УДК 004.522:004.424.23

ОПРЕДЕЛЕНИЕ ОБЪЕМА КОНТРОЛЬНОЙ ВЫБОРКИ В УСЛОВИЯХ АПРИОРНОЙ НЕОПРЕДЕЛЕННОСТИ ПО ПРИНЦИПУ ГАРАНТИРОВАННОГО РЕЗУЛЬТАТА

В. В. САВЧЕНКО

Нижегородский государственный лингвистический университет

e-mail: svv@lunn.ru

Предложен новый подход к расчету объема выборки в условиях априорной неопределенности – по принципу гарантированного результата в отношении точности и надежности статистической оценки вероятности случайного события. Рассмотрены примеры его применения. Показано, что благодаря предложенному подходу в ряде актуальных случаев на практике объем выборки сокращается в несколько раз по сравнению с известными оценками.

Ключевые слова: теоретическая информатика, статистическая оценка, статистическая выборка, объем выборки, проблема малых выборок, проблема априорной неопределенности.

В связи с повсеместным распространением информационных технологий математические методы синтеза и анализа сложных систем все шире проникают в различные сферы человеческой деятельности. В наибольшей мере это относится к методам теории вероятностей и математической статистики, распространение которых особенно сильно возросло в последние годы как в области технического, так и гуманитарного знания. И в этой связи даже в теории явно обозначился определенный разрыв между потребностями исследователей в эффективном математическом аппарате, с одной стороны, и их ограниченными, часто интуитивными представлениями о его обоснованности и методике применения. Сказанное в полной мере относится к проблеме определения объема контрольной выборки наблюдений, которая на практике решается, как правило, путем заведомо (и многократно) завышенных оценок. И этим сильно ограничиваются возможности статистических методов в условиях малых выборок и априорной неопределенности, характерных, например, для большинства задач в области речевых технологий [1, 2]. Исследованию путей ее решения и посвящена настоящая статья.

Доминирующий подход к определению требуемого (по минимуму) объема выборки в математической статистике основан на расчете длины доверительного интервала значений $[\theta_1; \theta_2]$ контролируемого параметра распределения $f(x,\theta)$ при заданном уровне значимости $\alpha = 1 - p = 0{,}025...0{,}1$, или заданной доверительной вероятности $p = 0{,}9...0{,}975$[3]. Тем самым «по умолчанию» задачу сводят к статистической (интервальной) оценке $\hat{\theta}$ $(X_1, X_2, ... X_n)$ некоторого параметра θ анализируемой (наблюдаемой) случайной величины X по выборке $X_1, X_2, ... X_n$ фиксированного объема $n>1$. Например, это может быть неизвестное, в общем случае, математическое ожидание случайной величины $M(X)$. Его состоятельная точечная оценка вычисляется по формуле средней арифметической величины (САВ) $\overline{X} = n^{-1}\Sigma_{i=1}^{n} X_i$.

На практике [2] объем выборки n жестко ограничен сверху требованиями к условиям наблюдений. Немаловажную роль при этом играют и причины экономического характера. В результате потенциальный максимум n обычно не превышает значения в несколько сотен и даже десятков единиц. Но и в таких, не самых благоприятных для статистического анализа условиях, со ссылкой на центральную предельную теорему исследователями повсеместно используется нормальная или гауссовская аппроксимация статистической оценки математического ожидания $M(X)$ и, вслед за ней, классическое выражение для половины длины ее доверительного интервала

$$\Delta = z_p \sigma / \sqrt{n} \qquad (1)$$

в роли количественной характеристики точности оценки по конечной выборке наблюдений. Здесь σ – СКО (среднеквадратичное отклонение) случайной величины X по результатам ее повторных наблюдений, z_p – коэффициент надежности или «доверия», определяемый корнем уравнения $\Phi\left(z_p\right)=p$ с интегралом вероятности нормального закона [3] в левой части. Переписав (1) относительно величины n, получим общеизвестное выражение

$$n \geq n^* = \left(\frac{z_p}{\Delta}\right)^2 \sigma^2 \qquad (2)$$

для определения минимального объема выборки n^* в зависимости от заданных (допустимых) уровней погрешности Δ и значимости α оценки математического ожидания по формуле САВ.

Например, при $\sigma = 1$, $\alpha = 0{,}05$ (соответствующая доверительная вероятность равна p=0,95) и допустимой погрешности $\Delta = 0{,}05$, или 5% относительно СКО, по таблицам нормального распределения находим $z_{0,95} \approx 1{,}96$. И, следовательно, получаем $n^* \approx 1537$, или, после округления, 1600 единиц – это стандартный объем выборки при социологических исследованиях.

Проблема состоит в том, что требование $n \geq 1600$ далеко не всегда осуществимо на практике. Для ее ослабления исследователи упрощают первоначальную формулировку задачи и переходят в (1) к бинарной случайной величине $X = (1;\ 0)$, или к дихотомии, т.е. к статистическому эксперименту с двумя возможными исходами испытаний по схеме Бернулли: противоположными случайными событиями **А** и $\overline{\mathbf{A}}$. И в этом приеме нет ничего ограничительного: специалисты хорошо понимают подчиненную роль понятия «случайная величина» по отношению к «случайному событию» в теории вероятностей. При этом выражение (2) преобразуется к виду

$$n^* = \frac{z_p^2 P_A (1 - P_A)}{\Delta^2}, \qquad (3)$$

где P_A – вероятность события **А**. Идея здесь состоит в том, чтобы радикальным образом ограничить дисперсию вариаций σ^2 из выражения (2). Нетрудно понять, что в варианте (3) дисперсия ограничена сверху на уровне 0,25. А достигаемый эффект иллюстрируется следующим примером. При той же, что и выше, доверительной вероятности p=0,95 и той же допустимой погрешности $\Delta = 0{,}05$ оценка вероятности P_A по формуле относительной частоты (или частости) $\tilde{P}_A = m_A / n$ случайного события **А** требует всего $n^* \approx 384$ испытаний. Здесь m_A - частота появления события **А** в серии из n независимых наблюдений. Как видим, благодаря дихотомии требуемый объем наблюдений сократился примерно в 4 раза. И это далеко не предел, что подтверждается результатами проведенного далее исследования, в котором идея дихотомии получила свое дальнейшее развитие в задачах с априорной неопределенностью.

Перепишем выражение (1) в терминах относительной длины доверительного интервала

$$\delta = \Delta / M(X) = z_p \sigma / \left[\sqrt{n} M(X)\right] \qquad (4)$$

с целью получения гарантированного результата вне зависимости от истинного распределения случайной величины *X*. И при учете очевидного равенства $M(X)=P_{\text{A}}$ при дихотомии из выражения (2) получим

$$n^{*}=\frac{z_{p}^{2}(1-P_{\text{A}})}{\delta^{2}P_{\text{A}}}=z_{p}^{2}\Big/\left(K_{\text{A}}\delta^{2}\right), \qquad (5)$$

где $K_{\text{A}}=P_{\text{A}}\big/(1-P_{\text{A}})$ – коэффициент обусловленности случайного события **A**. Отметим, что в отличие от (3) в выражении (5) отражена естественная асимметрия результата вычислений объема выборки относительно вероятностей двух альтернативных исходов **A** и $\overline{\mathbf{A}}$ каждого отдельного испытания.

Следуя полученному выражению (5), в рамках предыдущего примера вычислений при равенствах $\delta=0{,}05$, $P_{\text{A}}=0{,}91$ и $K_{\text{A}}=10$ будем иметь $n^{*}\approx 154$, или в 2,5 раза меньше, чем на основе классического подхода с использованием выражения (3). А при уменьшении требований к точности оценки до $\delta=0{,}1\ldots0{,}15$ при том же коэффициенте обусловленности $K_{\text{A}}=10$ приходим к еще более радикальному сокращению требований к объему выборки: до $n^{*}\approx 38$ и ниже. Для сравнения, при тех же условиях известный подход дает согласно (3) существенно худший результат, а именно: $n^{*}\approx 96$. При этом особо отметим, что даже в предельном варианте полученный объем выборки $n^{*}\approx 38$ по-прежнему хорошо согласуется с условиями центральной предельной теоремы, положенной в основу выражения (4). А это, в свою очередь, подтверждает обоснованность принципа гарантированного результата в формулировке (5). При этом достигаемый эффект объясняется использованием дополнительной информации о степени обусловленности события **A.**

На первый взгляд, здесь возникает острый вопрос в отношении точности и обоснованности такого рода информации. Однако положение спасет простая логика рассуждений. Даже при полном отсутствии априорной информации об истинном значении коэффициента K_{A} можно использовать наше знание в отношении разновидности поставленной перед исследователем задачи, а также беспрецедентных особенностей зависимости $K_{\text{A}}(P_{\text{A}})$ по области ее определения: она плавно затухает до нуля слева от точки $P_{\text{A}}=0{,}9$ на оси абсцисс и, напротив, резко возрастает до бесконечности справа от нее. Для многих решаемых с использованием статистических методов задач [1-3], значение $P_{\text{A}}=0{,}9$ может рассматриваться в качестве порогового уровня при тестировании работы исследуемой информационной системы. Порогового том смысле, что по условиям задачи вероятность успеха P_{A} для эффективных технических решений в данной области исследований не может опускаться ниже уровня 0,9. Поэтому мы можем изначально, не обладая достоверной априорной информацией, переписать критерий (5) в его предельно упрощенном виде

$$n^{*}=0{,}1\,z_{p}^{2}\big/\delta^{2}, \qquad (6)$$

которым, тем не менее, гарантируется необходимый результат в отношении точности δ и достоверности p оценки вероятности P_{A} в условиях априорной неопределенности.

В самом деле, для систем с неизвестной истинной вероятностью успеха мы при условии $P_{\text{A}}\geq 0{,}9$ согласно выражению (6) будем иметь завышенную оценку объема выборки с гарантированно высокой эффективностью статистического анализа. В системах

же с относительно низкой вероятностью успеха $P_{\text{A}} < 0{,}9$, которые по определению не представляют собой практического интереса, требования к точности и надежности оценок их эффективности могут быть существенно понижены. Как видим, критерий (6) гарантирует необходимый результат в рабочем диапазоне значений вероятности успеха P_{A}.

Физическим объяснением достигнутого эффекта могут служить особенности ряда задач из практики статистического анализа данных. К ним, главным образом, относятся задачи проверки статистических гипотез с явной (по своему физическому смыслу) асимметрией в отношении степени априорной обусловленности тестируемых гипотез. Классический пример – цифровые системы связи, при применении которых вероятность безошибочного обнаружения сигнала редко опускается ниже уровня $P_{\text{A}} = 0{,}9$. В этом случае вероятность пропуска сигнала не превышает значения $P_{\overline{A}} = 0{,}1$.

Примером может служить метод фонетического декодирования слов [4] из области речевых технологий. Данный метод характеризуется повышенной точностью и надежностью среди своих аналогов за счет предусмотренной в нем автоматической настройки на голос диктора. Для его экспериментального тестирования в работе [5] использовалась выборка суммарным объемом 2000 слов, составленная из аудиозаписей 50 речевых команд диктора, т.е. по 40 реализаций на каждую команду – точно в соответствии с результатами предыдущих вычислений. При этом была достигнута вероятность безошибочного распознавания каждого слова из используемого словаря команд в диапазоне значений от 0,93 и выше. И эти данные были подтверждены десятью разными дикторами.

Нетрудно подсчитать, что суммарный объем экспериментального словаря составил в данном случае 20 тысяч слов, а трудоемкость его наполнения – из расчета минимальных затрат порядка 5 секунд на запись одной реализации речевой команды от каждого диктора – примерно 27,8 часа. Это большая, но вполне практически реализуемая работа силами небольшого исследовательского коллектива. Для сравнения, при применении известного выражения (3) в рамках классического подхода к статистическому анализу трудоемкость исследования того же объекта составила бы почти трое суток непрерывной работы. Более того, если учесть, что в ряде случаев, как, например, в той же работе [5], оцениваемая по выборке конечного объема n^* вероятность успеха находится в пределах $P_{\text{A}} = 0{,}95$ и выше, то в выражение (6) вместо множителя 0,1 следует подставить множитель 0,05. Это означает, что объем выборки сократится в данном случае до $n^* = 20$ и ниже, а трудоемкость статистического эксперимента – примерно до 14 часов в течение одного рабочего дня.

Полученный результат подтверждается следующими несложными вычислениями: при больших значениях вероятности успеха $P_{\text{A}} = 0{,}95...0{,}99$ в серии из 20 последовательных испытаний по схеме Бернулли вероятность появления более одного неуспеха $1 - 20 \cdot P_{\text{A}}^{19} \cdot (1 - P_{\text{A}}) - P_{\text{A}}^{20}$=0,02...0,09 весьма близка к нулю при том, что 1 неуспех в данной серии – это как раз гарантированная нами точность (на уровне δ =5%) статистической оценки вероятности P_{A}.

Таким образом, по результатам проведенного исследования предложен новый подход к определению объема контрольной выборки, рассчитанный на широкий класс задач проверки статистических гипотез в условиях априорной неопределенности. Благодаря предложенному подходу во многих случаях удается существенно понизить требования к организации и условиям статистического эксперимента и, тем самым, сделать эксперимент значительно более доступным и реализуемым силами небольших исследовательских коллективов.

Литература

1. Жиляков Е.Г., Белов С.П. Обнаружение звуков речи на фоне шумов // Научные ведомости БелГУ: Серия «История. Политология. Экономика. Информатика». 2012. Т. 22. № 7-1. С. 182-189.

2. Савченко В.В., Васильев Р.А. Анализ эмоционального состояния диктора по голосу на основе фонетического детектора лжи // Научные ведомости БелГУ: Серия «История. Политология. Экономика. Информатика». 2014. № 21 (192). Вып. 32/1. С. 186-195.

3. Мюллер П., Нойман П., Шторм Р. Таблицы по математической статистике: Пер. с нем. М.: Финансы и статистика, 1982. 278 с.

4. Савченко В.В., Савченко А.В. Разработка быстродействующих алгоритмов автоматического распознавания голосовых команд с регулируемой точностью и надежностью на основе принципов слоговой фонетики русского языка и метода фонетического декодирования слов в информационной метрике Кульбака-Лейблера // Тезисы докладов Всероссийской научно-технической конференции и выставки, посвященной итогам реализации федеральной целевой программы «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы». Москва, 24–26 сентября 2013 г. М.: Изд-во МИСиС, 2013. С.184–185.

5. Савченко В. В., Савченко А. В. Метод фонетического декодирования слов в информационной метрике Кульбака-Лейблера для систем автоматического анализа и распознавания речи с повышенным быстродействием // Информационно-управляющие системы. 2013.№2. С. 7-12.

THE DETERMINATION OF SAMPLE SIZE IN CONDITIONS OF A PRIORI UNCERTAINTY ON THE PRINCIPLE OF GUARANTEED RESULT

V. V. SAVCHENKO

Nizhny Novgorod state linguistic university

e-mail: svv@lunn.ru

A new approach to calculation of volume of sampling in the conditions of aprioristic uncertainty – by the principle of the guaranteed result concerning accuracy and reliability of a statistical estimate of probability of a casual event is offered. Examples of its application are reviewed. It is shown that thanks to the offered approach in a number of actual cases in practice the volume of sampling is reduced several times in comparison with known estimates.

Keywords: theoretical informatics, statistical estimate, statistical samples, volume of sampling, problem of small volume sampling, problem of aprioristic uncertainty.