования", т.е. метрику, позволяющую измерить различие между группами. Например, весьма популярна среди исследователей разность групповых средних. Однако, вычислив для каждой комбинации среднее значение в первой группе, мы автоматически имеем всю необходимую информацию для расчета среднего во второй группе. Использование в тесте среднего (или просто суммы значений) только для первой группы привело бы к тому же самому заключению, что и для разности групповых средних, и избавило бы нас от лишних вычислений. Точно к таким же результатам мы придем, используя *t*-статистику, т.е. разделив разность средних на стандартную ошибку разности. Все четыре перечисленных статистики являются *эквивалентными* относительно процесса рандомизации, поскольку дают совершенно одинаковый результат вероятности того, что измерения в группе 1 не

отличаются от таковых в группе 2. Будем, однако, использовать в нашем примере *t*-статистику, чтобы проследить аналогию с параметрическими медодами.

Поскольку все необходимое для анализа определено, выполним 5000 итераций перемешивания двух выборок с помощью бесплатной программы, которую можно найти на сайте Д. Хауэла www.uvm.edu/~dhowell/StatPages/Resampling/ResamplingPackage.zip. Результаты расчетов на рис. 1 содержат восстановленный график распределения плотности вероятности всех возможных значений *t*-статистики в этом примере, а точнее - гистограмму относительных частот, полученную в ходе 5000 случайных перераспределений оригинальных наблюдений по двум группам.

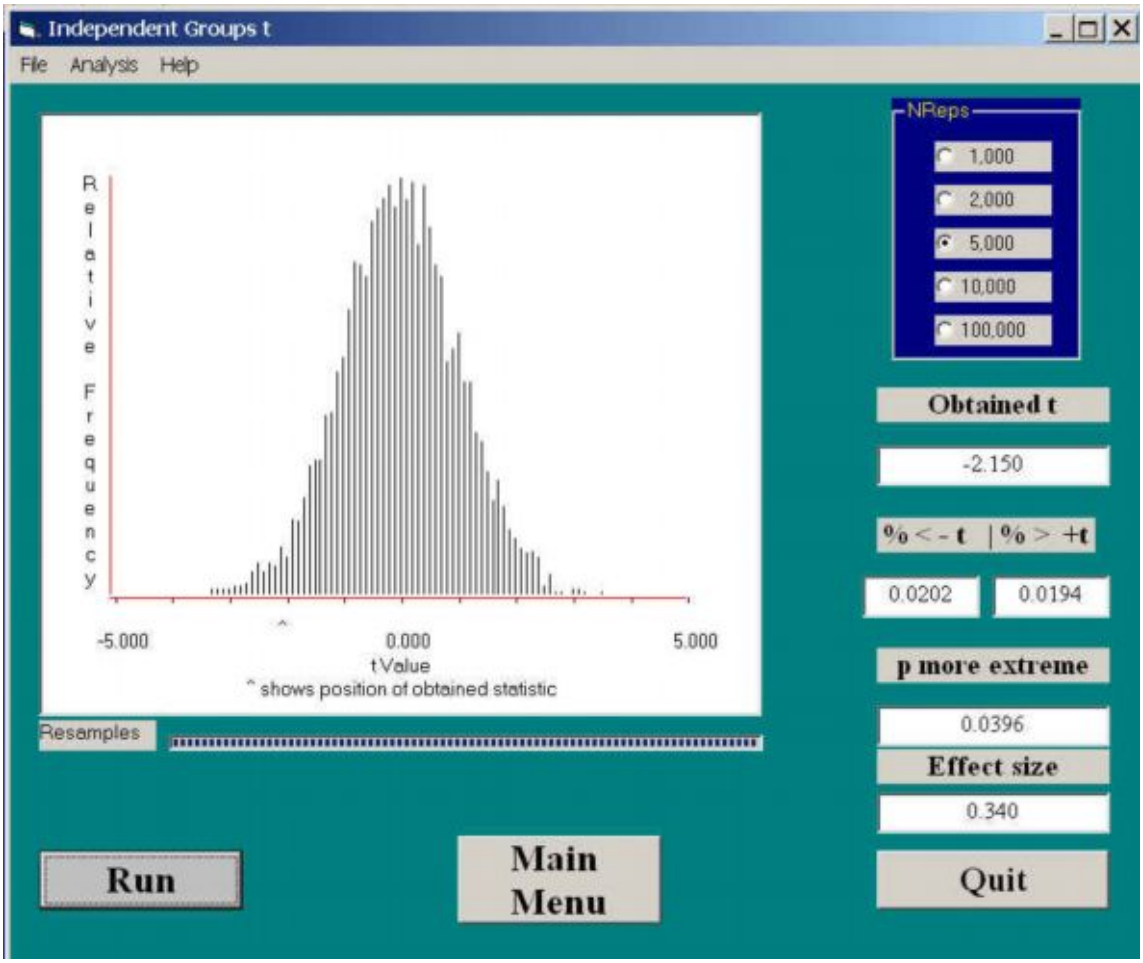


Рис.1

Поскольку данные случайно перетасовывались между двумя группами, мы получили распределение тестовой статистики из предположения, что H_0 верна (т.е. **нуль-модельное распределение**). Теперь зададимся основным вопросом «Насколько вероятно в этих условиях получение того значения статистики (здесь, t =-2.15), что имела место для эмпирических данных?». Гистограмма на рис. 1 показывает, что вероятность появления *t*-значения, столь же или еще более экстремального чем наш, составляет только 0.0396. На традиционном уровне значимости p = 0.05 нулевая гипотеза может быть отклонена и сделано заключение, что водители действительно задерживаются на стоянке более длительное время, когда видят, что кто-ожидает занять их место.

Представляет интерес рассмотреть, как изменятся результаты тестрования, если эмпирические данные содержат выбросы – аномально высокие или низкие значения. Пусть в нашем примере найдется особенно грубый или беспечный водитель, который заставляет очередь ожидать, пока он ведет длительные разговоры на своем сотовом телефоне. Чтобы смоделировать это, заменим последнее наблюдение во второй группе

(42.06 сек.) на экстремальную величину (242.06 сек.). Поскольку рандомизационный тест только переприсваивает данные вместо того, чтобы получать их обновленные порции, это необычно высокое значение будут появляться в каждой перестановке поочередно то в одной, то в другой группе, что приведет к специфичному бимодальному распределению испытательной статистики (рис. 2).

.

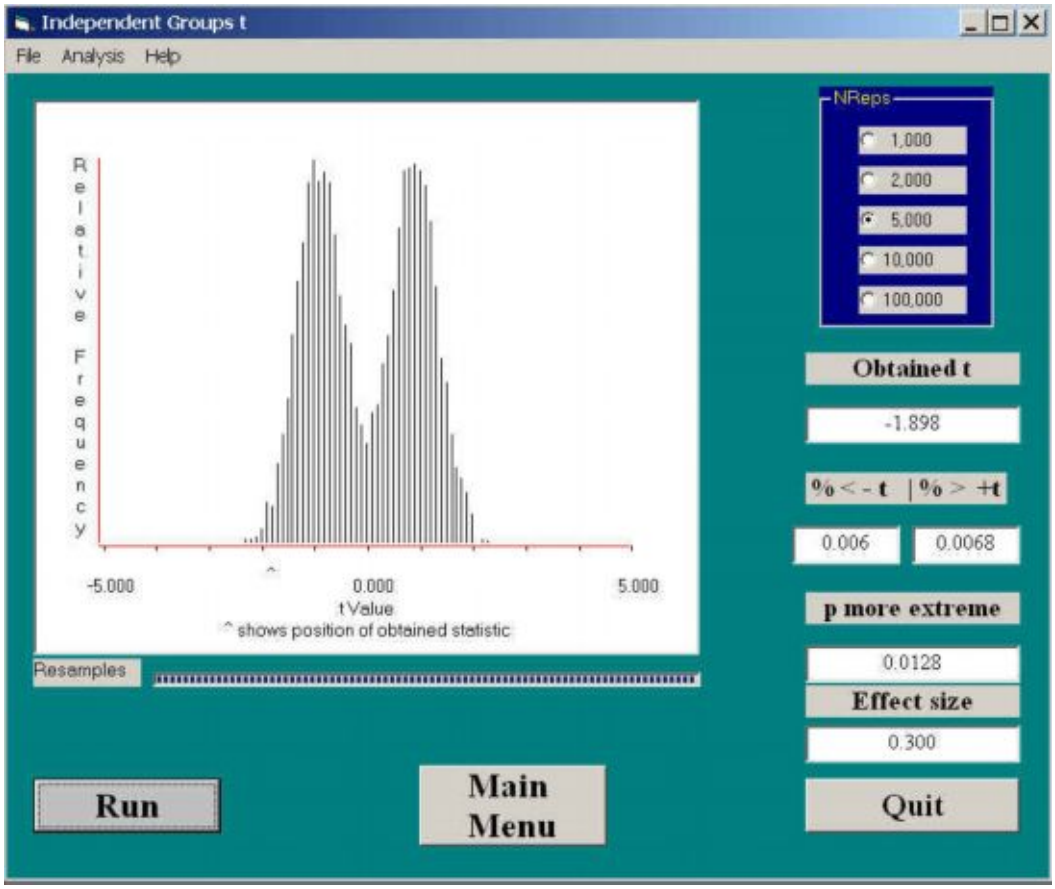


Рис. 2.

Полученное распределение *t*-статистики отчетливо отличается от распределения, смоделированного Госсетом–”Стьюдентом”. Оно является симметрично бимодальным со средними приблизительно +0.9 и -1.1, а некоторое смещение центров распределения в отрицательную область связано с более длительными значениями времени в группе с очередью. Другая необычная особенность состоит в том, что величина t = 1.898, полученная для фактических выборок, значительно ниже традиционного предела для статистически значимого *t*. Однако, если проанализировать результаты, вероятность ее появления (и еще большего значения) составляет 0.0128, что позволяет отклонить нулевую гипотезу. Это – хорошая иллюстрация того, что *t*-распределение, полученное на эмпирических выборках при условии справедливости нулевой гипотезы, может весьма отличаться от стандартного *t* распределения как по форме, так и по величине критических значений.

Следует отметить, что в этих условиях результат использования традиционного теста Стьюдента (как правило, но не всегда) также приводит к более низкой величине *t*-статистики, но и более низкой вероятности отклонения нулевой гипотезы: t = -2.086, p = 0.051 для варианта без выброса и t = 1.946, p = 0.067 в условиях аномально высокого значения. Забегая вперед, сходные результаты получаются при использовании бутстрепа, хотя в этом случае восстанавливаемое *t* распределение будет более плоским и более широким, но снова с уменьшенной вероятностью отклонения нулевой гипотезы.

Представленные результаты иллюстрируют сложность и неоднозначность проблемы использования выбросов. С одной стороны, тест рандомизации для второго примера был выполнен на совершенно законых основаниях, поскольку не нарушались исходные предположения и использовались вполне объясняемые реальные данные. Можно чувствовать себя очень уверенными в заключении , что задержка водителя на стоянке не зависит от существования очереди, с вероятностью p = 0.0128. Однако вызывает опасения тот факт, что наличие только одного специфического наблюдения имеет столь важное значение для итогового вывода. Мы предположили, что выброс имел место в условиях очереди. Но одинаково вероятным можно предположить, что кто-то заходит в автомобиль, видит, что его никто не ждет, и здраво намеревается позвонить по телефону, чтобы не делать этого в движении. Наличие выброса в группе без очереди привело бы к выводу о статистической незначимости отличий между группами. Полагаться на результаты только одного наблюдения – признак плохой экспериментальной методологии. Однако, несмотря на то, что многие теоретические руководства по статистике говорят о необходимости отбракови аномальных значений, на практике этот совет используется не столь часто.

2.3. Рандомизация: сравнение сопряженных пар наблюдений

Параметрический *t*-тест для сопряженных пар наблюдений сводится к анализу выборки, составленной из разностей: если H_0 верна, то средняя разность между парами измерений статистически значимо не отличается от нуля. На такой же простой идее основывается и рандомизационный тест: если исследуемый фактор не имеет никакого влияния на характер данных, то с равной вероятностью величина показателя, измеренного у любого объекта **после** воздействия, будет больше или меньше значения показателя у того же объекта **до** нанесения воздействия. Другими словами, если нулевая гипотеза верна, то перестановка данных в пределах любой *пары* множества равновероятна и приводит к одинаковому итоговому результату.

Если зафиксировать друг с другом все пары измерений и менять местами измерения ДО и ПОСЛЕ воздействия в одной или нескольких случайно выбранных парах, вычисляя каждый раз значение тестовой статистики, то после многократных перестановок можно восстановить ее нуль-модельное вероятностное распределение (другое название - reference distribution). На основе этого распределения оценивается, какую вероятность составляет получение величины этой статистики для эмпирически измеренных данных.

Например, использование когнитивной терапии поведения (Cognitive Behavior Therapy - Everitt, 1994) при лечении анорексии может сопровождаться изменением массы тела пациентов (фрагмент данных представлен ниже):

До СВТ	80.50	84.90	81.50	82.60	79.90	88.70	94.90	76.30	81.00	80.50
После СВТ	82.20	85.60	81.40	81.90	76.40	103.6	98.40	93.40	73.40	82.10
До СВТ	85.00	89.20	81.30	76.50	70.00	80.40	83.30	83.00	87.70	84.20
После СВТ	96.70	95.30	82.40	72.50	90.90	71.30	85.40	81.60	89.10	83.90
До СВТ	86.40	76.50	80.20	87.80	83.30	79.70	84.50	80.80	87.40	
После СВТ	82.70	75.70	82.60	100.4	85.20	83.60	84.60	96.20	86.70	

Использование рандомизационного теста в ходе 5000 комбинаций перестановок данных привело к результату на рис. 3. При этом с использованием *t*-метрики проверялась нулевая гипотеза, что средний прирост массы равен нулю. Отметим также, что аналогичный результат был бы получен с использованием среднего или суммы привеса. Можно увидеть, что вычисленное значение *t*-статистики для эмпирических данных составляет 2.216. Из 5000 различных перестановок только 3.64 % от их числа

привели к величине *t*, столь же экстремальной, как для фактических измерений. Таким образом, мы можем отклонить нулевую гипотезу и заключить, что при CBT-терапии действительно имеет место эффект увеличения веса. Отметим, что в данном случае стандартный параметрический *t*-тест Стьюдента оценил бы статистическую значимость роста массы тела при $p = 0.035$, что чрезвычайно близко к результату, который мы нашли с использованием рандомизации.

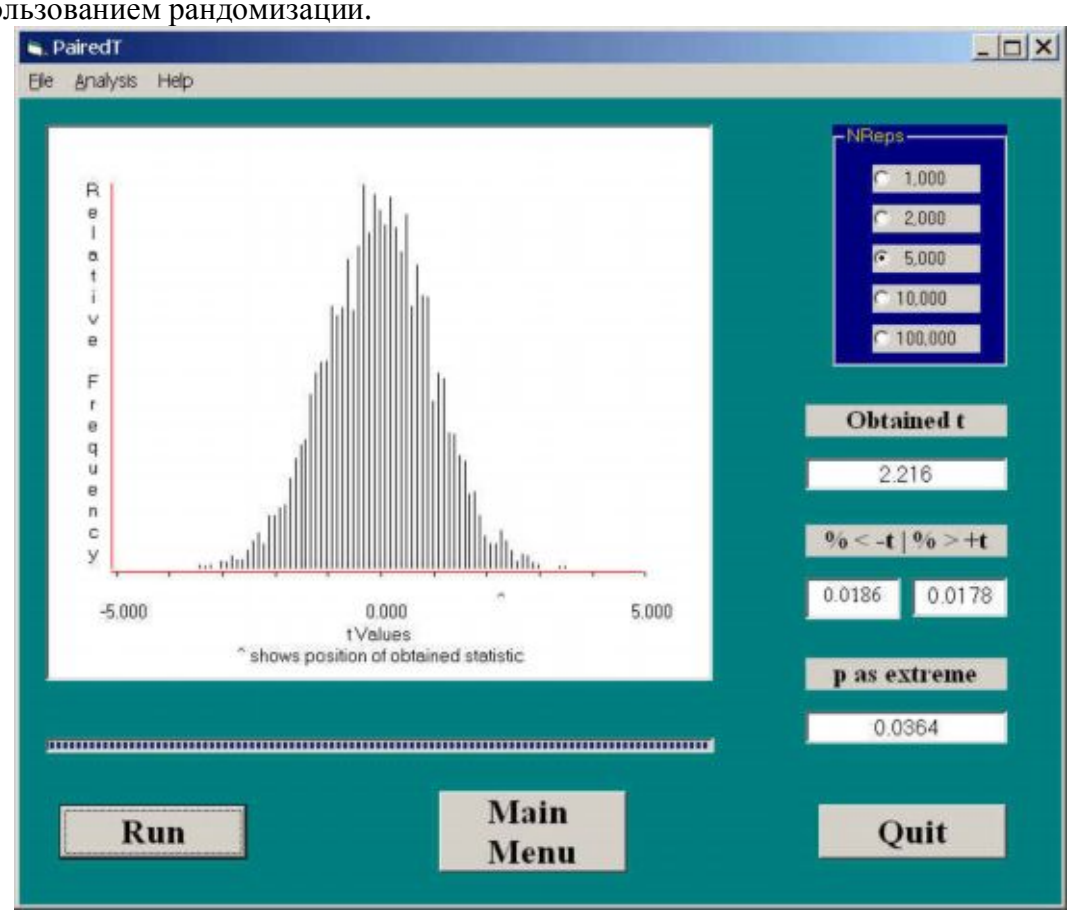


Рис. 3.

Наряду с естественным воодушевлением, что благодаря CBT-терапии можно толстеть, питаясь лишь познанием (cognitive – познавательный), приведенный пример является ярким образцом некорректной экспериментальной методологии. Классическая концепция проведения эксперимента такова: любой управляемый эксперимент должен иметь повторности, причем группы экспериментальных единиц формируют случайным образом и для каждой из них также случайно должны быть назначены различные уровни изучаемого воздействия, включая обязательную контрольную группу (Hurlbert, 1984). Но в нашем примере все пациенты получали одно и то же терапевтическое воздействие, контрольной группы укомплектовано не было, поэтому нет никакого основания утверждать, что увеличение массы тела произошло вследствие CBT-терапии, а не по причине каких-то иных факторов (например, пациентов просто хорошо кормили). Из-за этого мы сомневемся, что читатель мог бы увидеть этот пример в серьезной литературе, в частности, книге Эджингтона (Edgington, 1987) по тестам рандомизации, где постоянно подчеркивается необходимость случайного назначения воздействий экспериментальным единицам.

2.4. Рандомизация: сравнение двух медиан

На первый взгляд сравнение медиан подобно тесту на сравнение средних. Однако здесь мы уже не можем использовать *t*-статистику, поскольку нет корректного способа вычислить стандартную ошибку. Альтернативой является восстановить распределение

медианных разностей и подсчитать число итераций, при которых эта разность была бы столь же большой (или больше), чем в эмпирических выборках.

Однако мы пойдем другим путем. Вычислим доверительные границы разности медиан и будем отклонять нулевую гипотезу, если эмпирическое различие окажется вне этого интервала. Преимущество этого подхода состоит в том, что он дает нам доверительный интервал, который всегда полезно иметь. С другой стороны, формально доверительные границы рассчитаны не для истинной разницы медиан, а в предположении, что верна нулевая гипотеза, и их познавательная ценность ограничена. Наконец, при таком подходе можно отклонить H_0, если разность медиан фактических выборок окажется вне доверительного интервала, но мы не получим точного p-значения вероятности такого различия при справедливости нулевой гипотезы.

Рассмотрим в качестве примера исследование Мирелла, изучавшего скорость прохождения лабиринта двумя группами мышей. Медиана продолжительности прохождения мышами первой группы под аккомпанемент музыки Моцарта составила 119 секунд, в то время мыши второй группы задумчиво пробирались к выходу из лабиринта в сопровождении музыки американской металл-группы Anthrax (русск. «сибирская язва») со среднемедианной длительностью 306 секунд.

Выполним рандомизационный тест из 5000 итераций на выборках 48 наблюдений по 24 измерения в каждой группе. Для каждой случайной перестановки вычислим медианы и их разность между группами и восстановим распределение вероятностей этого показателя при условии, что H_0 верна. На основе полученной гистограммы (см. рис. 4) вычислим границы доверительного интервала способом, который Б.Эфрон называет "методом процентилей", т.е. найдем значения, соответствующие 97.5 % и на 2.5 % найденного распределения.

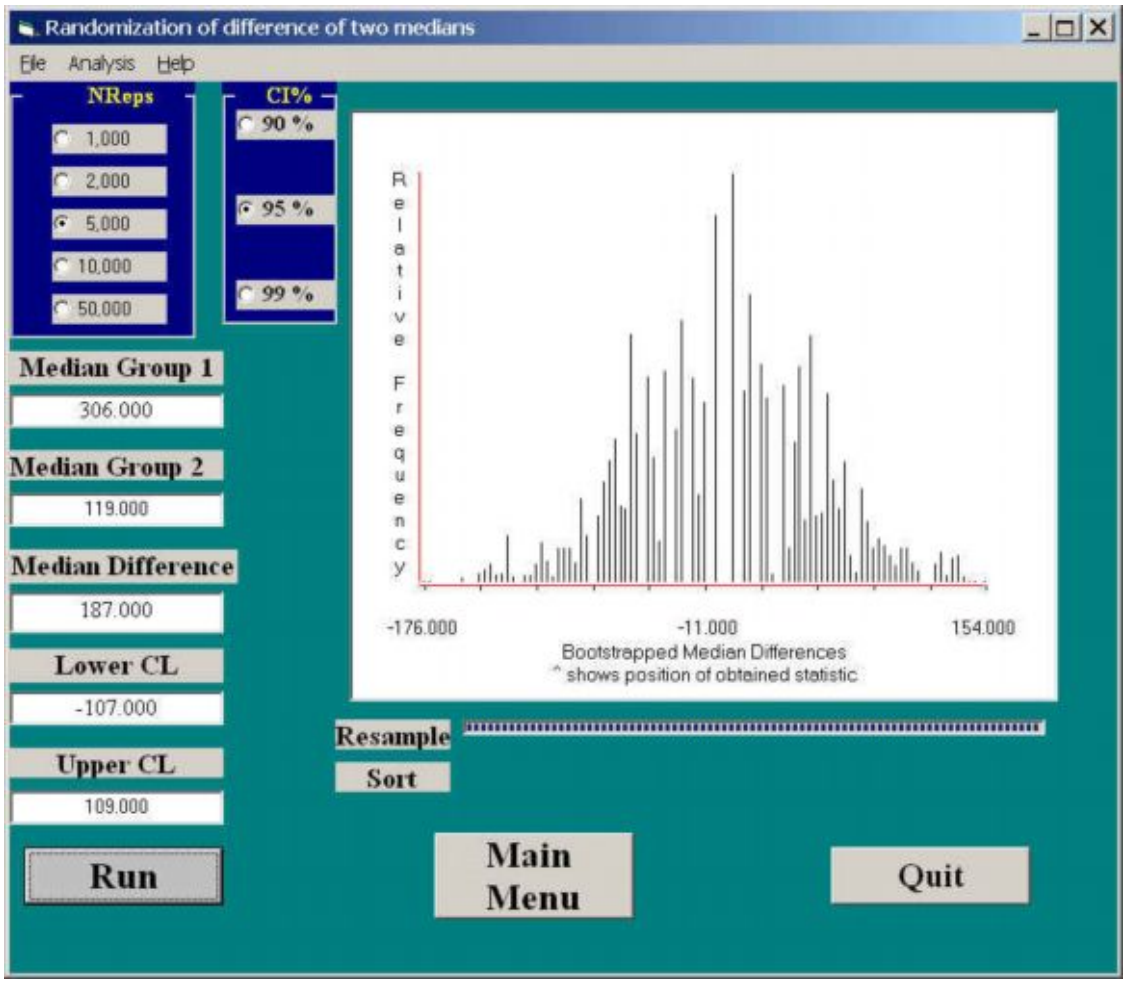


Рис. 4.

Можно увидеть, что 95%-ый доверительный интервал разности медиан находится в пределах от -107 до +109, т.е. он почти симметричен относительно 0. Можно также легко установить, что эмпирическая разность, равная 187, отчетливо выпадает из этого интервала, что дает нам основания отклонить нулевую гипотезу и заключить, что среднее время отклика в условиях треш-метала «Сибирской язвы» значительно больше, чем под музыку Моцарта.

В дальнейшем мы вернемся к проблеме сравнения медиан, но на иной концептуальной основе, когда будем рассматривать использование для этой же цели бутстреп-метода. При рандомизации мы находим доверительные пределы разности медиан (от -107 до +109) при условии, что *нулевая гипотеза верна*. Не останавливаясь на технических способах реализации, отметим, что в случае с бутстрепом мы имеем возможность установить доверительные пределы *истинной медианной разности* для конкретных выборок, которые при 95%-ом уровне доверия будут в границах от 146 до 226 (см. рис. 5). Поскольку этот интервал не включает 0, мы также можем отклонить H_0 , но следует отметить, сколь различны эти два подхода.

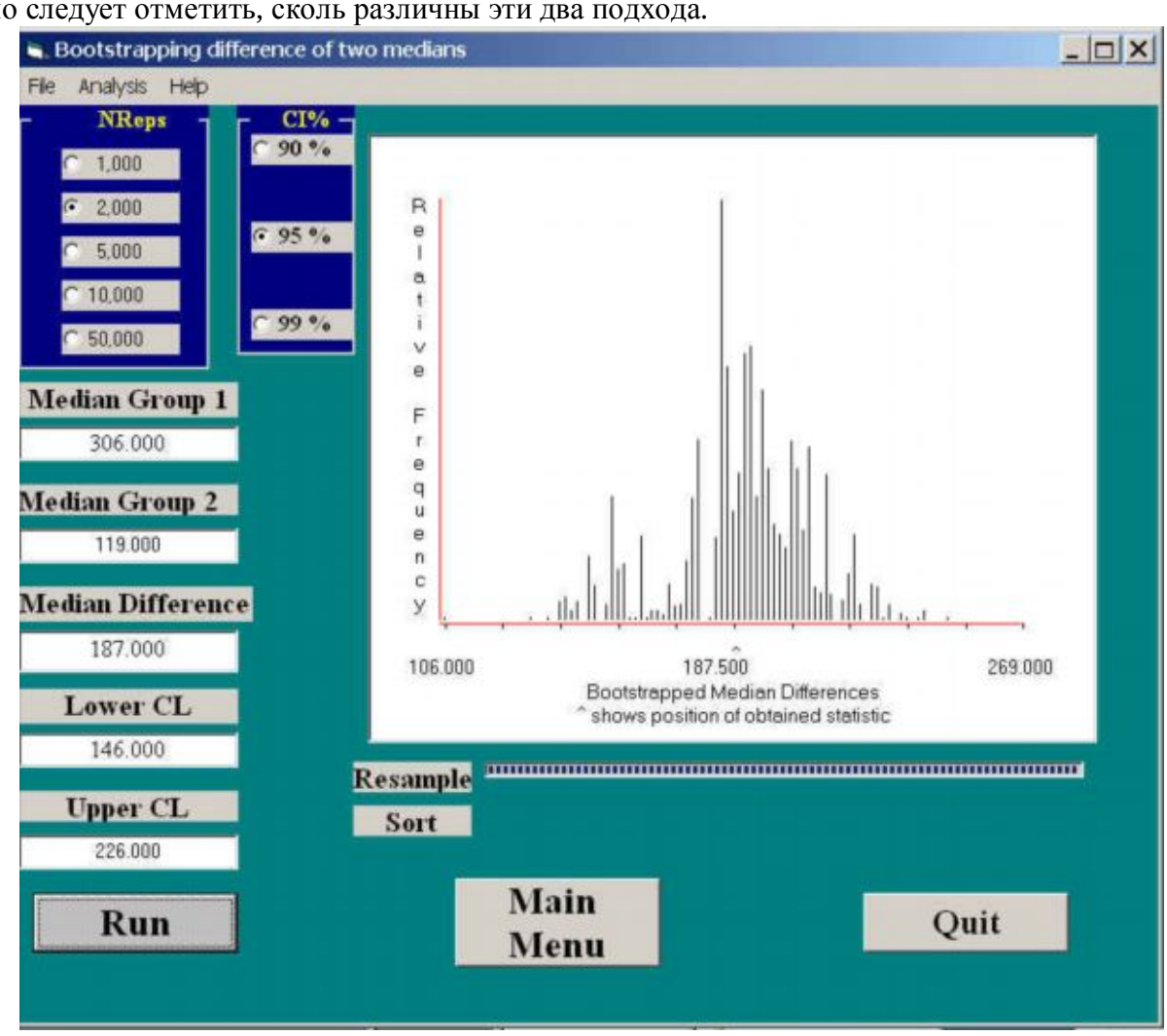


Рис. 5.

2.5. Рандомизация: оценка зависимости двух переменных

Применение рандомизационной процедуры к оценке линейной связи двух переменных обычно сводится к тому, что проверяется нулевая гипотеза о равенстве нулю коэффициента корреляции $\rho = 0$. Для этого фиксируются значения одной переменной (например, *X*) и случайным образом перемешиваются значения другой переменной *(Y)* относительно постоянного вектора **X**. Поскольку каждый x_i *беспорядочно* связан со значениями *Y*, ожидаемый коэффициент корреляции *r* рандомизированных данных равен

0. Повторяя этот процесс большое количество раз, можно восстановить распределение случайной величины *r* для представленного комплекта данных, исходя из предположения, что истинное значение силы связи двух переменных ρ = 0. На основе этого рассчитывается доверительный интервал для ρ или подсчитывается количество случаев, когда иммитируемый коэффициент корреляции для рандомизированных комбинаций превысил значение *r* для эмпирически полученных выборок.

Обратимся для примера к исследованию американских психологов (Katz et al., 1990), которые вознамерились показать, что результаты стандартного тестирования студентов колледжа (аналог нашего ЕГЭ) в первую очередь зависят от приобретенного навыка сразу изолировать и отклонять маловероятные ответы. Они попросили испытуемых не читать содержания вопросов и оценивали коэффициент корреляции между результатами тестирования (*Y*) и количеством экзаменационных тестов, которым студенты подвергались ранее (*X*):

X	58	48	48	41	34	43	38	53	41	60	55	44	43	49
Y	590	590	580	490	550	580	550	700	560	690	800	600	650	580
X	47	33	47	40	46	53	40	45	39	47	50	53	46	53
Y	660	590	600	540	610	580	620	600	560	560	570	630	510	620

На рис. 6 показано, что коээфициент корреляции на оригинальных данных составил r_{obs} = 0.532. Нуль-модельное распределение *r* является приблизительно симметричным относительно 0.0 и 17 значений r_{sim} из 5000 комбинаций рандомизации превысила +0.532. Это дает нам оценку вероятности H_0 для эмпирических данных p = 0.003, позволяющую сразу отклонить нулевую гипотезу. Поскольку распределение симметрично для ρ = 0, здесь был применен двухсторонний тест, но итоговый вывод будет аналогичен и для одностороннего теста.

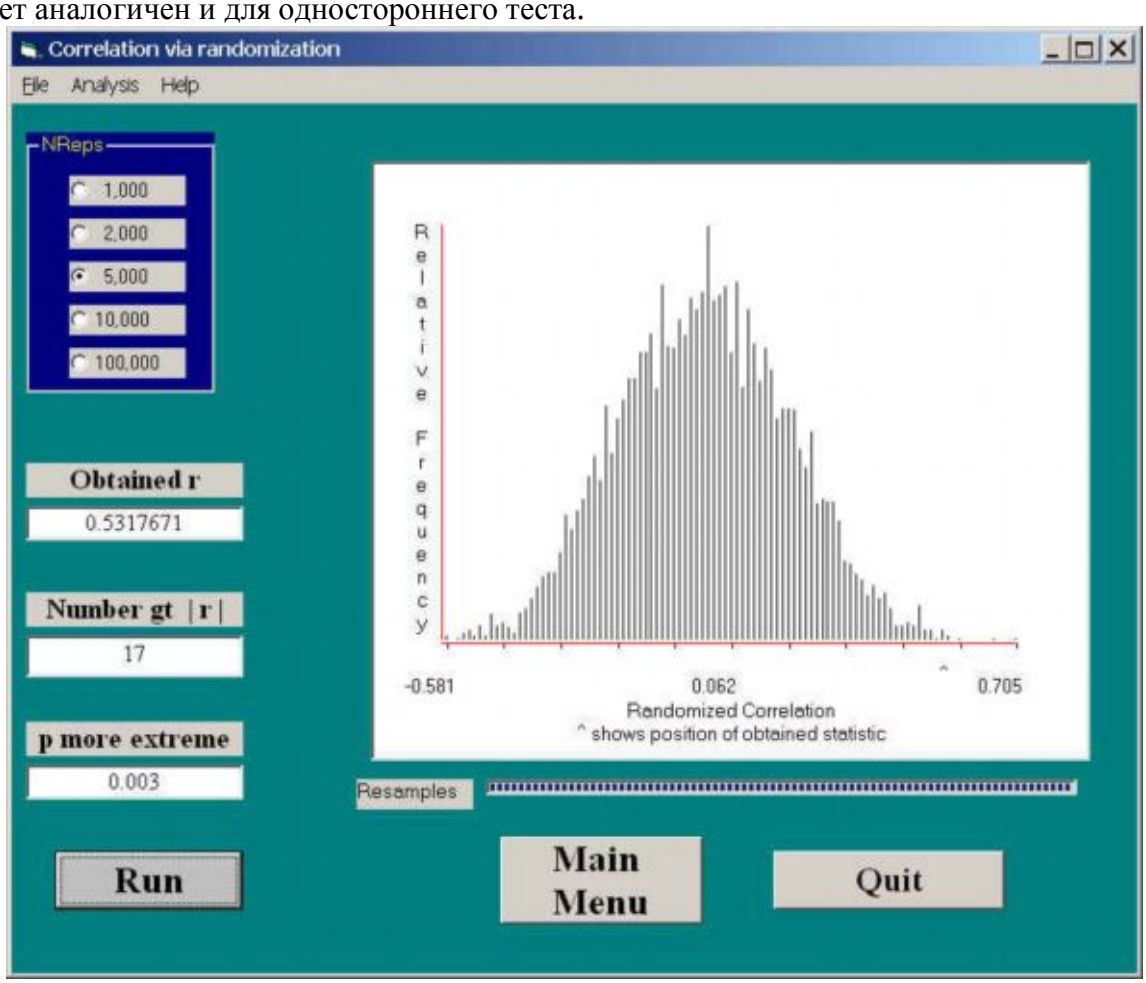


Рис. 6.

И опять мы имеем хороший предлог, чтобы акцентировать различия между рандомизацией и бутстрепом. При рандомизации мы восстанавливаем распределение коэффициента корреляции при условии справедливости H_0 и ожидаем, что центр этого распределения будет близок к 0. В случае бутстрепа все комбинации пар x_i – y_i зафиксированы и мы только случайно укомплектовываем повторную выборку из этих пар с заменой и возвращением. Это означает, что оценка корреляции r_{sim} между X и Y для любой псевдо-выборки будет корреляцией в оригинальных данных и мы только «улучшаем» истинное значение параметра ρ и уточняем его доверительные пределы.

2.6. Рандомизация: тестирование нескольких независимых групп

Обсудим проблему, как распространить тест рандомизации для двух независимых групп на более общий случай однофакторного дисперсионного анализа при нескольких группах. Здесь уже нельзя использовать в качестве тестовой статистики сумму значений для первой группы, межгрупповую разность средних или *t*–значение. Чтобы принять во внимание различия средних для всех групп, в качестве эквивалентных статистик можно использовать сумму квадратов отклонений групповых средних от глобального среднего ($SS_{between}$) или традиционную *F*-метрику. Во всех остальных деталях рандомизационная процедура нам уже вполне знакома:

1. Вычисляем *F*-значение для эмпирических данных (обозначим как F_{obs}).
2. Генерируем *B* псевдовыборок, каждая из которых – результат случайной перестановки исходных данных между группами.
3. На каждой иттерации:
    - случайным образом перемешиваем весь комплект данных
    - назначаем первые n_1 наблюдений в первую группу, следующие n_2 наблюдений во вторую группу, и так далее.
    - вычисляем F_{sim} для этих данных, и, если $F_{sim} > F_{obs}$, увеличиваем счетчик *S*;
4. После завершения рандомизации вычисляем $p = (S+1) / B$, которая представляет собой вероятность получения *F* такой же величины (или более), что была найдена на экспериментальных данных, если верна нулевая гипотеза.
5. Отклоняем или принимаем нулевую гипотезу.

Примером традиционного однофакторного дисперсионного анализа является сравнение методов лечения жертв насилия (Foa et al., 1991). Обрабатывались данные по четырем группам: первая группа получила терапию снятия стресса (SIT - Stress Innoculation Therapy), вторая – продолжительное коллективное обсуждение (PE - Prolonged Exposure), , третья - стандартную Благосклонную Рекомендацию (SC - Supportive Counseling), (SC) и последняя группа была контрольной (WL):

Группа	Количество пациентов	Среднее	Стандартное отклонение
SIT	14	11.07	3.95
PE	10	15.40	11.12
SC	11	18.09	7.13
WL	10	19.50	7.11

На рис. 7 можно увидеть распределение вероятности значений *F*, где отмечено местоположение значения $F = 3.046$ для эмпирических данных, которое может быть получено также любой стандартной программой дисперсионного анализа. Уровень значимости нулевой гипотезы $p = 0.038$ мы нашли путем подсчета числа итераций ресамплинга с *F*, большим, чем 3.046. Это хорошо согласовывается с вероятностью, полученной из стандартного *F* распределения с 3 и 41 степенями свободы, что далеко не всегда имеет место.

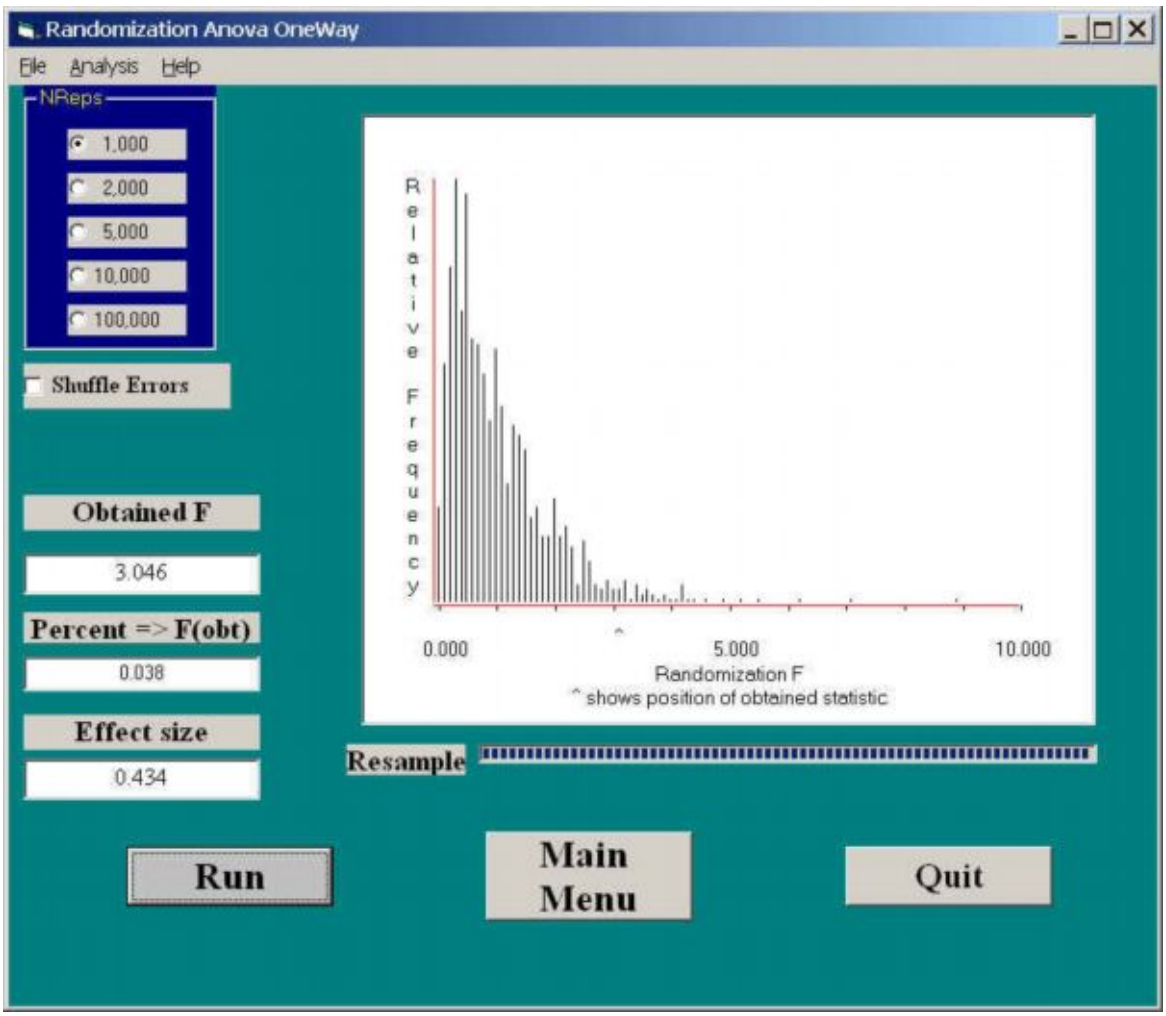


Рис. 7.

Некоторую специфику составляет схема рандомизации применительно к дисперсионному анализу с повторными измерениями (тестирование эффекта для отдельных экспериментальных единиц**)**. В этом случае имеются определенные ограничения на перебор: обмен осуществляется только между отдельными последовательностями наблюдений для каждого отдельного объекта, тогда как перемещение данных между объектами не происходит. Если отсутствует какой-либо эффект воздействия или временной тренд, то для любого испытуемого набор данных является генерацией некоторого стохастического процесса. Как и при однофакторном дисперсионном анализе на независимых выборках, при повторных измерениях можно использовать традиционную *F*-статистику, однако бывают случаи, когда ее величина может вводить в заблуждение. В частности, если предположения о сложной симметрии и сферичности не выполняются или наблюдается коррелированность дисперсий и средних, дисперсионный анализ может давать ошибочные результаты.

Еще один пример основан на оценке психологического состояния калифорнийских студентов (Nolen-Hoeksema, Morrow, 1991). Авторам повезло и они делали свое обследование за две недели до землетрясения 1987 г. Потом они возвратились и в течение долгого времени исследовали характер изменения депрессии у одних и тех же студентов. Ниже приведены результаты обработки их данных для пяти сессий обследования, сделанных через каждые три недели, выполненных программой SPSS:

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	Sphericity Assumed	94.288	4	23.572	2.726	.034
	Greenhouse-Geisser	94.288	2.467	38.215	2.726	.063
	Huynh-Feldt	94.288	2.772	34.018	2.726	.055
	Lower-bound	94.288	1.000	94.288	2.726	.112
Error(TIME)	Sphericity Assumed	830.112	96	8.647		
	Greenhouse-Geisser	830.112	59.215	14.019		
	Huynh-Feldt	830.112	66.521	12.479		
	Lower-bound	830.112	24.000	34.588		

Заметим, что F_{obs} = 2.726, если использовать нескорректированное число степеней свободы *df*, предполагающее сферичность данных, является статистически значимым с p = 0.034. Однако, если выполнить коррекцию Гринхауса-Гайссера или Гина-Фельда, то это может отразиться на итоговых выводах и не позволит отклонить H_0 при уровне значимости α = 0.05.

При использовании рандомизационного теста мы получаем (рис. 8, p = 0.033) вероятность того, что нуль-модельное F-значение превысит наблюдаемое, близкую к найденной, исходя из предположений о сферичности, однако мы не располагаем аргументами, чтобы прокомментировать это обстоятельство.

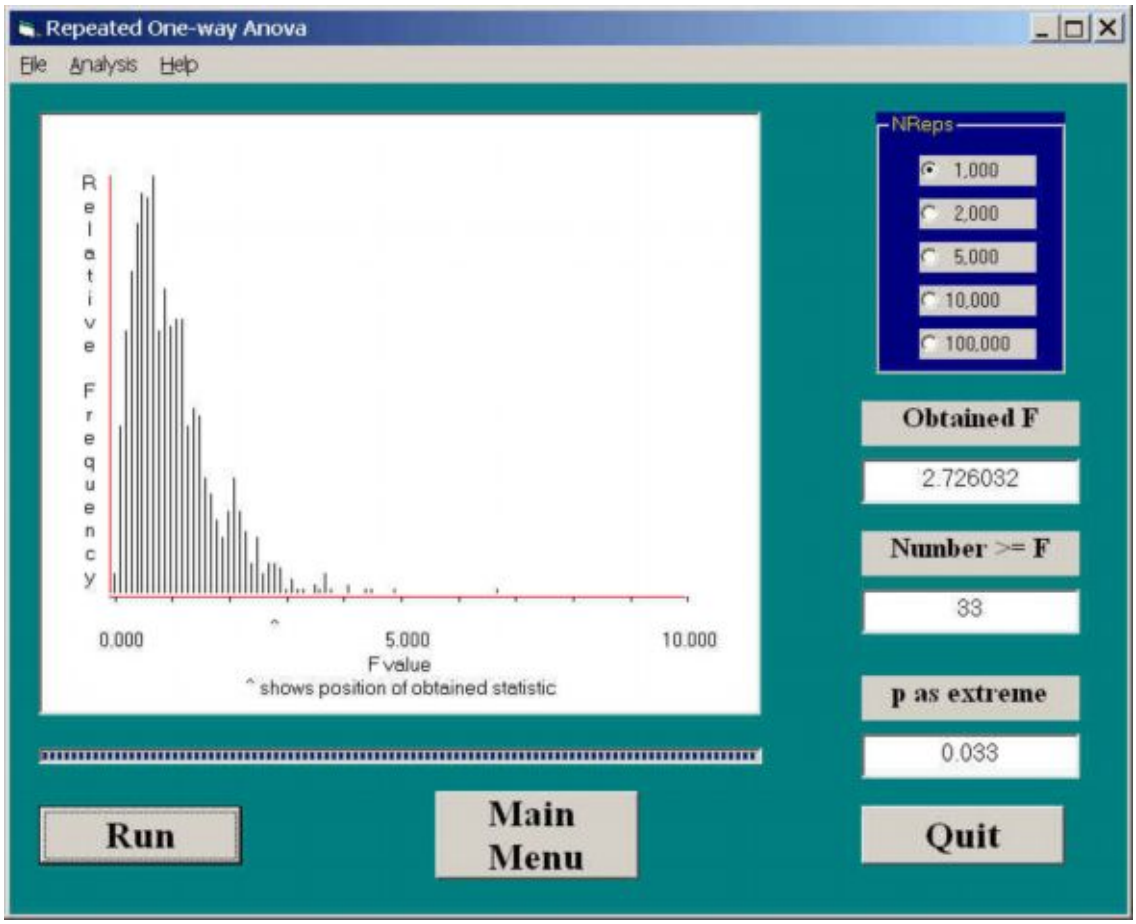


Рис. 8.

2.7. Рандомизация: дисперсионный тест Шлютера

Основные положения рандомизационного теста выглядят вполне убедительными, если речь идет об анализе одномерных выборок. Однако для многомерного случая эта методология сталкивается с проблемой неопределенности выбора ограничений на рандомизацию, т.е. исследователю необходимо предварительно оценить степень «вольности», с которой будут перемешиваться данные, и задать механизм перестановок, адекватный поставленной задаче.

Пусть мы имеем матрицу наблюдений **X**, столбцами являются значения *s* различных показателей X_i (i = 1, 2, …, s), измеренные для каждого из m объектов. В частности, этими показателями могут быть различные симптомы или иные характерные признаки для некоторой выборки из m испытуемых объектов. Ограничимся также тем, что признаки измерены в шкале альтернативных высказываний: 1 – наличие симптома, 0 – его отсутствие. Зададимся целью проверить нулевую гипотезу о случайном характере формирования матрицы **X**, т.е. между симптомами отсутствует какая-либо взаимосвязь (т.е. вероятность совместного появления или взаимного исключения соответствует стохастическому процессу). Альтернативная гипотеза сводится к предположению, что метаструктура таблицы наблюдений определенным образом детерминирована и конфигурация ее отдельных фрагментов не может быть интерпретирована как случайность. Наиболее характерные типы структурной организации матрицы представлены на рис. 8.

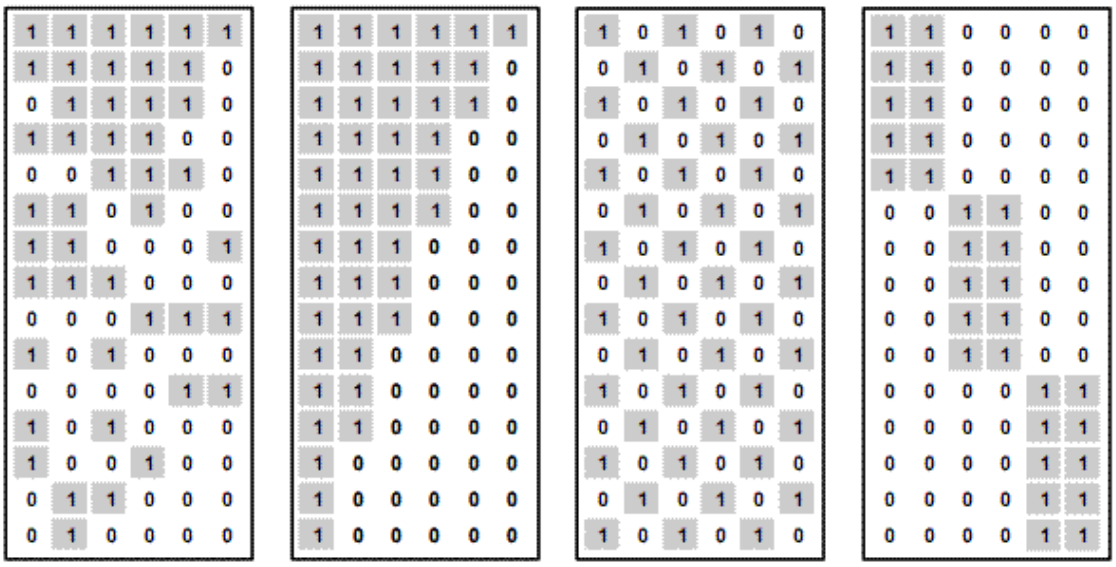


Рис. 8.

Рассмотрим одну из возможных метрик, которую мы будем использовать для проверки нулевой гипотезы. Если использовать известное положение теории вероятности о том, что общая суммарная дисперсия нескольких случайных величин равна сумме дисперсий каждой из них плюс удвоенная сумма ковариаций, то на основе данных в столбцах таблицы можно записать $D(X) = \sum_{i=1}^{s} D(X_i) + 2\sum_{i<l}^{s} D(X_i, X_l)$. Компоненты разложения дисперсии в общем случае неизвестны, но могут быть рассчитаны по результатам наблюдений.

Пусть выборочная оценка $D(X_i)$ равна $\sigma_i^2 = p_i(1-p_i)$, где p_i – средняя встречаемость i-го симптома у обследованных пациентов, $p_i = n_i/n$, а оценка $D(X)$ общей

дисперсии появления всех признаков – $\sigma_X^2 = \left(\sum_{i=1}^{S} (p_i - \bar{p}) \right) / s$, где $\bar{p}$ – наблюдаемое средняя встречаемость симптома у одного пациента.

Нулевая гипотеза (H_0) об отсутствии сопряженности между признаками справедлива, если сумма ковариаций $D(X_i, X_j)$ равна нулю. Это будет верно, когда симптомы независимо распределены среди экземпляров обследованной группы, но также может иметь место, если положительные и отрицательные ковариации уравновешивают друг друга. Если проверять H_0 против альтернативной гипотезы, что есть чисто положительная или чисто отрицательная зависимость между показателями, то, при справедливости H_0, имеет место отношение $E(\sigma_X^2 \mid p_1, p_2, ..., p_s) = \sum_{i=1}^{S} \sigma_i^2$. Следовательно, выражение $V = \sigma_X^2 / \sum_{i=1}^{S} \sigma_i^2$ может служить обобщенным «индексом взаимозависимости» признаков в выборке (Schluter, 1984).

Если H_0 верна, то ожидаемое значение $V = 1$. Значение V, большее или меньшее 1, указывает, что между показателями в группе есть статистическое положительное или отрицательное взаимодействие. Термин «статистическое» употреблен нами, чтобы подчеркнуть частный характер таких связей: например, на фоне «нейтральных» взаимоотношений между большинством признаков может оказаться одна или несколько пар «антагонистов», редко встречающихся совместно.

Если предположить, что для данных обследования справедлива центральная предельная теорема, то при достаточно больших значениях m и s последовательности p_i можно интерпретировать как независимые случайных величины, приблизительно распределенные по нормальному закону. Тогда индекс взаимозависимости показателей V при справедливости H_0 будет иметь χ^2-распределение с m степенями свободы, а критические значения для того, чтобы отклонить нулевую гипотезу определяются пределами табличных значений, например: $\chi^2_{m,\ 0,05} \le mV \le \chi^2_{m,\ 0,95}$.

Анализ статистических закономерностей структурой организации можно также выполнять, не прибегая к теоретико-вероятностным представлениям. Например, индекс заполнения шахматной доски (Checkerboard score – Stone, Roberts, 1990), оценивает среднее число подматриц $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ и $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ размерностью 2×2 для произвольной пары показателей i и j: $CS = \sum_{ij} (n_i - n_{ij})(n_j - n_{ij})$, где n_i и n_j – частоты их встречаемости, n_{ij} – одновременная встречаемость этой пары признаков. Индекс изменяется в диапазоне от 0 до $\frac{2}{s(s-1)} \sum_{ij} n_i n_j$);

Предположения о нормальности распределения вероятностей появления симптомов привели к тому, что *дисперсионный тест Шлютера* в его оригинальной версии оказался чрезвычайно чувствителен к признакам с высокой частотой и склонен к гипердиагностике взаимосвязи. Для решения этих проблем были разработаны имитационные процедуры рандомизации путем многократного случайного перемешивания матриц наблюдения (Gotelli, 2000).

Существует достаточно большой набор взглядов на способы конструирования алгоритмов рандомизации матриц, поэтому существенной проблемой статистического анализа статистик V и CS является выбор вычислительной процедуры генерации нуль-модели с теми или иными ограничениями на перебор. «Равновероятные» **EE** (Equiprobable) алгоритмы осуществляют перестановку значений в пределах исходной матрицы без каких-либо ограничений и сохраняют минимальное количество информации,

содержащейся в исходных данных. В противоположность этому, наименее «либеральная» дважды фиксированная модель **FF** (Fixed-Fixed) требует, чтобы общая встречаемость «единиц» в строках и столбцах нуль-матрицы соответствовала бы наблюдаемым значениям в эмпирической матрице. Комбинированная модель **EF** (Equiprobable-Fixed) сохраняет неизменным число симптомов каждого участника, но позволяет частотам появления признаков (т. е. общему количеству единиц в строках) изменяться беспорядочно и равновероятно. Модель **FE** (Fixed-Equiprobable) делает то же самое в отношении столбцов (т. е. объектов). Разработана также коллекция пропорциональных (Proportional) моделей **P**, которые в процессе перебора отдают предпочтение тем или иным признакам согласно вероятности их встречаемости, а также пропорционально их абсолютным значениям, либо иным популяционным параметрам.

В теоретическом плане становится все более ясным, что статистические тесты, которые используют полностью равновероятные нуль-модели **EE**, не вполне соответствуют практическому смыслу стохастичности (Gotelli, McGill, 2006). Идеальная нуль-модель должна обладать, как хорошей мощностью идентифицировать истинные закономерности, если взаимосвязи между признаками имеют неслучайный характер, так и не обнаруживать статистической значимости эффекта в матрицах, где распределение показателей сгенерировано стохастическим процессом.

В качестве примера рассмотрим оценку взаимодействий между видами донных организмов: наличие конкуренции за пищевые ресурсы, отношения «хищник – жертва», кооперация или мутуализм. Предположим, что в разных точках водоема сделано 53 пробы со дна реки Сок и при этом было обнаружено 205 разных видов бентоса (моллюски, личинки комаров, стрекоз и других насекомых, различные черви и проч.) Если между видами есть взаимодействие, то эмпирическая матрица по выбранному критерию E статистически значимо отличается от нуль-модельных матриц, в которых комбинации видов хаотически перемешаны. Проверка статистических гипотез ($\alpha = 0.05$) проводилась с использованием Z-критерия $Z = (E_{\text{obs}} - \hat{E}_{\text{sim}}) / SD_{\text{sim}}$ и на основе границ интервалов, соответствующих 95%-ной доверительной вероятности.

Приведенные ниже результаты показывают, что итоговые выводы сильно зависят принятых ограничений на рандомизацию и выбора нуль-модели: для **EF, FE** и **EE** нулевая гипотеза об отсутствии межвидовых взаимодействий отклоняется, в то время как они принимаются статистически незначимыми в некоторых случаях использования моделей **P** и **FF**.

Наименование тестовой статистики E	E_{obs} по эмпирическим данным	Тип нуль-модели	Среднее E_{Sim} из 100 итераций	Стандартное отклонение E_{Sim}	Z-статистика	Нижний уровень 95% ДИ	Верхний уровень 95% ДИ
Заполнение шахматной доски (CS)	7.29	(P) Пропорционально частотам	6.44	0.19	4.45	6	6.78
		(FF) Фиксируются суммы строк и столбцов	7.22	0.04	1.51	7.15	7.32
		(FE) Фиксируются суммы строк	7.67	0.04	-9.51	7.58	7.74
		(EF) Фиксируются суммы столбцов	9.4	0.05	-45.08	9.3	9.49
		(EE) Равновероятная модель	9.66	0.04	-61.08	9.58	9.72
Дисперсионный тест Шлютера (V)	7.96	(P) Пропорционально частотам	8.59	0.79	-0.8	7.18	10.46
		(FF) Фиксируются суммы строк и столбцов	7.96	0	0	7.96	7.96
		(FE) Фиксируются суммы строк	8.05	0.01	-18.45	8.04	8.07
		(EF) Фиксируются суммы столбцов	54.05	5.94	-7.75	44.17	67.81
		(EE) Равновероятная модель	54.57	5.07	-9.19	45.07	63.3

Все представленные выше имитационные процедуры с различными схемами генерации нуль-моделей можно ранжировать в ряд по мере снижения ограничений на рандомизацию: **P > FF > (EF, FE) > EE**. В этом же направлении увеличивается уровень ошибки 1-го рода (т. е. анализ становится менее консервативным) и нулевая гипотеза будет отклоняться, даже если взаимосвязь между признаками выглядит весьма сомнительно. Однако в этом же ряду уменьшается ошибка 2-го рода (принятие ложной нулевой гипотезы), поэтому предпочтение часто отдается моделям типа **EF**. Согласно другим рекомендациям в условиях пассивного формирования выборок более надежные результаты дают модели **FF**.

3. Оценка параметров случайной величины с использованием бутстрепа

3.1. Общее описание алгоритма

Применение рандомизационного теста оправдано, если ставится задача оценить степень упорядоченности структуры данных или взаимосвязи между отдельными ее фрагментами. При подсчете же обычной выборочной статистики (например, среднего) порядок следования элементов выборки не имеет значения и каждая итерация рандомизации будет возвращать одну и ту же величину, поскольку сами по себе данные не изменяются. В этом заключается фундаментальное различие между рандомизацией и бутстрепом: если рандомизация формирует распределение тестовой статистики при справедливости H_0, то бутстреп используется для получения наиболее корретной оценки **параметров распределения** случайной величины.

Пусть дана выборка $x_1, x_2, \ldots, x_n$ и предполагается, что это – набор независимых и одинаково распределенных реализаций случайной величины, извлеченных из генеральной совокупности X. Задача заключается в изучении свойств некоторой статистики f_n (x_1, x_2, …, x_n), которую мы трактуем как выборочную оценку произвольного параметра $\hat{\theta}$ распределения X. Обычно мы имеем некоторый сдвиг $b = E(\theta - \hat{\theta})$ вычисленного значения параметра $\hat{\theta}$ относительно его истинной величины θ, который вызывается многими причинами. Во-первых, выборочные значения имеют погрешность измерений, во-вторых, нет особенных гарантий, что выборка состоит из независимых и случайных значений, и, наконец, при оценке параметра мы обычно задаемся какими-то предположениями о законе распределения X.

Например, в случае нормального распределения X оценкой меры положения случайной величины является арифметическое среднее, а несмещенной оценкой дисперсии σ^2 − квадрат стандартного отклонения s^2. Однако так ли это на самом деле и справедливо ли наше предположение? Один из способов проверить вычисления заключается в том, чтобы извлекать из нашей генеральной совокупности все новые и новые повторные выборки, пересчитывать на этой основе оценки параметров и анализировать дрейф $\hat{\theta}$. Но нет ли иного более экономного способа, позволяющего обойтись без дополнительных измерений?

Метод "складного ножа" (jackknife) первого порядка состоит в том, чтобы из одной выборки сделать n новых, исключая каждый раз по одному наблюдению. Для каждой из сгенерированных выборок объемом $(n - 1)$ можно рассчитать псевдозначение интересующей нас статистики: $\hat{\theta}_{n-1,\,k} = f_{n\text{-}1}(x_1, x_2, x_3, \ldots, x_{k\text{-}1}, x_{k+1}, \ldots, x_{n\text{-}1}, x_n)$. Среднее из всех возможных псевдозначений дает нам оценку $S_{\text{-}1}$ «складного ножа» первого порядка $\hat{\theta}_{-1}$, а разность $\left(\hat{\theta} - \hat{\theta}_{-1}\right)$ определяет полное смещение первоначальной выборочной оценки $\hat{\theta}$, т. е. отражает возможное искажение оценки параметра, если, например, наше распределение X не слишком «похоже» на гауссиану.

Основная идея бутстрепа по Б. Эфрону состоит в том, что методом статистических испытаний Монте-Карло многократно извлекать повторные выборки из эмпирического распределения. А именно, берется конечная совокупность из n элементов исходной выборки $x_1, x_2, x_3, \ldots, x_{k\text{-}1}, x_k, x_{k+1}, \ldots, x_{n\text{-}1}, x_n$ и из нее на каждом шаге итерации «замены с возвращением» с помощью датчика случайных чисел формируется любое, сколь угодно большое число размноженных выборок. Например, при $n = 8$ одна из наших псевдовыборок могла бы иметь вид $x_4, x_2, x_8, x_2, x_1, x_2, x_4, x_5$, т.е. отдельные элементы могут повторяться. Как и в случае «складного ножа», в результате легкой модификации частотного распределения реализаций исходных данных можно ожидать, что каждая следующая бутстреп-выборка будет возвращать значение параметра, немного отличающееся от вычисленного для первоначальной выборки. Образующийся разброс показателя дает возможность построения доверительных интервалов анализируемой статистики (Manly, 2007).

Итак, в основе бутстреповского подхода лежит действительно очень простая идея, что истинное распределение статистик можно получить эмпирически, не опираясь ни на какие предварительные предположения. Возможность гибкой настройки и использование идей самоорганизации выгодно отличает бутстреп от метода «складного ножа» с его плоским и менее интенсивным вычислительным подходом, а существование самостоятельных выборочных процессов для разных уровней воздействия внешних факторов делает бутстреп менее зависимым от выборочных эффектов

Трудности начинаются, когда мы практически столкнемся с тонкостями отдельных ситуаций, чтобы качественно устранить сдвиг и/или скомпенсировать неустойчивость конкретного параметра. Например, есть много способов развития идеи размножения выборок (Орлов, 2006), внеся в переборную стохастику некоторое зерно детерминизма. Можно, например, по исходной выборке построить эмпирическую функцию распределения, а затем тем или иным образом от кусочно-постоянной функции перейти к непрерывной функции распределения, например, соединив точки $[x(i); i/n]$, $i = 1, 2, \ldots, n$, отрезками прямых. Другой вариант построения размноженных выборок – к исходным данным добавляются малые независимые одинаково распределенные погрешности (при таком подходе одновременно соединяются вместе идеи устойчивости и бутстрепа).

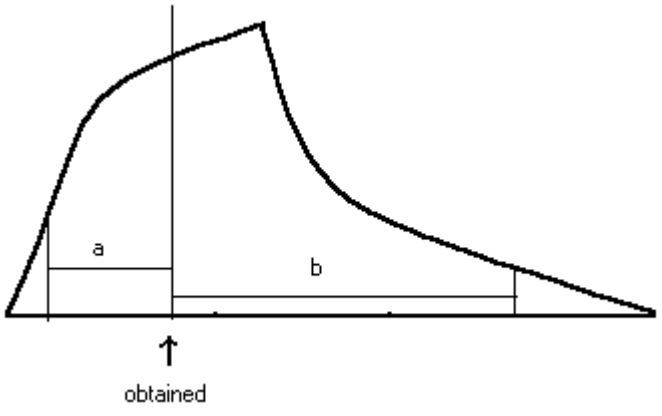
Рассмотрим процесс получения бутстреповских статистик на примерах.

3.2. Бутстреп-оценка доверительных интервалов среднего

Существует несколько различных методов, чтобы скомпенсировать сдвиг единственного параметра. Самым простым является метод процентилей. Предположим, что имеется выборка из 20 элементов с выборочным средним, равным 15. Сгенерируем из эмпирических данных 1000 псевдовыборок, используя алгоритм замены с возвращением, вычислим для каждой из них оценку математического ожидания и восстановим функцию распределения выборочного среднего. В методе процентилей в качестве границ 95%-го доверительного интервала среднего будут приняты считанные с гистограммы 25-ое (0.025·1000) и 975-ое (0.975·1000) значения статистических величин, полученных в ходе имитации.

Метод процентилей кажется разумными, но фактически мы находим здесь доверительные интервалы выборочных реализаций среднего, а не доверительные интервалы для искомого параметра. Иными словами, если эти границы окажутся равными 11 и 19, мы можем сделать заключение, что среднее из любой комбинации наших эмпирических данных с вероятностью 95% укладывается в пределах между 11 и 19, тогда как мы собирались оценить уверенность относительно значения параметра μ.

Эта оговорка не составляет серьезной проблемы для оценки математического ожидания, поскольку для разумных эмпирических выборок распределение этого параметра приблизительно симметрично. Однако предположим для общности рассуждений, что распределение выборочного среднего, полученное бутстреппингом, имеет отчетливую асимметрию:



Здесь *a* представляет собой расстояние от меры положения, за которую мы принимаем среднее для оригинальной выборки, до нижней 2.5-ой процентили, и *b* – аналогичное расстояние до верхней 97.5-ой процентили, причем *a* <> *b*.

К.Лунненборг (Lunneborg, 2000) анализирует эту ситуацию и показывает, что, если использовать среднее $\overline{X}$ как анализируемую статистику, её доверительные пределы будут от $\overline{X}$ - (0.975%-е значение - $\overline{X}$) до $\overline{X}$ + ($\overline{X}$ - 0.025%-е значение). Иными словами, при заданной доверительной вероятности оценка параметра будет статистически значима в пределах от ($\overline{X}$ - *b*) до ($\overline{X}$ + *a*), тогда как с использованием метода процентилей эти границы вычисляются с точностью до наоборот, т.е. от ($\overline{X}$ - *a*) до ($\overline{X}$ + *b*). Еще раз отметим, что для симметрично распределенных параметров *a* = *b* и эта проблема не имеет практического значения, однако в иных случаях на нее следует обратить самое пристальное внимание.

Сконцентрируемся однако на классическом понятии доверительных интервалов относительно оцениваемого параметра μ математического ожидания. При использовании параметрических методов выборочные оценки стандартных доверительных границ могут быть найдены как $CI_{\alpha/2} = \overline{X} \pm t_{\alpha/2} S_{\overline{X}}$, т.е. пределы симметричны (*a* = *b*) и вычисляются как произведение критического значения *t*-критерия на стандартную ошибку среднего. Заметим, что Госсет первоначально получил *t*-распределение при условии, что анализируемая выборка распределена нормально, поэтому истинные доверительные границы могут иметь некоторый сдвиг, пропорциональный тому, насколько конкретная эмпирическая выборка отклоняется от этого предположения.

Попробуем как-то скомпенсировать этот сдвиг, отказавшись от предположения о нормальности эмпирического ряда, и скорректировать критические значения $t_{\alpha/2}$. Выполним *B* итераций бутстрепа, вычисляя для каждой *i*-й сгенерированной псевдовыборки значения среднего $\overline{X}_i^*$ и стандартного отклонения S^*. На основе этих статистических данных мы можем вычислить бутстреппированные значения $t_i^* = (\overline{X}_i^* - \overline{X}) / S_{\overline{X}}^*$ и восстановить функцию распределения t^*, не использующую предположения о нормальности. Нам теперь остается только найти по гистограмме характерные значения t^* для 97.5 % и 2.5 %-х вероятностей и заменить ими критическую величину $t_{\alpha/2}$ в традиционной формуле. Мы получаем доверительные границы $CI_{\alpha/2} = \overline{X} + t_{0.975} S_{\overline{X}}$ и $\overline{X} + t_{0.025} S_{\overline{X}}$. Заметим, что мы поменяли местами 2.5 и 97.5-ые процентили t^* по той же причине, по которой это сделал Лунненборг.

Надо сказать, что мы привели не самый лучший вариант решения по компенсации сдвига. Эфрон 20 лет посвятил этой проблеме и разработал процедуру BCA (bias correction and acceleration) коррекции доверительных границ, которая учитывает различные выбросы, дрейф стандартной ошибки среднего и другие факторы. Процедура слишком громоздка, чтобы обсуждать ее здесь, но ясно представлена в одном из самых полных учебных пособий по бутстрепу (Efron, Tibshirani, 1993). А мы завершим наши рассуждения конкретным примером.

К. Маккойли (Macauley, 1999) с использованием нейробихавиорального теста когнитивной способности (Neurobehavioral Cognitive Status Examination) оценила значение ментального статуса у 123 пожилых людей в возрасте от 60 до 95 лет. Распределение характеризуется сильной асимметрией, поскольку некоторая часть обследованных пациентов серьезно потеряла способность к логическому мышлению (см. рис. 9).

Маккойли проводила оценку доверительного интервала для медианы, но мы используем этот пример, чтобы сформировать 95%-ые доверительные интервалы для среднего (рис. 10). Найденные бутстреппированные граничные значения симметричны относительно оценки математического ожидания: нижняя граница 8.65 на 0.489 единиц меньше среднего, а верхняя граница 9.60 превысила его на 0.465 единиц. Можно также

увидеть, что бутстреп-распределение оценок среднего приблизительно нормально, а а его стандартное отклонение (эквивалентное стандартной ошибке среднего) равно 0.241.

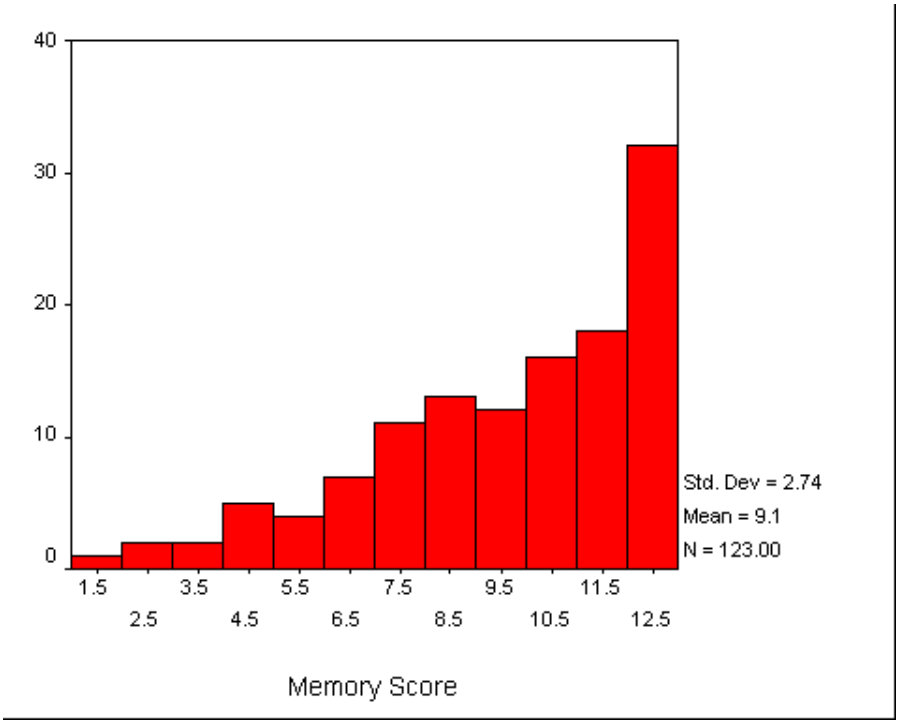


Рис. 9

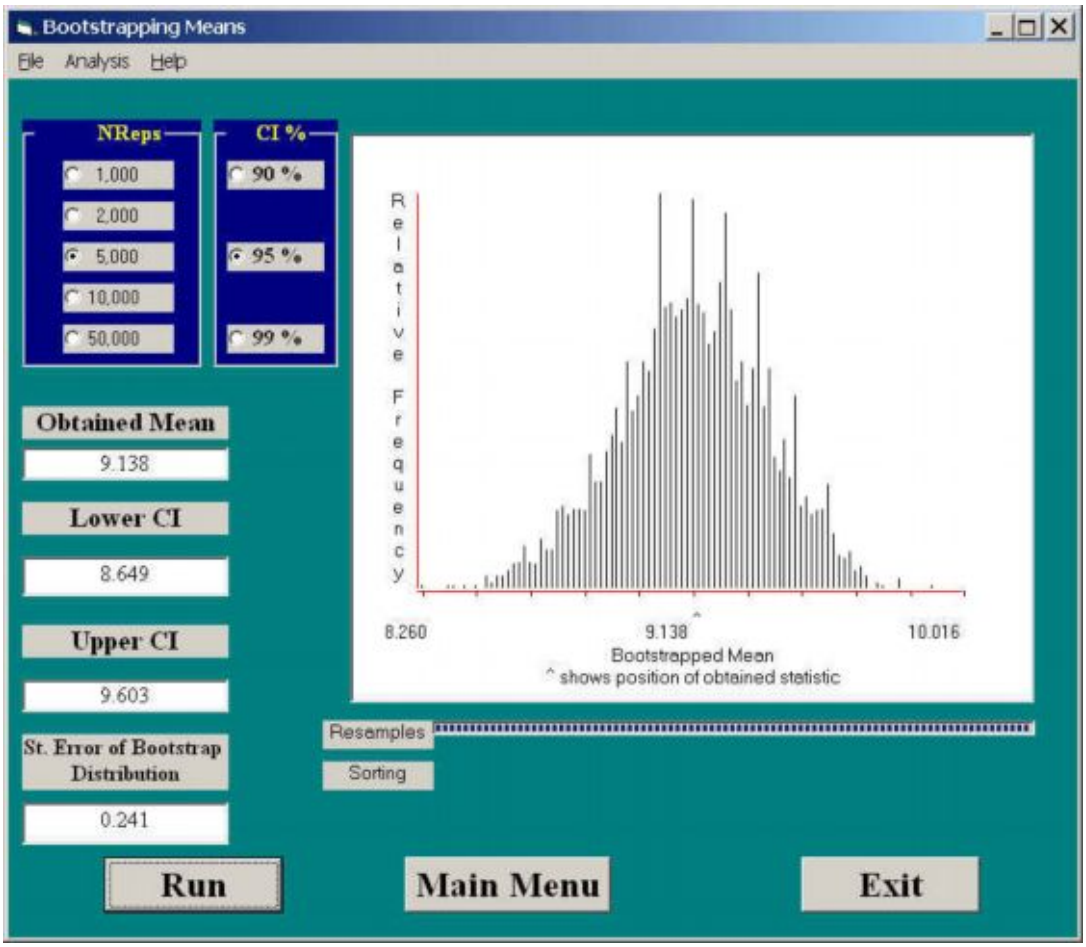


Рис. 10

Если игнорировать тот факт, что распределение измеренных данных имеет сильную отрицательную асимметрию, то использование параметрических формул дает результаты, весьма близкие к полученным бутстрепом: стандартную ошибку 0.247 и 95%-ые доверительный интервал от 8.65 до 9.63. Приходится признать, что, если мы имеем относительно большие выборки, нормальная асимптотика работает даже тогда, когда распределение эмпирического ряда заметно отличается от гауссового.

Многие специалисты, работающие в этой области, не видят никакого особого преимущества использования бутстрепа для оценки доверительных интервалов среднего. Наши результаты хорошо подтверждают эту убежденность. Однако мы в дальнейшем увидим, что это не всегда будет одинаково верно, когда речь дойдет до оценки более сложных статистических параметров.

3.3. Бутстреп-оценка доверительных интервалов медианы

Значительная часть того, что было сказано о бутстреп-оценке среднего, относится и к медиане, поэтому нет необходимости подробно повторять алгоритм процедуры. Для каждой итерации бутстрепа вычисляется выборочная медиана Med^* и после генерации заданного числа B псевдовыборок эти значения сортируются по возрастанию Med^*, т.е. восстанавливается функция распределения плотности вероятности.

Если смоделированное распределение симметрично, то доверительные интервалы $CI_{\alpha/2}$ для заданной вероятности α легко и быстро находятся как $\alpha/2$- и $(1 - \alpha/2)$-процентили. Если это распределение асимметрично, то нижняя граница $CI_{\alpha/2}$ = (Med - b), а верхняя – $CI_{\alpha/2}$ = (Med + a). Здесь a и b – соответственно расстояния от эмпирической медианы Med_{obs} до $B\cdot \alpha/2$ и $B\cdot (1 - \alpha/2)$-го порядкового значения в ранжированном ряду имитированных статистик.

У нас нет адекватной формулы для оценки стандартной ошибки, когда мы говорим о медиане, поэтому параметрический метод оценки ее доверительных интервалов отсутствует. Однако мы можем вычислить бутстреп-оценку t^* для медианы, как это делали для среднего, если сумеем предварительно преодолеть по крайней мере две проблемы.

Во-первых, мы могли бы заменить медианами Med_{obs} и Med^* числитель выражения для t^*, но чем мы собираемся восполнить отсутствие эмпирического стандартного отклонения подмножества медиан? Выход видится в бутстрепе второго уровня относительно основного бутстрепного цикла. Иными словами, псевдозначение Med^*, полученное на на i-й итерации наших вычислений, позиционируется как “эмпирическое” и для него запускается новая бутстрепная процедура, позволяющая найти стандартную ошибку Med^* и стьюдентизировать разность (Med_{obs} - Med^*).

Эфрон рекомендует B= 1000 – число итераций бутстрепа, которые мы проводим во внешней петле, и b = 50 – число итераций для подсчета стандартной ошибки, выполняемых для каждого внешнего цикла. Ничего священного в этих числах нет, но они дают общее представление об объеме вычислительной работы. Нужно найти всего $B\cdot b$ = 50 000 медиан и это – довольно много, но компьютеры никогда не умели спать или отдыхать, поэтому ничего невозможного для них нет.

И еще один шаг. При генерации псевдовыборок основного бутстрепного цикла, описанного выше, мы получаем B значений t^*, для которых мы найдем $\alpha/2$- и $(1 - \alpha/2)$-процентили. Но нам нужна еще стандартная ошибка медианы, которая соответствует стандартной ошибке среднего в традиционной формуле. И мы можем получить ее оценку S_{med}, просто вычисляя стандартное отклонение распределения B псевдозначений медиан Med^*. Тогда границами наших доверительных интервалов становятся $CI_{\alpha/2} = Med + t_{0.975} S_{Med}$ и $Med - t_{0.025} S_{Med}$.

Этот подход мы обсуждаем на примере медианы, но его можно использовать для широкого набора статистических параметров. Мы можем не иметь никакого представления о формуле для расчета стандартной ошибки, но непринужденно оцениваем

ее значение с помощью “грубой силы”, и в данном случае это – очень привлекательная и гуманная идея.

Рассмотрим пример оценки математического ожидания времени реакции: Стернберг (Sternberg, 1966) показывал испытуемым ряд чисел, после чего предъявлял им тестовое число. Респонденты должны были в течение максимально короткого времени нажать одну из двух клавиш, в зависимости от того, видят ли оно это число в ряду.

Распределение времени реакции имеет немного асимметричный и бимодальный характер. Если вычислить 95%-ые доверительные интервалы среднего, то программа SPSS нашла бы их как 61.01 и 68.19, но ничего не сказала бы о доверительных границах для медианы (рис. 11). Найдем их с помощью бутстреп-процедуры (рис. 12).

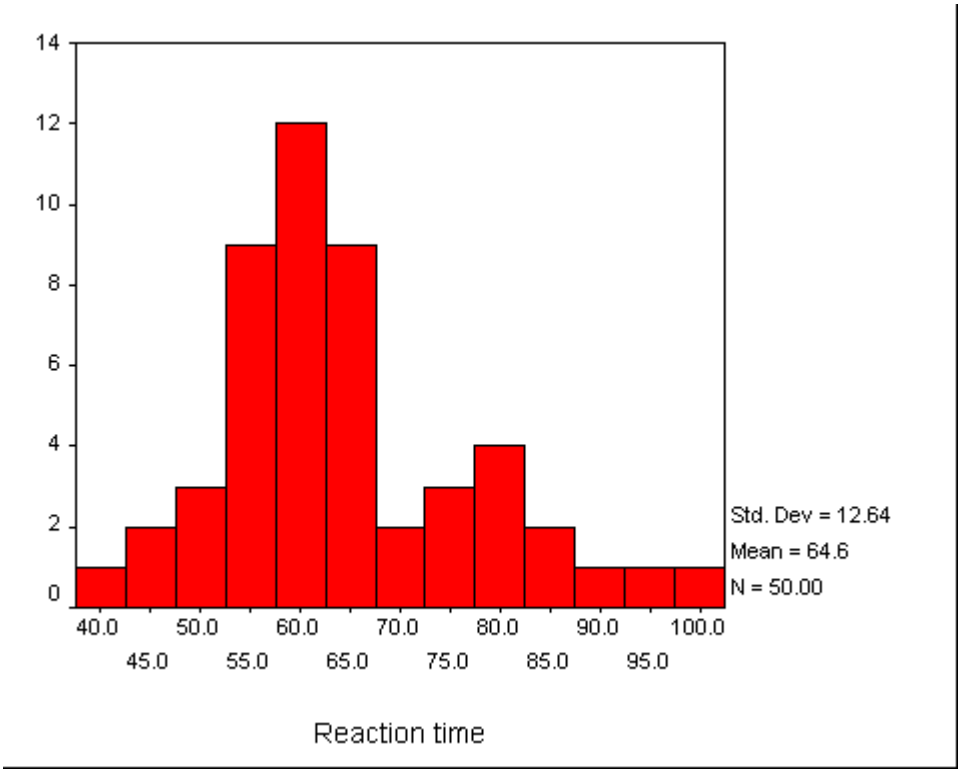


Рис. 11

Заметим, что для медианы доверительный интервал является несколько более узким (от 57.5 до 65.0), смещен влево и расположен асимметрично относительно выборочной медианы. Число различающихся значений модельных медиан существенно меньше, чем в примерах со средними, поскольку медиана должна совпадать с одной из величин, полученных эмпирически.

Здесь уместно высказать одно предостережение. Вы будете разочарованы, если пытаетесь применить процедуру бутстрепа в случае, когда эмпирические данные состоят из небольшого количества уникальных значений. Предположим, например, что данные образуют ряд 3 5 4 5 7 6 5 4 5 8 3 5 4 6 5 с медианой, равной 5. Когда будет генерироваться много выборок, распределение медиан сведется к трем значениям: 4, 5, или 6. С низкой вероятностью могло бы встретиться и число 7, но лишь в фантастической комбинации, когда значение 8 будет случайно извлечено 8 раз.

Более убедительные результаты можно получить, если выборка состоит из разумного числа различающихся значений. Это обстоятельство особенно важно, если ставится задача сравнивать медианы двух независимых выборок.

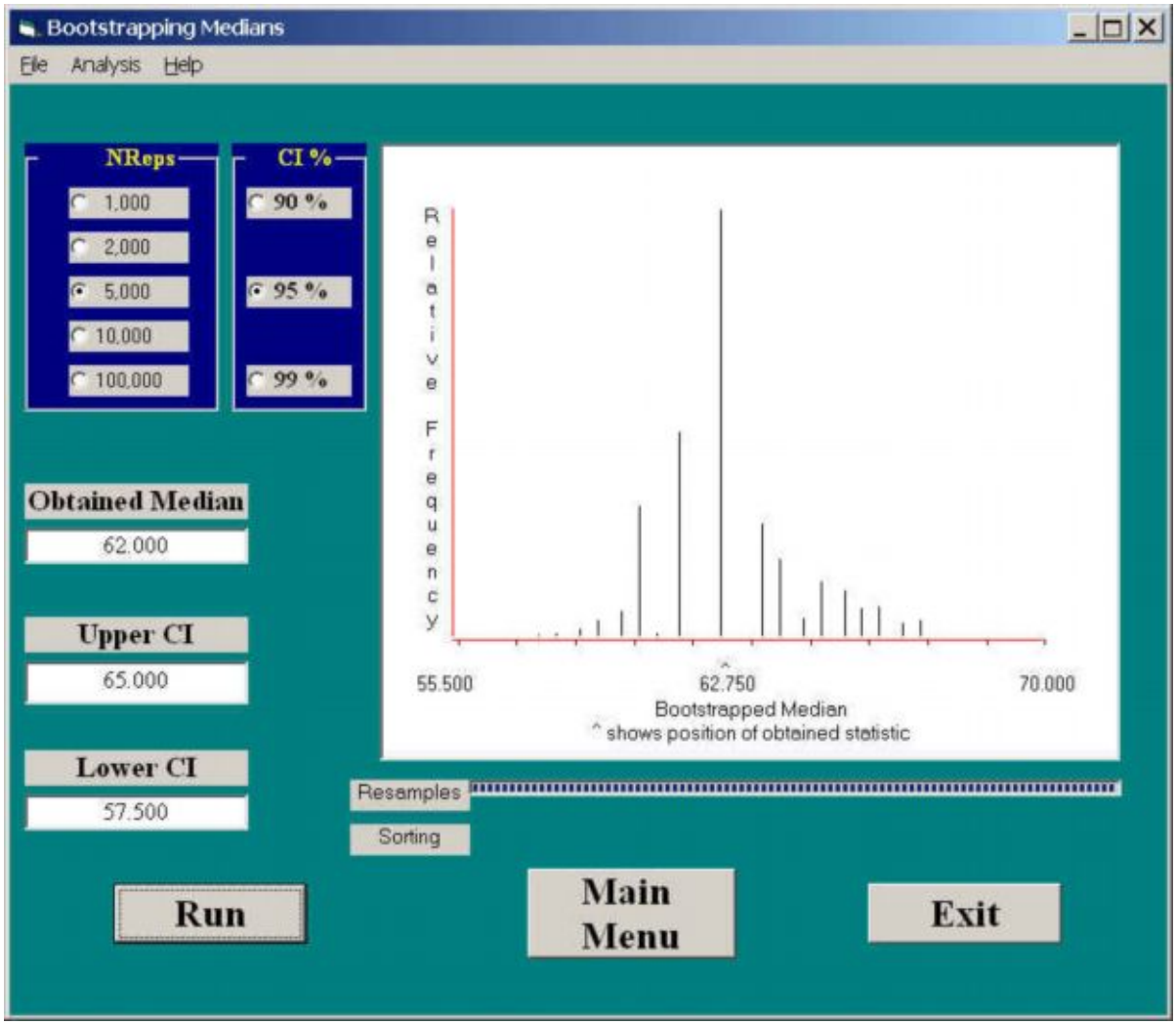


Рис. 12

3.4. Бутстреп: сравнение двух медиан

Основная процедура бутстрепа при сравнении медиан двух независимых выборок достаточно прямолинейна: вычисляется медианная разность двух псевдовыборок и формируется вариационный ряд этих разностей. Доверительный интервал различий может быть найден методом процентилей. Это означает, что при 1000 итераций бутстрепа и $\alpha = 0.05$ его границы соответствуют 25-му и 975-му значению в отсортированном ряду разностей. Если доверительный интервал не включает 0, можем отклонить нулевую гипотезу об отсутствии различий между медианами двух совокупностей.

Мы уже ранее рассматривали результаты использования бутстрепа на примере с исследованием Миррелла на скорость прохождения крыс в лабиринте (см. раздел 2.3, рис. 5), в котором была отклонена нулевая гипотеза об одинаковом воздействии музыки Моцарта и группы «Сибирская язва». Отметим здесь лишь то обстоятельство, что, хотя распределение медианных разностей не является симметричным с более низкой частотой результатов слева, границы доверительных интервалов оказались достаточно симметричными относительно математического ожидания медианной разности: 41 единиц ниже 187 сек. и 39 единиц выше 187 сек.

В другом примере, иллюстрирующем высказанное выше предостережение, коллектив психологов (Werner et al, 1970) опрашивал матерей нормальных и шизофреничных детей, которые рассказывали родителям истории по предъявляемым сериям из 10 картинок с неоднозначным смыслом. Психологи хотели показать, что нормальные дети, вероятно, будут рассказывать менее «страшные» истории, чем шизофреничные, т.е. отношения между родителями и детьми имеют более позитивный

характер. В конечном итоге исследователи нашли, что искали, хотя оказались не в силах отделить причину и следствие. Результаты этого анализа представлены на рис. 13.

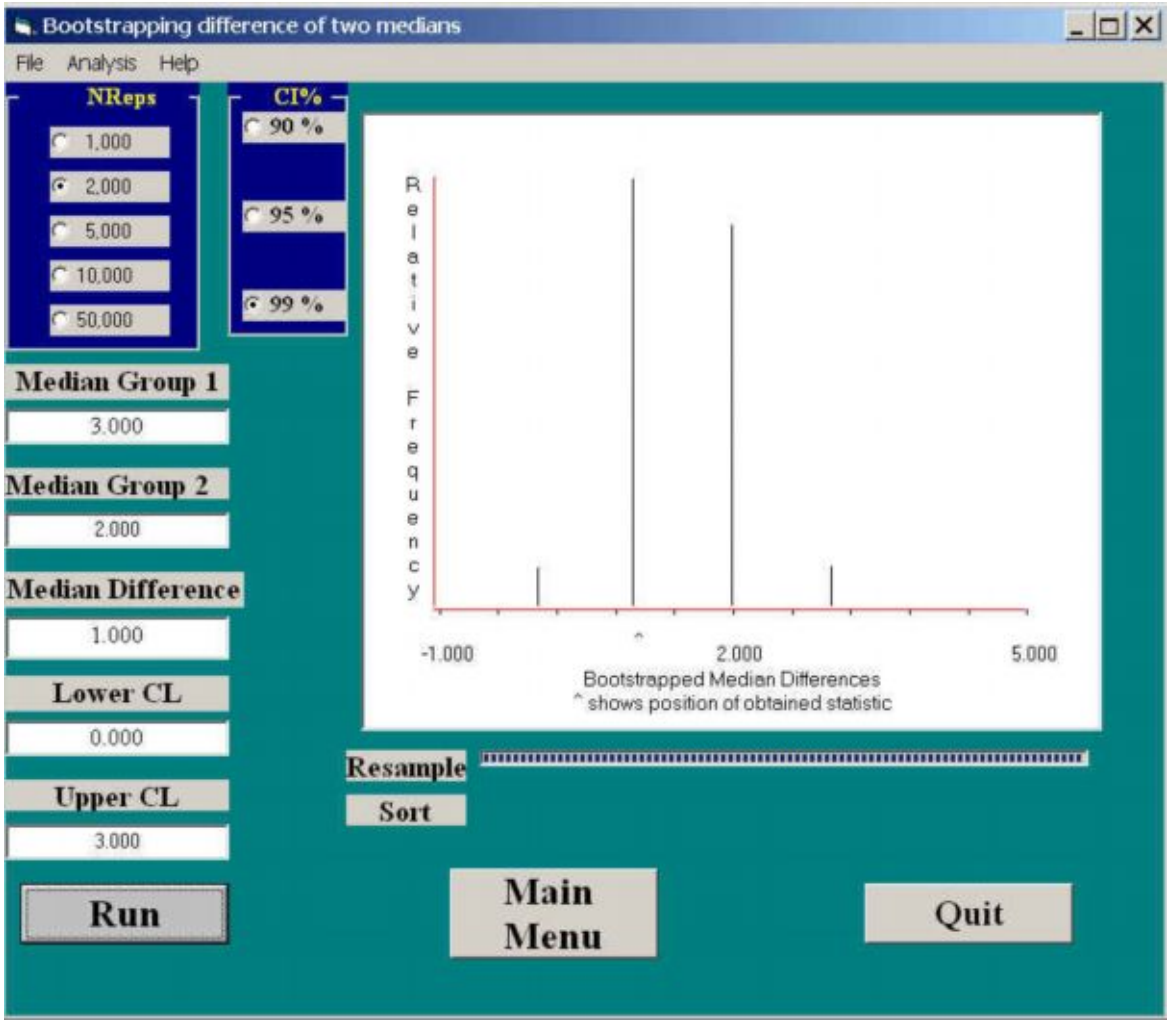


Рис. 13

Здесь можно увидеть, что медианы двух групп были 3 и 2 с медианной разностью, равной 1, для которой границы 95%-ых доверительных интервалов составили 0.0 и 3.0. Поскольку интервал включает 0.0, мы не можем отклонить нудевую гипотезу, что число позитивных трактовок картинок одинаково в обоих группах. В этом же примере параметрический двусторонний *t*-тест различия между средними 3.55 и 2.1 отклонил бы H_0 с $t = 2.662$, $p = 0.011$ и границами доверительных интервалов $0.347 \div 2.55$.

3.5. Бутстреп: оценка зависимости двух переменных

Оценка статистической значимости и доверительного интервала коэффициента корреляции *r* является важной и одновременно наиболее трудной задачей статистики. В общем случае логика проста: корреляция должна существенно отличаться от нуля. Однако насколько велики должны быть эти отличия? Например, мы можем предполагать, что между индексом интеллекта (IQ) и последующим карьерным успехом есть очевидная связь. Но если при этом коэффициент корреляции равен 0.2, значит ли это, что интеллект никак не помогает найти счастья в жизни? Ответы на этот вопрос может дать корректная оценка доверительных интервалов.

Напомним, что значение параметра $\rho_{x,y}$ связи двух переменных рассчитывается, исходя из предположения об их двухмерном нормальном распределении. Для того, чтобы построить доверительные интервалы для ρ, нужно знать выборочное распределение *r*, которое неизвестно, а его параметрическая аппроксимация достаточно сложна. Более

простой метод предложил Р.Фишер, который нашел, что величина $Z = 0.5 \ln\left(\frac{1-r}{1+r}\right)$ распределена приблизительно по нормальному закону с математическим ожиданием $E(Z) \approx \text{arcth}(\rho)$ и $\sigma^2_Z \approx (n-3)^{-2}$.

Ранее в разделе 2.5 мы рассматривали пример американских психологов (Katz et al., 1990), которые анализировали зависимость результатов экзамена от благоприобретенных навыков тестирования и с использованием процедуры рандомизации отклонили H_0 об отсутствии связи этих переменных при эмпирической величине $r_{obs} = 0.532$. С использованием аппроксимации Фишера можно получить оценки доверительных интервалов: $CI_{0.95} = Z \pm t_{0.95} \cdot S^2_Z = 0.594 \pm 1.96$; $0.202 \le Z \le 0.986$, где $Z = 0.594$, $S^2_Z = 25^{-2} = 0.25$. Конвертируя Z в r, получаем доверительные интервалы параметра $0.2 \le \rho \le 0.756$.

Самым простым способом оценить бутстреппированные доверительные интервалы коэффициента корреляции является использование аналога "метода процентилей". Будем генерировать на основе эмпирических данных большое число псевдовыборок парами, т.е формируя случайную комбинацию из значений *x*-ов, выбираем также соответствующие им значения *y*-ов. На каждой итерации вычисляем r_j^*, $j = 1, 2, \ldots, B$, где B – общее число итераций и после завершения процесса сортируем получившийся вариационный ряд. Значения, соответствующие 2.5 и 97.5 процентилям r_j^*, принимаем за 95%-ые доверительные интервалы. Отметим, что этот тест не использует какие-либо предположения о распределении ρ или двухмерной нормальности исходных данных.

Результаты на рис.14, полученные после 5000 итераций, дали нам немного более узкий и несколько смещенный относительно r_{obs} диапазон оценок доверительного интервала $0.269 \le \rho \le 0.722$, по сравнению с параметрическим подходом. Это – весьма частое явление, позволяющее сделать вывод о тенденции к гипердиагностике эффекта при аппроксимации Фишера, радующей глаз практических исследователей, привыкших обнаруживать взаимосвязь «под любым фонарем». Заметим также, что полученные нами доверительные границы не включают 0.0, что подтверждает статистическую значимость корреляции в ситуации с экзаменационными тестами.

Эфрон и другие исследователи долго искали дополнительные способы улучшить бутстреп-решение для оценки значимости коэффициента корреляции. Особенно это касается развития методов функционального преобразования исходных данных, чтобы добиться линейной зависимости между переменными. Новые подходы были работоспособны в некоторых частных случаях, но далеко не всегда. В любом случае, более подробный экскурс в литературные источники по сути этой непростой проблемы весьма поучителен, хотя не особенно утешителен.

3.6. Бутстреппинг при однофакторном дисперсионном анализе

Обсуждавшиеся выше процедуры бутстрепа были сосредоточены прежде всего на оценке параметра с использованием доверительных интервалов. Хотя доверительные интервалы действительно могут обеспечивать проверку гипотез, это не является их основной задачей. В случае дисперсионного анализа проверка гипотез - генеральная цель, а не второстепенный эпизод.

Оба подхода ресамплинга в случае дисперсионного анализа отличаются лишь в одной детали: если при рандомизации одни и те же исходные выборочные данных только объединяются, перемешиваются и случайно перераспределяются по группам, то при использовании аналогичной бутстреп-процедуры объединенная выборка перед распределением по группам частично модифицируется по схеме «замены с возвращением». Но при любом из подходов мы избегаем необходимости делать какие-либо предположения о нормальности распределения данных или гомоскедастичности.

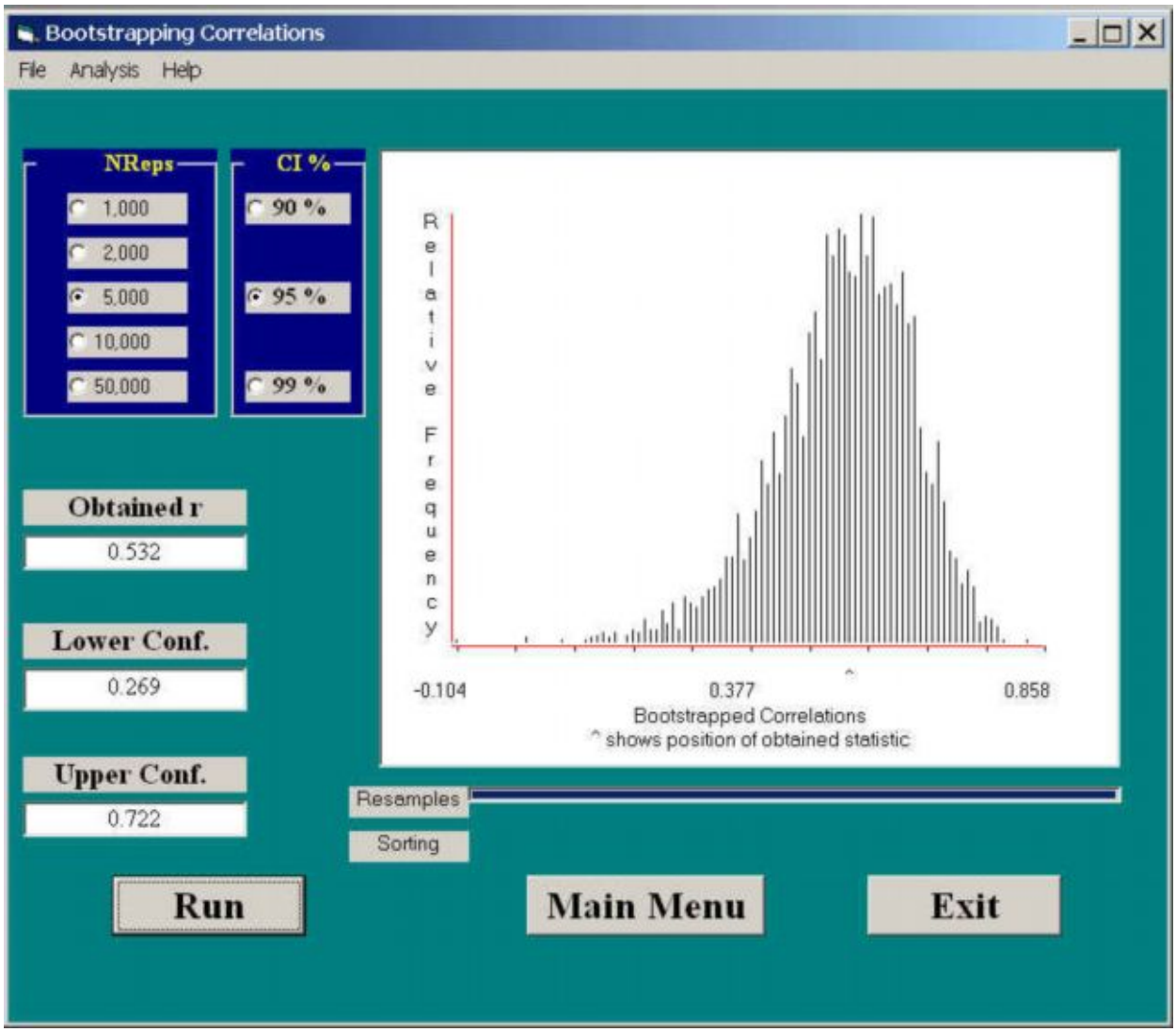


Рис. 14

Например, ранее (раздел 2.6, рис. 7) нами рассматривался пример сопоставление методов лечения жертв насилия. Стандартный однофакторный дисперсионный анализ на этих данных привел бы к отклонению нулевой гипотезы при $F = 3.046$, $p = 0.039$. Однако со столь небольшим объемом данных в каждой ячейке у нас нет хорошего способа убедить оппонентов, что нормальность распределения их ошибок и однородность дисперсий является разумным предположением. Что это не так, достаточно сравнить стандартные отклонения для групп SIT и PE.

При использовании рандомизации или бутстрепа никаких предварительных предположений делать уже не надо, как и нет необходимости заботиться о том, существенно ли они повлияют на окончательные выводы. Можно отметить также, что оба алгоритма ресамплинга дали между собой настолько близкие результаты, что мы сочли излишним дублировать идентичные рисунки.

Предположим, однако, что между двумя группами с различными методами терапии есть очень серьезное различие (например, для SIT и PE – значение показателя около 4, а для SC и WL – более 30). Когда мы объединяем все группы в одну большую псевдовыборку, ее частотное распределение будет бимодальным, и это создаст нам некоторые проблемы, поскольку мы исходим из предположения, что псевдовыборка сгенерирована при справедливости нулевой гипотезы. Эти проблемы неразрешимы для рандомизации, но в случае бутстрепа есть очень простое решение – использовать не сами исходные данные, а их отклонения от групповых средних ($X_{ij} - \overline{X}_j$). Ряд подобных

приемов и алгоритмов определяют гибкость бутстрепа по сравнению с другими методами ресамплинга.

4. Программная реализация ресамплинга

Главная трудность недостаточного практического использования процедур рандомизации и бутстрепа сводится к возможности иметь необходимое программное обеспечение. Хотя сами методики расчета концептуально очень просты, у вас должен быть доступный программный модуль, позволяющий осуществить генерацию повторных выборок. В частности, нам не удалось найти внятных возможностей использовать бутстреп-процедуры во многих важных пунктах анализа данных в необычайно, хотя и незаслуженно популярном пакете STATISTICA 8.

Ф. Гуд (Good, 2006) в своем практическом руководстве по применению методов ресамплинга при анализе данных рассматривает два подхода:

- использование макроопределений статистических пакетов, таких как MatLab, SAS, Stats, а также непосредственное программирование в средах Visual Basic, C++, Delphi;
- использование программ, управляемых с помощью меню, таких как S-Plus, Stata, StatXact, SYSTAT или Testimate.

В ряде случаев анализ реализуется комбинированным способом: через меню и с использованием системы команд, что является наиболее рациональным.

Гуд приводит коллекцию избранных текстов модулей на языках C++ и SAS, список макросов обращений к внутренним компонентам в средах Eviews, MatLab, Resampling Stats, R, S-Plus, Stata, а также дает описание расширения Resampling Stats Excel add-in. На сайте http://www.spsstools.ru обсуждается использование использование командного языка SPSS и реализация бутстреп-процедур для следующих методик анализа: построение доверительных интервалов среднего, медианы и дисперсии, восстановление выборочного распределения коэффициента корреляции, сравнение средних для *k* выборок, бутстреп-оценки коэффициентов линейной регрессии по МНК, анализ таблиц сопряженности и т.д.

Поскольку перечисленные программные комплексы весьма недешевы, рекомендуем обратить внимание на компактные версии очень удобных, а и часто бесплатных программ, созданных усилиями следующих авторов:

- Д. Хауэлла (D. Howell, http://www.uvm.edu/~dhowell/StatPages/Resampling/);
- П. Ядвижчака (P.Jadwiszczack, http://pjadw.tripod.com/);
- Дж.Саймона и П.Брюса (J.Simon and P.Bruce, www.resample.com);
- Р. Пикэлла и П. Смайса (R. Peakall and P. Smouse, http://www.anu.edu.au/BoZo/GenAlEx);
- Б. Рипли (B.D. Ripley), К. Халворсена (K. Halvorsen), А. Канти (A.J. Canty) и других разработчиков пакетов для бесплатной статистической среды R (www.r-project.org);
- Б. Манли (B. Manly, RT - randomization testing, http://www.west-inc.com);
- Н. Пеладью (N. Peladeau) и других разработчиков программы Simstat (http://www.provalisresearch.com).

ЛИТЕРАТУРА

Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия // Вып. 1.Пер. с англ. Ю. Н. Благовещенского. - 1982. - 320 с. Вып. 2.Перевод с англ. Б. Л. Розовского. - 1982. - 240 с.

Эфрон Б. Нетрадиционные методы многомерного статистического анализа. – Москва: Финансы и статистика, 1988. – 263 с.

Chernick M. R. Bootstrap methods, a practitioner's guide. Wiley Series in Probability and Statistics, 1999. – 369 p.

Chernick M.R., Fritis R. Introductory biostatistics for the health sciences: modern applications including bootstrap. – Wiley Series in Probability and Statistics, 2003. – 406 p.

Davison A. C., Hinkley D. V. Bootstrap methods and their application. (8th ed.). Cambridge: Cambridge University Press, 2006. – ___ p.

Edgington E. S. Randomization tests (3rd ed.). New York: Marcel Dekker, 1995. – 341 p.

Efron B., Tibshirani R. J. An introduction to the bootstrap. New York: Chapman & Hall, 1993. – 436 p.

Good P. Resampling Methods: a practical guide to data analysis (3nd ed.). New York: Springer, 2006. – 218 p.

Good, P. Permutation, parametric and bootstrap tests of hypotheses. New York: Springer, 2005. – 315 p.

Gotelli N.J., McGill B.J. Null versus neutral models: what's the difference? // Ecography. 2006. V. 29. P. 793-800.

Lunneborg C. E. Data analysis by resampling: Concepts and applications. Pacific Grove, CA: Duxbury, 2000. – ___ p.

Manly B. F. J. Randomization, bootstrap and Monte Carlo methods in biology (3nd ed.). London, UK: Chapman & Hall, 2007. – 445 p.

Mooney C Z , Duval R D. Bootstrapping. A nonparametric approach to statistical inference. Newbury Park, CA: Sage, 1993. – ___ p.

Rubinstein R.Y., Kroese D. P. Simulation and the Monte Carlo Method, 2nd Ed. John Wiley & Sons, 2003. – 336 p.

Schluter D. A variance test for detecting species associations, with some example applications // Ecology. 1984. V. 65. P. 998-1005.

Simon J. L. Resampling: the new statistics. Resampling Stats, 1997. – ___ p.

Stone L., Roberts A. The checkerboard score and species distributions // Oecologia. 1990. V. 85. P. 74-79.