Анализ влияния дихотомических признаков на дожитие: конечно-геометрический и информационный подходы

Смирнов Иван Борисович, гр. 522

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н. Алексеева Н.П.
Рецензент: м.н.с. Грачева П.В.

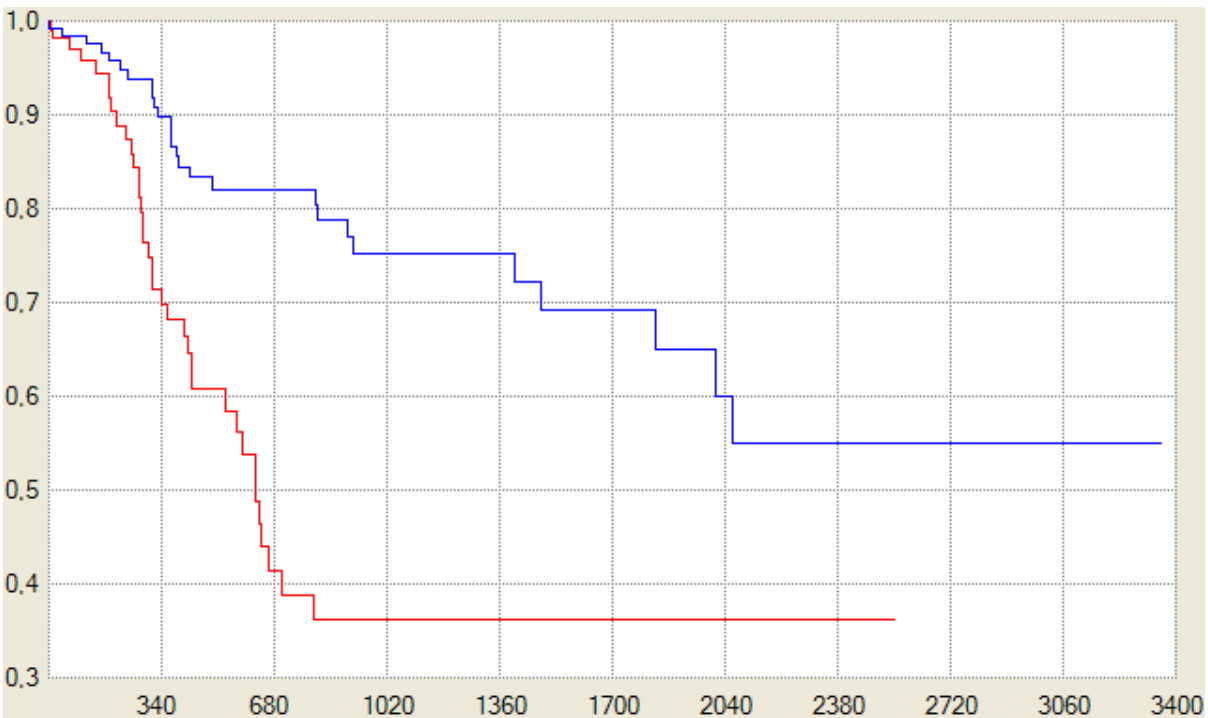
Санкт-Петербург
2009 г.

Постановка задачи

Дожитие

- τ – случайное время до наступления отказа
- Цензурирование: вместо τ_i наблюдается $(\tilde{\tau}_i, c_i)$
  $c_i = 0$, если $\tilde{\tau}_i = \tau_i$; $\quad c_i = 1$, если $\tilde{\tau}_i > \tau_i$
- Кривая дожития: $S(t) = P(\tau > t)$, $S(t) = 1 - F(t)$

Задача: исследование влияния признаков на дожитие



Значимое отличие между группами $X = 0$ и $X = 1$
$p = 0.0001$

Основные методы

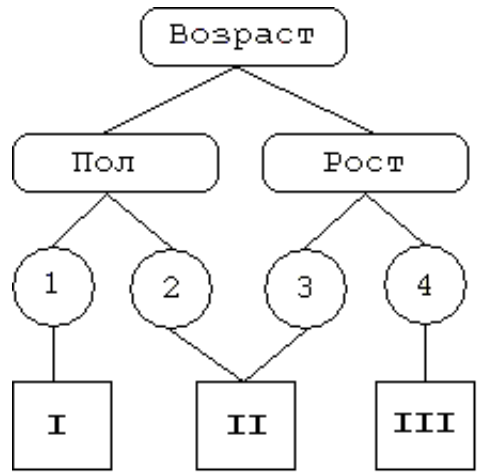
Регрессионная модель Кокса (Cox, 1972)

- Риск: $h(t) = \lim\limits_{\Delta \to 0} \frac{P(t<\tau<t+\Delta|\tau>t)}{\Delta}$
- Модель: $h(t|z) = h_0(t)exp(\beta^T z)$
- Функция правдоподобия:

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta^T z_i)}{\sum\limits_{j \in R_{t_i}} \exp(\beta^T z_j)} \right)^{c_i}$$

$$R_{t_i} = \{j : t_j \geq t_i\}$$

Пошаговый алгоритм разбиения (Gordon, Olshen, 1985)



Проблемы в использовании

- Модель
- Пропуски
- Интерпретация
- Объём выборки
- Пошаговая процедура разбиения

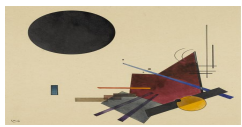
Конечно-геометрический подход: симптомы

$X = (X_1, \ldots, X_n)^T$ – вектор дихотомических признаков

Определение 1

Симптом: $X_\tau = A_\tau X \pmod 2$
$\tau \subset \{1, \ldots, n\}$, $A_\tau = (a_1, \ldots, a_n)$, где $a_i = \mathsf{I}(i \in \tau)$

$X_1 = 0$
$X_2 = 0$



$X_1 = 0$
$X_2 = 1$

$X_1 = 1$
$X_2 = 0$

$X_1 = 1$
$X_2 = 1$

Конечно-геометрический подход: симптомы

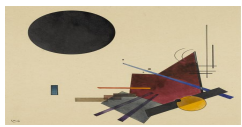
$X = (X_1, \ldots, X_n)^T$ – вектор дихотомических признаков

Определение 1

Симптом: $X_\tau = A_\tau X \pmod 2$
$\tau \subset \{1, \ldots, n\}$, $A_\tau = (a_1, \ldots, a_n)$, где $a_i = \mathsf{I}(i \in \tau)$

$X_1 = 0$
$X_2 = 0$

$X_{12} = 0$



$X_1 = 0$
$X_2 = 1$

$X_{12} = 1$

$X_1 = 1$
$X_2 = 0$

$X_{12} = 1$

$X_1 = 1$
$X_2 = 1$

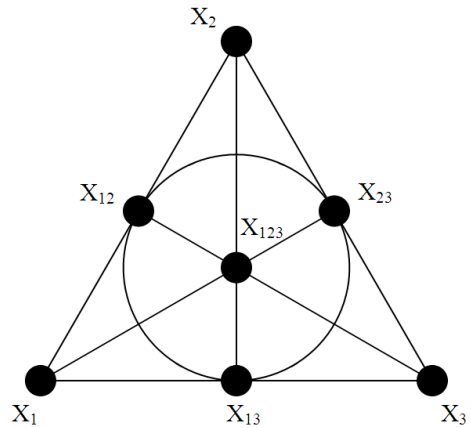
$X_{12} = 0$

Новый признак характеризует взаимодействие исходных, является дихотомическим, и может пониматься как «непропорциональность».

Конечно-геометрический подход: синдромы и обобщённые симптомы

Определение 2.1

Синдром $\Delta_k = \{\beta_0 X_{\tau_0} + \ldots + \beta_k X_{\tau_k}\}$, где $X_{\tau_0}, \ldots, X_{\tau_k}$ – базовые симптомы

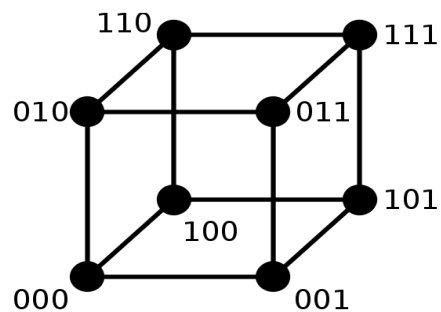


$PG(2,2)$

$\Omega_k = (X_{\tau_0}, \ldots, X_{\tau_k})$, $|\Omega_k| = 2^{k+1}$
$\Omega_k = B(X_{\tau_i} = 0) \bigsqcup B(X_{\tau_i} = 1)$

Определение 2.2

X_Z – обобщённый симптом
$\Omega_k = B(X_Z = 0) \bigsqcup B(X_Z = 1)$,
$|B(X_Z = 0)| = |B(X_Z = 1)|$



$EG(2,2)$

Применение в статистике, теорема о рангах синдрома

Распределение синдрома Δ_k – совместное распределение базовых симптомов $(X_{\tau_0}, \ldots, X_{\tau_k})$. Не зависит от их выбора.

Определение 3.1

Ранг симптома $|X_\tau| = |\tau|$

Определение 3.2

m-ранг синдрома $|\Delta_k|_m = \min\{|X_{\tau_0}| + \ldots + |X_{\tau_{m-1}}|\}$

Теорема

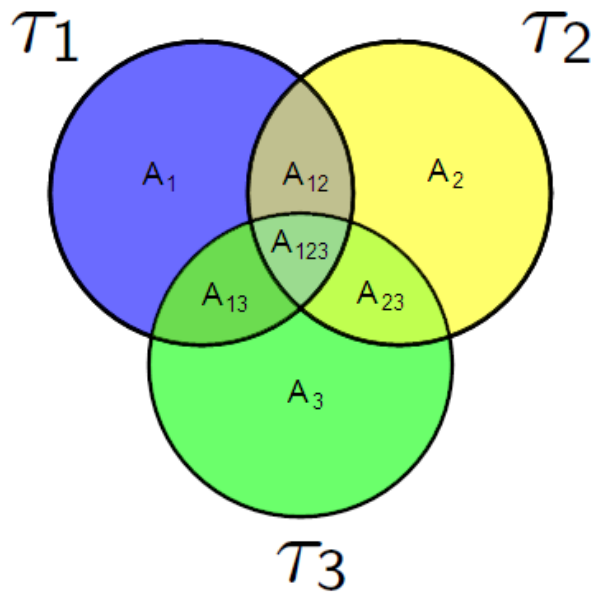
$$|\Delta_k|_m \leq \frac{2^k nm}{2^{k+1} - 1}$$

Практическое значение – исключение дублирующих вычислений

Идея доказательства в частном случае

$$\Delta_2 = (X_{\tau_1}, X_{\tau_2}, X_{\tau_3}, X_{\tau_1\tau_2}, X_{\tau_1\tau_3}, X_{\tau_2\tau_3}, X_{\tau_1\tau_2\tau_3})$$

$$\begin{aligned}
|X_{\tau_1}| &= A_1 + A_{12} + A_{13} + A_{123} \\
|X_{\tau_2}| &= A_2 + A_{12} + A_{23} + A_{123} \\
|X_{\tau_3}| &= A_3 + A_{13} + A_{23} + A_{123} \\
|X_{\tau_1\tau_2}| &= A_1 + A_2 + A_{13} + A_{23} \\
|X_{\tau_1\tau_3}| &= A_1 + A_3 + A_{12} + A_{23} \\
|X_{\tau_2\tau_3}| &= A_2 + A_3 + A_{12} + A_{13} \\
|X_{\tau_1\tau_2\tau_3}| &= A_1 + A_2 + A_3 + A_{123}
\end{aligned}$$



$$X = BA, \qquad (BA)_i \geq |\Delta_2|_1, \qquad e^T A = n$$

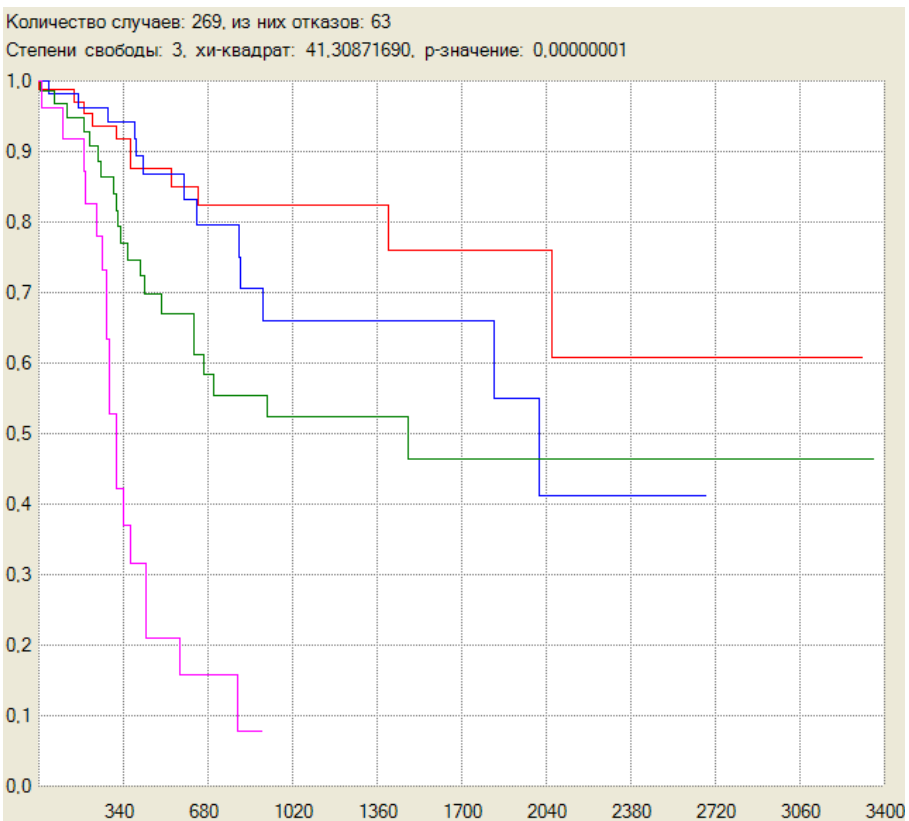
- $|\Delta_2|_1 \leq \frac{4}{7} n$
- $|\Delta_2|_m \leq \frac{4}{7} mn$
- $|\Delta_k|_m \leq \frac{2^k}{2^{k+1}-1} mn$

Применение

Данные о послеоперационном дожитии 272 пацинетов с глиомой *(Военно-медицинская академия, кафедра нейрохирургии)*:

9 дихотомических признаков, возраст, индекс Карновского, размеры опухоли, тип операции

$\Delta_2 = (X_{36}, X_{1457}, X_{128})$, Z_{0135} и Z_{0123}



Группа 4

Больший возраст ($p = 0,007$)
Худшее состояние ($p = 0,001$)

Группа 1

Лучшее состояние ($p < 0,001$)
Меньше опухоль ($p = 0,002$)

Группы 2 и 3

Тип операции ($p = 0,018$)

Меры сходства и различия

- A, B – группы наблюдений, X – признак, $X \in \overline{1, s}$

  $|A| = n = n_1 + \ldots + n_s$, $n_i = |\{a \in A | X(a) = i\}|$

  Информационное разнообразие: $I_X = n \ln n - \sum n_i \ln n_i$, $I = \sum I_{X_i}$

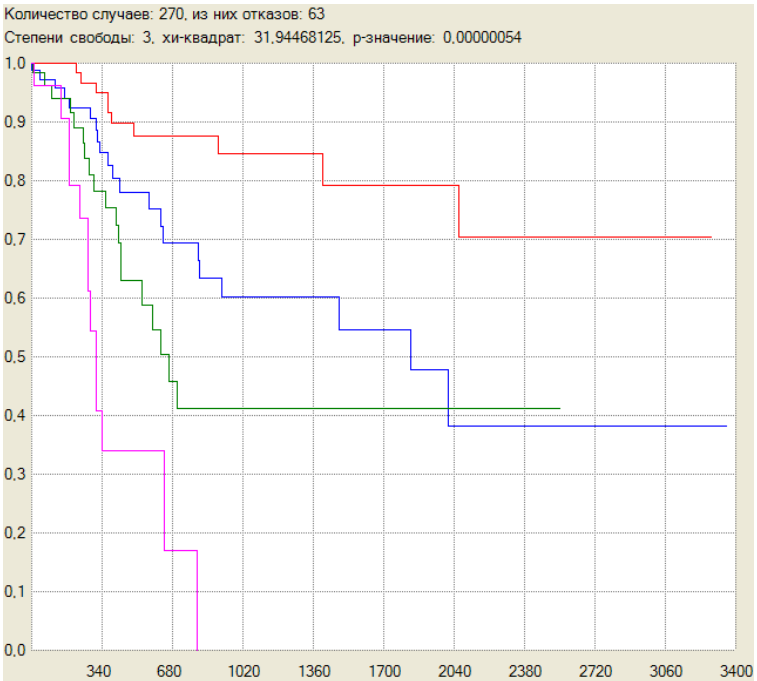
  Информационный выигрыш от объединения:
  $\Delta I(A, B) = I(A \bigcup B) - I(A) - I(B)$ – мера различия
- Выделенные на первом шаге кластеры объединяем, используя в качестве меры сходства p-значение критерия Вилкоксона-Гехана

Идентификация кластеров

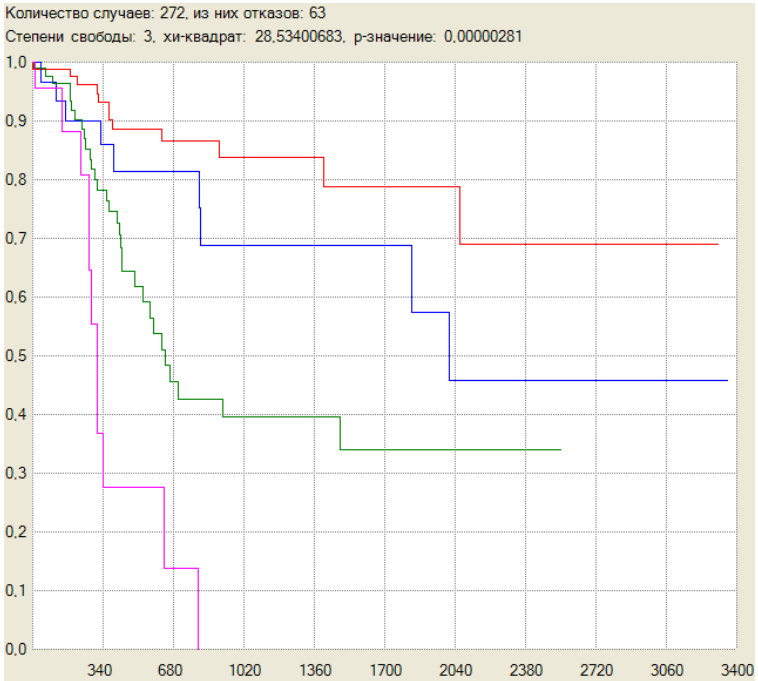
Коэффициент неопределённости: $R_\xi = \frac{H(\eta) - H(\eta|\xi)}{H(\eta)}$

Сравнение методов, выводы

Пошаговый алгоритм разбиения



Кластерный анализ



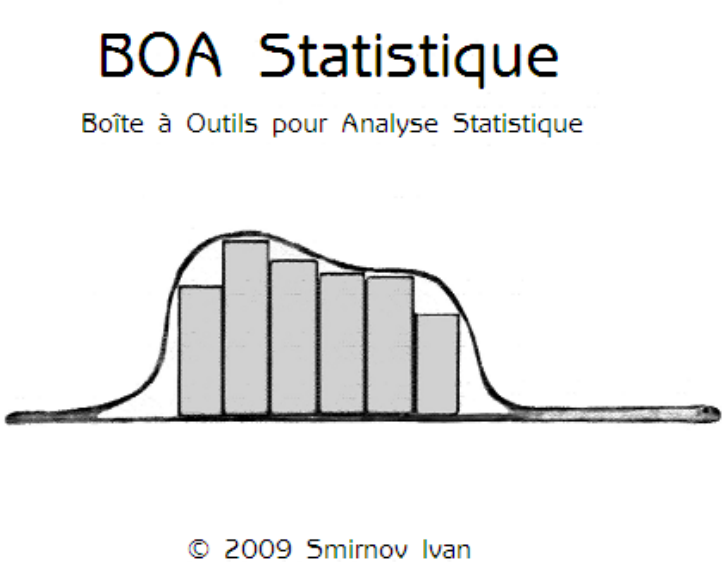
Совпадение результатов

Группа 3: 95,9%, Группа 0: 89,3%, Группа 1 и Группа 2: 87,5%

Выводы

- Много наблюдений: алгоритм пошагового выбора
- Мало наблюдений: симптомный анализ
- Много признаков: кластерный анализ

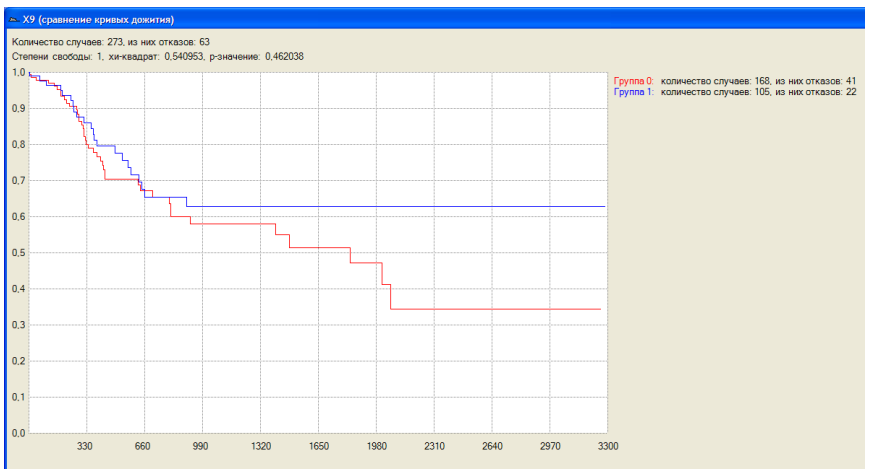
Приложение

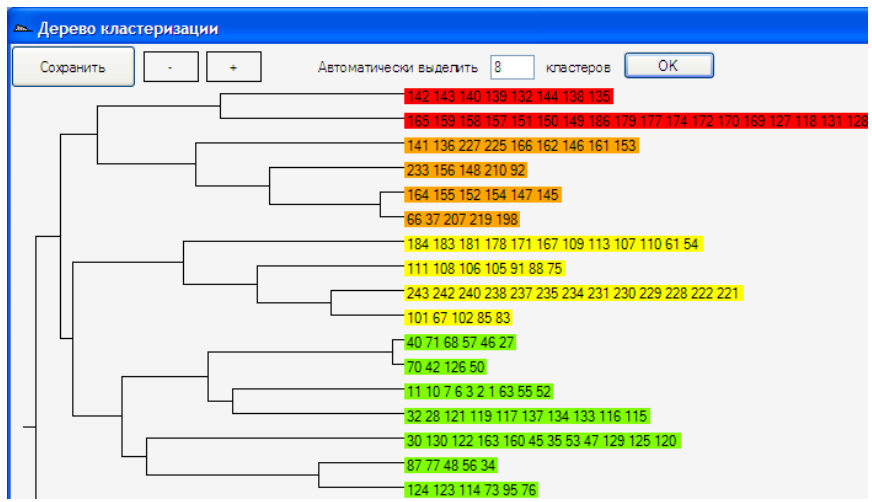


Результаты

Старт | Пауза

First	Second	Third	Chi-square	df	p-value
X1 X2 X4	X1 X2 X3 X5	X2 X3 X6	34,37498214	7	0,00001465
X1 X2 X3 X4	X2 X5	X1 X2 X3 X6	32,21408114	6	0,00001485
X1 X2 X3 X4	X2 X5	X3 X6	29,79420306	5	0,00001619
X1 X2 X4	X2 X5	X1 X2 X6	32,01286139	6	0,00001623
X1 X2 X4	X5	X1 X6	29,49678708	5	0,00001852
X1 X2 X4	X2 X5	X3 X6	28,56321877	5	0,00002824
X4	X2 X5	X1 X2 X6	28,40563233	5	0,00003032
X1 X2 X3 X4	X5	X1 X6	28,37112707	5	0,00003080
X1 X2 X3	X4	X2 X5 X6	32,35000214	7	0,00003497
X3 X4	X2 X5	X3 X6	30,25491390	6	0,00003516
X1 X2	X3 X4	X1 X5 X6	32,25482388	7	0,00003643
X1 X2 X3	X5	X1 X6	30,13446580	6	0,00003706
X1	X2	X5 X6	27,90592009	5	0,00003797
X1	X2 X3 X5	X2 X3 X6	30,05408066	6	0,00003839
X1 X2 X3 X4	X5	X1 X3 X6	27,87218682	5	0,00003855
X3 X4	X1 X2 X3 X5	X2 X3 X6	32,11581865	7	0,00003866
X2 X4	X1 X2 X3 X5	X1 X2 X3 X6	32,08282153	7	0,00003920
X1	X2 X3 X4	X5 X6	29,97879301	6	0,00003968





Результаты

- Предложены два дополнительных метода для выявления факторов, влияющих на дожитие
- Сформулирована и доказана теорема, позволяющая существенно сократить время вычислений
- Создано приложение, позволяющее выполнять все необходимые статистические процедуры
- Выполнено биометрическое исследование реальных данных с использованием предложенных методов