



УДК 004.93:519.2

## НОВАЯ КОНЦЕПЦИЯ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ИНФОРМАЦИИ НА ОСНОВЕ ПРОГНОСТИЧЕСКОЙ ФУНКЦИИ ТЕОРИИ ВЕРОЯТНОСТЕЙ

### THE NEW CONCEPT OF THE SOFTWARE OF STATISTICAL INFORMATION PROCESSING ON THE BASIS OF PREDICTIVE FUNCTION OF PROBABILITY THEORY

**В.В. Савченко**  
**V.V. Savchenko**

*Нижегородский государственный лингвистический университет, Россия, 603155, Нижний Новгород, ул. Минина, 31-а  
Nizhny Novgorod state linguistic university, 31-a Minin St, Nizhny Novgorod, 603155, Russia*

*e-mail: svv@lunn.ru*

*Аннотация.* На основе прогностической функции теории вероятностей предложена новая концепция учебного курса в расчете на самую широкую категорию пользователей современного программного обеспечения статистической обработки информации, в том числе, в области социально-экономических и гуманитарных исследований. В статье дан анализ недостатков традиционного подхода к изучению теории вероятностей и определены уточненные понятия ее предмета, цели и задач. При этом особое внимание уделено прикладному значению статистических методов.

*Resume.* On the basis of predictive function of probability theory the new concept of a training course counting on the broadest category of users of the modern software of statistical information processing, including, in the field of social and economic and humanitarian researches is offered. In article the analysis of shortcomings of traditional approach to studying of probability theory is given and the specified concepts of its subject, the purpose and tasks are defined. Thus the special attention is paid to applied value of statistical methods.

*Ключевые слова:* статистическая обработка информации, теоретико-вероятностный подход, прогностическая функция теории вероятностей.

*Keywords:* statistical information processing, probability-theoretic approach, predictive function of probability theory.

В связи с повсеместным распространением информационных технологий в мире все шире проникают в самые разные сферы человеческой деятельности математические методы обработки информации и, прежде всего, статистические методы, в том числе в области социально-экономического и гуманитарного знания. В этой связи на практике явно обозначился определенный разрыв между потребностями исследователей в универсальном математическом аппарате, с одной стороны, и их ограниченными, часто интуитивными представлениями о его возможностях и методике применения, с другой. И этим сильно ограничивается эффективность современного программного обеспечения. Для преодоления указанного противоречия должны быть внесены коррективы в традиционную концепцию, а также в структуру учебного курса современной теории вероятностей (ТВ), усилив роль физического подхода ко всем ключевым понятиям и определениям. Решению данной задачи и посвящена настоящая статья.

В Большой советской энциклопедии ([1], см. по ссылке <http://www.big-soviet.ru/99/14680/%D0%92%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C>) «вероятность» определяется как «характеристика степени объективной возможности определённого события в будущем». И там же уточняется: «окончательное познавательное значение для ТВ ... имеют лишь вероятности со значениями в непосредственной близости к единице». БСЭ вторит автор современного учебного пособия: «Цель ТВ – предсказать, появится ли изучаемое событие при осуществлении некоторого наперед заданного комплекса факторов (явлений-причин)» [2, с.12]. Правда, чуть ниже он же утверждает и прямо противоположное: «предсказать результат единичного испытания можно лишь для достоверных или невозможных событий» [2, с.14], явно путая при этом задачу прогнозирования с предсказанием. Предсказание – это, скорее, из области астрологии. К науке оно отношения не имеет. В результате несложных рассуждений приходим к неизбежному выводу: задача прогнозирования вероятного развития наблюдаемых явлений в будущем – центральная задача современной ТВ. По этому поводу в конспекте лекций [3, с. 12] сказано следующее: «Выявление и изучение статистических закономерностей и ... прогнозирование их дальнейшего развития – главная цель ТВ».



Таким образом, неубывающий интерес человечества к своему будущему – главный мотив существующего в мире спроса на математический аппарат ТВ.

Математическую основу современной ТВ образует абстрактная булева алгебра, точнее, алгебра событий [2, с. 19]. События  $A, B, \dots, C$  – универсальная форма качественного описания вероятного развития (поведения) явления в будущем: состоялось или не состоялось. При этом норма

$P(A)$  события  $A$  со свойствами

А)  $0 \leq P(A) \leq 1$ ;

Б)  $P(\bar{A}) = 1 - P(A)$ ,

В)  $P(A+B) = P(A)+P(B)$  при  $AB = O$ ,

где  $\bar{A}$  – противоположное событие, в ТВ [1] получила название «вероятности события». И вместе с этим названием приобрела свой ясный физический смысл как объективной меры степени возможности каждого конкретного события в будущем. При равенстве  $P(A)=1$  здесь по традиции можно говорить о достоверном событии, при  $P(A)=0$  – о невозможном событии. Но в промежуточном случае  $0 < P(A) < 1$  мы будем говорить о вероятном (в будущем) событии  $A$ . Это и есть истинный объект исследования в рамках современной ТВ. А ее предметом исследования служат статистические (т.е. основанные на выборке в ретроспективе) методы прогнозирования вероятных событий в будущем или, в качестве синонима, вероятностные методы. Понятие «случайное событие», как и «случайное явление», из теории исключаются как атавизмы. И вслед за этим утрачивается всякий практический смысл понятие массового случайного явления.

Основу строгого, статистического определения вероятности события  $A$  в будущем

$$P(A) = \lim_{n \rightarrow \infty} \left( \frac{m_A(n)}{n} \right) \tag{1}$$

образует фундаментальная гипотеза [4, с. 28] об идеальной устойчивости относительной частоты  $w_A = m_A(n)/n$  его появления на бесконечно большом временном интервале в ретроспективе. Здесь  $n$  – объем наблюдений или выборки. Понятно, что всякая асимптотика – всегда абстракция. Как и неявное предположение об однородности выборки и независимости наблюдений на бесконечном временном интервале. Поэтому абстрактный характер имеет и представленное определение вероятности. Однако ее физический смысл от этого, отнюдь, не утрачивается. Вероятность, по-прежнему, играет роль объективной меры возможности события  $A$  в будущем. По существу, именно рассмотренное выше выражение определяет в наиболее общем виде строгое решение центральной задачи прогнозирования – в вероятностной или, говорят, статистической (т.е. по выборке в ретроспективе) форме  $P(A)$ . На это, кстати, явно указывает и этимология слова «вероятность» [1].

Сделанный вывод усиливается фундаментальной теоремой Бернулли: в условиях схемы независимых испытаний с двумя возможными исходами  $A$  и  $\bar{A}$  с вероятностями  $P(A)=p$  и  $P(\bar{A})=1-p=q$  соответственно при любом, сколь угодно малом значении константы  $a>0$  справедливо соотношение вида

$$P \left\{ \left| \frac{m_A(n)}{n} - P(A) \right| > a \right\} \leq \frac{pq}{na^2} \leq \frac{0,25}{na^2}. \tag{2}$$

Иными словами, всегда найдется такое достаточно большое значение  $n$ , при котором относительная частота  $w_A$  события  $A$  практически не будет отличаться от его вероятности  $P(A)$ . И при любом конечном объеме выборки  $n < \infty$  константа  $a$  из этого соотношения характеризует точность статистической оценки прогнозирования случайного события  $A$  по формуле его относительной частоты с доверительной вероятностью  $\beta \geq 1 - \frac{0,25}{na^2} > 0$ . Этим дается строгое обоснование статистических прогнозов.

Таким образом, вероятностные методы прогнозирования будущих явлений и событий – суть статистические методы. И, наоборот, говоря о вероятных событиях, мы всегда подразумеваем определенную степень их обусловленности в будущем. Вероятность события – количественная характеристика степени такой обусловленности. При этом, как указывалось выше со ссылкой на БСЭ, практическое значение имеют, прежде всего, близкие к единице значения вероятности – согласно самой логике прогнозирования.



Для сравнения, в вышеупомянутом фундаментальном объемном (576 с.) учебнике профессора Е.С. Вентцель понятие «прогноз» встречается всего лишь 5 раз, причем безотносительно к целям и задачам ТВ, а лишь как иллюстрация ее приложений. Однако в более позднем (2006 г) учебном пособии [5, с. 6-198] оно упоминается уже 35 раз, причем, всего на 193 страницах. Правда, и в этом случае закон больших чисел по традиции представлен лишь в середине учебного курса. Кроме того, не обошлось без подробного освещения элементов комбинаторики и, главное, мирно соседствуют два противоречивых взгляда автора на понятие «вероятность»: «классическое» определение вероятности события по Лапласу (там же, с.16) и ее статистическое – единственно строгое – определение. А это недопустимо. Общеизвестная формула Лапласа – всего лишь равенство, справедливое в очень частном случае построения статистического эксперимента по схеме с равновероятными, несовместными и образующими полную группу событий исходами.

Случайная величина – ключевое понятие ТВ. Традиционно, ее почему-то противопоставляют, явно или неявно, понятию случайного события, как некую «количественную характеристику случайного явления». На наш взгляд, это еще одна методическая ошибка. И первое, и второе понятия – тесно связаны между собой, как связаны, например, понятия «интегральная функция распределения»  $F(x) = P(X < x)$  и «плотность вероятности случайной величины  $X$ »  $f(x) = \frac{dF(x)}{dx}$ . Это всего лишь разные формы математического описания (задания) закона распределения случайной величины. Так и понятие случайной величины – универсальная и весьма продуктивная форма математического описания вероятного события в будущем.

В нашем случае ценность данного понятия еще более возрастает для теории, поскольку возникает в самом начале учебного курса – при доказательстве рассмотренной выше теоремы Бер-

нулли. В самом деле, вводя обозначение случайной величины  $X_i = \begin{cases} 1 \\ 0 \end{cases}$  – с ее равенством 1 при

появлении события  $A$  и 0 – в отсутствии  $A$  в  $i$ -м наблюдении, по результатам  $n$  повторных испытаний будем иметь

$\sum_{i=1}^n X_i = m_A$ ;  $\frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n = \frac{m_A}{n}$ ;  $M(X_i) = M(X) = 1 \cdot p + 0 \cdot q = p$ ;  $D(X_i) = D(X) = pq$ . Здесь симво-

лами  $M(\cdot)$  и  $D(\cdot)$  обозначены математическое ожидание и дисперсия случайной величины соответственно. Их этих выражений со ссылкой на теорему Чебышева

$$P\left\{\left|\bar{X}_n - M(X)\right| > a\right\} \leq \frac{D(X)}{a^2 n} \quad (3)$$

приходим к формулировке теоремы Бернулли (2), что и требовалось доказать.

Особо отметим, что понятие случайной величины возникло в настоящем исследовании практически сразу вслед за понятием вероятности события  $A$ . Причем, одновременно возникла необходимость и в понятиях математического ожидания и дисперсии случайной величины. Все перечисленные понятия объединяются в законе больших чисел (3). И это, подчеркнем, в самом начале учебного курса. Ничего подобного в известных курсах не до сих пор встречается. Однако простая логика подсказывает: здесь кроется глубокий смысл. И он действительно существует – в классическом выражении для вычисления вероятности случайного события (1) через плотность вероятности случайной величины согласно известному выражению  $P(A) = \int_A f(x) dx$  [1]. Здесь инте-

грирование ведется по области определения события  $A$ . При заданной плотности вероятности  $f(x)$  мы этим получили второй вариант статистического решения задачи прогнозирования, который легко обобщается [4, с.159-184] на случай многомерного закона распределения:  $P(A) = \int_A f(\vec{x}) dx_1 dx_2 \dots dx_m$ . Здесь  $\vec{x} = (x_1, x_2, \dots, x_m)$  –  $m$ -мерная переменная величина в качестве век-

торного аргумента многомерной плотности вероятности. Последнее выражение традиционно изучается в рамках завершающего раздела ТВ – теории случайных процессов.

Случайный процесс – математическая модель для развивающихся во времени (динамических) явлений. Благодаря такой модели в ТВ формализуется универсальная схема последовательных наблюдений над сложным, динамическим явлением, причем, возможно с коррелированными отсчетами. В результате ее использования были открыты принципиально новые возможности решения задачи статистического прогнозирования. Так, отталкиваясь от критерия минимума дисперсии погрешности прогнозирования, получено в теории общее выражение для экстраполяции



данных по формуле условного математического ожидания  $\xi_{n+1} = M(x_{n+1} / x_n, x_{n-1}, \dots, x_{n+1-m})$ ,  $n=1, 2, \dots$ ,  $m$ -го порядка. Во многих случаях на практике, в частности, при обработке временных рядов в экономике, такой тип прогноза выглядит предпочтительнее по сравнению с его вероятностной формой общего вида.

Проиллюстрируем сказанное на примере обработки стационарного временного ряда с гауссовским распределением и с обратной автоковариационной матрицей  $K^{-1}$  ленточной структуры. Для этого случая оптимальный прогноз приобретает линейный вид

$$\xi_{n+1} = a_1 x_n + a_2 x_{n-1} + \dots + a_m x_{n+1-m}, \quad (4)$$

где  $(a_1, a_2, \dots, a_m)$  – вектор коэффициентов линейной авторегрессии  $m$ -го порядка. В теории пока-

зано [5, с.57], что с точностью до знака этот вектор определяется первой строкой матрицы  $K^{-1}$ . И по индукции последнее выражение легко преобразуется к рекуррентному виду

$$\xi_{n+k} = \sum_{i=1}^k a_i \xi_{n+k-i} + \sum_{j=k}^q a_j x_{n+k-j}, \quad (5)$$

предназначенному для прогнозирования временного ряда на произвольное число шагов (отсчетов)  $k \geq 1$  в будущее.

Последняя зависимость охватывает (4) как частный случай при равенстве  $k=1$ . Она определяет ближайшую (краткосрочную) перспективу поведения явления в будущем. В качестве примера несколько иного рода можно сослаться на работу автора [6], в которой оптимальная обработка речевого сигнала осуществляется по методу обеляющего фильтра, основанного на авторегрессионной оценке прогнозирования (4), (5) на интервале длительностью в одну минимальную речевую единицу.

В основе рассмотренного примера ключевое значение имел закон распределения случайного процесса. И в этом смысле нельзя обойти вниманием краеугольную проблему ТВ – априорной неопределенности. По существу, именно в ней заключен и смысл, и сама идея статистического подхода. В самом деле, в подавляющем большинстве случаев на практике исследователь не знает априори точного закона распределения случайной величины. Единственный выход для него – восстановление или оценивание закона по выборке данных в ретроспективе. Указанная задача относится к области математической статистики. Таким образом, с позиций прогностической функции ТВ математическая статистика является ее неотъемлемой частью. Причем, именно этот раздел теории, как никакой другой, способен (и должен) адаптировать ее универсальный математический аппарат под конкретную профессиональную подготовку студентов вуза. В результате примерная структура учебного курса примет следующий вид:

1. Аксиоматика нормированной алгебры событий и введение в ТВ;
2. Понятие случайной величины и ее закон распределения вероятностей;
3. Закон больших чисел и введение в математическую статистику;
4. Актуальные (по профилю подготовки специалистов) главы математической статистики;
5. Актуальные главы теории случайных процессов.

Цель первого раздела теории – аксиоматическое введение в понятие «вероятность события в будущем» и его статистическое (строгое) определение. Здесь же излагаются основные теоремы ТВ, формулы полной вероятности, Байеса и другие. Во втором разделе через понятие «возможного в будущем события» вводятся понятия случайной величины и ее закона распределения. Назначение последнего – вычисление математическими методами вероятности возможного события в будущем. Далее определяются все основные статистические моменты вероятностного распределения: от математического ожидания и дисперсии до центрального корреляционного момента. Затем рассматриваются примеры распространенных распределений, подробно изучается нормальный или гауссовский закон, его вероятностные характеристики. Третий раздел теории играет роль своеобразного буфера или введения в проблему априорной неопределенности. Здесь впервые возникает понятие статистической выборки, в частности, репрезентативной. Под предлогом ее минимизации в четвертом разделе излагаются основные формулировки статистических задач, статистические критерии, статистические оценки распределений и наиболее эффективные из статистических методов обработки информации. И, наконец, в заключительном, пятом разделе теории рассмотренные ранее статистические критерии и методы распространяются на многомерный случай в расчете на их дальнейшее применение при анализе динамических процессов в экономике, социологии, лингвистике и в других областях человеческой деятельности.

Таким образом, благодаря проведенному исследованию с акцентом на прогностическую функцию вероятности в представленной статье удалось не только уточнить цель, объект и предмет



исследований в рамках современной ТВ, но и существенным образом видоизменить структуру учебного курса в расчете на широкую аудиторию обучающихся и в соответствии с объективным процессом резкого расширения сферы практических приложений данной научной дисциплины в области информационных технологий. На междисциплинарный характер исследования указывает и то немаловажное обстоятельство, что большинство современных наук, в которых применяются статистические методы, таких как статистическая теория связи, диагностика, социология, лингвистика и другие, по существу использует тот же прогностический потенциал понятия «вероятность события в будущем», что и сама ТВ, поскольку рассчитано на устойчивость своих выводов и оценок в обозримой перспективе. Зачем, иначе, проводить исследование статистическими методами? Например, никто не измеряет площадь местности под новую застройку в вероятностном смысле. А вот показатель качества продукции предприятия в пределах малой выборки наблюдений при определенных условиях [7] распространяют (прогнозируют) на всю партию товара. Также и в социологии. Поэтому и теория статистических оценок, и проверка статистических гипотез с их бесконечными модификациями на практике по существу представляют собой ту или иную разновидность вероятностного подхода к прогнозированию.

### Список литературы References

1. Большая советская энциклопедия: 3-е изд. [Электронный ресурс]. – М.: ЭКСМО, 2008. – URL: <http://www.big-soviet.ru/> (23.12.2014).
2. Bol'shaja sovetskaja jenciklopedija: 3-e izd. [Elektronnyj resurs]. – М.: JeKSMO, 2008. – URL: <http://www.big-soviet.ru/> (23.12.2014).
3. Фирсов А. Н. Теория вероятностей. Часть 1: Уч. пособие. – С.-Пб.: Изд-во С.-Пб. гос. ун-та, 2005. – 112 с.
4. Firsov A. N. Teorija verojatnostej. Chast' 1: Uch. posobie. – S.-Pb.: Izd-vo S.-Pb. gos. un-ta, 2005. – 112 s.
5. Савченко В. В. Теория вероятностей и математическая статистика. Конспект лекций: Уч. пособие. 2-е изд. – Нижний Новгород: Изд-во Нижегород. гос. лингв. ун-та, 2003. – 104 с.
6. Savchenko V. V. Teorija verojatnostej i matematicheskaja statistika. Konspekt lekcij: Uch. posobie. 2-e izd. – Nizhnij Novgorod: Izd-vo Nizhegorod. gos. lingv. un-ta, 2003. – 104 s.
7. Вентцель Е. С. Теория вероятностей: Учеб. для вузов. – 5-е изд. – М.: Высшая школа, 1998. – 576 с.
8. Ventcel' E. S. Teorija verojatnostej: Ucheb. dlja vuzov. – 5-e izd. – M.: Vysshaja shkola, 1998. – 576 s.
9. Савченко В. В. Информационная теория колебаний биржевых котировок в динамике // Информационные технологии. – 2011. – № 3. – С. 57-63.
10. Savchenko V. V. Informacionnaja teorija kolebanij birzhevyh kotirovok v dinamike // Informacionnye tehnologii. – 2011. – № 3. – S. 57-63.
11. 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.
12. Savchenko V.V., Vasil'ev R.A. Analiz jemocional'nogo sostojanija diktora po golosu na osnove foneticheskogo detektora lzhi // Nauchnye vedomosti BelGU: Serija «Istorija. Politologija. Jekonomika. Informatika». 2014. № 21 (192). Vyp. 32/1. S. 186-195.
13. Савченко В.В. Определение объема контрольной выборки в условиях априорной неопределенности по принципу гарантированного результата // Научные ведомости БелГУ: Серия «История. Политология. Экономика. Информатика». 2015. № 212(193). Вып. 33/1. С. 184-192.
14. Savchenko V.V. Opredelenie ob#ema kontrol'noj vyborki v uslovijah apriornoj neopredelen-nosti po principu garantirovannogo rezul'tata // Nauchnye vedomosti BelGU: Serija «Istorija. Politologija. Jekonomika. Informatika». 2015. № 212(193). Vyp. 33/1. S. 184-192.
15. Жилияков 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.
16. Zhilyakov E.G., Belov S.P. Obnaruzhenie zvukov rechi na fone шумов // Nauchnye vedomosti Bel-GU: Serija «Istorija. Politologija. Jekonomika. Informatika». 2012. T. 22. № 7-1. S. 182-189.
17. Лисьев В. П. Теория вероятностей и математическая статистика: Уч. пособие. – М.: Изд-во Моск. гос. ун-та экономики, статистики и информатики, 2006. – 199 с.
18. Lis'ev V. P. Teorija verojatnostej i matematicheskaja statistika: Uch. posobie. – M.: Izd-vo Mosk. gos. un-ta jekonomiki, statistiki i informatiki, 2006. – 199 s.