Русский орнитологический журнал
The Russian Journal of Ornithology

Издается с 1992 года

Том XI

Экспресс-выпуск • Express-issue

2002 № 205

СОДЕРЖАНИЕ

Редактор и издатель А.В.Бардин
Кафедра зоологии позвоночных
Санкт-Петербургский университет
Россия 199034 Санкт-Петербург

Дата опубликования: 23 декабря 2002

Русский орнитологический журнал
The Russian Journal of Ornithology
Published from 1992
Volume XI
Express-issue
2002 № 205

CONTENTS

A.V.Bardin, Editor and Publisher
Department of Vertebrate Zoology
S.Petersburg University
S.Petersburg 199034 Russia

ISSN 0869-4362
*Русский орнитологический журнал 2002, Экспресс-выпуск **205**: 1083-1100*

Математические методы для интеллектуальных баз данных в биологии. 5. Элементы стохастического подхода к задачам обработки данных. Классические критерии

Э.А.Тропп, В.А.Егоров, Ю.Г.Морозов

Физико-технический институт им. А.Ф.Иоффе РАН, Санкт-Петербург, 194021, Россия

Поступила в редакцию 31 мая 2002

В биологических задачах, в частности в задачах орнитологии, приходится сталкиваться с большими объёмами достаточно сложно организованных данных. Грамотная современная обработка таких данных является, как правило, совместной работой биологов, математиков и программистов. Это связано с тремя этапами работы: сбор и осмысление данных, создание математической модели наблюдений, выбор или создание алгоритмов обработки и собственно обработка данных на ЭВМ. Следует отметить, что на всех этих этапах желательна совместная работа специалистов. Например, при создании математической модели требуется как знание биологического содержания задачи, так и умение разбираться в соответствующих математических вопросах; выбрать же математическую модель требуется таким образом, чтобы для решения задачи можно было выбрать алгоритм, реализуемый на ЭВМ за разумное время. Кроме того, в настоящее время делать какие-либо обоснованные новые выводы, например в орнитологии, можно только на основе большого числа правильно организованных наблюдений, для которых было бы легко создать соответствующие базы данных.

За последнее время в связи со значительным ростом возможностей ЭВМ и бурным развитием математических статистических методов стало возможным применять новые математические модели и создавать под них новые базы данных. Например, в течение более ста лет считалось, что наблюдения в большинстве случаев подчиняются нормальному закону или, в крайнем случае, являются простейшими функциями от нормально распределённых случайных величин. Такое широкое распространение нормального закона часто обосновывалось: математиками — ссылкой на прикладников, имеющих дело непосредственно с наблюдениями, прикладниками — ссылкой на математиков, которые выводят нормальный закон из теории, например, из асимптотической нормальности усреднённых наблюдений. На самом деле применение методов, основанных на предположении нормальности наблюдений, часто было связано с отсутствием других методов обработки, которые можно было бы реализовать вручную или при помощи примитивных вычислительных средств. В настоящее время появилась реальная возможность вооружить учёных современными методами обработки и анализа материала наблюдений.

Ниже для иллюстрации методов проверки статистических гипотез мы приводим ряд простейших примеров, в которых для вычислений не требуется прибегать к помощи ЭВМ. В реальных условиях приходится этими и

другими методами обрабатывать значительные массивы данных, обладающих специфическими особенностями и образующих сложную структуру данных. Поэтому для решения реальных задач необходимо использование ЭВМ и соответствующих баз данных.

Как отмечалось в наших предыдущих статьях (Тропп и др. 2002а,б,в,г), стохастические математические модели биологических процессов, наряду с детерминированными моделями, имеют широкое распространение. В данной статье мы будем использовать более узкое понятие математической модели исследуемых явлений. Мы будем предполагать, что исследуемый объект (или определённые числовые характеристики объекта) с достаточной степенью адекватности могут быть описаны в виде случайной величины, случайного вектора или случайной функции. Заключения о свойствах объекта мы делаем на основе наблюдений (выборки), которые представляют собой n независимых копий объекта или n независимых траекторий случайного процесса, описывающего изучаемый объект. Число n здесь называется объёмом выборки. Иногда предполагается, что наблюдается одна, но достаточно длинная траектория соответствующего случайного процесса. В этом случае в предположении эргодичности этого процесса необходимую нам информацию об объекте можно извлечь из одной траектории.

Остановимся на некоторых дополнительных особенностях изучаемых нами моделей. Одной из причин появления случайности в моделях исследуемых объектов является наличие стохастических погрешностей в детерминированных уравнениях, описывающих аналитические связи наблюдений с некоторыми известными переменными. Таким образом, считается, что "идеальные наблюдения" при отсутствии нежелательных погрешностей не включали бы в себя случайности.

Даже всем известный закон Ньютона при конкретной опытной проверке имеет погрешности, часто объясняемые отклонениями от идеальных условий проведения эксперимента. Принято считать, что при идеальных условиях этот закон выполняется точно. Исследование явлений, включающих подобного рода случайность, часто проводится методами математической статистики. При этом с точки зрения статистики закономерности, устанавливающие определённые связи между наблюдениями и некоторыми переменными, переходят в закономерности между стохастическими характеристиками наблюдений (например, математическими ожиданиями) и этими переменными.

Наряду с этими иногда рассматриваются модели, в которых случайность играет не только роль нежелательных погрешностей, но присутствует для объяснения причин наблюдаемых явлений. Широко известно, что случайный процесс Винера, или процесс броуновского движения, первоначально был открыт английским ботаником Броуном, наблюдавшим под микроскопом хаотические движения мелких частиц в некотором химическом растворе. Траектории движения этих частиц были столь замысловатыми, что первоначально зародилось предположение о "свободе воли" и, следовательно, о биологическом происхождении этих частиц. Позднее выяснилось, что наблюдавшиеся Броуном траектории являются типичными для траекторий виннеровского процесса, поскольку само движение частиц

вызвано их столкновениями с хаотически движущимися молекулами жидкости. Таким образом, положение частицы определяется суммой большого числа независимых между собой малых случайных сдвигов. Такая сумма, как известно, должна приближаться к виннеровскому процессу. Похожие рассуждения позволили Л.Башалье (Bachelier 1900) использовать винеровский процесс для описания модели эволюции стоимостей акций на рынке ценных бумаг. Это была одна из первых моделей в финансовой математике, использующих математическую теорию случайных процессов.

В теории вероятностей имеется много моделей явлений природы, описывающих явления в виде случайных процессов или случайных функций. Часто эти случайные функции являются решениями определённых стохастических дифференциальных, интегральных или интегро-дифференциальных уравнений.

Эти модели принципиально отличаются от моделей с детерминированными дифференциальными уравнениями, поскольку обладают совершенно новыми свойствами. Например, для стандартного винеровского процесса w_t

$$\int_0^t w_s dw_s = \frac{w_t^2 - t}{2},$$

в то время как для детерминированной функции w_t

$$\int_0^t w_s dw_s = \frac{w_t^2}{2}.$$

Таким образом, для стохастических процессов возникают дифференциальные уравнения, решения которых отличаются от решений уравнений, описывающих детерминированные явления. Различие в приведённых выше формулах вызвано, грубо говоря, тем, что за малое время Δt приращение винеровского процесса пропорционально $\sqrt{\Delta t}$, а не t, как для детерминированной функции (см.: Вентцель 1996).

Наконец отметим, что стохастические модели иногда применяются и для анализа чисто детерминированных сложных систем, зависящих от большого числа факторов. В этом случае сначала выделяют небольшое количество главных факторов, от которых зависимость выходного значения наибольшая, а суммарное влияние остальных факторов рассматривают как случайную ошибку. К этому подходу примыкает также точка зрения выдающегося советского математика А.Н.Колмогорова (1987), рассматривавшего случайную числовую последовательность как последовательность, имеющую максимальную сложность. Под сложностью он понимал минимальную длину программы для ЭВМ, написанной на специальном алгоритмическом языке, с помощью которой можно было бы воссоздать данную случайную последовательность. Удивительно, но так определённые случайные последовательности обладают многими свойствами обычных случайных последовательностей, построенных на основе теории меры.

Классический подход и общие принципы проверки статистических гипотез

В статистическом анализе обычно рассматриваются аспекты анализа данных, не являющиеся специфическими для какой-либо частной области исследований. Они представляют собой общие идеи и методы, которые после соответствующих видоизменений могли бы применяться в различных областях приложений. Поэтому разработка того или иного статистического анализа начинается с выбора достаточно формализованной математической модели, в рамках которой рассматриваются различные задачи проверки гипотез, классификации наблюдений, оценивания и т.д. Тем не менее, каждая область приложений имеет свои особенности в интерпретации данных и предлагаемых статистических решений. Более того, существует разумная точка зрения, согласно которой статистические процедуры только преобразуют данные в удобную для принятия решений форму. Окончательное же решение (хотя бы на этапе выбора уровня значимости) принимает специалист в конкретной содержательной области науки.

Модели могут быть самые разнообразные, рассчитанные как на чисто статистический, так и на логический анализ данных. Вопрос об адекватности выбранной модели исследуемым данным является трудным, и обычно основывается на большом опыте исследований в конкретной области и на чисто житейском опыте. Выбор модели фиксирует некоторые "априорные" знания об изучаемом объекте, которые, безусловно, всегда имеются у исследователя. Ограничения модели могут иметь разную форму. Например, они могут приводить к предположениям: 1) о форме зависимости данных от некоторых факторов; 2) об ограничениях на теоретические распределения данных; 3) о независимости определённых признаков и т.д.

Например, если мы наблюдаем популяцию птиц одного пола и примерно одного возраста, то скорее всего, их размеры и веса подчиняются нормальному распределению. Если в популяции присутствуют особи разного пола и возраста, то можно ожидать, что распределение данных является смесью нормальных распределений и т.д.

Некоторые предположения можно сделать на основе общетеоретических рассуждений, другие делаются на основе многолетнего опыта работы с конкретным материалом.

Иногда можно предполагать, что наблюдаемые данные представляют собой логические следствия некоторых известных случайных или детерминированных явлений. Такие ограничения приводят к логическому анализу данных или к смешанному логически-статистическому анализу. Например, при исследовании мигрирующих птиц следует учитывать погодные условия на пути их миграции. При этом можно использовать известные связи между погодными условиями и наблюдениями, либо рассматривать погодные условия как часть наших статистически изучаемых данных.

Подчеркнём, что выбор стохастической модели явления является важной, если не самой важной, частью анализа данных. Модель должна быть достаточно простой, хорошо объяснимой с точки зрения области приложения этой модели и не должна содержать большого числа параметров. Следует учитывать, что практически для любых реальных данных стохастическая модель является лишь приближением к действительности. Насколько

хорошее приближение выбрано иследователем, зависит во многом от его знаний и интуиции. К сожалению, неадекватные стохастические модели и неправильная интерпретация данных часто приводят к неверным статистическим выводам, а иногда и просто к парадоксам. Так, например, известный советский математик академик Фоменко провёл сложное сравнительно статистическое исследование правителей древнего мира и средневековья. Сравнивались генеалогические деревья правителей, учитывались длительность правления, причины смерти, число детей и пр. Вывод был парадоксальный — сравниваемые массивы данных идентичны. Это означало, что либо древнего мира, либо средневековья не существовало. Естественно, всё сообщество историков восприняло это исследование как противоречащее фактам, например, фактам археологии. Тем не менее Фоменко в своих исследованиях использовал допущения, приёмы и методы, часто используемые при статистической обработке данных.

Вернёмся к формальной постановке задачи.

Простейшая классическая постановка задачи проверки статистических гипотез такова. Согласно выбранной модели данных, имеются наблюдения, называемые в статистике выборкой, которые представляют собой одномерные или многомерные обычно независимые случайные векторы. Обозначим всю выборку одной буквой X. Имеется основная интересующая исследователя гипотеза $H(0)$ и совокупность альтернативных гипотез H, а также набор возможных решений D, из которых исследователь выбирает своё решение. Для проверки основной гипотезы требуется на основе выборки наиболее оптимальным способом выбрать правило принятия решения, т.е. выбрать функцию $d = d(X)$ со значениями из D. Обычно D = {да, нет}, где "да" означает принятие испытуемой гипотезы, "нет" — её отвержение. Можно использовать и другие наборы D, например, можно считать, что D — это числа интервала [0,1] и интерпретировать d из D как вероятность, с которой нужно отвергнуть гипотезу $H(0)$ при получении наблюдений X, т.е. окончательное решение в этом случае принимается с помощью соответствующей процедуры случайного выбора. Можно также к D добавить значки, означающие недостаточность информации для принятия определённого детерминированного решения, значки, указывающие на противоречивость данных в рамках выбранной модели и т.д. Всё это влияет на логическую структуру процедуры принятия решения и на сам характер принимаемых решений. Например, если в выборке появилось резко выделяющееся наблюдение, не похожее на остальные, то можно поступать по-разному. Согласно одной модели обработки, это наблюдение не следует учитывать вовсе. Согласно другой модели, на основании наших наблюдений следует выбрать распределение данных с "тяжёлыми хвостами", которое допускает такую структуру наблюдений. Наконец, согласно третьей модели, при соответствующем выборе набора возможных решений D и способа принятия решений следует сигнализировать об определённых противоречиях в данных.

Оптимальность выбора функции от наблюдений $d = d(X)$ также можно понимать по-разному. В классической статистике обычно предполагается, что имеется некоторая параметрическая модель наблюдений, т.е. что теоре-

тическая функция распределения данных имеет вид $P(t,X)$, где t — некоторый вектор неизвестных параметров (например, можно предполагать, что теоретическое распределение нормально, считая параметрами математическое ожидание и дисперсию). В этом случае задача принятия решений состоит в выборе истинного значения параметра t или в проверке какой-либо гипотезы о значении этого параметра (например, в случае нормальности наблюдений можно оценить математическое ожидание или проверить гипотезу о том, что это математическое ожидание превосходит заданный порог). В этом случае естественно считать, что функция $d(X)$ принимает значения из области возможных значений параметра t, и равенство $d(X) = t$ означает, что согласно принятому решению, значения параметра модели равно t.

Для формализации задачи принятия решения в случае параметрической модели наблюдений вводят специальную функцию, называемую функцией потерь, $L(t, d(X))$, которая описывает потери (например, в денежном выражении), если мы приняли значение параметра равным $d(X)$ в то время, как истинное значение параметра равно t. Например, если все неправильные решения неприемлемы в одинаковой степени, то можно считать, что потеря от любого неправильного решения одна и та же и равна, к примеру, единице, а правильное решение ведёт к отсутствию потерь. В этом случае функция $L(t, d(X)) = 1$, если принято неправильное решение, и $L(t, d(X)) = 0$, если принято правильное решение.

Оптимизация состоит в выборе решения $d = d(X)$, минимизирующего функцию риска $R(t,d) = EL(t, d(t))$. Здесь буква E обозначает математическое ожидание, соответствующее теоретической функции распределения $P(t,X)$, т.е. математическое ожидание берётся, грубо говоря, в предположении, что истинное значение параметра есть t.

Стихийно на бытовом уровне такой подход применяется довольно часто, хотя и не осознанно. Например, если мы предпринимаем поездку на поезде, то обычно приходим к нему за 15-20 минут до отправления, если же мы собираемся лететь на самолёте, то оставляем гораздо больший запас времени. Это связано с тем, что мы сознаём, что потери (в том числе и финансовые) при опоздании на самолёт значительно больше. Поэтому мы готовы пойти на дополнительные затраты, связанные с более ранним приездом в аэропорт. В обоих случаях может случиться случайный набор обстоятельств, ведущий к опозданию, но при путешествии на самолёте в силу выбранной нами стратегии поведения вероятность этого стечения обстоятельств меньше, поэтому математическое ожидание потерь от возможности опоздания не велики.

К сожалению, минимизировать функцию двух переменных $R(t,d)$ за счёт выбора решающего правила d одновременно по всем значениям t в большинстве случаев невозможно. Можно, конечно, использовать минимаксный подход, рассчитанный на наихудший случай, т.е. выбирать решение $d(X)$, минимизирующее максимум по t функции $R(t,d)$, но такой подход считается слишком осторожным.

Для иллюстрации рассмотрим следующий пример. Пусть данные представляют собой измерения длины крыла изучаемых птиц. Предположим, что эти наблюдения имеют теоретическое нормальное распределение с ма-

тематическим ожиданием m и дисперсией σ^2. В этом случае $t = (m, \sigma^2)$. Пусть, согласно правилу принятия решений d(X), верны равенства $m = m_0$, $\sigma^2 = \sigma_0^2$, где $m_0 = m_0(X)$, $\sigma_0^2 = \sigma_0^2(X)$ — известные функции от наблюдний X. В этом случае, например, разумно выбрать функцию потерь вида $L = |m - m_0| + |\sigma - \sigma_0|$. Тогда, конечно, не все неправильные решения равнозначны. Здесь, грубо говоря, чем дальше теоретическое распределение от предполагаемого по правилу принятия решений, тем потери больше.

Функция риска в этом случае равна

$$R = R(m, \sigma^2, d) = \int p(X)(|m - m_0| + |\sigma - \sigma_0(X)|dX,$$

где p(X) — плотность распределения нашей нормальной выборки с параметрами m и σ^2. Оптимизация в этой задаче сведётся к выбору функций m_0 и σ_0^2 от наблюдений X, на которых достигается минимум функции R. Разумеется, минимум для всех значений m и σ^2 одновременно достигается только в исключительных случаях. Минимаксный подход в этом случае сводится к выбору таких функций m_0, σ_0^2, для которых максимум функции R по всем допустимым значениям переменных m, σ^2 будет наименьшим.

Иногда разумно предполагать, что сам параметр t случаен и имеет некоторое "априорное" распределение. Тогда можно минимизировать средние потери вида $E = E(d) = ER(t,d)$, поскольку после взятия математического ожидания зависимость от t исчезнет, и E теперь является функцией только решающего правила $d(X)$. Такой подход к проверке статистических гипотез называется байесовским. Например, если мы наблюдаем некоторую популяцию птиц, и множество значений параметра t состоит из двух значений M и F, где M соответствует мужскому полу особи, а F — женскому, то в качестве априорных вероятностей для M и F можно взять частоты количества самцов и самок в предыдущих аналогичных исследованиях.

Рассмотрим случай, когда имеются только две гипотезы: нулевая $H(0)$ и альтернативная $H(1)$,— и эти гипотезы являются простыми. Это означает, что они однозначно определяют соответствующие теоретические распределения наблюдений, поэтому D = {да,нет}, где "да" соответствует высказыванию в пользу нулевой гипотезы, "нет" — высказыванию против неё. Будем считать, что в этом случае параметр t принимает только два значения: 0 и 1, причём $t = 0$ означает справедливость гипотезы $H(0)$, а $t = 1$ означает справедливость гипотезы $H(1)$. Тогда можно поступить следующим образом. Сначала ограничиться классом функций $d(X)$, для которых $R(0,d) < \alpha$, для некоторого достаточно малого числа α (например, $\alpha = 0.01$), называемого уровнем значимости статистического критерия, затем в этом классе минимизировать функцию $R(1,d)$.

Наиболее просто это правило проверки гипотез выглядит в теории Неймана-Пирсона (см.: Леман 1977), в которой конкретизируется функция потерь L. В теории Неймана-Пирсона предполагается, что $L = 1$, если согласно выбранному правилу принятия решений, принимается ошибочное решение и $L = 0$, если решение правильное. При таком выборе функции L задача сводится к условной минимизации вероятности ошибки второго рода при условии, что вероятность ошибки первого рода $p(1)$ — это вероятность отвергнуть $H(0)$, когда она верна. Вероятность ошибки второго рода

$p(2)$ — это вероятность принять нулевую гипотезу, когда она не верна. В этом случае, согласно теории Неймана-Пирсона, оптимальный критерий состоит в том, что $d(X)$ = "нет" (т.е. нулевая гипотеза отвергается), если $T(X) > C$, где Т — отношение правдоподобия: $T = p(1,X)/p(0,X)$; $p(1,X)$ и $p((0,X)$ — теоретические плотности распределения при справедливости первой и нулевой гипотез, соответственно; C — постоянная, определяемая уровнем значимости α с помощью уравнения $P_0(T > C) = \alpha$, где P_0 — условная вероятность при справедливости нулевой гипотезы.

Назовём $R = 1 - p(2)$ мощностью критерия. Тогда построенный в теории Неймана-Пирсона критерий сводится к условной максимизации R и называется наиболее мощным. Часто такие критерии являются равномерно наиболее мощными по широкому классу альтернативных гипотез $H(1)$ (Леман 1977).

Для иллюстрации рассмотрим пример из книги Н.Бейли (1962). Предположим, что мы поймали 3 птиц, средняя масса тела которых оказалась 89.33 г. Спрашивается, значимо ли различие между этой величиной и известным теоретическим средним, равным 95.61 г, если среднеквадратичная ошибка известна и равна 4.52 г, и распределение массы предполагается нормальным.

Для решения этой задачи рассмотрим разность выборочного и теоретического средних. Обозначим её буквой m. После вычислений получим, что $m = -6.28 < 0$. Поэтому в нашем случае можно положить: $H(0) = \{m = 0\}$, $H(1) = \{m = m(1)\}$, где $m(1) < 0$ — некоторое число, вообще говоря, нам неизвестное. Квадратичное отклонение от этого распределения равно 4.52. Наблюдённая разность, выраженная в долях квадратичного отклонения, равна $6.28/4.52 = -1.39$. Согласно теории Неймана-Пирсона, нулевая гипотеза отвергается, если $T(X) > C$. Последнее неравенство может быть переписано в виде $m / (4.52) < C'$, где взяв, например, уровень значимости $\alpha = 0.05$, получим из таблиц нормального распределения $C' = -1.7$. Это означает, что рекомендуется отвергнуть нулевую гипотезу только если $m / 4.52 < -1.7$, а вычисленное нами значение этого отношения -1.39 говорит о том, что наблюдения соответствуют предполагаемому распределению. Отметим, что построенный в этом примере критерий является равномерно наиболее мощным для всех альтернатив против $m < 0$.

Интуитивно на этот критерий можно посмотреть следующим образом. Если $T(X) > C$, то следует признать, что произошло очень маловероятное событие (вероятности, не превосходящей числа α) и нулевая гипотеза верна, или что нулевая гипотеза не верна. Считается, что наблюдатель обычно не доверяет маловероятным событиям и отвергает в этом случае нулевую гипотезу.

Логика построения критериев в этом классическом примере присутствует практически во всех конструкциях критических областей, в том числе и при построении критериев для испытания сложных гипотез. (Здесь под сложной гипотезой мы понимаем гипотезу, не определяющую истинное распределение полностью, а определяющую лишь семейство распределений, к которому истинное распределение принадлежит.) Сначала определяется некоторое зависящее от выборки событие A, вероятность которого

при нулевой гипотезе не превосходит заданного малого уровня значимости, а при нарушении нулевой гипотезы эта вероятность (мощность критерия) должна принимать большие значения. Затем решение о справедливости нулевой гипотезы принимается в зависимости от того, произошло или нет событие A. Конечно, вычисление вероятностей происходит в рамках выбранной вероятностной модели наблюдений. Определение этой модели выходит, как правило, за рамки работы математика и является результатом совместной деятельности математика и специалиста в соответствующей прикладной области исследований.

Таким образом, конструкция множества A определяется как исходными предположениями (выбор вероятностно-математической модели), так и характером решаемой задачи.

Рассмотрим предыдущий пример в ситуации, когда среднеквадратическое отклонение не известно *a priori*, а вычислено по тем же наблюдениям, на основе которых принимается решение. В этом случае задачу нельзя решить в рамках теории Неймана-Пирсона, поскольку гипотезы зависят от неизвестного среднеквадратического отклонения и, следовательно, перестают быть простыми. Тем не менее, критерий можно построить на основе приведённой выше логики их построения. Действительно, теория показывает, что в данном случае можно пользоваться той же самой критической областью, выбирая постоянную C не из таблиц нормального распределения, а из таблиц распределения Стьюдента с двумя степенями свободы.

Приведённое выше обсуждение касалось, в основном, построения параметрических статистических процедур, поскольку предполагало, что теоретическое распределение наблюдений принадлежит какому-то известному параметрическому семейству наблюдений. Более детальное обсуждение этого вопроса можно найти в книге Э.Лемана "Проверка статистических гипотез" (1977).

Дисперсионный анализ

Остановимся на одной широко используемой в приложениях модели организации данных — на дисперсионном анализе. Для простоты рассмотрим случай двухфактороного дисперсионного анализа без взаимодействия факторов. Случай многофакторного дисперсионного анализа рассматривается аналогично. При полном двухфакторном анализе предполагается, что наблюдения имеют следующую форму:

$$X_{i,j} = \alpha + \alpha_i + \beta_j + \varepsilon_{i,j}, \quad i = 1, 2, \ldots, I, \quad j = 1, 2, \ldots, J, \quad \sum_{i=1}^{I} \alpha_i = 0, \quad \sum_{j=1}^{J} \beta_j = 0.$$

Здесь α называется общим средним, α_i — i-м уровнем фактора α, β_j — j-м уровнем фактора β, $\varepsilon_{i,j}$ представляют собой независимые между собой погрешности. (Иногда предполагают погрешности зависимыми с известной ковариационной матрицей.) Проверяются гипотезы об отсутствии влияния фактора α (т.е. все α_i равны нулю) или об отсутствии влияния фактора β (т.е. все β_j равны нулю).

Например, $X_{i,j}$ могут обозначать массы пойманных птиц, α — фактор, связанный с полом птицы, β — фактор одного из J мест обитания пойман-

ных птиц. Можно проверять гипотезы о влиянии пола на массу тела птицы или о влиянии на неё условий местообитания.

Конечно, приведённая выше схема дисперсионного анализа проверяет лишь гипотезу о линейной зависимости математических ожиданий наблюдений от определённых факторов. В работе Г.Шеффе (1963) анализируются также некоторые виды нелинейных зависимостей математических ожиданий от соответствующих факторов.

Для решения задач дисперсионного анализа строится определённое выражение, называемое дисперсионным отношением, и решение принимается в зависимости от того, превосходит ли это выражение заданный уровень или нет. Отметим, что для задания зависимости этого уровня от уровня значимости критерия необходимо конкретизировать распределение погрешностей. Обычно эти погрешности предполагаются нормально распределёнными, хотя на качественном уровне предлагаемому критерию можно придать определённый разумный геометрический смысл и без уточнения вида распределения погрешностей.

Описанная общая конструкция построения критериев позволяет строить большое их число. Возникает проблема выбрать среди них наиболее оптимальный, причём оптимальность может пониматься с различных точек зрения. Для параметрического случая эта задача хорошо исследована, и для многих ситуаций приемлемые критерии найдены. Для непараметрических критериев, о которых речь пойдет в следующей части нашей статьи, дело обстоит значительно хуже. Количество разработанных и опубликованных к настоящему времени критериев насчитывает несколько сотен наименований. Некоторые из этих критериев относятся к однотипным статистическим задачам. Поэтому задача эффективного сравнения непараметрических критериев является и сейчас актуальной и для многих случаев нерешённой. С ситуацией по этому вопросу на сегодняшний день можно ознакомиться в монографии Я.Ю.Никитина (1995).

Классификация и оценки максимального правдоподобия

Вернёмся к критерию Неймана-Пирсона проверки простой гипотезы против простой альтернативы. Переформулируем задачу в параметрической постановке.

Имеется выборка X и семейство возможных теоретических плотностей выборки $p(X,\Theta)$. Простая нулевая гипотеза имеет вид H_0: $\Theta = \Theta_0$, а простая альтернативная гипотеза H_1: $\Theta = \Theta_1$. Тогда, согласно приведённой в начале статьи лемме Неймана-Пирсона, мы вибираем гипотезу H_0, если $p(X,\Theta_0) > Cp(X,\Theta_1)$, и выбираем гипотезу H_1, если выполняется противоположное неравенство. Постоянную C мы выбираем в соответствии с выбранным уровнем значимости критерия.

В монографии С.Р.Рао (1968) проведено обобщение критерия Неймана-Пирсона на задачи классификации. В отличие от задачи Неймана-Пирсона, в задаче классификации *a priori* предполагается справедливость не одной из двух, а одной из k попарно несовместных гипотез:

$$H_i: \Theta = \Theta_i,\ i = 1, 2, \ldots, k.$$

Задача состоит в оптимальном выборе одной из этих гипотез. Не уточняя понятия оптимального выбора гипотез, приведём решение этой задачи в байесовской постановке. Последнее означает, что при принятии решения мы располагаем априорными вероятностями гипотез, т.е. вероятностями

$$p_i = P(\Theta = \Theta_i), \quad i = 1, 2, \ldots, k,$$

вычисленными заранее без учёта наблюдений, на основании которых выбирается гипотеза. В этом случае оптимальное правило выбора, минимизирующее математическое ожидание числа ложных классификаций (при многократном применении этого правила) состоит в выборе той гипотезы, на которой достигается максимум произведения $p_i P_i(X)$, $i = 1, 2, ..., k$. В том случае, когда априорные вероятности гипотез отсутствуют, разумно предполагать все рассматриваемые гипотезы *a priori* равновероятными, т.е. предполагать $p_i = 1/k$, $i = 1, 2, ..., k$. Тогда мы придём к принципу максимального правдоподобия, согласно которому выбирается та гипотеза, на которой функция правдоподобия $p(X, \Theta)$, $\Theta = \Theta_i$, $i = 1, 2, ..., k$, достигает наибольшего значения. В теории классификации функции от выборки, на основании которых производится классификация, называются дискриминантными информантами. Монотонные преобразования множества дискриминантных информантов (т.е. преобразования, при которых большему значению дискриминантного иформанта соответствует большее значение преобразования) дают нам снова дискриминантные информанты.

Например, наблюдатель измеряет m характерных признаков пойманных птиц (масса тела, длина крыла и т.д.) и на основании анализа аналогичных наблюдений за много лет он хочет построить дискриминантную функцию, относящую пойманную птицу к тому или иному виду. Предполагается, что наблюдаемые признаки имеют многомерное нормальное распределение с математическими ожиданиями, зависящими от вида птицы, и с одинаковыми для всех видов ковариационными матрицами. В этом случае, записанные в матричной форме, дискриминантные информанты будут иметь вид

$$S_i = \left(\mu_i^T R^{-1}\right) X - \frac{1}{2} \mu^T R^{-1} \mu + \log p_i,$$

где R — общая ковариационная матрица наблюдений, μ_i, $i = 1, 2, ..., k$, векторы математических ожиданий признаков для каждого вида, p_i — априорные вероятности поимки птицы i-го вида. В качестве этих априорных вероятностей можно взять частоты поимки птиц каждого вида за многолетние предшествующие наблюдения, а в качестве ковариационной матрицы и векторов математических ожиданий можно взять стандартные статистические оценки, вычисленные на основе предшествующих наблюдений. Таким образом, для классификации нам достаточно на основе наблюдений вычислить линейные функции S_i и выбрать гипотезу Hj, на которой достигается максимум S_i.

Похожие выводы можно сделать и на основе информационного подхода. В этом случае вводится специальное информационное расстояние между распределениями, и выбирается то из гипотетических распределений, которое ближе к выборочному распределению. По этому поводу — см. монографию С.Кульбака (1967).

Полученный выше принцип максимального правдоподобия пригоден и для случая бесконечного множества гипотез Θ. В этом случае, естественно, решение представляет собой оценку параметра Θ, называемую оценкой максимального правдоподобия. Функция $p(X,\Theta)$, рассматриваемая как функция параметра Θ, называется функцией правдоподобия, а оценка максимального правдоподобия есть то значение Θ, на котором достигается максимум функции правдоподобия.

Метод максимального правдоподобия является наиболее распространённым методом оценивания в математической статистике. Он продолжает общие традиции, идущие, вероятно, из физики, сведения многих научных проблем к экстремальных задачам. Этот метод обладает хорошими асимптотическими свойствами. В частности, при определённых условиях оценки максимального правдоподобия асимптотически нормальны. Практически это позволяет для многих распределений при больших объёмах выборки предполагать, что исходные данные имеют нормальное распределение. Однако в последние годы выяснилось, что, казалось, вырожденные ситуации, когда асимптотической нормальности оценок нет, представляют для статистиков также очень большой интерес.

Метод наименьших квадратов

Известно, что в случае нормальности наблюдений метод максимального правдоподобия превращается в другую экстремальную задачу — в метод наименьших квадратов. Мы рассмотрим этот метод на примере одномерных регрессионных задач.

Предположим, что для оценки p неизвестных параметров t_1, t_2, ..., t_p используется n независимых наблюдений y_1, y_2, ..., y_n, причём эти величины связаны соотношениями

$$y_i = \sum_{j=1}^{p} x_{i,j} t_j + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

где $x_{i,j}$ суть известные значения контролируемых наблюдателем переменных, а ε_i — независимые нормальные случайные величины с нулевым математическим ожиданием и одинаковой неизвестной дисперсией σ^2. Отметим, что качественные выводы на основе геометрических свойств метода наименьших квадратов можно делать и без предположения нормальности погрешностей.

Оценки наименьших квадратов t_j, $j = 1, 2, \ldots, p$,— это те значения параметров t_j, на которых достигается минимум выражения

$$Q = \left(\sum_{i=1}^{n} y_i - \sum_{j=1}^{p} x_{i,j} t_j \right)^2.$$

Таким образом, методом наименьших квадратов мы определяем приближённую линейную зависимость между контролируемыми переменными и нашими наблюдениями.

Приравнивая производные функции Q по переменным t_j к нулю, легко получим систему линейных уравнений, из которой определяются оценки

наименьших квадратов. Эти уравнения, называемые нормальными уравнениями, просто записываются в матричном виде. Действительно, введём: вектор-столбец $\boldsymbol{y}$, состоящий из всех наблюдений y_j, $j = 1, 2, ..., n$, вектор-столбец $\boldsymbol{t}$, состоящий из параметров t_i, $i = 1, 2, ..., p$, матрицу $\boldsymbol{X}$, имеющую n строк и p столбцов, состоящую из величин $x_{i,j}$. Тогда система нормальных уравнений будет иметь вид

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{t} = \boldsymbol{X}^T\boldsymbol{y}.$$

Здесь $\boldsymbol{X}^T$ — транспонированная матрица для матрицы $\boldsymbol{X}$.

Известно, что если матрица $\boldsymbol{X}^T\boldsymbol{X}$ не вырождена, то оценки наименьших квадратов определяются однозначно по формуле:

$$t = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

В случае вырожденности матрицы $\boldsymbol{X}^T\boldsymbol{X}$ система нормальных уравнений имеет бесконечное множество решений, каждое из которых доставляет минимум функции Q.

Приведённая схема метода наименьших квадратов носит общий характер. Остановимся на некоторых более частных схемах.

Пусть наблюдения линейно с точностью до случайных погрешностей ε зависят от некоторой детерминированной переменной z, которая может быть временем, температурой, высотой местности над уровнем моря или какой-нибудь другой содержательной характеристикой условий, при которых получено соответствующее наблюдение. Это можно записать в форме обычной линейной зависимости

$$y = \beta_0 + \beta_1 z + \varepsilon,$$

или в форме представления выборки

$$y_i = \beta_0 + \beta_1 z_i + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

Таким образом, в рассмотренной выше схеме мы взяли $t_1 = \beta_0$, $t_2 = \beta_1$, $x_{i,1} = 1$, $x_{i,2} = z_i$, $i = 1, 2, ..., n$. Вычисления показывают, что решая нормальные уравнения для этого случая, мы получаем следующие оценки:

$$\beta_1 = \frac{\sum_{i=1}^{n}(z_i - z)(y_i - y)}{\sum_{i=1}^{n}(z_i - z)^2}, \quad \beta_0 = y - \beta_1 z.$$

Здесь $y = \frac{1}{n}\sum_{i=1}^{n} y_i$, $z = \frac{1}{n}\sum_{i=1}^{n} z_i$.

Нетрудно понять, что подобные вычисления можно проводить для квадратичной зависимости наблюдений от известной детерминированной переменной. Более того, можно рассматривать полиномиальную зависимость любой степени. Конкретные вычислительные формулы для оценки коэффициентов будут различны, но все они, конечно, будут решениями соответствующих нормальных уравнений.

Тем не менее, предаваться эйфории по поводу простоты использования метода наименьших квадратов преждевременно. Трудности часто появляются на этапе выбора модели. Например, возникает вопрос, какую зависи-

Тем не менее, предаваться эйфории по поводу простоты использования метода наименьших квадратов преждевременно. Трудности часто появляются на этапе выбора модели. Например, возникает вопрос, какую зависимость выбрать, линейную или квадратичную. С одной стороны, чем выше степень многочлена, тем точнее можно этим многочленом приблизить наши наблюдения. С другой стороны, чем выше степень многочлена, тем больше параметров нужно оценивать. Это влияет на надёжность соответствующих статистических выводов.

Один из способов убедиться в правильности выбранной модели заключается в следующем. Выберем, например, квадратичную зависимость. Затем проверим методами дисперсионного анализа гипотезу о равенстве нулю коэффициента при квадратичном члене. Если эта гипотеза принимается, то можно остановиться на линейной зависимости, а если отвергается, то следует таким же образом сравнить квадратичную зависимость с зависимостью, описываемой многочленом третьей степени.

Можно качество выбранной модели оценивать величиной остаточной дисперсии, равной

$$Q_0 = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{i,j} t_j \right)^2 .$$

В случае адекватности модели эта величина, делённая на дисперсию наблюдений, должна иметь χ^2-распределение с $(n\text{-}p)$ степенями свободы. Поэтому большие значения этой статистики говорят либо о большой исходной дисперсии наблюдений, либо о неадекватности модели.

В рассмотренных выше примерах мы использовали полиномиальную зависимость данных от детерминированной переменной. Но бывают случаи, когда такое предположение только усложняет нахождение истинной зависимости. Например, наблюдения могут зависеть от переменной периодически. Это часто происходит, когда на наблюдаемую переменную влияет температура окружающего воздуха, время дня, время года и т.д. В этом случае разумная модель имеет вид

$$y = a\cos(bz) + c\sin(dz) + \varepsilon,$$

где постоянные b, d известны.

Полиномиальная модель в этом случае даст хорошее приближение только при высокой степени аппроксимирующего полинома. Это приведёт к сложностям и, в конечном счёте, к не очень надёжным выводам. Поэтому выбор аппроксимирующей функции требует глубокого знания изучаемого материала.

Робастная регрессия

Другая сложность, связанная с методом наименьших квадратов, состоит в наличии в выборке аномальных элементов. Присутствие таких элементов может быть обусловлено как природой наблюдаемого явления (например, не нормальными распределениями ошибок или сложным видом регрессионной зависимости), так и редкими грубыми отклонениями от методики проведения наблюдений. Последнее приводит к тому, что теоретическое

рактеристиками. Борьбу с такими аномальностями осуществляют так называемые робастные оценки.

Термин "робастность", введённый впервые П.Хьюбером в 70-х годах прошлого века, означает устойчивость статистических выводов по отношению к отклонениям от "идеальных" априорных предположений. Так, например, обычно наблюдения предполагаются распределёнными по нормальному закону. Тем не менее наблюдатель никогда не может иметь полной уверенности в справедливости этого предположения. Оценки же наименьших квадратов очень чувствительны к отклонениям от нормальности. Например, присутствие в выборке небольшого числа аномальных наблюдений (которые ошибочно попали в выборку и имеют характеристики, резко отличающиеся от характеристик основной массы наблюдений) может сделать непригодными оценки наименьших квадратов. Если же применять описываемый ниже метод наименьших модулей, то полученные согласно этому методу оценки испортятся значительно меньше. В этом случае оценки метода наименьших модулей считаются более робастными, чем оценки наименьших квадратов.

Первоначально Хьюбер термину "робастность" придал чёткий математический смысл, который, конечно, соответствует тому описанию робастности, которое мы привели выше. Позднее Хампель использовал этот термин в несколько более общем математическом смысле. В более поздних, особенно прикладных, публикациях термин "робастность" стал широко применяться не как математический термин, а как синоним слова "устойчивость".

Часто одной из рекомендаций при обработке наблюдений является требование визуального контроля качества выборки. Следование этой рекомендации приводит, в частности, к тому, что выявленные аномальные наблюдения изымаются их выборки, и обработке подвергаются оставшиеся наблюдения. В итоге мы получаем стихийные робастные оценки, в которых резко выделяющиеся наблюдения не учитываются, так сказать, волюнтаристски, поскольку наблюдатель на глазок решает, какие наблюдения аномальны, а какие нет. При таком подходе наблюдатель имеет возможность получать желаемые выводы, например, отбрасывая данные, находящиеся в противоречии с выбранной им формой зависимости. Особенно это относится к ситуации, когда обрабатывается сравнительно небольшое число однородных данных.

Рассмотрим пример из работы П.Хьюбера (1984). Пусть данные зависимости y от x соответствуют таблице:

Номер точки	1	2	3	4	5	6
x	-4	-3	-2	-1	0	10
y	2.48	0.73	-0.04	-1.44	-1.32	0

Предполагая линейную зависимость y от x, легко получим из метода наименьших квадратов

$$y = 0.41 - 0.077x.$$

Если относится к этому выводу формально, то его вполне можно принять, поскольку остаточная дисперсия небольшая, да и дисперсионный

Если относится к этому выводу формально, то его вполне можно принять, поскольку остаточная дисперсия небольшая, да и дисперсионный анализ даёт приемлемые результаты. Мы можем в целях повышения точности попытаться найти приближение с помощью квадратичной функции, что приведёт нас к параболе. Эта парабола уже при не очень больших положительных x даёт значения, сильно отличающиеся от значений на прямой. Наконец, мы можем обнаружить, что шестое наблюдение сильно отличается от остальных и отбросить его. Тогда мы получим прямую, близкую к прямой $y = -2 - x$, которая даёт очень маленькую остаточную дисперсию. Отметим, что данные этого примера получены моделированием. Первые 5 точек взяты на прямой $y = -2 - x$ и изменены добавлением случайных нормальных погрешностей (со средним 0 и стандартным отклонением 0.6), а шестая точка взята сознательно аномальной. Этот пример показывает, насколько путает карты исследователю даже одно аномальное наблюдение.

Теперь обсудим кратко подход Хьюбера (1984) к робастному оцениванию. Предположим сначала, что выборка имеет представление

$$y_i = \beta + \varepsilon_i,\ i = 1, 2, ..., n,$$

где погрешности ε_i независимы и имеют функцию распределения вида

$$F(x) = (1 - \delta)H(x) + \delta G(x).$$

Здесь $H(x)$ — функция распределения известного (например, нормального) распределения; $G(x)$ — функция распределения аномального распределения, которое может быть полностью или частично известным, параметр загрязнения выборки δ считается известным малым числом. Таким образом, в выборке присутствует в среднем δn аномальных наблюдений. Предполагается, что распределения H, G симметричны. Требуется найти оценку параметра β, на точность которой не влияют "плохие" свойства аномальных наблюдений.

В первой части нашей работы было сформулировано одно необходимое требование к математической модели — её корректность. Это требование состояло в том, что малые изменения модели должны мало влиять на получаемые из неё выводы. В нашей модели разумно считать, что присутствие малой доли аномальных наблюдений является малым изменением модели. Поэтому свойство робастности оценок является крайне желательным.

Простейший подход к получению приемлемой оценки состоит в том, что в качестве оценки берут выборочную медиану m. Она для нечётных n определяется следующим образом. Все наблюдения упорядочиваются по возрастанию, и в качестве m берётся элемент с номером $(n+1)/2$ среди всех этих упорядоченных элементов выборки. Для чётных n берётся среднее арифметическое элементов с номерами $n/2$ и $n/2+1$. Поскольку элементы выборки вблизи выборочной медианы концентрируются достаточно тесно, то легко видеть, что наличие небольшого числа аномальных наблюдений мало влияют на значение выборочной медианы.

Это означает, что выборочная медиана обладает хорошими робастными свойствами. Хорошо известно, что при слабых предположениях выборочная медиана является асимптотически нормальной. Выражение для её

вычислять асимптотическую эффективность оценивания с помощью приведённого метода. Отметим, что она достаточно высока.

Известно, что точно так же, как на среднем арифметическом элементов выборки достигает минимума функция

$$Q(a) = \sum_{i=1}^{n} (y_i - a)^2,$$

на выборочной медиане достигает минимума функция

$$R(a) = \sum_{i=1}^{n} |y_i - a|.$$

Поэтому можно ожидать, что в задаче о линейной регрессии мы получим более робастную оценку, если вместо метода наименьших квадратов будем использовать метод наименьших модулей. Это означает, что в качестве оценки мы выбираем те допустимые значения параметров t_j, на которых достигает минимума выражение

$$Q_1 = \sum_{i=1}^{n} |y_i - \sum_{j=1}^{p} x_{i,j} t_j|.$$

В отличие от метода наименьших квадратов, получить простые аналитические формулы для оценок t_j в этом случае не удаётся, но оценки нетрудно получить численно с помощью ЭВМ.

По аналогии с методом наименьших квадратов и методом наименьших модулей Хьюбер рассмотрел выражение вида

$$Q_2 = \sum_{i=1}^{n} \rho\left(y_i - \sum_{j=1}^{p} x_{i,j} t_j \right).$$

Здесь ρ — некоторая известная выпуклая функция. Для метода наименьших квадратов $\rho = x^2$, а для метода наименьших модулей $\rho = |x|$. В качестве оценок параметров он предложил рассматривать те их значения, на которых достигает минимума выражение Q_2. Такие оценки Хьюберт назвал М-оценками, поскольку они являются оценками максимального правдоподобия в случае, когда погрешности ε имеют плотность распределения вида

$$p(y) = A \cdot \exp\{-B\rho(y)\},$$

где A, B — положительные постоянные. В монографиях П.Хьюберта (1984) и Ф.Хампеля с соавторами (1989) исследованы асимптотические свойства М-оценок, доказана их асимптотическая нормальность, найдены выражения для асимптотических дисперсий. Это позволяет сравнивать М-оценки между собой, добиться выбора наиболее робастной оценки. Отметим, что функцию ρ можно, в частности, подобрать таким образом, что аномальные элементы выборки практически не участвуют в формировании соответствующей М-оценки.

Литература

Бейли Н. 1962. *Статистические методы в биологии.* М.: 1-260.
Вентцель А.Д. 1996. *Курс теории случайных процессов.* М.: 1-399.

Колмогоров А.Н. 1987. Таблица случайных чисел //*Теория информации и теория алгоритмов.* М.: 204-213.
Кульбак С. 1967. *Теория информации и статистика.* М.: 1-408.
Леман Э. 1977. *Проверка статистических гипотез.* М.: 1-498.
Никитин Я.Ю. 1995. *Асимптотическая эффективность непараметрических критериев.* М.: 1-238.
Рао С.Р. 1968. *Линейные статистические методы и их приложения.* М.: 1-547.
Тропп Э.А., Егоров В.А., Морозов Ю.Г. 2002а. Математические методы для интеллектуальных баз данных в биологии. 1. Математические методы в биологии. Общий анализ //*Рус. орнитол. журн.* Экспресс-вып. **177**: 163-171.
Тропп Э.А., Егоров В.А., Морозов Ю.Г. 2002б. Математические методы для интеллектуальных баз данных в биологии. 2. Уровни организации живого, математические языки их описания и корректность постановки задач математического моделирования //*Рус. орнитол. журн.* Экспресс-вып. **190**: 631-642.
Тропп Э.А., Егоров В.А., Морозов Ю.Г. 2002в. Математические методы для интеллектуальных баз данных в биологии. 3. Математические модели экологических систем //*Рус. орнитол. журн.* Экспресс-вып. **193**: 723-735.
Тропп Э.А., Егоров В.А., Морозов Ю.Г. 2002г. Математические методы для интеллектуальных баз данных в биологии. 4. Математические модели экологических систем //*Рус. орнитол. журн.* Экспресс-вып. **201**: 951-966.
Хампель Ф., Рончетти Э., Рауссеу П., Штаэль В. 1989. *Робастность в статистике.* М.: 1-512.
Хьюбер П. 1984. *Робастность в статистике.* М.: 1-303.
Шеффе Г. 1963. *Дисперсионный анализ.* М.: 1-625.
Bachelier L. 1900. Theorie de la speculation //*Annals scientifiques de l'Ecole Normale Superieure* **17**: 21-86.

ISSN 0869-4362
*Русский орнитологический журнал 2002, Экспресс-выпуск **205**: 1100-1102*

О *Riparia riparia dolgushini* Gavrilov et Savchenko, 1991

Э.И.Гаврилов

Центр мечения животных, Институт зоологии МОиН РК, Казахстан

Поступила в редакцию 9 декабря 2002

Проведя ревизию азиатских береговых ласточек *Riparia riparia* (Linnaeus, 1758) и *Riparia diluta* (Scharp et Wyatt, 1893), В.М.Лоскот и Э.Диккинсон (Loskot, Dickinson 2001) свели *R. riparia dolgushini* Gavrilov et Savchenko, 1991 в синоним *R. riparia innominata* Zarudny, 1916. Это заключение я считаю глубоко ошибочным, поскольку авторы не обратили внимания на происхождение коллекционных материалов Н.А.Зарудного (1916) и приводимые им размеры птиц.

В своей работе Н.А.Зарудный анализирует три формы береговых ласточек из Русского Туркестана.

1. *Riparia riparia riparia*, "которая обыкновенна на пролётах и на гнездовье (Аральское море, низовые части Сыр-дарьи и Аму-дарьи)" (с. 30). В

приводимом им списке мест добычи коллекционных экземпляров фигурируют с. Бугунь и Ташкент. Для юго-восточного Казахстана (Семиречья) приводится *Cotile riparia* (= *R. riparia*), без указания подвида (Зарудный, Кореев 1905), или только *R. riparia diluta* (Шнитников 1949). Точка зрения, что на юго-востоке Казахстана гнездится *R. riparia diluta*, а на остальной территории Казахстана — *R. riparia riparia*, существовала до последнего времени (Бородихин 1970). Трудно сказать, почему южную береговую ласточку, относящуюся к виду *R. riparia*, не указывали для юго-восточного Казахстана. Может быть, это результат недостатка коллекционных материалов из этих мест, или следствие её расселения из южного Казахстана в восточном направлении. Во всяком случае, в 1970-1980-е годы она была многочисленна на гнездовье в окрестностях Алма-Аты и в Алакольской котловине. Южная граница *R. riparia riparia* проводится через Закавказье, долины Мургаба и Теджена (Мекленбурцев 1954). Таким образом, совершенно очевидно, что коллекционные экземпляры Н.А.Зарудного, которые он отнёс к номинативному подвиду *R. riparia riparia*, происходят из районов обитания *R. riparia dolgushini.*

2. *Riparia riparia diluta*, по Н.А.Зарудному (1916, с. 34), "в большом числе гнездится в Бухарских владениях. По-видимому, главным образом она же выводится в бассейне среднего течения Сыр-дарьи. На Аральском море, а также в дельтах и самых низовых частях Сыр-дарьи и Аму-дарьи всецело заменена типичной формой". Сейчас мы знаем, что эта форма гнездится шире — в юго-восточном Казахстане к северу по крайней мере до Алакольской котловины. Расхождений здесь нет. При её описании, которое изложено очень кратко, Н.А.Зарудный ни разу не возразил Р.Шарпу, что его экземпляры отличаются от описанного им подвида. Отсюда следует, что эта форма — настоящая *diluta* в современном понимании.

3. *Riparia riparia plumipes*, по Н.А.Зарудному (с. 34), "во множестве встречается на пролётах в восточной части Сыр-дарьинской области и в прилегающих сюда частях областей Ферганской и Самаркандской. Вполне возможно её гнездование в Ферганской области". Это полностью совпадает с ареалом предыдущей формы *diluta.* Приводимый им диагноз также укладывается в рамки индивидуальной изменчивости *diluta*, причём, по моим наблюдениям, некоторые экземпляры имеют очень светлую перевязь на груди с сильно размытыми краями, которая практически не отличается от имеющейся у *Riparia paludicola* (Vieillot, 1817). Единственное различие — обильное оперение задней и внутренней сторон плюсны, которое у *diluta* и *riparia* выражено лишь маленьким пучком перьев над основанием заднего пальца. Я считаю, что пёрышки на цевке у береговой ласточки — наследие от предковой формы, имевшей оперённую плюсну.

Рассмотрим размеры береговых ласточек (см. таблицу). Несмотря на то, что Н.А.Зарудный измерял длину крыла по хорде, а мы — на выпрямленном крыле, средняя длина крыла у *R. riparia riparia* и *R. riparia dolgushini* практически одинакова, различия статистически незначимы. В то же время *R. riparia riparia* значительно крупнее, чем *R. riparia diluta* или *R. riparia plumipes* ($P < 0.01$). Последние две формы по длине крыла почти не отличаются от *R. diluta diluta.*

Длина крыла (мм) береговых ласточек (коллекционные экземпляры)

Южный Казахстан (Зарудный 1916)				Юго-восточный Казахстан (Гаврилов, Савченко 1991)			
Форма	lim	$X \pm S.E.$	*n*	Форма	lim	$X \pm S.E.$	*n*
R. r. riparia	102-112.5	106.3±0.46	36	*R. r. dolgushini*	103-109	106±0.73	10
R. r. diluta	95-106	100.5±0.79	17	*R. diluta diluta*	95-103	99.7±0.94	10
R. r. plumipes	97-104.6	100.2±0.37	24				

Таким образом, изложенное показывает, что *R. riparia dolgushini* = *R. riparia riparia* у Зарудного, а *R. diluta diluta* = *R. riparia diluta* = *R. riparia plumipes*. Ошибка В.М.Лоскота и Э.Диккинсона как раз и заключается в том, что они свели *R. riparia dolgushini* в синоним *R. riparia diluta* (= *innominata*). В заключение замечу, что Н.А.Зарудный прекрасно знал птиц и замечал малейшие особенности их окраски. Именно поэтому многие из описанных им подвидов отошли в синонимы. Даже сама мысль о том, что он мог спутать *riparia* с *diluta*, мне кажется кощунственной.

Литература

Бородихин И.Ф. 1970. Семейство ласточковые //*Птицы Казахстана*. Алма-Ата, **3**: 161-193.

Гаврилов Э.И., Савченко А.П. 1991. О видовой самостоятельности бледной ласточки (*Riparia diluta* Sharpe et Wyatt, 1893) //*Бюл. МОИП*. Отд. биол. **96**, 4: 34-44.

Зарудный Н.А. 1916. О некоторых ласточках из Русского Туркестана //*Орнитол. вестн.* 1: 25-38.

Мекленбурцев Р.Н. 1954. Семейство ласточковые //*Птицы Советского Союза*. М., **6**: 685-750.

Loskot V.M., Dickinson E.C. 2001. Nomenclatural issues concerning the common sand martin *Riparia riparia* (Linnaeus, 1758) and the pale sand martin *R. diluta* (Sharpe & Wyatt, 1893), with a new synonymy //*Zool. Verh.* Leiden: 167-173.

ISSN 0869-4362
*Русский орнитологический журнал 2002, Экспресс-выпуск **205**: 1102-1104*

О гнездовании погоныша *Porzana porzana* в Калининграде

Е.Л.Лыков

Кафедра экологии и зоологии, факультет биоэкологии, Калининградский университет, ул. Университетская, 2, Калининград, 236040, Россия. E-mail: elykov@mail.ru

Поступила в редакцию 15 декабря 2002

В Калининградской области погоныш *Porzana porzana* в настоящее время распространён спорадично. Сохраняются тенденции к точечной локализации мест гнездования (даже на обширных заболоченных территориях) и

сокращению численности (Гришанов 1994). В 1880-е погоныш гнездился на Верхнем пруду в Кёнигсберге (ныне Калининград), но в первую половину XX века исчез (Tischler 1941). Он был снова обнаружен в Калининграде в гнездовой период во время проведения наблюдений для составления атласа гнездящихся птиц Калининграда (1991-1995), на этот раз в междуречье Старой и Новой Преголи (Гришанов 1999). Однако гнёзд погоныша в послевоенное время в городе не находили.

На участке между посёлком Первомайский и улицей Гайдара (Ленинградский р-н Калининграда) в 1997-1998 мы нашли два гнезда погоныша, располагавшиеся на заболоченном лугу. В мае 1997 сначала слышали брачные крики самца, а затем обнаружили пару погонышей возле пустого гнезда. В последующие дни гнездовая постройка была затоплена водой.

В 1998 на том же участке первый самец зарегистрирован 22 апреля (по голосу). Через месяц (20 мая) найдено гнездо с полной насиженной кладкой, состоявшей из 13 яиц. Оно располагалось на кочке осоки *Carex acuta* среди воды (глубина воды у гнезда 15 см) в 35 м от места расположения прошлогоднего гнезда. Гнездо было сделано из сухой травы и имело следующие размеры, см: диаметр гнезда 17-18, диаметр лотка 12-13, глубина лотка 3. Окраска яиц: глинисто-желтоватый фон с равномерно покрывающими яйцо расплывчатыми коричневыми и неясными расплывчатыми серыми пятнами, пятнышками и крапинками. Из кладки были изъяты 3 яйца; их размеры, мм: 36.1×24.2, 34.9×23.7, 33.3×23.2, в среднем 34.77×23.70. Судя по срокам вылупления, кладка началась в первой декаде мая. Первый птенец вылупился 27 мая, второй и третий 28 мая, четвертый 30 мая, пятый и шестой 31 мая, седьмой 1 июня, восьмой 2 июня. Таким образом, вылупление первых восьми птенцов происходило в течение 7 сут. С учётом того, что полная кладка состояла из 13 яиц, можно предположить, что время вылупления всех птенцов заняло бы не менее 9 сут. Это значительно больше, чем указывалось другими авторами: 2-3 сут (Спангенберг 1951), 1-5, реже 8 сут (Курочкин, Кошелев 1987). Растянутость вылупления у погоныша связана с тем, что плотное насиживание начинается до завершения полной кладки (Курочкин, Кошелев 1987; Коблик 2001; Рябицев 2001).

Следует отметить, что оба гнезда погоныша располагались в 200-220 м от многоэтажного жилого дома. Примыкающую территорию жители интенсивно используют для прогулок и выгула собак. В последующие годы погоныш здесь ни на пролёте, ни на гнездовании не обнаружен.

Литература

Гришанов Г.В. 1994. Гнездящиеся птицы Калининградской области: территориальное размещение и динамика численности в XIX-XX вв. I. Non-Passeriformes //*Рус. орнитол. журн.* **3**, 1: 83-116.

Гришанов Г.В. 1999. Орнитофаунистическая карта г. Калининграда //*Экологический атлас Калининграда.* Калининград.

Коблик Е.А. 2001. *Разнообразие птиц (по материалам экспозиции Зоологического музея МГУ).* М., **2**: 1-400.

Курочкин Е.Н., Кошелев А.И. 1987. Погоныш //*Птицы СССР: Курообразные. Журавлеобразные.* Л.: 389-400.

Рябицев В.К. 2001. *Птицы Урала, Приуралья и Западной Сибири: Справочник-определитель.* Екатеринбург: 1-608.

Спангенберг Е.П. 1954. Погоныш //*Птицы Советского Союза.* М., **5**: 663-671.
Tischler F. 1941. *Die Vogel Ostpreussens und seiner Nachbargebiete.* Konigsberg; Berlin, **1/2**: 1-1304.

ISSN 0869-4362
*Русский орнитологический журнал 2002, Экспресс-выпуск **205**: 1104-1106*

Биология малой мухоловки *Siphia parva* в юго-западной части Литвы

А.Алексонис

*Второе издание. Первая публикация в 1976**

Малая мухоловка *Siphia parva* до настоящего времени остаётся практически неизученным видом птиц Литвы. Это обстоятельство и побудило меня специально заняться изучением биологии этой птицы.

Исследования проведены в 1957-1975 годах преимущественно в лесном массиве Рудшилис (Шакяйский р-н). Здесь, в т.н. Северной Судуве (левобережье Нямунаса), малая мухоловка обитает в смешанных насаждениях ели, чёрной ольхи и берёзы: по берегам родников, поросших елью и сосной. Иногда встречается по берегам речек, протекающих по окраинам лесного массива, поросших серой ольхой и черёмухой, стволы которых часто обвиты хмелем.

В лесу Рудшилис на площади 100 га смешанного леса из ели, чёрной ольхи, берёзы и сосны гнездится в среднем 6 пар малой мухоловки. Здесь же в 1960 обнаружены даже 4 гнезда этой птицы на участке площадью примерно 15 га, причём расстояние между ближайшими гнёздами не превышало 70 м.

Весенний прилёт малых мухоловок наблюдается в первой половине мая (6-13, в среднем 9 мая). Улетают они, как правило, уже в середине сентября; лишь в 1975 одна птица была отмечена 8 октября.

Строительство гнёзд начинается в конце мая и наблюдается в течение всего июня. Гнёзда мухоловки устраивают на дне разрушенных дупел, в щелях стволов деревьев, за отставшей корой, иногда прямо среди ветвей. В лесах Северной Судувы из 115 найденных гнёзд малой мухоловки 17 располагались на "одноствольных" деревьях, 6 — на "двуствольных" деревьях, 28 (24.3%) — на деревьях с повреждённым стволом, 4 — на сухих деревьях, 57 (49.5%) — на деревьях с обломанной вершиной, 2 — на ветвях поваленных деревьев, 1 — в коряге. По породам деревьев эти 115 гнёзд распределялись следующим образом: на ели *Picea abies* — 22, на сосне *Pinus sylvestris* — 11, между молодыми елями и стволами других — 5, на чёрной

* Алексонис А. 1976. Биология малой мухоловки в юго-западной части Литвы //*Материалы 9-й Прибалт. орнитол. конф.* Вильнюс: 3-6.

ольхе *Alnus glutinosa* — 61, на серой ольхе *Alnus incana* — 2, на берёзе *Betula pendula* — 12, на осине *Populus tremula* — 1, на бересте *Ulmus foliacea* — 1 гнездо. Гнёзда располагались на разных высотах: от самой земли до 13 м. Больше всего гнёзд было найдено на высоте 1.5-2.0 м (табл. 1).

Таблица 1. Высота расположения гнёзд малой мухоловки

Высота над землёй, м	Число гнёзд	%
ниже 0.5	1	0.8
0.5-1.0	20	17.4
1.1-1.5	21	18.3
1.6-2.0	24	20.9
2.1-3.0	17	14.8
3.1-4.0	11	9.6
4.1-5.0	8	7.0
5.1-6.0	5	4.3
6.1-7.0	1	0.8
7.1-8.0	3	2.6
8.1-9.0	1	0.9
11.1-12.0	2	1.7
12.1-13.0	1	0.9

Гнёзда малые мухоловки строят из мха, выстилая изнутри волосом (например, косули *Capreolus capreolus*), тонкими корешками и даже кусочками фольги. Размеры гнёзд, см: наружный диаметр 7-10, диаметр лотка 4.5-5.5, глубина лотка 3.0-3.8. Бывают случаи, когда на одном и том же дереве мухоловки гнездятся 3-4 года.

Величина кладки от 4 до 7, в среднем 5.0 яиц (n = 50). 4 яйца было в 11 гнёздах (22%), 5 — в 26 (52%), 6 — в 10 (20%) и 7 — в 3 гнёздах (6%).

Начало откладки яиц у малых мухоловок (115 гнёзд) по пятидневкам распределяется следующим образом (табл. 2). Наиболее ранняя дата появления первого яйца — 21 мая 1971, а самая поздняя — 4 июня 1975. Большая часть кладок появляется в последней декаде мая - первой декаде июня. После гибели кладки птицы обычно гнездятся повторно, устраивая новое гнездо. Бывают, правда, случаи, когда после уничтожения кладки первое яйцо новой кладки появляется в том же гнезде. Хотя малая мухоловка обычно выводит только один выводок в сезон, но в случае гибели первых гнёзд наблюдаются относительно поздние повторные кладки.

Насиживание начинается с откладки предпоследнего яйца и продолжается 14-15 дней. С откладки первого яйца до вылупления птенцов проходит 16-18 дней, причём продолжительность насиживания прямо пропорциональна величине кладки: когда в гнезде 4 яйца — 16 дней, при 5 — 17 и при 6 яйцах — 18 дней. В период насиживания самку кормит самец. Он предупреждает её голосом и об опасности.

Птенцы вылупляются не одновременно. Последний как правило вылупляется более чем на сутки позже первых. В выводке обычно 4 или 5 птенцов, в среднем 4.1 птенца (n = 36). Распределение числа птенцов в выводке: 1 птенец — 1 выводок (2.8%), 2 — 4 (11.1%), 3 — 4 (11.1%), 4 — 12 (33.3%), 5 — 12 (33.3%) и 6 птенцов — 3 выводка (8.4%).

Таблица 2. Сроки появления первого яйца в гнёздах малой мухоловки

Пятидневка	Число гнёзд	%
21-25 V	6	5.2
26-30 V	18	15.7
31 V -4 VI	9	7.8
5-9 VI	24	20.9
15-19 VI	19	16.5
20-24 VI	15	13.1
25-29 VI	9	7.8
30 VI - 4 VII	3	2.6

В 5-сут возрасте вес птенцов составляет 5.6-5.8 г, в 7-8-сут — 10 г. На 14-15-е сут птенцы покидают гнездо. Ранние выводки вылетают в конце июня, поздние — в последние дни июля (табл. 3).

Хотя к 10-му дню после оставления гнезда слётки сами начинают отлавливать насекомых,

Таблица 3. Сроки сезонных явлений у малой мухоловки

Год	Появление весной	Начало кладки	Вылет птенцов		Самая поздняя встреча осенью
			Начало	Конец	
1957	—	1 июня	2 июля	—	—
1959	9 мая	(4 июня)	(6 июля)	—	1 сентября
1960	13 мая	—	—	1 августа	19 сентября
1961	—	—	—	25 июля	9 сентября
1962	—	—	—	—	19 сентября
1963	7 мая	(4 июня)	(5 июля)	29 июля	4 сентября
1964	—	30 мая	30 июня	26 июля	27 сентября
1965	—	(8 июня)	(10 июля)	—	15 сентября
1966	8 мая	27 мая	28 июня	27 июля	19 сентября
1967	12 мая	27 мая	27 июня	26 июля	6 сентября
1968	8 мая	25 мая	26 июня	—	3 сентября
1969	7 мая	—	—	—	8 сентября
1971	9 мая	21 мая	26 июня	—	12 сентября
1972	8 мая	25 мая	25 июня	—	—
1973	11 мая	23 мая	27 июня	22 июля	8 сентября
1974	6 мая	(1 июня)	—	28 июля	11 сентября
1975	11 мая	28 мая	29 июня	4 августа	8 октября
В среднем	9 мая	26 мая	28 июня	28 июля	13 сентября

родители окончательно перестают их подкармливать спустя 15-20 дней после вылета. При наблюдении за гнездом малой мухоловки в лесу Рудшилис 28 июня 1959 установлено, что кормить птенцов они начали в 4 ч 35 мин, а закончили в 22 ч 25 мин. За время наблюдения в течение 9 ч 45 мин родители покормили птенцов 155 раз. За весь "рабочий день", длившийся 17 ч 50 мин, они могли покормить птенцов приблизительно 284 раза.

В Северной Судуве в 1958-1975 из находившихся под наблюдением 110 гнёзд малой мухоловки погибло 54 (49%). От врановых птиц погибли 27, от четвероногих хищников — 17 и от человека — 2 кладки. Из-за болезней погибли 2 выводка птенцов; 6 кладок оказались брошенными. Имеются основания полагать, что в действительности доля разорения гнёзд ещё выше, так как половина наблюдавшихся гнёзд была найдена уже с птенцами.

ISSN 0869-4362
*Русский орнитологический журнал 2002, Экспресс-выпуск **205**: 1107*

Грач *Trypanocorax frugilegus* L. — полный альбинос

Э.В.Шарлеман

*Второе издание. Первая публикация в 1916**

Грач *Trypanocorax frugilegus* — полный альбинос — был подстрелен в июне 1909 близ села Куракина Куракинской волости Орловской губернии. Птицу, как это видно из письма С.Г.Горбачёва, хотели сохранить живою, для чего подрезали ей крылья, но она вскоре околела. Экземпляр этот — молодая птица. Окраска оперения, клюва и лап белая, слегка грязноватая. На вершинной части некоторых перьев лба, темени, спины и плечей заметна примесь кремового цвета. Маховые, рулевые и кроющие крыла палево-белые, слегка буроватые. На рулевых заметна неясная теневая полосатость (такая, как, например, у речной кобылочки *Potamodus fluviatilis* Wolf.). Перья верхних частей (головы, спины и пр.) с интенсивным маслянистым блеском. Низ слабо блестящ, местами даже матовый. Размеры: общая длина (по шкурке) 325, крыло 280, хвост 120, плюсна 50, клюв (длина от лба) 47 мм.

ISSN 0869-4362
*Русский орнитологический журнал 2002, Экспресс-выпуск **205**: 1107*

Новая для России овсянка

С.А.Бутурлин

Второе издание. Первая публикация в 1916†

4 июля 1913 (речка Тумень-ула, устье, Кангуй) А.И.Черский добыл двух овсянок, при проверке сборов оказавшихся взрослым самцом и молодым самцом этого, на первый взгляд, похожего на *Emberiza schoeniclus*, но вполне самостоятельного вида *E. yessoensis* (Swinhoe, 1874), известного из Японии и найденного на Курильских островах, но для наших пределов и, сколько помню, вообще для материка Азии не указанного. Ноги этой птицы светлее, а клюв чернее, чем у *E. schoeniclus pyrrhulinus* Swinh. Зашеек, как и поясница, не беловато-сероватый, но очень рыжий; мелкие же кроющие крыла не рыжи, как у *E. schoeniclus*, но серы, как у *E. pallasi* Cab. У взрослой птицы чёрный цвет ограничивается зобом, не идя на переднюю часть груди.

---

* Шарлеман Э.В. 1916. Грач (*Trypanocorax frugilegus* L.) — полный альбинос //*Орнитол. вестн.* 7, 3: 190-191.

† Бутурлин С.А. 1916. Новая для России овсянка //*Орнитол. вестн.* 7, 2: 103.