Программное обеспечение оценки и анализа показателей выживаемости и смертности

Долматкин А.Н.
Факультет информатики и робототехники
Уфимский государственный авиационный технический университет
Уфа, Россия
flitenym@gmail.com

Прокудина Е.И.
Факультет информатики и робототехники
Уфимский государственный авиационный технический университет
Уфа, Россия
preliv@gmail.com

Аннотация[1]

В статье для анализа выживаемости применены аналитические модели смертности. Разработано программное обеспечение, реализующее данный подход. Проведен сравнительный анализ различных моделей смертности для описания продолжительности жизни человека произвольного возраста.

Ключевые слова: программное обеспечение, анализ выживаемости, аналитические модели смертности.

1. Введение

Для анализа выживаемости, моделирования процессов наступления терминальных событий используются разнообразные методы, описанные, в частности, в [1]. Не вызывает сомнений актуальность применения и, в случае необходимости, разработки адекватного специализированного программного обеспечения. Также важен вопрос о том, насколько определенный метод анализа выживаемости подходит для решения конкретной задачи.

Исследуем целесообразность применения одного из параметрических методов анализа выживаемости – аналитических моделей смертности для описания продолжительности жизни человека.

2. Постановка задачи

Необходимо разработать программное обеспечение, реализующее использование аналитических моделей смертности для описания продолжительности жизни человека произвольного возраста, позволяющее выполнять оценку качества данных моделей, оценивать на основе статистических данных различные функциональные и числовые характеристики продолжительности жизни,

Труды **Восьмой всероссийской научной конференции "Информационные технологии интеллектуальной поддержки принятия решений", 26-29 мая, Уфа-Ставрополь, Ханты-Мансийск, Россия, 2020**

2. Математические методы

Предполагается, что продолжительность жизни человека представляет собой непрерывную случайную величину, для описания которой используются различные известные распределения. Параметры распределений можно оценить на основе статистических данных с помощью метода максимального правдоподобия [2].

В качестве исходных данных используется массив из n времен жизни: $t_i,\ i = \overline{1.n}.$ Цензурирование данных в работе не учитывается.

Для четырех выбранных распределений (экспоненциального, Рэлея, Гомпертца, Вейбулла) в таблице 1 представлены: функция выживания $s(x)$ (вероятность, что новорожденный доживет до возраста x), на основе которой можно вычислить различные показатели выживаемости и смертности, и плотность распределения $f(x)$, которая также называется кривой смертей. На основе функции выживания $s(x)$ и кривой смертей $f(x)$ легко получить формулы для функции $f(x)/s(x)$, которая называется интенсивностью смертности или функцией риска.

Таблица 1. Характеристики распределений

	$s(x)$	$f(x)$
Эксп.	$e^{-\lambda x},\ x \geq 0,\ \lambda > 0$	$\lambda e^{-\lambda x}$
Рэлея	$1 - e^{-\frac{x^2}{2\beta^2}},\ \ x \geq 0,\ \beta > 0$	$\frac{x}{\beta^2} e^{-\frac{x^2}{2\beta^2}}$
Комп.	$e^{\beta\left(1-e^{\alpha x}\right)/\alpha},\ x \geq 0,\ \beta > 0,\ \alpha \neq 0$	$\beta e^{\alpha x + \beta\left(1-e^{\alpha x}\right)/\alpha}$
Вейб.	$e^{-\lambda x^{\gamma}},\ x \geq 0,\ \lambda > 0\ \ \gamma > 0$	$\lambda\gamma\gamma^{\gamma-1} e^{-\lambda x^{\gamma}}$

Формулы для вычисления параметров распределений приведены в таблице 2. Чтобы найти параметры α и γ для распределений Гомпертца и Вейбулла соответственно необходимо решить указанные в таблице нелинейные уравнения.

Таблица 2. Параметры распределений

	Параметр 1	Параметр 2
Эксп.	$\lambda = n / \sum_{i=1}^{n} t_i$	–
Рэлея	$\beta = 4n^2 / \left(\sum_{i=1}^{n} t_i^2 \right)^2$	–
Гомп.	$\alpha \sum_{i=1}^{n} t_i - n \frac{(1-\alpha^2) \sum_{i=1}^{n} e^{\alpha t_i} + n}{\sum_{i=1}^{n} e^{\alpha t_i} - n} = 0$	$\beta = \frac{\alpha n}{\sum_{i=1}^{n} e^{\alpha t_i} - n}$
Вейб.	$\frac{n}{\gamma} + \sum_{i=1}^{n} \ln t_i - \frac{n \sum_{i=1}^{n} t_i^{\gamma} \ln t_i}{\sum_{i=1}^{n} t_i^{\gamma}} = 0$	$\lambda = \frac{n}{\sum_{i=1}^{n} t_i^{\gamma}}$

Вычислить логарифм функции максимального правдоподобия $\ln L$ для рассматриваемых распределений можно по формулам, приведенным в таблице 3.

Таблица 3. Логарифм функции максимального правдоподобия

	$\ln L$
Эксп.	$\ln\lambda - \lambda / \sum_{i=1}^{n} t_i$
Рэлея	$\sum_{i=1}^{n} \ln \frac{t_i}{\beta^2} - \frac{1}{2\beta^2} \sum_{i=1}^{n} t_i^2$
Гомп.	$n \ln\beta + \alpha \sum_{i=1}^{n} t_i + \frac{\beta}{\alpha} \left(n - \sum_{i=1}^{n} e^{\alpha t_i} \right)$
Вейб.	$n \ln(\lambda\gamma) + (\gamma - 1) \sum_{i=1}^{n} \ln t_i - \lambda \sum_{i=1}^{n} t_i^{\gamma}$

Выбор распределения, наилучшим образом описывающего исходные данные, осуществляется на основе информационного критерия Акаике. Лучшим считается распределение с минимальным значением функции [3,4]:

$$AIC = -2\ln L^* + 2k + \frac{2k(k+1)}{n-k-1},$$

где $\ln L^*$ – оценка логарифма функции максимального правдоподобия при найденных значениях параметров распределения. k – число параметров распределения.

3. Программное обеспечение и вычислительный эксперимент

Разработано программное обеспечение, позволяющее на основе статистических данных находить параметры распределений, оценивать показатели выживаемости и смертности. Программа написана на платформе .NET и технологии WPF на языках C# и XAML с использованием паттерна MVVM (Model-View-ViewModel) и TAP (Task-based asynchronous programming), также реализована связь между ViewModel через общий класс Singleton. В программе реализованы CRUD (create, read, update, delete) – операции с локальной базой данных SQLite посредством библиотеки Dapper. Загрузка данных из Excel выполняется через IronXL. Графики отображаются с помощью LiveCharts. Для визуализации компонентов применяются библиотеки MaterailDesign, Dragablz и MahApps.

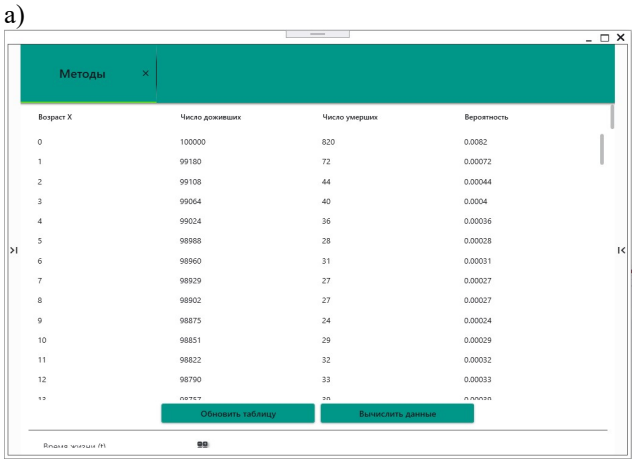
Программное обеспечение можно обновить, используя ссылку, которая содержит zip-архив с новой версией (см. рисунок 1).

Рис. 1. Пример работы программного обеспечения

Исходные данные для вычислительного эксперимента (массив времен жизни t_i, $i = \overline{1.n}$) (см. рисунок 2а), сгенерированы на основе данных о количестве живых в каждом возрасте из таблицы смертности для мужского населения России за 2014 год [5].

На основе формул из таблицы 2 найдены значения параметров моделей. Аналогичные вычисления проведены для исходных данных (см. рисунок 2б), полученных на основе таблицы смертности для мужского населения России за 2001 год

а)



б)

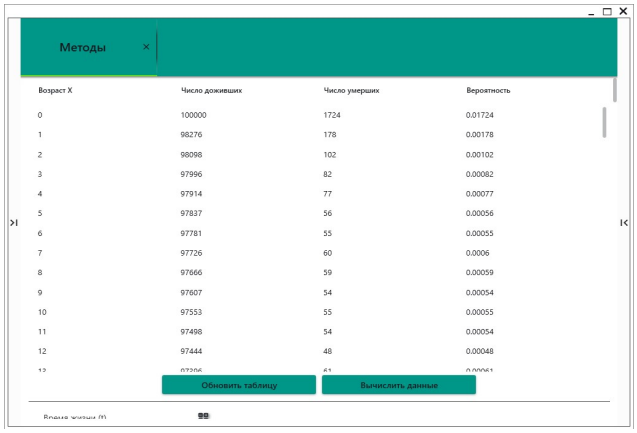
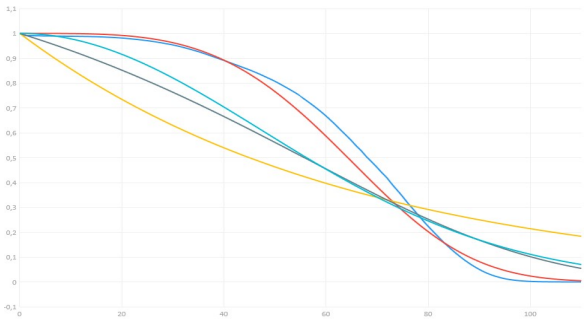


Рис. 2. Исходные данные для 2014 (а) и 2001 (б) годов соответственно

На рисунке 3 представлены графики функции выживания, соответствующей табличным данным, и функций выживания для каждой из 4 рассматриваемых моделей.

Как видно из рисунка 3, наиболее близко друг к другу графики всех функций находятся в интервале возрастов от 70 до 80 лет, т.е. вероятности дожития до возраста из данного интервала для рассматриваемых распределений близки. Для данных 2001 года аналогичный интервал возрастов немного смещен влево.

а)



б)

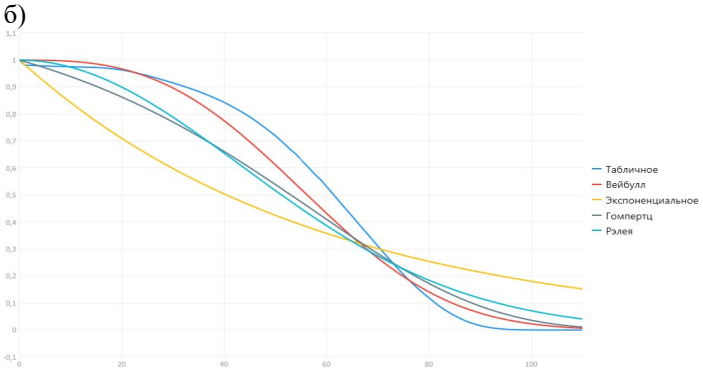
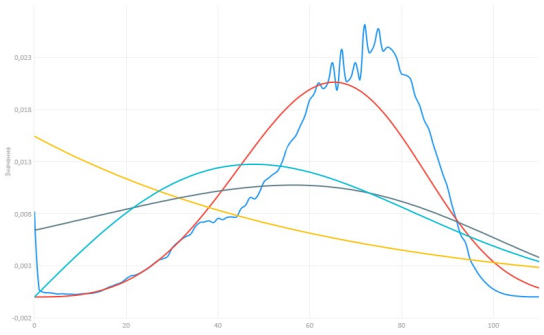


Рис. 3. Функция выживания $s(x)$ для исходных данных 2014 (а) и 2001(б) годов соответственно

На рисунке 4 приведены графики плотности распределения $f(x) \approx s(x) - s(x+1)$ (кривой смертей) для табличных данных и для каждой из 4 рассматриваемых моделей.

а)



б)

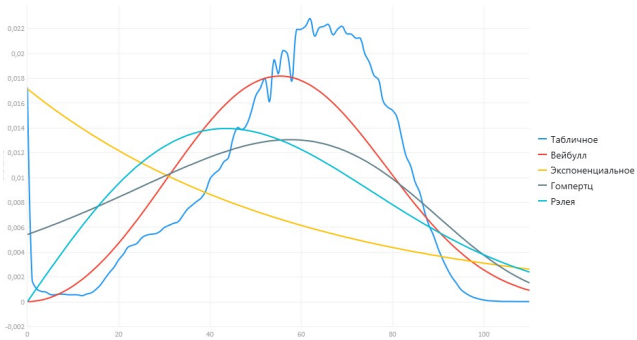
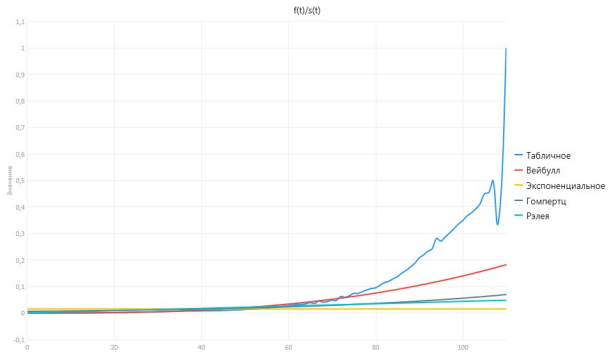


Рис. 4. Плотность распределения $f(x)$ для исходных данных 2014 (а) и 2001 (б) годов соответственно

График плотности распределения имеет две точки максимума, первая соответствует младенческой смертности, вторая примерно среднему времени жизни: 65,56 лет для данных 2014 года и 58,89 лет для данных 2001 года. Лучшее приближение к табличным данным демонстрирует распределение Вейбулла.

На рисунке 5 представлены графики интенсивности смертности (функции риска) $f(x)/s(x)$ для табличных данных и для каждой из 4 рассматриваемых моделей.

а)



б)

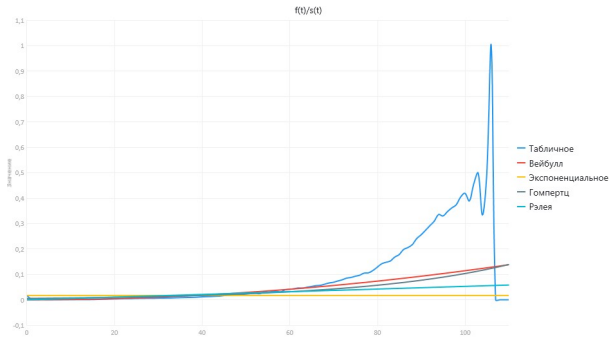


Рис. 5. Интенсивность смертности (функция риска) *f*(*x*)/ *s*(*x*) для исходных данных 2014 (а) и 2001 (б) годов соответственно

Качество моделей оценивается следующим образом: на основе критерия Акаики, а также с помощью двух метрик d_1 и d_2:

$$d_1 = \sum_{x=0}^{\omega} \left| g(x) - g_{таб}(x) \right|,$$

$$d_2 = \sqrt{\sum_{x=0}^{\omega} \left(g(x) - g_{таб}(x) \right)^2},$$

где $g(x)$ – значение функции выживания или плотности распределения для одной из рассматриваемых моделей, $g_{таб}(x)$ – значение соответствующей функции, найденное на основе табличных данных, ω – предельный возраст человека.

Оценка качества (см. рисунок 6) с помощью критерия Акаики, а также двух метрик для функции выживания на данных за 2014 год показывает, что распределение Вейбулла лучшим образом описывает исходные данные.

Оценка	Вейбулла	Рэлея	Гомпертца	Экспоненциальное
Акаики	8.85956	9.30417	9.28777	10.34392
Первая метрика	3.17845	12.52583	14.47885	23.48945
Вторая метрика	0.41569	1.41276	1.58676	2.48613

Рис. 6 Оценка качества моделей для исходных данных 2014 года

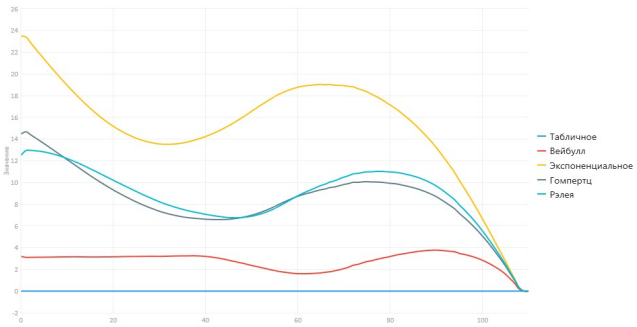
На данных за 2001 год (см. рисунок 7) модель Гомпертца оказалась лучше модели Вейбулла по критерию Акаике, но в данном случае значения критерия отличаются менее, чем на 0,1, в то же время оценка качества моделей с помощью метрик показывает, что лучшим является распределение Вейбулла.

Оценка	Вейбулла	Рэлея	Гомпертца	Экспоненциальное
Акаики	9.10491	9.27678	8.99221	10.13435
Первая метрика	4.6286	10.33119	9.58527	21.27852
Вторая метрика	0.57039	1.18223	1.10901	2.26822

Рис. 7. Оценка качества моделей для исходных данных 2001 года

Разработанное программное обеспечение позволяет получить графики функций выживания, кривой смертей и интенсивности смертности для человека в произвольном возрасте *x*, а также сравнить эти функции для каждой из рассматриваемых моделей с функциями, построенными по табличным данным на основе метрик d_1 и d_2 на заданном временном интервале. На рисунке 8 представлены графики зависимости метрики от возраста для функции выживания при исходных данных 2014 года.

а)



б)

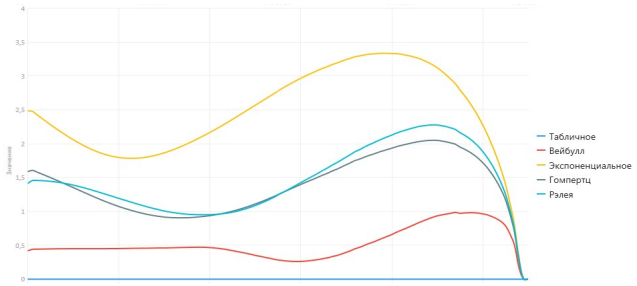


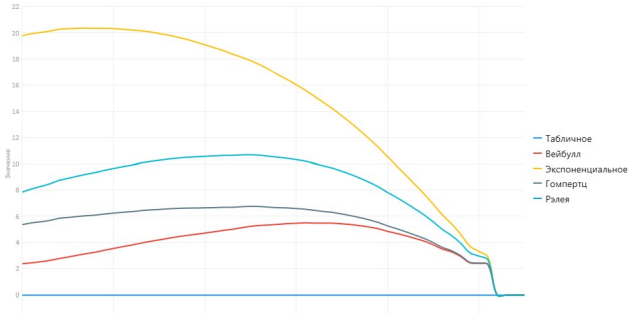
Рис 8. Оценка качества на основе первой (а) и второй (б) метрик для исходных данных 2014 года

Из рисунка 8 видно, что лучше всего исходные данные описывает распределение Вейбулла, причем до 40 лет значения метрик для данного распределения постоянны, принимают минимальное значение для 60 лет, далее увеличиваются. Т.е.

распределение Вейбулла для возрастов более 80 лет описывает исходные данные хуже, чем в интервале от 0 до 80 .

В то же время для исходных данных 2001 года такой особенности нет (см. рисунок 9). Значения метрик увеличиваются с возрастом, т.е. при увеличении возраста человека его остаточная продолжительность жизни описывается рассматриваемыми распределениями менее точно.

а)



б)

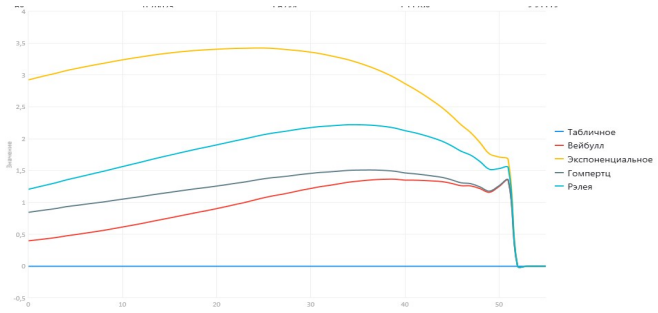


Рис 9. Оценка качества на основе первой (а) и второй (б) метрик для исходных данных 2001 года

4. Заключение

- Разработано программное обеспечение, реализующее один из методов анализа выживаемости – применение аналитических моделей смертности.
- Проведено исследование целесообразности применения распределений Вейбулла, Рэлея, экспоненциального и Гомпертца к описанию продолжительности жизни человека. В вычислительном эксперименте показано, что исходные данные о смертности мужского населения России 2001 и 2014 года лучше всего описывает распределение Вейбулла, а также то, что при увеличении возраста человека качество описания его остаточной продолжительности жизни рассматриваемыми распределениями падает.

Acknowledgments (благодарности)

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00780.

Список используемых источников

1. Прокудина Е.И. Обзор математических моделей и методов анализа данных о выживаемости/ Е.И. Прокудина, Э.Э. Абсаттарова // Информационные технологии интеллектуальной поддержки принятия решений (ITIDS'2019):: VII Всероссийская научная конференция (с приглашением зарубежных ученых). Уфа: УГАТУ, 2019. т. 2. С. 241-244.
2. Гайдышев И.П. Подгонка распределений в параметрическом анализе выживаемости. // Вестн. Ом. ун-та. 2016. №4. С. 10–15.
3. Burnham K. P. Model selection and multimodel inference. A practical information–theoretic approach / K. P. Burnham, D. R. Anderson –USA: Springer, 1998. – 488 p.
4. Akaike H. A Bayesian analysis of the minimum AIC procedure // Annals of the Institute of Statistical Mathematics. 1978. Vol. 30. Part A. № 1. P. 9–14.
5. Таблица смертности населения России для календарных лет 1959-2014.[Электронный ресурс]. – Режим доступа: http://www.demoscope.ru/weekly/ssp/rus_ltmenu.php (дата обращения: 15.02.2020).