Проект статьи для журнала Аналитическая Химия

Статистические методы в аналитической химии

(Пока это скорее эссе В.А. Дементьева, предназначенное для двух читателей – М.К. Беклемишева и Л.А. Грибова)

Постановка задачи

Когда ставится новая задача (а для меня она совершенно новая), то принято поискать в литературе, что сделано в данной узкой области науки. Правда, бывают редкие исключения. Энрико Ферми так охолаживал своих молодых сотрудников – «Не надо бежать в библиотеку. Сначала мы решим эту задачу, а потом уже посмотрим, что думают о ней другие».

Я обратился в Scopus за списком подходящих статей. Поиск по запросу «Analytical Chemistry AND Mathematical Statistics» по всей мировой периодике за все годы дал список из 650 статей. Оказалось, что, в основной массе, это касается производственных лабораторий, обслуживающих исследования по загрязнениям окружающей среды, качеству продуктов питания, медицину и фармакологию. Хорошо видно место статистических методов в таких работах. Это а) ссылки авторов, какие методы статистики они использовали при проведении конкретных анализов [например, 1]; б) отчеты о поисках новых статистических приемов, если хорошо известные не дают приемлемых результатов [2]; в) рекомендации вышестоящих организаций, обществ и комитетов [3], а также явно учебная литература для производственников [4].

Мне показалось, что такие производственные работы неинтересны для нас, занятых фундаментальными исследованиями в аналитической химии. В отечественной литературе можно легко найти метрологические рекомендации и отчеты по их исполнению в конкретных аналитических работах. А статьи, посвященные повышению статистической культуры аналитических лабораторий, у нас появились даже раньше зарубежных. Работая на заводе, я в 1959 году читал цикл увлекательных статей В.В. Налимова в журнале Заводская Лаборатория, где изложение постепенно поднималось от репрезентативной статистики до планирования эксперимента.

Особняком стоит набор статей по использованию методов хемометрии в аналитической химии. Эти методы нельзя считать в строгом смысле методами математической статистики, хотя они на нее, конечно, опираются. Математическая статистика имеет дело со сравнительно простыми ситуациями, в которых проявляются случайности. Таких типичных ситуаций в аналитической практике не так уж много. Зато в связи с наплывом сложных приборных методов исследования, дающих огромные массивы данных сложной структуры, возникла необходимость как-то разбираться в этой структуре, чтобы выделять ясные аналитические сигналы. А так как систематических рецептов математическая статистика здесь предложить не может, исследователями делаются попытки изобрести различные эвристические подходы к отделению полезных аналитических сигналов от бесполезного информационного шума. Вот эти попытки и составляют новую отрасль искусства первичной обработки экспериментальных данных. Ясно, что эта отрасль нехарактерна для исследований в фундаментальной аналитической химии. Как и всякая фундаментальная наука, эта ветвь химии старается упрощать изучаемые природные и лабораторные ситуации, а не усложнять их. С идеями хемометрии лучше всего ознакомиться, получив обзор из первых рук от редактора журнала Chemometrics and Intellectual Laboratory [5].

Поняв, что упомянутые области применения математической статистики не могут заинтересовать фундаментальных исследователей, я сократил поиск и запросил систему Scopus: «Analytical Chemistry AND Mathematical Statistics» по страницам только журналов Analytical Chemistry и Analytica Chimica Acta за все годы. Нашлось всего 23 статьи, куда попали и упомянутые выше.

Я задал себе вопрос – почему улов столь беден. Ответ я мог искать, только опираясь на свой опыт работы в Докторском диссертационном совете по аналитической химии в ГЕОХИ. Слушая работы наших аспирантов и их руководителей, будущих докторов наук, я мог себе представить круг интересов исследователей фундаментальных проблем этой науки. И мне стало понятно, почему отечественная и мировая аналитическая химия столь мало заинтересована в данном инструменте обработки данных. Повторю еще раз сказанное выше иными словами. Фундаментальная наука ищет либо новые природные закономерности, либо средства прорыва в решении давно назревших и нерешенных проблем. Такие поиски требуют предельного возможного упрощения ситуаций. Этого можно достичь либо с помощью изощренного эксперимента, который устранит все неважные факторы, либо с помощью глубоко продуманных теоретических моделей. Эти средства заранее элиминируют источники случайностей, оставляя лишь какой-то неустранимый их минимум. И тогда сложные статистические методы исследователю просто не нужны. Можно обойтись простыми стандартными методами.

Есть еще одна причина слабого интереса ученого к методам математической статистики. Эта причина еще более важная. Я знаю ее проявление на собственном исследовательском опыте, хотя начало моей научной карьеры складывалось в такой области знаний, где без статистических методов обойтись просто невозможно. Это радиометрия, а мой опыт был отражен в книге [5]. Дело в том, что упорно работая над сложной проблемой, на переднем крае узкой области знания, исследователь очень глубоко погружается в сущность получаемых данных. Никто не может знать особенности поведения этих данных лучше, чем сам исследователь. В том числе, природы этих данных не может так же глубоко знать никакой специалист по методам математической статистики. Поэтому исследователю сложной проблемы обычно достаточно самых простых средств репрезентативной статистики, чтобы грамотно представить свои результаты, чтобы коллеги могли однозначно понять полученные автором данные, попытаться воспроизвести новые рекордные результаты и сопоставить их с результатами автора. Конечно, при таком сопоставлении необходима объективная оценка точности данных, полученных разными авторами в не совсем одинаковых лабораторных или природных условиях. Вот тут и помогает информация о статистических свойствах получаемых данных, причем достаточно выразить эту информацию в простой форме доверительных интервалов при объявленной их надежности. Вот когда на основе прорыва в решении проблемы (например, резкого снижения предела обнаружения вещества в пробе) строятся новые аналитические методики, приходится сталкиваться с дополнительными факторами неясной статистической природы. Тогда-то и приходится подбирать адекватные статистические рецепты, либо разрабатывать новые хемометрические приемы обработки данных.

После всего сказанного я ставлю перед собой задачу угадать, какие статистические методы могут оказаться полезными в исследовательской, а не производственной деятельности химика-аналитика. Я надеюсь, что мне в этом поможет некоторый положительный опыт, накопленный мной в совместных работах со специалистами чужих для меня областей знания.

Одномерная статистика

При изучении даже очень сложных явлений, когда мы собираем о таком явлении обширную информацию, нам хочется представить наше конечное понимание каким-то одним числом. К сожалению, наши представления всегда несколько расплывчаты. Поэтому добросовестный исследователь нуждается в средствах, позволяющих представить главный результат в форме доверительного интервала. Такие средства он находит в одномерной математической статистике. Это главный раздел статистики для любого исследователя. Его рецептами исследователь обязан владеть в полной мере.

Напоминаю, что одномерная статистика пытается описать поведение результата измерения единственной характеристики *x* объекта в таких постоянных условиях, когда повторяющиеся измерения дают отличающиеся друг от друга значения. Такое поведение считается полностью описанным, если его удается сформулировать в виде закона распределения значений *x.* Такой закон может быть сформулирован самыми различными способами.

В руководствах по статистике закон распределения характеристики *x* чаще всего задается математической формулой. По такой формуле можно рассчитать либо вероятность появления точного значения *x*, либо вероятность попадания значения *x* в интервал от *x* до *x* + *dx*. В первом случае закон задается формулой

$$W = W(x,a)\,, \tag{1}$$

где *a* – вектор параметров, характеризующих условия измерения.

Во втором случае закон задается формулой

$$w = w(x,a)\,, \tag{2}$$

где *w* – плотность вероятности, которая затем используется для расчета вероятности $W(x,dx) = w(x,a)dx.$

Я не знаю примеров работ, где исследователь, проводивший измерения, приводил бы установленные им конкретные формы законов (1, 2) в качестве отчета перед научной общественностью. Установление одного из этих законов является обязанностью исследователя. Она сводится к выбору конкретной формулы и определению вектора параметров *a*, характеризующих условия эксперимента. После этого на основе установленного конкретного закона вычисляются различные производные величины, которые коротко и ясно описывают поведение объекта. Обычно это среднее значение $x_{ср}$, среднеквадратичное отклонение Δ*x* и надежность *Q* доверительного интервала $x = x_{ср} \pm \Delta x$. Если соблюдать правила хорошего тона, то надо упомянуть название того закона, на основе которого получен доверительный интервал. Обычно этого не делают. Закон просто подразумевается. В работах по радиометрии считают каждый отсчет детектора редким событием. Тогда (1) есть закон Пуассона. Из него следуют формулы для подсчета величин $x_{ср}$, Δ*x* и *Q*. Исследователю остается определить по результатам эксперимента величину единственного параметра *a* = λ. Это средняя скорость поступления импульсов за время одного измерения. Если же измеряется не дискретная, а непрерывная характеристика объекта, то обычно полагают, что (2) это нормальный закон распределения, имеющий два параметра – математическое ожидание μ и дисперсию σ^2. Все величины, входящие в выражение для доверительного интервала, легко вычисляются. Исследователю остается определить по результатам эксперимента величины параметров μ и σ^2. Все эти положения и вычислительные рецепты хорошо известны. Известны и неприятности, связанные с применением

этих рецептов на практике. Пусть мы совершенно уверены, что в эксперименте имеем дело с проявлением именно нормального закона распределения. Очень хорошо. Беда в том, что мы никогда не знаем и не узнаем конкретных значений параметров этого закона μ и σ^2. Мы вынуждены довольствоваться эмпирическими оценками этих параметров. Вместо μ мы по данным эксперимента находим среднее значение

$$x_{\text{ср}} = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \tag{3}$$

где *i* – номер измерения; *n* – число измерений. Вместо σ^2 мы по данным эксперимента находим эмпирическую оценку дисперсии

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_{\text{ср}} - x_i)^2 . \tag{4}$$

Обычно нас интересуют статистические свойства не результатов измерений x_i, а случайной величины $x_{\text{ср}}$. С математическим ожиданием ее все в порядке, оно тоже равно μ, независимо от числа измерений. А вот дисперсия ведет себя хуже, она равна s^2/n, то есть зависит от числа измерений. Обратим внимание на то, что сумму в (4) полагается делить не на *n*, а на число степеней свободы $k = n-1$. Это плата за то, что в (4) мы используем не теоретическую величину μ, а ее оценку $x_{\text{ср}}$,, которая связывает линейной связью все величины x_i. Возникает неприятность – мы утверждаем, что нам твердо известен проявившийся в эксперименте закон распределения, но не знаем точные параметры этого закона.

Есть еще неприятность. Если бы мы знали точную форму проявившегося у нас нормального закона, то мы могли бы использовать его следствие, что нормированная случайная величина $Z = \frac{\mu - X}{\sigma}$ распределена также по нормальному закону с параметрами 0, 1. Это было бы очень удобно при построении доверительного интервала для отдельного результата измерения $x = x_{\text{ср}} \pm \Delta x$. Мы могли бы написать $\Delta x = t\sigma$ с надежностью *Q*, которую нашли бы из таблиц нормального распределения в зависимости от выбранного по вкусу коэффициента *t*. Или наоборот, задаться надежностью *Q*, а это автоматически привело бы к выбору коэффициента *t*. Однако оказывается, что как раз из нормального распределения для интересующей нас случайной величины $x_{\text{ср}}$ получено следствие: нормированная случайная величина $z = \frac{\mu - x_{\text{ср}}}{s/\sqrt{n}}$ распределена уже не по нормальному закону, а по закону Стьюдента. С помощью этого закона доверительный интервал строится в такой же форме $\Delta x_{\text{ср}} = t\frac{s}{\sqrt{n}}$, но при той же величине надежности *Q* коэффициент *t* получается увеличенным. И он тем больше, чем меньше число степеней свободы *k*. Нам повезло только в том, что при очень большом *k* распределение Стьюдента быстро приближается к нормальному. Однако нам некогда проводить сотни измерений одной и той же величины. У нас в лаборатории есть и другие дела. Мы любим сделать три измерения, найти среднее и найти $\Delta x_{\text{ср}} = \frac{s}{\sqrt{3}}$. С какой же надежностью *Q* мы утверждаем, что $x_{\text{ср}}$ лежит в интервале $x = x_{\text{ср}} \pm \Delta x_{\text{ср}}$? Оказывается, *Q* = 57,7%. Маловато для приличного результата.

Однако это далеко не главные неприятности. Имеет смысл еще раз остановиться и оглянуться.

Сосредоточимся, к примеру, на законе (2). Этот закон на своем специфическом языке описывает условия эксперимента. Нравы и обычаи естествознания требуют, чтобы исследователь располагал исчерпывающей информацией об условиях своего эксперимента. Следовательно, он обязан не только упомянуть в отчете, какой закон распределения командует поведением измеряемой величины, но и самостоятельно установить форму этого закона. Никто, как правило, этого не делает, поскольку два гения – Гаусс и Чебышев – уже выяснили и объяснили нам два важных положения.

1. Нормальный закон распределения Гаусса является чрезвычайно удобным для вывода формул всех вычислительных рецептов, используемых в практике обработки результатов измерений. Другого такого удобного по форме закона просто нет.
2. Если Природа подает на вход измерительного прибора величину, искаженную бесконечным количеством мелких случайных факторов, а исследователь проводит бесконечно большое число измерений этой величины, то на выходе прибора наблюдается ряд случайных чисел, подчиняющихся именно нормальному закону распределения. (Это следует из закона больших чисел П.Л. Чебышева и из теоремы А.М. Ляпунова).

Все мы это поняли и привыкли пользоваться. Но научная мысль не стоит на месте, возникают недоумения и недовольства. Этот ропот недовольства можно заметить и в том наборе текстов, который я получил от системы Scopus. Значит, и аналитики оказались среди недовольных.

Кто виноват, понятно – Гаусс и Чебышев. Что делать, тоже более или менее понятно – надо всё-таки прояснять условия эксперимента. В том числе и в части закона распределения измеряемой величины, раз мы этот закон используем. Вот третий кардинальный вопрос – как мы это будем делать – на Руси почему-то задавать не принято. Конечно, специалисты по методам математической статистики (а они числят себя членами команды естествоиспытателей) задавались этим вопросом, но не очень громко отрапортовали, что у них получилось с ответом. А получилось, что самая трудная задача статистики, практически не решаемая на практике, состоит как раз в установлении конкретной формы закона (2). Или хотя бы в проверке гипотезы, что результаты серии измерений согласуются с нормальным законом. Отсюда следует, что тут мы не можем ждать помощи от профессионалов. Мы это будем делать, но мы пойдем своим путем. В специальном разделе я расскажу, как мы будем это делать.

Сейчас я расскажу, почему мы имеем право быть недовольными законом Гаусса. В этом законе

$$w(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \tag{5}$$

всё прекрасно, за исключением трудностей с физической интерпретацией тех прогнозов, которые он дает на своих крыльях, там, где $|x-\mu| \gg \sigma$. Закон предсказывает крайне малую вероятность обнаружить такую величину *x* в эксперименте. Но эта вероятность не равна нулю. А это означает, что мы имеем возможность наблюдать хотя бы несколько таких случаев, если запасемся терпением и будем проводить эти измерения миллионы и миллиарды раз. Однако любой

экспериментатор в неформальной обстановке скажет «Этого не может быть, потому что этого не может быть НИКОГДА». И в таком контексте эта знаменитая фраза из Письма ученому соседу А.П. Чехова не покажется смешной. Я готов на любом семинаре обосновать это НИКОГДА законами физики. Но это можно сделать и математически. Обратите внимание на обязательное условие практической выполнимости нормального закона – требуются БЕСКОНЕЧНОЕ число измерений и БЕСКОНЕЧНОЕ количество случайных посторонних влияний, чтобы реальный закон распределения сходился к нормальному. А мы знаем – если в физических рассуждениях и выводах появляется БЕСКОНЕЧНОСТЬ, то это надежная примета, что физик здесь чего-то не понимает. Математик имеет право вообразить себе бесконечность и исследовать ее проявления в математических объектах. Но мы ведь исследуем физическую реальность, природные объекты, где бесконечности не место.

Итак, нам следует понимать, что закон (5) является не законом Природы, а нашим модельным представлением о неизвестном нам законе. Им можно с удобством пользоваться, но с оглядкой на пределы его применимости. Мы так поступаем с теми эмпирическими законами, которые открываем и формулируем в естествознании. Почему закон (5) должен быть исключением? Если не помнить, что это наши модельные представления о поведении измеряемых величин, и относиться к нему сугубо серьезно, то возможны очень грубые ошибки интерпретации экспериментальных результатов. На такие собственные и чужие ошибки исследователи натыкались. Отсюда и возникло недовольство этим законом. Можно с уверенностью предсказать дальнейшее развитие этой неприятной ситуации: раз все вычислительные рецепты одномерной статистики выводятся из закона (5), то со временем возникнет недовольство и другими формулами математической статистики.

Здесь я должен сказать несколько слов в оправдание всей славной когорты специалистов, развивавших начальные идеи Гаусса и Чебышева. Дело в том, что математическая статистика не является наукой. Она даже не является частью математики. Сошлюсь на частное сообщение профессионального математика В.М. Имайкина, питомца школы Колмогорова: «Колмогоров, в конце концов, нашел математические основания теории вероятностей. Но даже он не смог найти математические основания статистики». Я не математик и не понимаю, что такое математические основания какого-то раздела математики. Однако я верю, что в этом заявлении есть что-то полезное для понимания наших взаимоотношений с математической статистикой. Так, я теперь понимаю, что виноваты не Гаусс и Чебышев, а мы, естествоиспытатели. Это видно из этапов развития математической статистики.

1. В геодезии возникла практическая трудная задача не допустить неконтролируемого накопления ошибок измерения, когда, двигаясь от репера к реперу, геодезисты добавляют одно к другому значения измененных и вычисленных расстояний. Потребовался гений Гаусса, чтобы угадать удачную форму одномерного закона распределения ошибок и придумать метод наименьших квадратов для равномерного размазывания всех ошибок по всем геодезическим реперам.
2. Взрыв производительности в сельском хозяйстве США поставил новые, уже многомерные статистические задачи. Был привлечен десант специалистов из Европы. Они решили эти задачи и оставили свои имена в истории. Очень скоро сходные задачи были осознаны в медицине. Были сформулированы стандарты обработки медицинских данных на основе методов, разработанных для сельского хозяйства. Этими стандартами сегодня пользуется весь мир.

3. Взрыв мировой банковской деятельности потребовал новых методов анализа очень непростых данных практической психологии. Пестрота этих данных навела на мысль об их статистической природе. В статистике появились имена людей, разработавших новые методы многомерной статистики. Оказалось, что эти методы могут быть реализованы только с использованием мощной вычислительной техники. Но к тому времени ЭВМ уже появились. А теперь уже хемометрия приспосабливает эти методы (факторный анализ, метод главных компонент) к проблемам обработки больших объемов экспериментальных данных.

Мы же, работники фундаментальных отраслей естествознания, не ставим перед специалистами таких острых задач. Следовательно, мы сами и виноваты в некоторых наших статистических неприятностях. Спасение утопающих – дело рук самих утопающих.

До сих пор говорил о серьезных проблемах, требующих нашего внимания. Теперь скажу не для публикации, а для сведений начальства. Предполагается, что методы одномерной статистики настолько давно и хорошо известны, что не нужно поднимать вопрос о статистической культуре наших коллег. К сожалению, это не так. Анализ многочисленных защит на нашем Аналитическом докторском совете показывает, что с этой культурой дело обстоит не столь благополучно. Все-таки в аспирантских работах иногда рассматриваются результаты измерений, полученные в разных повторностях и в разных группах опытов. Такие результаты полагается обрабатывать с привлечением дисперсионного анализа, который выясняет, меняется ли вариабельность результатов от группы к группе. За много лет я увидел один единственный пример использования схемы дисперсионного анализа. Зато очень часто можно увидеть такие графики, где прямая проводится по трем экспериментальным точкам. Я могу подробно доказать, что это ошибка и объяснить, как надо делать. Но достаточно сказать известную вещь. Через три точки можно не только провести прямую и увидеть отклонения от предполагаемого линейного закона. Через три точки с еще большим успехом можно провести кривую второго порядка. И никаких отклонений не будет.

Кто виноват, не знаю. Что делать, понятно – нужен практический статистический ликбез для аспирантов вместо какого-то другого кандидатского экзамена. Причем, это должен быть курс ликвидации непонимания смысла самых начальных статистических идей и рецептов. Примерно, как я проанализировал ситуацию с нормальным законом распределения ошибок. Как это делать, должно знать начальство. Но если это будет сделано, авторы работ станут грамотно использовать статистические рецепты, упакованные в сервисные компьютерные программы. И тогда мы перестанем получать статистические нарекания от оппонентов. К сожалению, такие нарекания уже поступают.

Многомерная статистика

В этом разделе пойдет речь о тех методах, которые могут оказаться полезными в исследовательской, а не производственной аналитической химии. Затем в специальном разделе я предложу, как можно реализовать эти непростые методы в нашей деятельности.

Простейший случай многомерной статистики известен почти всем, кто проводит измерения двух величин одновременно с целью выяснения, как одна величина зависит от другой. Такую зависимость, когда она уже выявлена, лучше всего выражать какой-то аналитической функцией, параметры которой определены как раз из проведенных измерений. Однако требуется пройти долгий путь, если такие измерения проводятся впервые. Из-за того, что обе величины измеряются

с ошибками, функциональная зависимость затемнена. Прежде всего, надо выяснить, есть ли какая-то зависимость между двумя величинами, насколько она тесная. Делается простейшее предположение, что зависимость похожа на линейную. Это всегда может быть правдоподобно, если обе величины изменяются в небольших пределах. Для оценки степени близости зависимости к линейной принято подсчитывать коэффициент корреляции. Такая характеристика изучаемого явления есть параметр двумерной плотности распределения одновременного появления двух измеряемых величин. Сосредоточимся на этом простейшем случае многомерной статистики, поскольку он позволяет достаточно ясно рассмотреть все преимущества и все трудности общего случая одновременного измерения многих величин.

Напомню, как возникает идея коэффициента корреляции между двумя случайными величинами *Y* и *X*. Пусть эти величины подчиняются каждая порознь своему нормальному распределению с параметрами μ_Y, σ_Y и μ_x, σ_x. Тогда совместное распределение этих величин, то есть плотность вероятности обнаружить одновременно величину *Y* в окрестности конкретного значения *y* и величину *X* в окрестности конкретного значения *x*, определяется двумерным нормальным распределением

$$w(y,x)=\frac{1}{2\pi\sigma_Y\sigma_X\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(y-\mu_Y)^2}{\sigma_Y^2}+\frac{(x-\mu_X)^2}{\sigma_X^2}-2\rho\frac{(y-\mu_Y)(x-\mu_X)}{\sigma_Y\sigma_X}\right]\right). \quad (6)$$

Теоретически зависимость между величинами *Y* и *X* характеризуется величиной ковариации cov(YX), являющейся аналогом дисперсии. Это математическое ожидание произведения $(Y-\mu_Y)(X-\mu_X)$. Коэффициент корреляции $\rho=\frac{\mathrm{cov}(YX)}{\sigma_Y\sigma_X}$ также характеризует связь между величинами *Y* и *X*, но на практике более удобен, чем ковариация, поскольку ρ есть безразмерная величина.

Устройство громоздкой формулы (6) легко понять, разобрав частные случаи степени зависимости между *Y* и *X*. Пусть они статистически независимы, то есть ρ = 0. Тогда правая часть (6) превращается в произведение двух нормальных плотностей.

Пусть *Y* и *X* связаны точной линейной зависимостью $Y=aX+b$. Тогда $\rho=\pm1$, и в правой части (6) появляется неопределенность типа $\infty\cdot0$. А мы знаем – как только в рассуждениях появилась бесконечность, значит кто-то чего-то не понимает. Мы тут не поняли вот чего. Раз *Y* и *X* связаны точной функциональной зависимостью, то нечего говорить о вероятностях и исследовать структуру формулы (6). Всё же спасибо математике, которая дает нам в трудных случаях понимание степени нашего непонимания.

Если $(-1<\rho<1)\ \&\ \rho\neq0$, то между *Y* и *X* имеется линейная регрессионная зависимость. Мы имеем счастливую возможность написать уравнение регрессии

$$y-\mu_Y=\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X). \quad (7)$$

К этому обычно и стремится каждый нормальный исследователь.

Итак, мы видим, что коэффициент корреляции есть полезнейшая вещь. Однако нас опять подстерегают неприятности. На этот раз и вовсе крупные. Сравните конструкцию (6) с простой

конструкцией (5), учитывая еще конструкцию для ρ. Ясно, что нас ждут еще большие неприятности, чем описанные в разделе об одномерной статистике. Как найти на практике величину *r*, являющуюся эмпирической оценкой теоретической величины ρ? Это не так сложно. Поступим так, как и в одномерной статистике – заменим в определении ρ все теоретические величины их эмпирическими оценками. Но каковы статистические свойства величины *r*? Вот тут беда. Свойства крайне скверные.

1. Значение *r* приближается к ρ лишь при огромном наборе пар измерений y_i, x_i. При этом оно всегда остается меньшим, чем ρ.
2. Дисперсия *r* зависит не только от числа измеренных пар, но и от значения самого коэффициента корреляции.

Последнее особенно неприятно, когда теоретическое значение $\rho \approx 0$. Тогда эмпирическое значение *r* может получиться практически каким угодно даже при большом числе *n*. Это видно из приближенной формулы

$$\sigma_r \approx \frac{1-\rho^2}{\sqrt{n}}. \qquad (8)$$

Мы уже подготовлены к тому, что теоретические дисперсии всегда меньше их эмпирических оценок. Следовательно, доверительный интервал для *r*, подсчитанный с помощью (8), будет сильно заужен. Пусть откуда-то известно, что в конкретных условиях ρ = 0. Сделаем три измерения пар y_i, x_i. Подсчитаем значение *r*. С высокой надежностью мы можем ожидать, что r окажется любым в пределах от – 0.58 до 0.58. Скорее всего, эти пределы могут быть еще более широкими. А мы склонны считать коэффициент корреляции *r* = 0.6 вполне приличным. Получается неприлично.

Всё, что сказано выше о неприятностях при нахождении коэффициента корреляции, относится к гипотетическому случаю следования двух величин *Y* и *X* двумерному нормальному закону распределения. Но если мы знаем, что в одномерных измерениях мы имеем мало шансов наблюдать проявления закона (5), то каковы шансы наблюдать в двумерных измерениях проявления закона (6)? Следует грустный вывод – раз все сложные формулы математической статистики выводятся только из закона (5), то на практике статистика не может нам ничего сказать о свойствах измеряемых нами коэффициентов корреляции. Не будем судить, кто виноват. А что и как нам делать, я объясню в специальном разделе этой работы.

Теперь о принципиальных методологических трудностях, подстерегающих нас при поиске корреляционных зависимостей между экспериментальными данными. Они возникают, когда мы одновременно измеряем множество величин, между которыми могут проявляться взаимные влияния. Самый распространенный случай это калибровка прибора, измеряющего концентрации компонент смеси веществ по спектральным признакам. Одни вещества здесь склонны подавлять спектральные проявления других веществ. И ясно, что поиском парных корреляций, описанным выше, здесь не обойтись. Имеют место множественные корреляции. Множественные корреляции мы просто не умеем измерять. Но нам крупно повезло. Многомерная математическая статистика умеет описывать множественные корреляции, опираясь на многомерный нормальный закон распределения совместного проявления случайных величин X. Здесь X = [$X1$, $X2$, ...X_N] есть матрица-строка, а ее элементы – случайные величины. Я не буду выписывать формулу для многомерного закона распределения. Она является обобщением (6) на *N*-мерный случай.

Разница в том, что в этом случае для описания связей между случайными величинами пользуются не коэффициентами корреляции, а парными ковариациями. В выражение этого закона входит квадратная матрица C, недиагональные элементы которой являются парными ковариациями $\mathrm{cov}(X_i X_j)$. Диагональные элементы $\mathrm{cov}(X_i X_i) = \sigma_i^2$ являются одномерными дисперсиями.

Матрица C является главным объектом исследования многомерной статистики. Некоторые из этих методов исследования и могут заинтересовать нас профессионально. Ради этого и написана данная работа.

Нас может и должен заинтересовать метод главных компонент. Исследуется матрица C или ее эмпирическая оценка по данным эксперимента. Ее подвергают диагонализации, то есть находят ее собственные числа eig(C) и собственные векторы. Как это делается – не наша забота, поскольку это стандартная операция компьютерных статистических пакетов программ. Нам важно, что собственные числа сортируют по убыванию и соответственно переставляют собственные векторы. Формально это есть переход от случайных величин X к другим случайным величинам Y. Переход совершается как раз с помощью матрицы собственных векторов. Каждая новая случайная величина в матрице Y является линейной комбинацией всех величин из X с коэффициентами, равными элементам подходящего столбца матрицы собственных векторов.

Оказывается, что разглядывать результат такого преобразования чрезвычайно полезно. Собственные числа матрицы C, то есть числа, стоящие на диагонали матрицы eig(C), являются одномерными дисперсиями новых, уже статистически независимых случайных величин Y. Матрица C обычно получается очень запутанной и трудно читаемой. А на диагонали eig(C) сразу видно, что новые величины Y обладают весьма различными статистическими свойствами. Первыми стоят крупные дисперсии. Это значит, что данные линейные комбинации исходных случайных величин подвергаются в данном эксперименте наибольшим колебаниям, вариациям. Они оказывают друг на друга наибольшие влияния. Такие комбинации и называются главными компонентами. Смотрим на столбец собственного вектора самой первой величины из Y. Вполне может оказаться, что в этом столбце некоторые элементы равны нулю или близки к нулю. Это значит, что соответствующие элементы матрицы X не оказывают влияния на вариабельность этой новой величины Y_1. Тем самым выясняется, между какими *X* существует наиболее сильное взаимное влияние. На них-то и надо обращать пристальное внимание при разработке аналитической методики.

Перебирая убывающие дисперсии в eig(C), дойдем, возможно, до нулевых или почти нулевых величин. Это значит, что некоторые линейные комбинации исходных величин вообще не подвержены вариациям, ведут себя как константы. Ну и очень хорошо. На них при разработке аналитической методики вообще не стоит опираться. Случается замечательно приятная вещь – сокращение числа случайных величин в данном эксперименте, сокращение мерности того пространства, в котором исследователь вынужден чувствовать себя скверно.

Вот здесь у меня конкретное предложение – что делать нам вместе. Я не уверен, что аналитикам, погруженным в фундаментальные проблемы своей родной науки, так уж просто освоить и применить в исследовательской рутине описанный пунктиром метод главных компонент. Я не уверен, что мы, физики, готовы дать коллегам-аналитикам конкретные методологические и методические советы, связанные со статистикой. Но я уверен, что было бы полезно затеять совместные работы, в которых мы нащупали бы, какие из известных нам методов математической

статистики адекватны исследовательской рутине аналитиков. Более того, какие методы могут оказаться палочкой-выручалочкой при решении проблем будущей аналитической химии.

Впрочем, такие поиски уже начались сравнительно давно [7-9]. Видны попытки математиков принять участие в этом процессе [10]. Есть работа, прямо иллюстрирующая смысл моего предложения [11].

Как преодолеть все замеченные статистические трудности

Я вижу единственный доступный нам способ. Опираться не на модельные представления о законах распределения наших экспериментальных данных. Как мы видели, это привычно, но очень вредно. Опираться не на вывод формул для вычисления статистических свойств наших сложных функций от этих данных. Это безумно трудно и очень вредно. А будем опираться на сами эти данные. И на те статистические свойства, которые они сами проявляют в данном конкретном эксперименте. Если же нам потребуются какие-то функции от наших данных, то будем вычислять эти функции, одновременно исследуя их статистические свойства.

Мы выяснили, что будем делать. Теперь я расскажу, как мы будем это делать. А если что-то сделаем не так, то сами и будем виноваты.

Основная идея прекрасно согласуется с методологией естествознания. Главное в естествознании это наблюдения. Обязательно многократные. Далее следуют размышления и попытки выразить результаты наблюдений количественно. Начнем опять со случая одномерной статистики. Несколько раз измерили одним прибором одну и ту же величину. Получили ряд несколько отличающихся чисел. Внимательно в них всмотрелись. Увидели, что они проявляют некоторые статистические свойства – похожесть и разброс. Наше понимание этих свойств надо выразить количественно и желательно в привычной для научной общественности форме. Пожалуйста – вычислим все привычные показатели так называемой репрезентативной статистики: среднее значение, среднее квадратичное отклонение. На это есть привычные определения и формулы. Они не зависят от закона распределения измеренной случайной величины. От этого закона зависят различные рецепты построения доверительного интервала для среднего значения измеренной величины. Но ведь полученный нами ряд чисел как-то проявляет присущий ему закон распределения. Пусть проявляет не очень четко.

Попробуем, не затрачивая значительного труда, имитировать поведение результатов измерения. Помня, что эти результаты случайны. Возьмем на заметку первое же попавшееся число из нашей экспериментальной выборки чисел. Могло оно не попасться нам в процессе измерений? Вполне могло. А соседнее число могло выпасть два раза подряд? Вполне могло. Совершенно законные предположения. Воспользуемся этим и приготовим новую выборку, убрав из экспериментальной выборки первое число и продублировав второе. Такая новая выборка вполне могла бы реализоваться в нашем измерительном процессе. Вычислим еще раз среднее значение измеренной величины. Оно будет несколько отличаться от первого. Очень хорошо – среднее значение начинает само по себе показывать нам свои статистические свойства.

Теперь ясно, как мы будем поступать дальше. Будем многократно повторять этот имитационный процесс, каждый раз назначая номер выбрасываемого числа из исходной выборки с помощью генератора случайных чисел. Каждый раз из новой несколько искусственной выборки будем вычислять среднее. Эти выборки только слегка искусственные. В основном, они несут все черты нашей единственной экспериментальной выборки. Объем *n* каждой выборки тот же самый,

состав чисел в выборке почти тот же самый, поскольку мы особо большого насилия над выборкой не совершили. Мы случайно вырезали одно число, зато добавили другое из той же компании (поэтому такой метод имитации, когда каждый раз просто вырезают из оригинальной выборки одно случайное измерение, называется jack-knife, а я показал его в чуть модифицированном виде, чтобы не изменять n). С каждым ударом ножа мы вычисляем среднее значение из имитационной выборки. Так мы можем набрать очень большой объем $n_{j\text{-}k}$ средних значений. Построим гистограмму этих значений. Она нам многое расскажет о статистическом законе, которому подчиняется уже не измеряемая величина, а ее среднее значение. А оно-то нам как раз и нужно. Ведь закон больших чисел Чебышева никто не отменял. Его проявления не зависят от закона распределения случайной величины. Он утверждает только, что с ростом $n_{j\text{-}k}$ среднее значение из средних приближается к математическому ожиданию случайной величины. Вычислив это среднее из всех средних, мы получим что-то близкое к математическому ожиданию среднего. Насколько близкое? Это мы сразу увидим из гистограммы средних. Задавшись надежностью Q, мы непосредственно по крыльям гистограммы определим границы соответствующего доверительного интервала. Готов отчет для научной общественности. Вот доверительный интервал. Вот надежность того, что неизвестное математическое ожидание результата измерения попадает в этот интервал. Оказалось, что даже не требуется вычислять эмпирическую оценку дисперсии.

Современному компьютерно продвинутому исследователю ясно, что я описал четкий алгоритм действий. Следовательно, нечего вообще затрачивать собственный труд, воспроизводя процесс jack-knife. Можно написать раз навсегда простенькую программу и поручить все это компьютеру. А скорее всего, эту программу можно найти уже в готовом виде в каком-то статистическом пакете. Сам я никогда не пользуюсь чужими программами, поэтому не могу посоветовать, где искать такую программу. Еще надо заметить, что я предпочитаю метод бутстреп (boot-strap), а не jack-knife. Бутстреп более интенсивно имитирует случайные процессы измерений, также создавая имитационные выборки, а потому работает быстрее. Известно, что готовая программа находится в статистическом tool-box MatLab. Однако я всегда пишу свою программу, учитывая особенности исходной выборки и вычисляемой функции.

Ясно, что таким же образом можно получить доверительный интервал и его надежность для любой другой функции от экспериментальной выборки, а не только для среднего значения. А мой опыт показывает – чем сложнее функция от результатов измерений, тем эффективнее работают имитационные методы статистики, основанные на интенсивном применении вычислительной техники [12].

В качестве иллюстрации бутстрепа как палочки-выручалочки, приведу фрагменты из работы [12]. К сожалению, у меня нет опыта совместной работы с химиками-аналитиками, поэтому я здесь воспользуюсь опытом работы с медиками.

… покажем, насколько различными могут быть интерпретации результатов, получаемых методами классической статистики и методом бутстрепа. Для этого воспользуемся данными биохимического анализа, известными нам из частного сообщения коллег. В нескольких возрастных группах пациентов в ходе обследований была проанализирована активность A некого фермента. Необходимо было выяснить, зависит ли эта активность от возраста. Все группы были малочисленными (не более 15 пациентов), но особенно малочисленной была некая возрастная группа из 6 пациентов. В этой же группе активность фермента в анализах оказалась самой низкой, она представлена следующим набором значений:

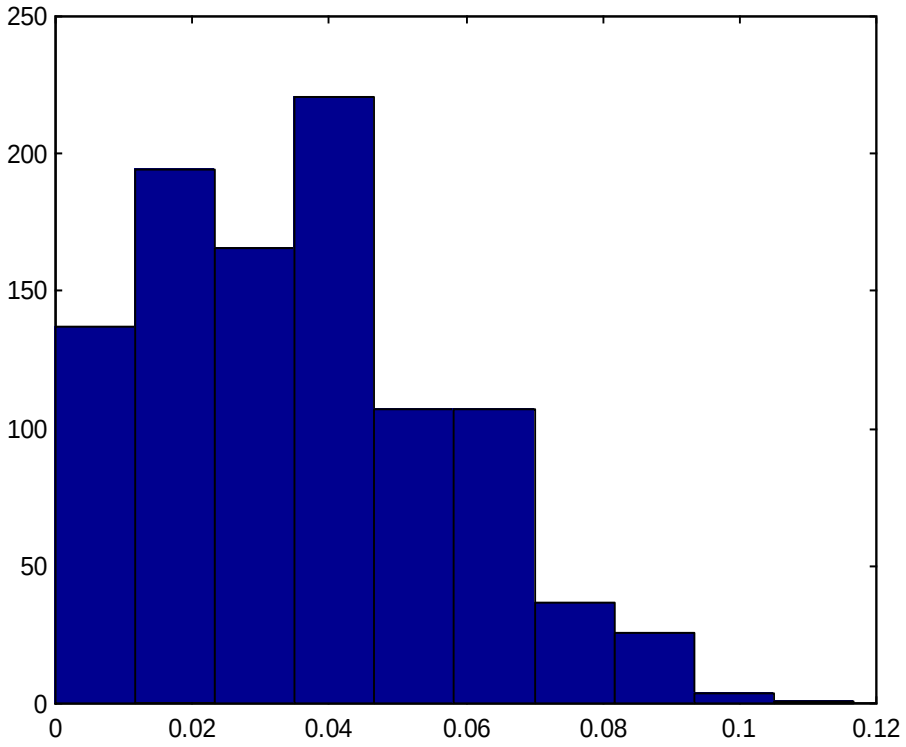
a_e = {0,0 0,0 0,0 0,04 0,14 0,0}.

Для того, чтобы сравнить показатели этой группы с другими, необходимо найти доверительный интервал, в который попадает среднее значение случайной величины *A* в данной группе. Потребуем, чтобы надёжность при этом составляла 95 %.

Рецепты классической статистики дают следующие результаты. Среднее значение a_{me} составляет 0,03; стандартное отклонение среднего составляет 0,025; коэффициент Стьюдента для надёжности 0,95 и числа степеней свободы 5 составляет 2,57. Отсюда получаем вывод:

среднее значение *A* с надёжностью 95 % лежит в интервале от -0,035 до +0,095. Это абсурдный вывод, поскольку биохимический смысл случайной величины *A* исключает возможность принимать отрицательные значения.

Бутстреп приводит к другому выводу. В выборке средних значений величины a_{mb} нет и не может быть отрицательных значений. Нулевые есть, поскольку вполне вероятно формирование таких случайных выборок бутстрепа, куда входят шесть нулевых экспериментальных значений величины *A*. Это хорошо видно на гистограмме, построенной из 1000 выборок бутстрепа.



Отбросим на гистограмме a_{mb} пятипроцентный хвост со стороны высоких значений и получим вывод:

среднее значение *A* с надёжностью 95 % лежит в интервале от 0,0 до +0,07. Это вполне приемлемый для специалиста вывод. Приятно и то, что ширина доверительного интервала по бутстрепу оказывается меньше, чем по классическому рецепту. Это позволяет надёжно зафиксировать меньшее различие между активностями фермента в двух близких возрастных группах.

Поясним, почему классические рецепты привели к неприемлемому результату. Дело в том, что в этом случае исследователь должен опираться на гипотезу - величины a_e следуют нормальному распределению с выборочными оценками: для математического ожидания $A_m = 0{,}03$; для стандартного отклонения $\sigma = 0{,}056$. Это значит, что с вероятностью 95 % в данной группе пациентов могут быть обнаружены значения A_e от -0,082 до +0,14. Мы видим, что по данному критерию мы не можем отбросить ни одно из экспериментальных значений активности фермента. Следовательно, принимая данную гипотезу, исследователь вносит в свои результаты дезинформацию. Поэтому он и получает артефакты в конечных результатах и вынужден дать им неверную интерпретацию.

Как в таких случаях поступает специалист? Скорее всего, он руководствуется своим богатым опытом, говоря - "Одномерная математическая статистика не способна видеть в моём экспериментальном материале ничего такого, чего я не видел бы невооруженным глазом. Дело статистики - лишь выработать численные оценки параметров для моей интерпретации наблюдаемых фактов. Если классическая статистика, намертво привязанная к закону нормального распределения, даёт мне абсурдную оценку нужных мне параметров, то тем хуже для этой статистики. Я буду действовать иначе и подберу другие рецепты, которые дадут мне разумные оценки".

Бутстреп как раз и является другим рецептом. Он не привязан не только к закону нормального распределения, но и ни к какому заранее предполагаемому распределению исследуемой случайной величины. Бутстреп пытается прощупать распределение этой случайной величины непосредственно в ходе вычислений. Он осматривает первичный материал настолько тщательно и подробно, что буквально каждый отдельный результат из экспериментальной выборки даёт свой весомый вклад в формирование конечного результата и его интерпретации. Бутстреп буквально выпячивает все статистические особенности первичного материала, а классическая статистика эти особенности сглаживает.

Проиллюстрируем сказанное ещё одним примером. Приводимый ниже материал возник при обработке анкетных данных группы пациентов, которым предлагалось субъективно оценить своё качество жизни Q перед прохождением лечения. В группе из 17 человек получились следующие суммарные оценки q_e.

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 (6) в виде отдельной строки для наглядности.

Мы хотим выяснить, что произойдёт с показателем Q в данной группе после курса лечения. Сравнивать мы будем доверительные интервалы для среднего значения Q в двух выборках, полученных в разные моменты времени. Сейчас сформируем доверительный интервал для приведенной выборки.

С помощью классической статистики получаем следующие результаты. Среднее значение q_{me} составляет 83,7; стандартное отклонение среднего составляет 6,1; коэффициент Стьюдента для

надёжности 0,95 и числа степеней свободы 16 составляет 2,12. Отсюда получаем вывод, опирающийся на предположении о применимости нормального распределения:

среднее значение *Q* с надёжностью 95 % лежит в интервале от 70,9 **до** 96,6. Как будто, всё в порядке. Но смущает наличие в выборке значения $q_{e7} = 0$. Согласуется ли это значение с нормальным распределением? Применим к этому значению критерий 3σ. Выборочное значение σ составляет 24,3. Это значит, что с вероятностью почти 100 % не может быть значений, меньших 10,8. Но q_{e7} существенно меньше этого предельного значения. И рецепты классической статистики требуют, чтобы мы отбросили это наблюдение как явно ошибочное, и обработали новый вектор

q_e =

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

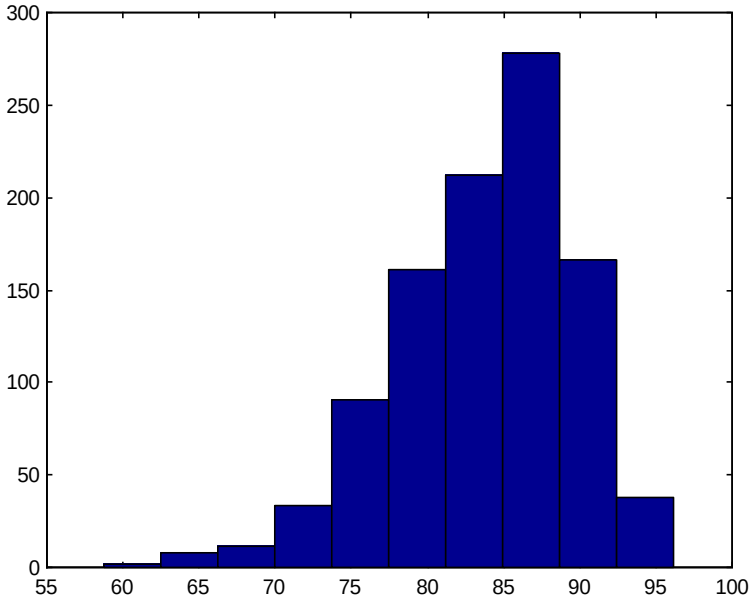
88,8 78,01 100 100 77,67 100 89,21 100 65,47 78,64}

Для этой выборки получим следующие результаты. Среднее значение q_{me} составляет 88,9; стандартное отклонение среднего составляет 3; коэффициент Стьюдента для надёжности 0,95 и числа степеней свободы 15 составляет 2,13. Отсюда получаем вывод, опирающийся на предположении о применимости к урезанной выборке нормального распределения:

среднее значение *Q* с надёжностью 95 % лежит в интервале от 82.7 **до** 95,3. Видно, что применение критерия 3σ существенно повлияло на результат. Показатель качества жизни заметно сдвинулся вверх, а доверительный интервал очень заметно сузился. Последнее увеличивает шансы исследователя заметить влияние проведенного лечения на изменение данного показателя в данной группе. Остаётся только предвкушать успехи выбранной терапии.

Однако зададимся вопросом - а насколько обосновано применения критерия 3σ. Ровно настолько, насколько обосновано предположение, что приведенные выборки по своей природе подчиняются нормальному распределению. Решать этот вопрос статистическими методами нерационально. Лучше спросить специалиста, какие свойства его пациентов могли привести к появлению нулевой субъективной оценки качества жизни. Или такая оценка могла появиться в результате неумения пациента обращаться с анкетами. Тут и выяснится, что среди пациентов нередко встречаются индивиды, склонные сильно преуменьшать свою оценку качества жизни перед курсом лечения. Если специалист хочет провести анкетирование с учетом такой возможности, то нет никаких резонов применения критерия 3σ, поскольку предварительная фильтрация в данном исследовании может быть проведена только на содержательном, а не на формальном уровне.

Выполним обработку исходной выборки по методу бутстрепа. Мы получим следующую гистограмму 1000 средних значений для этой выборки.



Обращает на себя внимание асимметрия гистограммы. Похоже, что большая выборка средних значений не подчиняется нормальному закону распределения. Это и не удивительно, поскольку исходная экспериментальная выборка явно не подчиняется этому закону, если мы интересуемся и пациентами, склонными занижать свои показатели. Мы не знаем, какой закон управляет распределением наших исходных данных. Мы даже не пытались его узнать. Но мы убеждаемся, что бутстреп разумно нащупывает следствия этого закона в распределении такой функции исходных данных, как их среднее. И ничто нам не мешает непосредственно на гистограмме отметить такой хвост низких средних показателей, число которых составляет 5 % от всей 1000 значений, сформированных бутстрепом. Отсюда мы получаем доверительный интервал для среднего показателя качества жизни в данной группе - от 74 до 96. Среднее из значений составляет 83,7.

Мы видим, что бутстреп даёт результаты, очень близкие к классическим. Разница в том, что классическая статистика понуждает нас заменить исходную выборку на урезанную, а мы не хотели бы этого делать.

В анализе двух вышеприведенных примеров ясно прослеживаются особенности метода бутстрепа, наталкивающие нас на два возможных обобщения, которые не были замечены авторами этого метода.

1. Бутстреп позволяет вычислять различные функции от экспериментальной выборки и непосредственно оценивать статистические свойства этих функций, не опираясь ни на какие законы распределения как аргументов, так и самих функций.
2. Бутстреп позволяет оценивать параметры тех неизвестных и сложных законов распределения, которым следуют результаты наблюдений за сложным поведением природных объектов.

Конец фрагмента работы [12]. В литературе уже можно найти примеры использования бутстрепа для проведения сложной калибровки в аналитической химии [13].

Этот раздел я закончу замечанием, что многомерный случай не вносит ничего нового в особенности применения имитационных методов статистики на основе применения вычислительной техники. Усложняется лишь вычисление функций от исходных выборок. Так, при

исследовании парных корреляций в выборке находятся пары результатов измерений. А на каждом обороте цикла бутстрепа здесь надо вычислять значение коэффициента корреляции. Затем надо строить гистограмму значений коэффициента. Компьютерная программа получается немногим сложней, чем в приведенных выше примерах. При исследовании множественных корреляций в выборке находятся векторы результатов измерений. А на каждом обороте цикла бутстрепа здесь надо вычислять собственные числа и собственные векторы случайной ковариационной матрицы. Тоже несложно. Даже скучно как-то.

Универсальная процедура оценки погрешности теоретического прогноза, используемого в аналитической методике

В этом разделе я вдруг вспоминаю, что мы, лаборатория молекулярного моделирования и спектроскопии ГЕОХИ, все же статистически причастны к исследовательским работам химиков-аналитиков. Мы прогнозировали спектральные проявления органических молекул, а химики находили с нашей помощью надежные аналитические сигналы в этих спектрах. Л.А. Грибов разрабатывает метод безэталонного спектрального анализа смесей органических соединений. Этот метод принципиально основан на использовании теоретических спектров сложных органических молекул. И без статистической оценки точности прогноза спектра смеси метод работать не может. Между тем, до сих пор в научной литературе не только не решен вопрос о статистических свойствах теоретического прогноза сложной многомерной функции от измеряемых или литературных значений параметров. Такой вопрос даже не ставился. Понятно почему, если учесть принципиальные трудности математической статистики, описанные в первых разделах данной работы.

Я разработал универсальный алгоритм, позволяющий для любой формы теоретического прогноза сложной функции сложного набора параметров найти доверительные интервалы для любой из выходных величин такой функции. Этот алгоритм основан на имитационной статистике, описанной в предыдущем разделе. Он опробован на материале молекулярного моделирования для спектроскопических приложений. Но поскольку он совершенно универсален, то им можно заинтересовать тех аналитиков, которые при разработке своих методик базируются на теоретическом прогнозе какой-то сложной характеристики сложного же явления. Это может быть прогнозирование хроматограммы в программах А.М. Долгоносова; прогнозирование поведения сложной термодинамической системы в работах Р.Х. Хамизова; прогнозирование реакции стали сложного состава на радиационное воздействие в работах В.П. Колотова. Однако я до сих пор не опубликовал этот алгоритм в периодике. Есть полный текст моей статьи на сайте постоянно действующей Интернет-конференции IVTN [14].

Пользуясь случаем, я хочу спросить совета у любезных читателей этого эссе, членов редколлегии журнала *Аналитическая химия* М.К. Беклемишева и Л.А. Грибова, не опубликовать ли описание этого алгоритма в журнале. Я могу дать краткое его описание здесь, когда эссе будет переработано в статью. А могу подать отдельную статью. Она готова, только была написана от имени Грибова и моего. Однако Л.А. был недоволен постановкой задачи, и я не стал тогда подавать статью в редакцию.

Конечно, теперь идея этого алгоритма совершенно ясна, она следует из описания имитационной статистики в предыдущем разделе. Я просто применил эту статистику к случаю очень сложного прогнозирования многомерной функции от многих измеряемых и литературных параметров. Но в

подготовленной статье я провел подробное исследование поведения результатов работы этого алгоритма. Они полезны для понимания, нужен ли кому этот алгоритм.

Радиометрия – раздел аналитической химии, который никак не может обойтись без математической статистики

Особенно это касается радиометрии малых активностей. Но там давно все выяснено. Странно только, что после моей давней книги [5] ничего существенно нового не появлялось. Хотя поиски чего-то нового ведутся. И совершаются ошибки, связанные с недопониманием сути тех специфических законов распределения, что царят в этой области. Например, приходилось слышать, что раз отсчеты от источника препарат+фон распределены по закону Пуассона, то и разность препарат+фон – фон также распределена по закону Пуассона. Но разность препарат+фон – фон в случае малых активностей может быть случайно отрицательной величиной. А закон Пуассона, закон распределения редких событий, никак не может описывать поведение отрицательного числа редких событий. Например, регистрации в данном эксперименте минус двух отсчетов детектора ядерных частиц.

Выводы

1. Сложные рецепты одномерной математической статистики не очень нужны в исследовательской рутине химика-аналитика. Он понимает статистические свойства своего экспериментального материала лучше, чем любой специалист в области математической статистики.
2. Простые рецепты одномерной математической статистики очень нужны в исследовательской рутине химика-аналитика. Их использование требует очень глубокого понимания статистического смысла этих рецептов. Иначе возможны досадные ошибки при представлении результатов на суд научной общественности.
3. Рецепты многомерной статистики могут оказаться чрезвычайно полезными именно в исследовательской деятельности химика-аналитика. Поиск адекватных аналитической химии методов этой статистики может потребовать постановки совместных работ с участием как химиков, так и специалистов в области математической статистики.
4. Методы имитационной статистики, базирующиеся на интенсивном использовании вычислительной техники, способны спасти исследователя от любых трудностей, связанных с принципиальными методологическими недостатками самой математической статистики.

Литература

1. Brown, R.J.C., Goddard, S.L., Brown, A.S. Using principal component analysis to detect outliers in ambient air monitoring studies. *International Journal of Environmental Analytical Chemistry*. 2010, 90 (10), pp. 761-772.
2. Lozano, V.A., Ibañez, G.A., Olivieri, A.C. Experimental design for the study and optimization of the effect of different surfactants on the spectrophotometric determination of sulfide based on phenothiazine dye production. 2010, *Analytical Chemistry* 82 (11), pp. 4510-4519.
3. Hubert, Ph., Nguyen-Huu, J.-J., Boulanger, B., Chapuzet, E., Cohen, N., Compagnon, P.-A., Dewé, W., (...), Rozet, E. Harmonization of strategies for the validation of quantitative analytical procedures: A SFSTP proposal. Part IV. Examples of application. 2008, *Journal of Pharmaceutical and Biomedical Analysis* 48 (3), pp. 760-771.

4. Coleman, D., Vanatta, L. Statistics in analytical chemistry: Part 32-detection limits via 3-sigma. 2008, American Laboratory 40 (20), pp. 60-62.
5. Hopke, P.K. The evolution of chemometrics. 2003, *Analytica Chimica Acta* **500 (1-2), pp. 365-377.**
6. Дементьев В.А. Измерение малых активностей радиоактивных препаратов. 1967, М., Атомиздат.
7. Frank, I.E., Pungor, E., Veress, G.E. Statistical decision theory applied to analytical chemistry. Part 1. The statistical decision model and its relation to branches of mathematical statistics. 1970, *Analytical Chemistry* 42 (3), pp. 358-365
8. McFarren, E.F., Lishka, R.J., Parker, J.H. Criterion for judging acceptability of analytical methods. 1970, *Analytical Chemistry* 42 (3), pp. 358-365
9. Sarbu, C. Application of informational analysis of variance in analytical chemistry. 1993, *Analytica Chimica Acta* 271 (2), pp. 269-274
10. Faber, N.M., Buydens, L.M.C., Kateman, G. Aspects of pseudorank estimation methods based on an estimate of the size of the measurement error. 1994, Analytica Chimica Acta 296 (1), pp. 1-20
11. Rao, R., Lakshminarayanan, S. Variable interaction network based variable selection for multivariate calibration. 2007, *Analytica Chimica Acta* 599 (1), pp. 24-35
12. 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.
13. Jones, G., Wortberg, M., Kreissig, S.B., Hammock, B.D., Rocke, D.M. Application of the bootstrap to calibration experiments. 1996, *Analytical Chemistry* 68 (5), pp. 763-770
14. Дементьев В.А. Оценка погрешностей прогнозирования в безэталонном молекулярном анализе с помощью имитационных вычислительных процедур. http://www.ivtn.ru/2008/pdf/d08_08.pdf