Отзыв

официального оппонента о диссертации Симушкина Дмитрия Сергеевича «Статистические критерии с ограничениями на d-риски», представленной на соискание ученой степени кандидата физико-математических наук по специальности 01.01.05 – Теория вероятностей и математическая статистика

В диссертационной работе Симушкина Д. С. решаются некоторые задачи байесовской статистики, связанные с вычислением апостериорных d-рисков – понятия, введенного Л. Н. Большевым и представляющего собой условное математическое ожидание потерь при условии решающей функции.

Байесовская статистика долгие годы остается одной из наиболее бурно развивающихся областей математической статистики, а байесовский подход позволяет решить многие проблемы математически корректной интерпретации результатов статистических исследований. Одним из наиболее актуальных вопросов в байесовской статистике является выбор априорного распределения, оказывающий существенное влияние на результат статистического исследования. В рамках эмпирического байесовского подхода априорное распределение оценивается с использованием имеющейся (архивной) информации. Идея непараметрического оценивания априорного распределения была предложена в работе Роббинса (Robbins, 1955). Параметрический подход к задаче оценивания априорного распределения активно развивался в работах Эфрона и Морриса в 70-е годы XX века. Дополнительный импульс развитию эмпирического байесовского подхода дало бурное развитие биоинформатики и генетических исследований с огромным объемом статистической информации. Эмпирический байесовский подход используется в различных областях биоинформатики и генетических исследований, включая задачу интерпретации результатов множества статистических тестов (множественного тестирования).

Важную роль в задаче множественного тестирования играет контроль ошибочных решений. Наиболее часто контроль касается ошибок I рода. Популярным показателем контроля надежности результатов множественного тестирования является ожидаемая доля неверно принятых альтернатив (FDR). В работе Сторея (Storey, 2002) был введен похожий показатель контроля надежности результатов множественного тестирования pFDR, отличающийся от FDR только тем, что ожидаемая доля неверно принятых альтернатив вычисляется при условии, что хотя бы в одном из тестов была принята альтернатива. В этой и в следующих работах автора развивается байесовский подход к контролю pFDR. Интересно отметить, что в байесовской парадигме d-риск I рода в точности совпадает с показателем pFDR. Статистическая процедура, позволяющая контролировать d-риски называется d-гарантийной. Таким образом, d-гарантийная процедура представляет собой байесовский аналог процедуры множественного тестирования, контролирующей pFDR. Следует отдельно отметить, что развитие теории d-рисков происходило задолго до появления методов контроля pFDR в задаче множественного тестирования. Существенный вклад в развитие теории d-рисков внесли работы казанской школы теории вероятностей и математической статистики И. Н. Володина, А. А. Новикова, С. В. Симушкина.

В работе Симушкина Д. С. акцент делается на три параметрических модели, в зависимости от распределения параметра модели и наблюдения при каждом фиксированном значении параметра: нормально-нормальная (N-N), гамма-показательная (G-E) и бета-биномиальная (B-B). Решаются задачи нахождения объема выборки, необходимого для достижения требуемых ограничений на d-риски I и II рода, построения последовательных d-гарантийных процедур и оценивания априорного распределения по архивным данным. Полученные d-гарантийные процедуры применяются для решения задач контроля качества выпускаемой продукции и поиска генетических ассоциаций.

Материал диссертации разделен на две главы. В разделе 1.2 изучается задача определения необходимого объема выборки (НОВ) для достижения определенных ограничений на d-риски в задаче различения односторонних гипотез. В случае использования оптимального критерия в рамках N-N модели получена точная формулы НОВ для d-гарантийной процедуры с одинаковыми фиксированными ограничениями на d-риски I и II рода, а также показано, что при увеличении дисперсии априорного распределения d-гарантийной процедуры НОВ стремится к единице. Для всех трех моделей получена асимптотика НОВ d-гарантийной процедуры при пропорциональном ужесточении ограничений на d-риски и в схеме стягивающегося априори, т.е. в случае сходимости априорного распределения к вырожденному распределению, сконцентрированному в пограничной точке между гипотезами. Раздел 1.3 посвящен изучению последовательных (адаптивных) d-гарантийных процедур. Рассматривается d-гарантийная процедура первого перескока, основанная на выходе апостериорной вероятности нулевой (или альтернативной) гипотезы из полосы, и ее аналог с усечением по НОВ, а также d-гарантийный последовательный критерий, основанный на статистике вклада. Показана связь процедуры первого перескока с последовательным критерием отношения вероятностей Вальда. Для всех трех моделей построены процедуры первого перескока и на базе статистики вклада. Изучается свойство конечности математического ожидания момента остановки последовательной процедур первого перескока и на базе статистики вклада. Показано, что для N-N модели момент остановки процедуры первого перескока имеет конечное математическое ожидание при каждом фиксированном значении параметра, за исключением пограничной точки между гипотезами. При этом отмечается, что математическое ожидание момента остановки процедуры первого перескока в байесовском смысле скорее всего бесконечно. Кроме того, показана конечность момента остановки процедуры на базе статистики вклада. В разделе 1.4 рассмотрена задача оценивания априорного распределения для трех моделей по архивным наблюдениям. Для оценивания предлагается использовать методы максимального правдоподобия и моментов, а также комбинированный метод. Обсуждается задача выхода оценки из допустимой области значений параметра. Определенное внимание в данном разделе уделяется непараметрической оценке плотности априорного распределения в рамках задачи деконволюции. Применение d-гарантийных критериев к задачам контроля качества и в генетических исследованиях (GWAS) описано в главе 2. Отдельно можно отметить оригинальную постановку задачи изучения экспрессии генов при наличии трех конкурирующих гипотез.

Результаты работы гармонично сочетаются друг с другом в рамках единой концепции исследования. Работа Симушкина Д. С. содержит 11 теорем, 17 лемм

и одно предложение. Список литературы содержит 70 наименований. Формулировки основных результатов вынесены во введение, что позволяет предварительно оценить суть полученных результатов, не вникая в доказательства.

В диссертации Симушкина Д. С. выявлены некоторые недостатки, не оказывающие существенного влияния на качество работы. Считаю, что еще во введении (с. 7) при обсуждении вопроса контроля ошибки следовало бы отметить связь последовательного алгоритма Беньямини—Хохберга [Benjamini & Hochberg, 1995] и метода Симса [Simes, 1986]. В основной части диссертации (с. 105) выражено ошибочное суждение, что процедура Беньямини – Хохберга/Симса гарантирует контроль FWER. В работе Беньямини – Хохберга [Benjamini & Hochberg, 1995] утверждается, что данный метод позволяет контролировать лишь более слабый критерий ошибки множественного тестирования FDR. Здесь же можно было бы отметить, что метод Симса не является последовательным, в отличие от процедуры Беньямини – Хохберга. Помимо этого, уместно было бы сказать об ограничениях на применение метода Беньямини – Хохберга и о дальнейших результатах в данной области. Отметим также, что теорема 2.1 (с. 16/111; автореферат с. 13) содержит в себе противоречие, т.к. пункты (i) и (ii) противоречат пункту (iii). Небольшие недостатки можно выявить в формулировках некоторых утверждений и в терминологии. В частности, в формулировке теоремы 1.2 (с. 9) уместно было бы сказать об ограничениях на d-риски; пояснение в скобках (с. 21, строки 11-12) не соответствует действительности; в замечании 5 (с. 36) следовало бы более четко указать, о какой формуле идет речь; следовало бы напомнить, что представляет собой $g(\theta)$ в равенстве (1.26) (с. 39); в пределе (с. 53, строка 6) не указано, к чему стремится τ; утверждение о том, что момент остановки «обладает большой вероятностью остановки на первых шагах обследования» требует дополнительного пояснения (с. 62, строка 12); теорема 1.6 не является подтверждением бесконечности v_{un} в байесовском смысле, т.к. $\theta = 0$ – единичное значение (с. 68 строки 9-10); область значений t, для которых $K(t) = 1$, вероятно, должна быть ограниченной (с. 90, строка -11); следовало бы пояснить, по отношению к чему усеченная процедура первого перескока уменьшает объем выборки (с. 95); в определение модели деконволюции, вероятно, следовало бы добавить независимость θ и ε (с. 98); формально, в определении FDR (FNR) не подразумевается зависимость от числа справедливых нулевых гипотез, поскольку данное значение неизвестно (с. 105); логично было бы пояснить необходимость выбора генов случайным образом и почему бы не рассмотреть все гены (с. 107, строка 11); в разделе 2.2.2 анонсируется две модели (с. 110, строка -7), а затем обсуждается только одна; в доказательстве теоремы 2.1 (с. 112) утверждение (iii) не следует из (i) и (ii); выводы 2.4.1/2.5.1 (с. 114/115) уместно делать в контексте числа сигналов в исходных данных, а результаты уместно сравнивать методами Storey, а не с процедурой Беньямини – Хохберга; на мой взгляд, недостаточно ясно, что такое минимаксный критерий в байесовской парадигме (с. 116, замечание 15); под предложением (с. 124, предложение 2.1) обычно понимают достоверное утверждение, в данном случае уместно назвать данное утверждение гипотезой. Имеются отдельные замечания по терминологии: термин «момент остановки замкнут», вероятно, требует определения (с. 65, строка -10); кажется не вполне обоснованным объединять FWER, PCER и FDR в одном понятии «общий уровень значимости», т.к. это принципиально разные характеристики ошибки

множественного тестирования (с. 104). Касательно использованных обозначений следует отметить, что, по всей видимости, $\pi_0(x^{(n)})$ в (1.50) и $\Pi_0(x^{(n)})$ в (1.51) совпадают, а обозначение Ψ используется для вероятности принятия определенного решения (с. 24) и для дигамма функции (с. 88). Также можно отметить, что не удалось обнаружить в тексте диссертации ссылки на работы [4],[9],[11],[13-16],[18-20],[23-24] из списка публикаций. В тексте работы имеется ряд опечаток (с. 6, строка -2; с. 7 строка -11; с. 17, строка -7; с. 78, строка -3; с. 91, строка -2; с. 124, строка 7; с. 125, строка 13). Кроме того, обнаружены некоторые неточности при оформлении ссылок в автореферате.

Отмеченные недостатки не влияют на высокую оценку диссертации. Автореферат диссертации правильно отражает ее содержание. Основные результаты работы являются новыми и математически строго доказанными. Результаты диссертации своевременно опубликованы в периодических изданиях и материалах конференций, 5 работ опубликованы в журналах, рекомендованных ВАК.

Считаю, что диссертация Симушкина Д. С. «Статистические критерии с ограничениями на d-риски» удовлетворяет всем требованиям, предъявляемым ВАК к диссертациям, представляемым на соискание ученой степени кандидата физико-математических наук по специальности 01.01.05 – Теория вероятностей и математическая статистика, а диссертант заслуживает присвоения ему ученой степени кандидата физико-математических наук.

17 сентября 2020 г.

Официальный оппонент

Ведущий научный сотрудник
лаборатории «Центр геномной
биоинформатики им.
Ф.Добржанского»
Санкт–Петербургского
государственного университета
кандидат физ.-мат. наук, доцент

Малов Сергей
Васильевич

199004, г. Санкт-Петербург,
Средний проспект 41А
тел.: +7(812)3636103
e-mail: malovs@sm14820.spb.edu

Личную подпись Малова С.в. заверяю
Документ подготовлен по личной инициативе
Текст документа размещён в открытом доступе на сайте СПбГУ по адресу http://spbu.ru/science/expert.html
специалист по кадрам Ю.Ю. Камолова