БУТСТРЕП НАМ СТРОИТЬ И ЖИТЬ ПОМОГАЕТ…

Рецензия на книгу В.К. Шитикова и Г.С. Розенберга «Рандомизация и бутстреп: статистический анализ в биологии и экологии с использованием R». Тольятти: Кассандра, 2014. 314 с.

BOOTSTRAP US HELPS TO BUILD AND LIVE

Book review Vladimir K. Shitikov and Gennady S. Rozenberg «Randomization and Bootstrap: Statistical Analysis in Biology and Ecology with R». Togliatti: Cassandra, 2014. 314 p.

Большинство известных методов статистической обработки результатов наблюдений, используемых в самых разных отраслях знаний (в том числе в биологии и экологии), базируются на ряде ограничений и допущений, нарушение которых приводит к некорректности использования выбранных методов и ошибочной интерпретации полученных результатов. При проверке предпосылок и допущений статистических методов предполагается решение комплекса задач, к которым относятся: оценка случайности фактора, оценка стационарности и эргодичности исследуемых процессов, проверка гипотезы на нормальность распределения ошибок, обнаружение выбросов, оценка автокорреляции остатков, проверка постоянства математического ожидания и дисперсии ошибок, выявление мультиколлинеарности и т. д.

При выполнении проверки предпосылок и допущений может оказаться, что рассматриваемая выборка не удовлетворяет основным математическим предположениям, положенным в основу метода. Весьма часто в такой ситуации исследователи просто не учитывали некорректность применения метода и получали с его помощью выводы, которые, в лучшем случае вызывали сомнение, а в худшем – оказывались неверными.

Такое положение дел имело место в докомпьютерный период, когда обработка данных требовала много времени и усилий и делался акцент на методы, которые позволили бы получить максимум информации при небольшом объеме вычислений. Общий подход был весьма прост: предполагалось, что структура полученных данных «похожа» на некоторую распространенную статистическую модель, после чего выборочные оценки параметров рассчитывались по относительно простым теоретическим формулам.

В настоящее время производительность современных компьютеров позволяет без особых усилий обрабатывать гигантские объемы данных. Такая возможность оказала влияние на развитие новых статистических методов обработки. Появился новый класс альтернативных компьютерно-интенсивных технологий, объединенных общим термином «численный ресамплинг». Эти методы, как правило, не требуют никакой априорной информации о законе распределения изучаемой случайной величины. Лежащая в основе «численного ресамплинга» многократная обработка различных фрагментов исходного массива эмпирических данных в большинстве случаев, когда применение методов, полученных теоретическим путем, некорректно, позволяет получить более достоверные и точные результаты.

Начиная со второй половины XX века, термин бутстреп стал все шире использоваться в науке, технологии и других областях человеческой деятельности. В 1979 г. Брэдли Эфрон из Станфордского университета США опубликовал работу по нетрадиционным методам статистического анализа, с переводом которой, понятие бутстреп проникло в русский язык. Впрочем, термин мог появиться в русском языке раньше, в середине шестидесятых, при переводе книги американского физика Дж. Чью [1, 2].

К сожалению, на сегодняшний день в русскоязычной литературе трудно встретить современные публикации, посвященные этой динамично развивающейся идеологии. Монография В.К. Шитикова и Г.С. Розенберга очень удачно заполняет образовавшийся пробел кратким и, вместе с тем, предельно понятным изложением сути основных методов ресамплинга.

В монографии емко представлена совокупность компьютерно-интенсивных методов, в широком смысле относящихся к семейству различных процедур Монте-Карло. Наиболее подробно рассмотрены процедуры численного ресамплинга, которые заключаются в различных методах генерации случайным образом повторных выборок. Из существующих методов генерации повторных выборок в монографии описаны наиболее известные алгоритмы, такие как перестановочный тест (permutation), бутстреп (bootstrap), метод «складного ножа» (jackknife) и кросс-проверки (cross-validation).

Показано, как с помощью процедур численного ресамплинга, можно корректно проверить статистическую гипотезу или получить несмещенные оценки характеристик: математического ожидания, дисперсии, доверительного интервала, коэффициентов модели и т. д. Особенно интересно, что возможности применения методов иллюстри-

руются на широком круге конкретных авторских примеров с использованием в качестве исходных материалов данных реальных наблюдений. Для большинства примеров выполнен сравнительный анализ полученных результатов с классическими асимптотическими методами, основанными на том или ином стандартном предельном распределении.

Важно отметить, что авторы знакомят читателя с использованием методов в решении конкретных исследовательских задач и соответствующей реализацией их на программном обеспечении. Основное внимание в этом вопросе уделялось доступной свободно распространяемой статистической среде R, которая постепенно становится общепризнанным мировым стандартом при проведении научно-технических расчетов.

Отличительными особенностями монографии является разумная умеренность в изложении теоретических аспектов представленных методов. Просто и понятно изложены важные элементы теории статистического анализа данных наблюдений, подробно описаны, как сами процедуры численного ресамплинга, так и примеры их применения. Большая часть решений рассмотренных задач, дополнена текстами несложных скриптов в кодах статистического пакета R. Такое дополнение позволит читателю с легкостью самостоятельно воспроизвести технику расчетов при выполнении своих собственных исследований. Благодаря этому, представляемая монография может рассматриваться также как справочник по реализации различных алгоритмов обработки данных для исследователей, которых привлекла эта инструментальная среда.

Следует отметить методическую ценность книги. В описании примеров авторы придерживались следующей последовательности: краткая содержательная постановка задачи, смысл алгоритма обработки данных с некоторыми расчетными формулами, основные полученные результаты и их возможная интерпретация. С учетом предметной области рассмотренных примеров монография может быть полезной в качестве учебного пособия по статистическим методам для студентов и аспирантов высших учебных заведений биологического профиля.

Несколько слов о процедурах численного ресамплинга. Это семейство процедур основывается на традиционных общих идеях статистического анализа. Фундаментальным остается рассуждение о соотношении между случайными повторностями эмпирических данных и генеральной совокупностью, причем никакие сверх интенсивные методы не являются панацеей от влияния неучтенных факторов или систематических погрешностей при плохо поставленном эксперименте. Статистические выводы также базируются на классических доверительных интервалах, основанных на выборочных распределениях используемых критериев. Ключевое отличие лишь в том, что повторности классической выборки извлекаются из генеральной совокупности, а псевдоповторности ресамплинга – из самой эмпирической выборки.

Из всех процедур численного ресамплинга в книге наибольшее внимание уделено бутстрепу. Авторы представили подробный сравнительный анализ его с другими, также хорошо известными метода этого же семейства. В монографии для каждого иллюстрационного примера, решенного классическим методом, представлено альтернативное решение (а в некоторых случаях и не одно), основанное на методе бутстреп, и сравнение результатов с рассмотрением плюсов и минусов применения того или иного метода. Хотелось бы отметить, что в перечень рассмотренных практических задач вошли элементарная статистика, проверка гипотез, различные подходы к оценке биоразнообразия, дисперсионный анализ, специальные формы регрессии и оценки информативного набора предикторов моделей, многомерные методы классификации, редукции и распознавания образов, процедуры, использующие байесовский подход, анализ временной или пространственной динамики и т. д.

Заметим, что универсального и строгого определения понятия «бутстреп» не существует. Можно встретить частные определения в рамках какой-то формальной, например, математической модели явления. По своей сути бутстреп – это практический компьютерный метод получения оценок числовых характеристик вероятностных распределений и определения точности полученных оценок. Основывается он на методе Монте-Карло. С его помощью по одной имеющейся выборке значений случайной величины, без проведения дополнительных серий наблюдений, генерируются новые, так называемые, повторные выборки, причем, в большом количестве (5-10 тыс.). Делается это следующим образом. Из выборки случайным образом с равной вероятностью извлекаются значения с последующим возвращением и включаются в новообразующуюся повторную выборку. Повторная выборка, достигшая такого же размера, что и исходная выборка, считается сгенерированной. В результате каждая повторная выборка представляет собой случайную комбинацию набора значений исходной выборки. В одной такой выборке некоторые исходные элементы могут встретиться несколько раз, тогда как другие отсутствовать. На следующем шаге получают псевдовыборку оценок интересующей характеристики распределения исследуемой случайной величины. Для этого по каждой повторной

выборки, известными из теории математической статистики методами, вычисляются оценки числовой характеристики распределения исходной случайной величины. В результате появляется возможность проанализировать распределение значений оценок числовой характеристики распределения исследуемой случайной величины, а также оценить их разброс и устойчивость.

Другими словами статистический бутстреп есть весьма искусственный метод размножения выборки, который позволяет получить примерный ответ на многие практические вопросы без модифицирования статистических методов анализа под конкретный случай, прибегая к помощи «грубой» компьютерной силе.

Данный метод находит широкое применение в различных областях, о чем свидетельствует обширный список литературы, доступный, например, в интернете по адресу http://www.resample.com. В то же время, строгое обоснование свойств бутстреп-оценок отсутствует, имеются лишь асимптотические оценки их поведения. Они не позволяют определить процедуру применения бутстреп-метода и его параметры для получения оценок требуемой точности. Поэтому встает проблема обоснования данного метода.

Исследования свойств бутстреп-метода в поисках ответа на вопрос: как связано количество повторных выборок на результаты оценивания, показали, что при малых количествах повторных выборок (до 1000) получающиеся бутстреп-оценки могут располагаться в значительном интервале вокруг истинного значения оцениваемого параметр. С ростом числа повторных выборок (от 5 до 10 тыс.) этот интервал уменьшается. Очевидно, что данное свойство определено связью между дисперсией бутстреп-оценки и количеством повторных выборок.

При выборе программного обеспечения для выполнения вычислительных операций статистических методов авторы учли, тот факт, что большинство пользователей для реализации расчетов оказывают предпочтение программам, управляемым с помощью графического интерфейса. Так для нескольких примеров в начальных главах использованы компактные версии удобных, бесплатных и «биологически ориентированных» компьютерных программ, таких как: «Resampling» (созданная Д. Ховелом, доступная на сайте http://www.uvm.edu), RundomPro 3.14 (автора П. Ядвижчака, сайт http://pjadw.tripod.com), пакет RT (randomization testing) из 11 программ (Б. Манли, http://www.westinc.com/computerprograms.html) и др.

Однако чтобы охватить широкий спектр исследовательских задач с применением методов статистического анализа авторам пришлось обратиться к мощному статистическому пакету с коротким названием R (доступному на сайте http://www.r-project.org). К сожалению многих пользователей, этот пакет требует навыков работы в командной строке консоли. R является и языком, и программным обеспечением. К его наиболее замечательным особенностям относится: эффективная обработка данных и простые средства для сохранения результатов, большая коллекция инструментальных средств для проведения статистического анализа, простой и эффективный язык программирования, который включает много возможностей. Авторы знакомят читателя с языком R без развернутого описания его стандартных функций, ограничиваясь только необходимым для понимания и грамотного самостоятельного использования. Такой подход позволил не увеличивать объем монографии до гигантских размеров и сэкономить читателю силы для освоения представленных технических приемов. Для более глубокого изучения статистической среды R читатели могут самостоятельно обратиться к документации пакета и многочисленным учебным пособиям по вычислениям в R.

В заключении хочется отметить, что, авторы на рассмотренных примерах убедительно показали, что бутстреп в ряде случаев имеет существенное преимущество перед классическими методами статистического анализа. К таким случаям относится анализ сложных систем, состоящих из большого количества взаимодействующих элементов, для которых практически невозможно выявить все причинно-следственные связи, а структура данных имеет существенное отличие от какого-либо теоретического распределения. Для подобных систем применение бутстреп освобождает исследователя от необходимой модификации классического метода для корректного применения его в исследовании системы.

При написании книги авторы достигли главной цели. Легко читаемое введение в процедуры и методы численного ресамплинга, изложенное простыми словами и не перегруженное теорией, обеспечивает читателю быстрое освоение необходимых для практики знаний. А богатый перечень рассмотренных примеров позволяет исследователю, практически всегда найти пути решения конкретной задачи количественного анализа и статистической обработки своих данных. В дополнение читатель имеет краткий справочник по реализациям метода бутстреп в среде R с широким спектром задач статистического анализа. Представленные тексты скриптов в кодах R, могут применяться как готовые решения статистического анализа в собственных исследованиях.

Последовательное, ясное и конкретное рассмотрение методов статистического анализа на

примерах биологического характера делает возможным применение монографии в качестве замечательного учебного пособия для дисциплин по статистическим методам в высших учебных заведениях биологического профиля.

СПИСОК ЛИТЕРАТУРЫ

1. *Чью Д.* Аналитическая теория S-матрицы. М.: Мир, 1966. 152 с.
2. *Chew G.F.* «Bootstrap»: a scientific idea? // Science. 1968. V. 161, No. 3843. P. 762-763.

Е.Я. Фрисман, К.В. Шлюфман
Институт комплексного анализа региональных проблем ДВО РАН, г. Биробиджан