СТАТИСТИЧЕСКАЯ ПРОЦЕДУРА БУТСТРЕП КАК СРЕДСТВО ОТ МНОГИХ ПРОБЛЕМ ИСПОЛЬЗОВАНИЯ КЛАССИЧЕСКИХ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ

Дементьев В.А.

Дементьев Василий Александрович – Доктор физмат наук, профессор по кафедре физики,
Кафедра аналитический отдел, лаборатория сорбционных методов,
Институт геохимии и аналитической химии им. В.И. Вернадского РАН,
г. Москва

Аннотация: *процедура реальной статистики бутстреп широко применяется к малым выборкам случайных величин. Она не только позволяет легко вычислять функции от таких выборок, но и получать достаточно ясные представления о законах распределения случайных величин в исследованиях реальных объектов.*

Ключевые слова: *реальная статистика, процедура бутстреп, преимущества, законы распределения.*

STATISTICAL BOOTSTRAP PROCEDURE AS A REMEDY FOR MANY PROBLEMS OF USING CLASSICAL DISTRIBUTION LAWS

Dementiev V.A.

Dementiev Vasily Alexandrovich - Doctor of Physics and Mathematics, Professor at the Department of Physics,
DEPARTMENT OF ANALYTICAL DEPARTMENT, LABORATORY OF SORPTION METHODS,
INSTITUTE OF GEOCHEMISTRY AND ANALYTICAL CHEMISTRY. IN AND. VERNADSKY RAS, MOSCOW

Abstract: *the bootstrap real statistics procedure is widely applied to small samples of random variables. It not only makes it easy to calculate functions from such samples, but also to get fairly clear ideas about the laws of distribution of random variables in the study of real objects.*

Keywords: *real statistics, bootstrap procedure, advantages, distribution laws.*

Постановка задачи

Любой исследователь реальной действительности, физик, биолог, биохимик, медик или социолог, сталкивается с очень неприятной проблемой, когда он должен представить научной общественности отчёт о результатах своей деятельности. Такой отчёт обязан содержать цифровые оценки изученного объекта или процесса в некой системе единиц. Это единственное, что позволяет сравнить полученные исследователем оценки с оценками других исследователей данного реального объекта.

Проблема состоит в том, что результат измерений или вычислений некого параметра изучаемого объекта обязательно являются случайной величиной, проявляющей вариабельность в одинаковых, казалось бы, условиях. Причина состоит в том, что в процессе исследования на результаты измерений действуют факторы, имеющие причудливый скрытный характер. В борьбе с этой неизбежной неприятностью исследователь неоднократно повторяет эксперимент или наблюдение и получает целый набор первичных результатов, называемый выборкой случайной величины. Обычно от этой выборки необходимо вычислить некие функции. Результаты вычислений опять-таки дают выборки случайных результатов вычислений.

Такие выборки обычно мало интересуют читателей. Читателям нужно удобное по форме обобщение данных из полученных выборок. Традиционно, такое обобщение представляется в виде интервала существенных значений измеренной или вычисленной величины и значение надёжности, с которой эти результаты могут попасть в указанный исследователем интервал. Этот интервал называется доверительным. Потребители сравнивают полученный доверительный интервал с ранее представленными в литературе. Только тогда они могут судить, насколько новые результаты отличны от ранее известных. Если доверительные интервалы не перекрываются, то с некой надёжностью можно утверждать, что новый результат существенно отличается от известных.

Как получить доверительный интервал, располагая выборкой данной случайной величины? Тут не обойтись без знаний или предположений о статистическом законе распределения данной случайной величины. Только из надёжно установленного закона распределения вытекают рецепты получения надёжного доверительного интервала.

Вот это знание, вернее, незнание и составляет проблему. Исследователь обычно не знает закона распределения представляемой им случайной величины. Но что-то предполагает. На основе либо своей образованности, либо, в более благоприятном для дела случае, на основе своего богатого предварительного опыта общения с данным объектом, либо с аналогичными объектами. На этом впервые в статистической литературе настаивает наша работа [1].

В настоящей работе мы предлагаем исследователям воспользоваться статистической процедурой бутстреп не только для получения привычных форм представления своих данных в форме репрезентативной статистики (среднее значение величины, погрешность представления среднего и надёжность утверждения, что данный интервал покрывает существенную для исследования долю полученных данных). Бутстреп

позволяет довольно просто выявлять закон распределения получаемой в измерениях или вычислениях величины. А располагать надёжными представлениями о законе распределения случайной величины, это – большое счастье для исследователя.

В настоящей работе мы делимся нашим опытом получения такого бесценного знания на примере реального исследования. Мы надеемся, что заинтересованные читатели последуют нашему примеру в своей работе. А возможно, даже введут в вузовские курсы математической статистики и показанные здесь элементы реальной статистики.

Почему исследователю неудобно обращаться к нормальному закону распределения или даже к распределению Максвелла-Больцмана в случае медицинских или социальных работ?

В классической математической статистике нормальное распределение случайной величины x имеет великолепные свойства. Плотность вероятности этого распределения аналитически изображается широко известной привлекательной формулой и красивой кривой, распространяющейся по оси абсцисс в диапазоне $-\infty < x < \infty$. Это всё можно найти в учебнике или в Википедии. Не будем загромождать текст.

Вот эта раскинутость плотности нормального распределения и составляет проблему для исследователя реальных объектов. Их размеры всегда лежат в ограниченных пределах. Это знает любой исследователь. Не математик, изобретающий полезные абстрактные конструкции, лишь со многими оговорками применимые к реальной действительности.

Нормальное распределение порождает многие полезные следствия, используемые в нашей исследовательской практике также с оговорками.

Пример. Пусть в результате исследования фигурируют две выборки величин X_1 и X_2, измеряемые в одинаковых единицах измерения. Пусть исследователь верит, что эти случайные величины каждая распределены нормально с параметрами: средние или медианы распределений равны μ_1, μ_2 и разбросы распределений σ1, σ2. Пусть исследователь задался вопросом, а как распределена производная вычисляемая величина $A = X_1 + X_2$? Классическая математическая статистика даёт надёжный и приятный ответ. A распределено также нормально с параметрами $\mu = \mu_1 + \mu_2$, $\sigma = \sigma1 + \sigma2$. Просто и красиво. А если исследователь заинтересуется распределением случайной величины $B = X_1 * X_2$? (Похоже на некую площадь.) Исследователя ждёт крупное разочарование. В учебных курсах ему ничего про это не говорили. Почему. Потому что классическая статистика Не Умеет вывести формулу для распределения случайной величины B. Скандал! Надо искать рецепт в какой-то реальной статистике. Исследователь, вооруженный процедурой бутстреп сам легко справится с такой задачей, имея при этом в виду, что он даже не знает, каким реальным распределениям следуют его X_1 и X_2.

Многие биологи верят, что случайные результаты измерений их объектов следуют распределению, очень похожему на распределение Максвелла для модуля скорости v молекул в идеальном газе:

$f_v(x) = Bx^2\exp(-\beta x^2)$, ($x \geq 0$, при $x < 0$ $f_v(x) = 0$).

Тоже красиво и аналитически и графически (смотрите графики в Википедии). Это распределение уже жёстко ограничено слева. Приятно. Однако большие величины x могут простираться до бесконечности.

Приведём пример несостоятельности этого утверждения для биологии. Пусть собрали в лесу листья деревьев одной породы. Хотим представить себе, как распределены случайные значения максимального диаметра листа дерева этой породы. Похоже ли наше представление на распределение Максвелла? Да, как-то похоже. Листьев длины 0 мы не найдём. Листьев с длиной, равной длине экватора планеты Земля, мы тоже не найдём. Природа ограничивает и этот параметр реального распределения. Но в пределах реальных размеров листьев получается очень похоже на распределение Максвелла. Очень малые листья почти не попались, очень длинные тоже не попались. А вот листьев средних размеров попалось много. Если мы построим гистограмму длин набранных листьев, то огибающая гистограммы будет очень похожей на среднюю часть кривой плотности распределения Максвелла.

В реальной статистике приходится работать не с кривыми плотности неизвестных распределений, а с гистограммами измерений или вычислений реальных случайных величин.

Ещё пример, ещё более существенный для реального исследователя реальных объектов, приведен в следующем разделе.

Как получить представление о реальном распределении случайной величины по добытой в исследовании её сравнительно малой выборке?

В настоящее время подробные сведения о процедуре бутстреп и методах её применения в реальной статистике можно почерпнуть из Википедии. А в год издания нашей с соавторами работы [2] бутстреп был некой экзотикой, и мы его в той работе подробно описали. Сейчас мы воспроизведём пример из области медицинской статистике, приведенный в [2]. Делаем это потому, что этот крайне удачный в методическом смысле пример не до конца был нами изучен. Здесь дано продолжение изучения полученных результатов.

В пульмонологическом отделении больницы им. Боткина врач Т.Г. Химочко предлагала пациентам бесплатно новые средства фармакологической фирмы США в 1999 году. Для отчёта перед этой фирмой она анкетировала выбранную группу пациентов на предмет их собственной оценки качества жизни до начала терапии и по её окончанию. Группа пациентов была немногочисленной. Вот данные в баллах для результатов опроса пациентов до начала терапии. Данные представлены вектором q_e (quality experimental).

q_e =

{100 100 100 78,64 89,21 78,01

0

88,8 78,01 100 100 77,67 100 89,21 100 65,47 78,64}

Странный показатель седьмого пациента $q_{e7} = 0$ выделен в векторе в виде отдельной строки для наглядности.

Исследователя результативности будущего лечения интересовало среднее значение q_e по данной группе пациентов. Это среднее было получено по методу бутстрепа. Заодно была построена гистограмма средних, полученных по многочисленным выборкам бутстрепа. Гистограмма показана на рисунке.

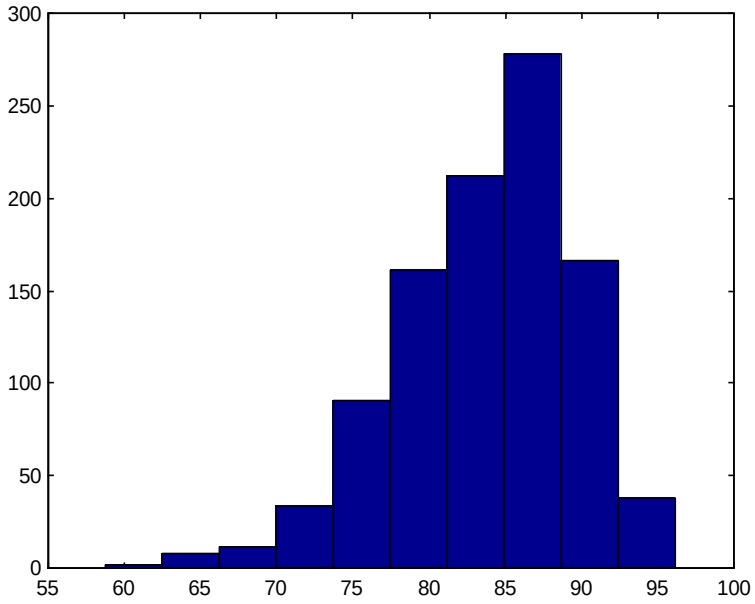

Рис. 1. Гистограмма значений среднего показателя качества жизни немногочисленной группы пациентов по 1000 выборкам бутстрепа.

По оси абсцисс отложены значения средних по бутстрепу в заданных интервалах, по оси ординат – количество случаев попадания среднего в данный интервал по выборкам бутстрепа.

Гистограмма почти не обсуждалась в работе [2]. Было лишь отмечено, что такая функция экспериментальной выборки, как среднее значение, совершенно не следует нормальному закону распределения.

Теперь рассмотрим возможности для специалиста, имеющего на руках данную гистограмму. Из неё специалист ясно увидит следующее.

Огибающая гистограммы очень напоминает кривую распределения Максвелла для некой принципиально положительной величины. Однако такая кривая похожа на зеркальное отражение закона распределения Максвелла. Отчего бы это? Когда известно, что многие биологические данные хорошо описываются именно этим распределением. Смотрите примеры в Википедии или в учебниках.

Специалисту совершенно ясно, что непосредственно из гистограммы легко получить эмпирическое генеральное среднее из всех 1000 выборок, генерированных процедурой бутстреп. Отсюда также получаем доверительный интервал для среднего показателя качества жизни в данной группе - от 74 до 96. Среднее из значений составляет 83,7. Обращаем внимание читателя на то, что этот интервал возможных значений среднего показателя качества жизни в данной группе пациентов и среднее значение качества исследователь получил из гистограммы, не делая никаких предположения о следовании средней величины какому-то определённому закону распределения. Это позволило специалисту работать с меньшим напряжением.

Если отбросить хвост малых значений среднего, обусловленный включением в процедуру бутстрепа странного нулевого значения качества жизни одного из пациентов, то получим тот же доверительный интервал для среднего показателя качества жизни в данной группе - от 74 до 96. Надёжность этого результата близка к 100%. При этом не пришлось обращаться ни к какой таблице коэффициентов Стьюдента, с которой приходится работать в предположении о следовании результатов измерений нормальному закону распределения.

При отбрасывании хвоста низких значений среднего хорошо видна роль специалиста, получавшего исходный вектор оценок качества q_e. Т.Г. Химочко на основе своего опыта знала, что среди пациентов нередко встречаются индивиды, склонные сильно преуменьшать свою оценку качества жизни перед курсом лечения.

Если мы захотим представить в отчёте более узкий интервал возможных значений среднего из вектора q_e, то необходимо представить заказчику работы и надёжность W утверждения, что среднее лежит в интервале $q_{\text{ср}} \pm \Delta q$. Это очень просто сделать, пользуясь непосредственно гистограммой. Специалисту здесь

всё понятно. Надо отложить справа-слева от $q_{ср}$ заданное значение Δq и подсчитать число выборок бутстрепа, попадающих в этот интервал. Это число надо разделить на полное число выборок, набранных по заданию специалиста в процедуре будстрепа (в нашем примере это 1000), и подать полученное W заказчику в процентах.

Более сложная вычислительная работа предстоит, если стоит задача представить доверительный интервал для среднего, попадающего на возрастающее или убывающее крыло гистограммы. Например, хотим выдать доверительный интервал для значения среднего $q_{ср} = 75 \pm 5$. Ясно, что в такой форме этого сделать нельзя, поскольку в расчёт W должно войти меньшее число выборок бутстрепа для $q_{ср} = 75 - 5$, и значительно большее число для $q_{ср} = 75 + 5$. Но и с этой задачей статистик, вооруженный вычислительными процедурами, например, МатЛаб, справится.

Заключение.

Отметим, что гистограммы, подобные гистограмме нашего примера, могут помочь специалисту в главном, о чём было сказано в преамбуле. Выполнив подобную работу для одной малой выборки данных об интересующей специалиста медицинской или биологической величины, специалист может быть вполне уверен, что результаты измерений родственной величины в сходных физических и социальных условиях будут следовать выявленному закону распределения, пусть совсем не похожему на распределения классической математической статистики. А это очень важно. Ведь предварительные (пусть приблизительные) знания о законе распределения интересующей специалиста величины дают ему большую свободу в обращении с объектом его профессионального интереса. Отсюда наш совет специалисту.

Не стесняйтесь изучать вычислительные процедуры современных статистических приложений для компьютеров, средства программирования системы МатЛаб. Это позволит создавать при небольших затратах труда собственные удобные средства анализа данных с включением в эти средства полезнейшей процедуры бутстреп. В Сети имеется великое множество описаний процедуры бутстреп и учебных пособий по её использованию в самых разных реальных исследованиях. Однако мы нигде не нашли ссылок на процедуру выявления закона распределения случайной величины с помощью бутстрепа.

Обсудив неприятные для исследователя свойства классических законов распределения, мы можем с облегчением сделать одно исключение. Оно предназначено для специалиста, вынужденного в своей работе использовать результаты измерений интенсивностей излучения радиоактивных препаратов. Такие интенсивности обычно представляют в форме числа отсчётов k детектора излучения за минуту. Особенно это характерно для малоактивных препаратов, возникающих в работах, использующих методы радиоактивных индикаторов. В таких работах МОЖНО непосредственно использовать распределение Пуассона, не опасаясь крупных неприятностей.

Распределение Пуассона определяет вероятность появления k отсчётов в одном измерении, если известно среднее значение λ интенсивности излучения от данного препарата:

$P(k) = (\lambda^k / k!)\, e^{-\lambda}$

Особенности получения следствий из дискретного распределения Пуассона подробно обсуждаются в книге [3].

Список литературы / References

1. *V.A. Dementiev*. Statistical Methods in Analytical Chemistry. Advances in Geochemistry, Analytical Chemistry, and Planetary Sciences, Springer, Chapt. https://doi.org/10.1007/978-3-031-09883-3_37pp 563–572
2. *Дементьев В.А., Химочко Т.Г., Сорока А.В*. Особенности применения метода бутстрепа при нахождении сложных статистических функций от малых выборок в биологических и медицинских исследованиях. Биомедицинская химия, Том 50, Приложение № 1, ГУ НИИ биомедицинской химии РАМН, М., 2004, с. 117-126.
3. *Дементьев В.А*. Измерение малых активностей радиоактивных препаратов. Издание второе, исправленное. URSS. Москва. ISBN 978-5-9716-9553-4 © ЛЕНАРД. 2022. 140 с.