Исследование однородности категориальных рядов с приложением в когнитивной биологии

Андреева Надежда Павловна

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

Научный руководитель: к. ф.-м. н., доц. Алексеева Н. П.
Рецензент: к. ф.-м. н., мл. науч. сотр. Ананьевская П. В.

15 июня 2015 г.

Цель и задачи

Цель:

- Сравнение категориальных рядов по встречаемости отдельных фрагментов, удовлетворяющих отрицательному биномиальному распределению.

Задачи:

- оценка параметров ОБР;
- проверка согласия с распределением;
- проверка однородности и анализ множественных сравнений.

Прикладная задача:

- Анализ клинических испытаний действия антидепрессантов на поведение животных.

Описание данных и модель

- Данные теста Порсолта о поведении крыс
  - категориальные ряды поведенческих актов
  - группы - физраствор, дезипрамин в дозах 5, 10, 20 мг/кг
  - объем выборки n = 42
- Фрагмент, встречаемость $\xi \sim NB(r, p)$ — слово

$$P(\xi = j) = \frac{\Gamma(r+j)}{\Gamma(j+1)\Gamma(r)} p^r (1-p)^j, \quad j = 0, 1, \ldots.$$

- Идентификация слов, по которым наблюдаются различия между группами.

Оценка параметров ОБР методом ОМП и сравнение с лингвистикой

Встречаемость $x_1, \ldots, x_n \sim NB(r, p)$.
Пусть $\psi(x) = \ln'(\Gamma(x))$, тогда $\hat{r}$ - решение уравнения

$$\sum_{i=1}^{n} \left(\ln(\frac{r}{r+\bar{x}}) + \psi(r+x_i) - \psi(r) \right) = 0,$$

$$\hat{p} = \frac{\hat{r}}{\hat{r}+\bar{x}}$$

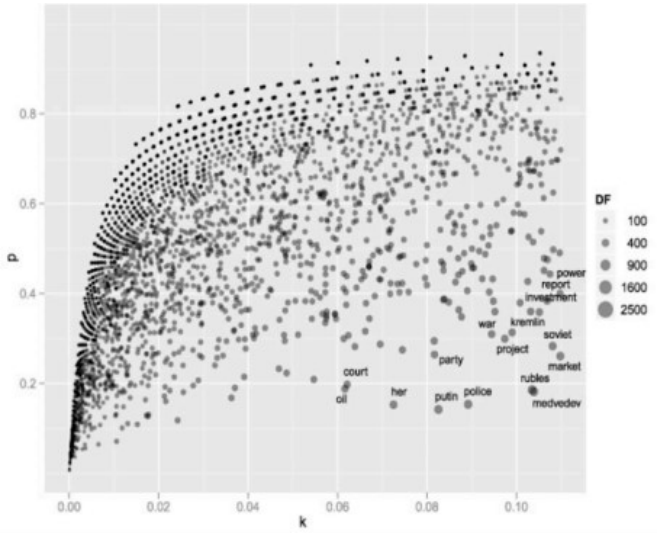


Рис.: Оценки параметров словоупотребления [Alexeyeva, Sotov, 2013]

Проверка адекватности модели

Критерии согласия

1. на основе производящей функции [Rueda, O'Reilly, 1999];
2. хи-квадрат (χ^2);
3. Крамера-фон Мизеса [Choulakian, Lockhart, Stephens, 1994];
4. Колмогорова-Смирнова;
5. Bootstrap [Szücs, 2008].

Критерий согласия на основе производящей функции

$$H_0 : F = F_0(\cdot;\theta), \qquad \phi(t;\theta) = E_\theta(t^x), \quad \text{для } |t| \le 1$$

$$x_1, x_2, \ldots, x_n \sim F\text{- дискретное}, \qquad \phi_n(t) = \frac{1}{n}\sum_{i=0}^{n} t^{x_i}$$

Утверждение [Klar B., 1999]

Эмпирический процесс

$$\xi_n(t;\theta) = \sqrt{n}\{\phi_n(t) - \phi(t;\theta)\} \xrightarrow{\mathscr{D}} \xi(t;\theta),$$

где, $\xi(t;\theta)$ - Гауссовский процесс в ℓ_1, такой что, для любых $s, t \in [0,1]$ и $\theta \in \Theta$, $E_\theta[\xi(t;\theta)] = 0$, ковариационная функция

$$\mathcal{L}_\theta(s;t) = \phi(st,\theta) - \phi(s;\theta)\phi(t;\theta)$$

Случай с неизвестными параметрами

Утверждение [Klar B., 1999]

Эмпирический процесс

$$\xi_n(t;\hat{\theta}_n) = \sqrt{n}\{\phi_n(t) - \phi(t;\hat{\theta}_n)\} \xrightarrow{\mathscr{D}} \xi(t;\hat{\theta}_n),$$

где, $\xi(t;\hat{\theta}_n)$ - Гауссовский процесс в ℓ_1, такой что, для любых $s,t \in [0,1]$ и $\theta \in \Theta$, $E_{\hat{\theta}_n}[\xi(t;\hat{\theta}_n)] = 0$, ковариационная функция

$$\hat{\mathcal{L}}_\theta(s;t) = \phi(st,\theta) - \phi(s;\theta)\phi(t;\theta) - \frac{\partial}{\partial\theta}\phi(s;\theta)\frac{\partial}{\partial\theta}\phi(t;\theta)\mathcal{I}^{-1}(\theta),$$

где $\mathcal{I}$- информационное количество Фишера

Ковариационная функция для ОБР:

$$\hat{\mathcal{L}}(s;t) = \left(\frac{p}{1-qst}\right)^r - \left(\frac{p^2}{(1-qs)(1-qt)}\right)^r - \frac{r(1-t)(1-s)p^{2r}q}{(1-qt)^{r+1}(1-qs)^{r+1}}$$

Статистика критерия на основе производящей функции

$$d_n(\hat{\theta}) = \int_0^1 \xi_n^2(t;\hat{\theta})dt$$

$$d_n(\hat{\theta}) = \frac{1}{n}(\underline{O} - \underline{e}(\hat{\theta}))^T M(\underline{O} - \underline{e}(\hat{\theta})),$$

где $\underline{O}$ - наблюдаемые частоты,

$\underline{e}(\hat{\theta})$ - теоретические частоты,

M - гильбертова матрица

$$M = \begin{pmatrix} 1 & 1/2 & 1/3 & \cdots \\ 1/2 & 1/3 & 1/4 & \cdots \\ 1/3 & 1/4 & 1/5 & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

Вычисление распределения статистики [Imhof, 1961]

$x = (x_1, ..., x_n)^T \sim N(0, \Sigma)$, $\mu = (\mu_1, ..., \mu_n)^T$ - вектор констант

$$Q = (x + \mu)^T A(x + \mu) = \sum_{r=1}^{m} \lambda_r \chi^2_{h_r;\delta_r},$$

где λ_r - не нулевые характеристические корни $A\Sigma$;

m - число не нулевых характеристических корней $A\Sigma$;

h_r - порядок кратности корней;

δ_r - некоторая линейная комбинация $\mu_1, ..., \mu_n$;

$\chi^2_{h;\delta} = (x_1 + \delta) + \sum_{i=2}^{h} x_i^2$

Функция в R: $imhof(t, \lambda, h, \delta)$ вычисление $P\{Q > t\}$.

Bootstrap тест

Для каждого слова был проделан следующий алгоритм:

- Получены оценки параметров $\hat{\theta}$;
- Проверена гипотеза о принадлежности распределению с $\hat{\theta}$ (с помощью CVM), получено $p\text{-}value\ P$;
- С оцененными параметрами $\hat{\theta}$ смоделировано 100 выборок;
- Для них оценены параметры $\tilde{\theta}$ и проверена гипотеза согласия. Получены вспомогательные значения $p\text{-}value\ P^*$;
- Вычислен процент вспомогательных P^*, меньших чем исходный P.

Применимость CVM и bootstrap

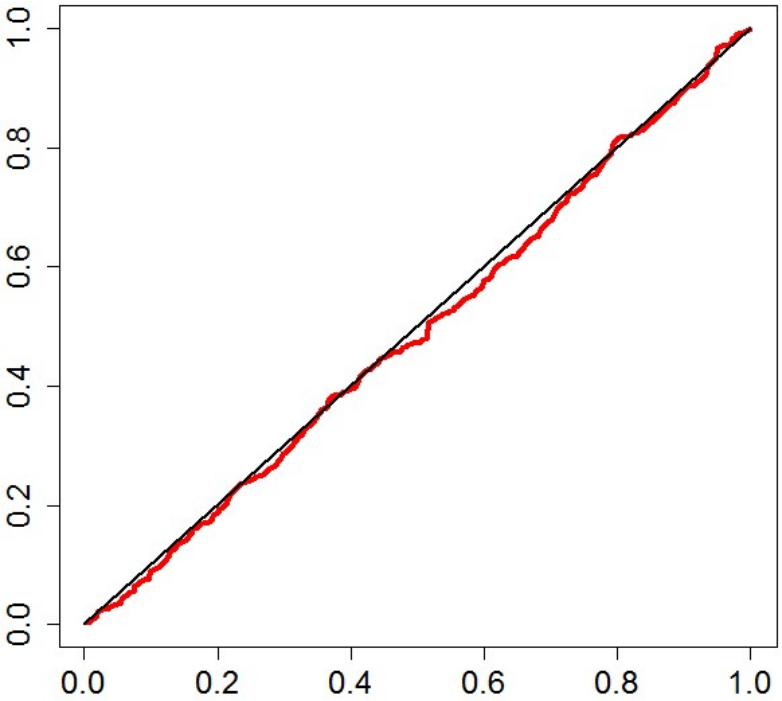


Рис.: CVM. Моделирование NB(0.5,0.4), проверка на NB с оценкой параметров

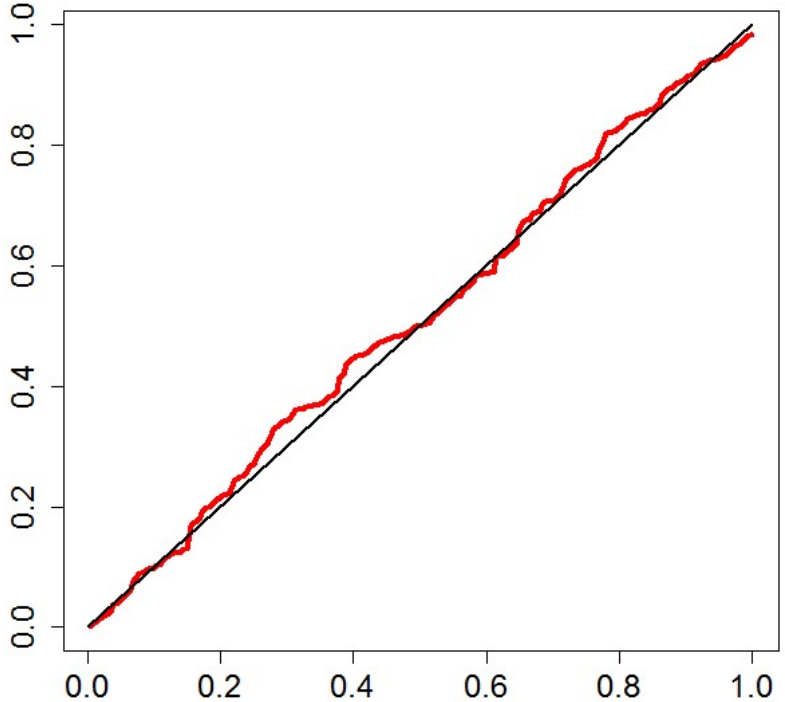


Рис.: Bootstrap. Моделирование NB(0.5,0.4), проверка на NB с оценкой параметров

Результаты проверки гипотез согласия с ОБР

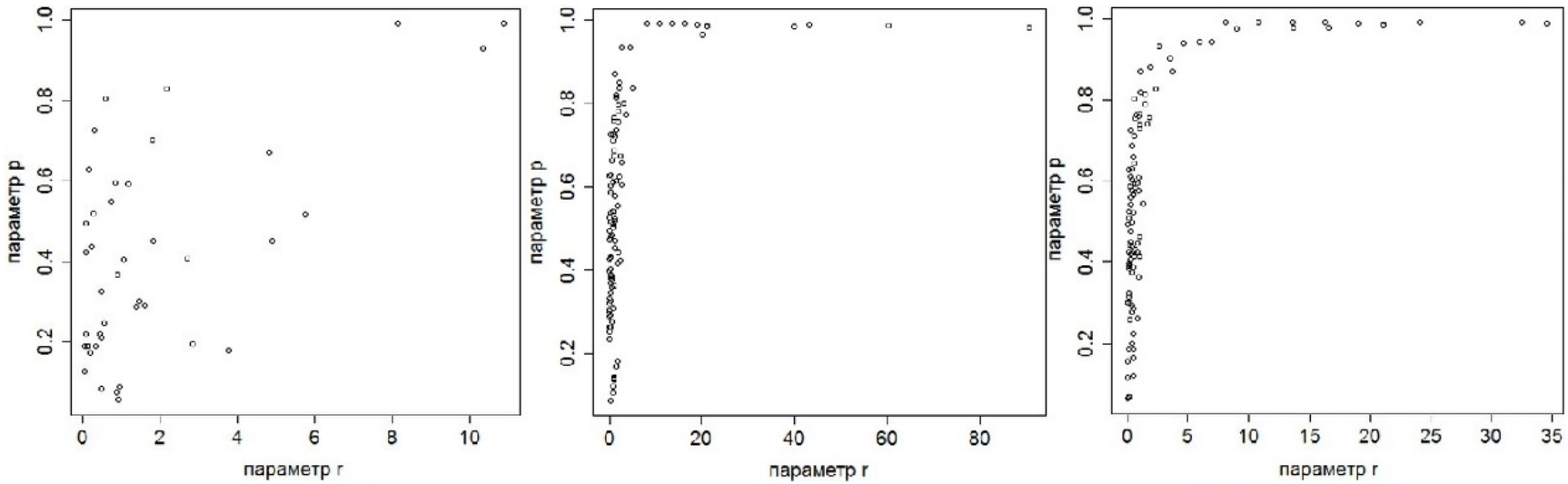


Рис.: Оценки параметров распределения для слов дли 2, 3 и 5

Длина фрагментов	1	2	3	4	5
Число фрагментов	8	48	140	282	383
Число не отвергнутых гипотез, CVM	7	48	139	273	368
Число не отвергнутых гипотез, CVM с bootstrap	7	38	106	230	334

Дисперсионный анализ по Краскелу-Уоллису

$$H = \frac{12}{n(n+1)} \sum_{i=0}^{r} \frac{R_i^2}{n_i} - 3(n+1),$$

где i — номер группы,

r — номер последней группы,

n_i — количество наблюдений в группе i,

n — общее число наблюдений,

R_i — суммы рангов в каждой группе.

Выявлены слова, по частоте появления которых группы крыс статистически значимо различаются:

Длина слов	1	2	3	4	5
Число слов	7	38	106	230	334
Число слов, по которым наблюдаются различия	1	5	15	20	20

Иллюстрация разделимости групп при увеличении длины слов

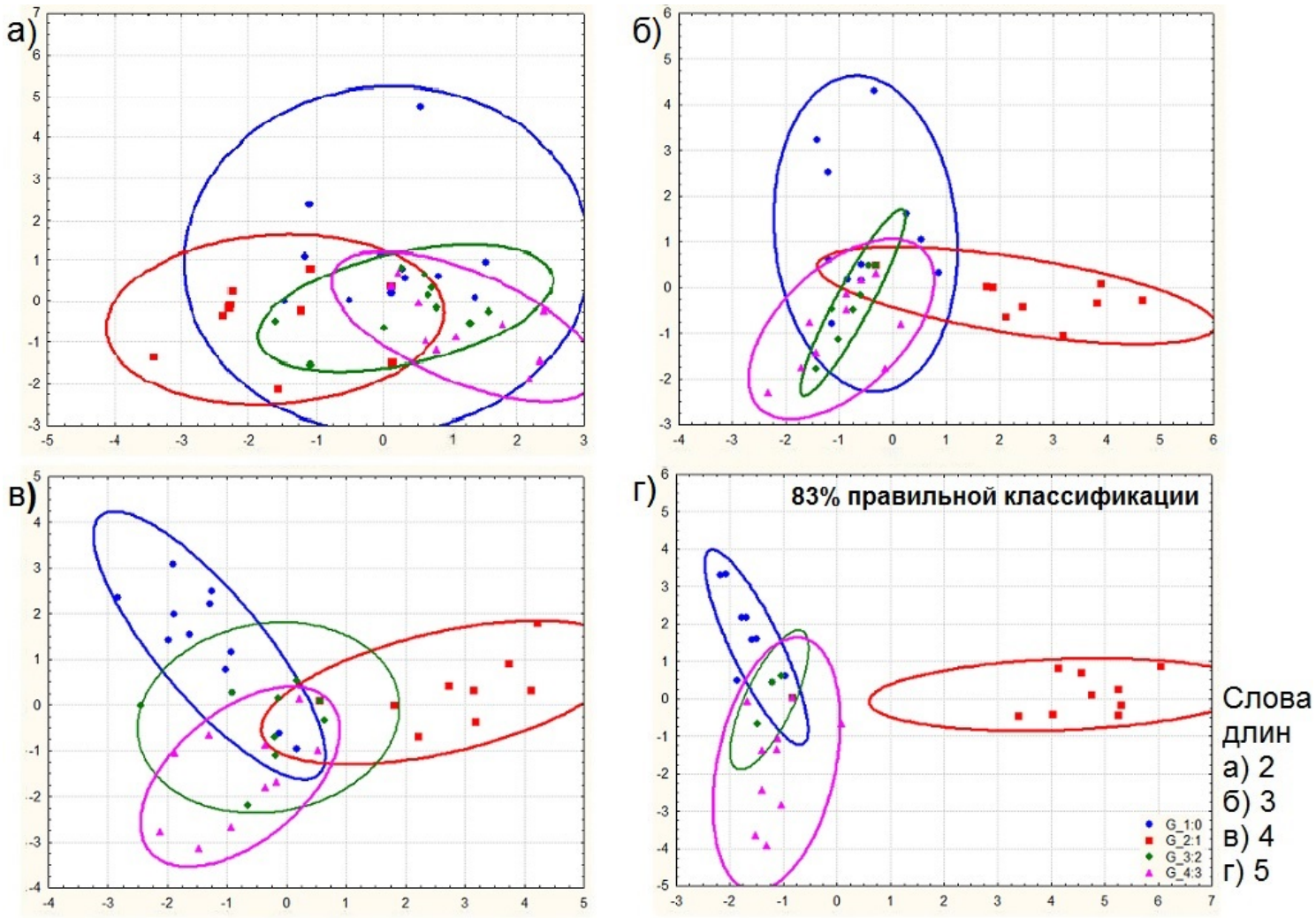
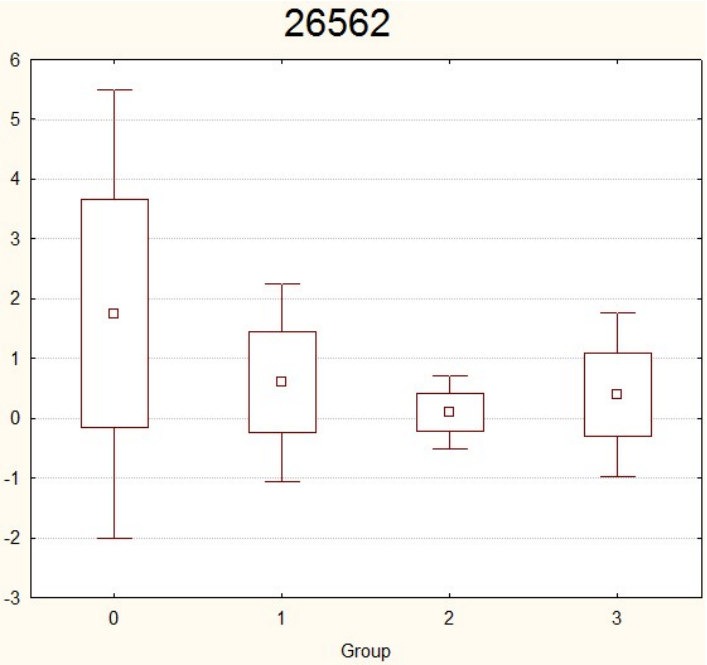


Рис.: Диаграммы рассеяния дискриминантных функций

Основные результаты дискриминантного анализа

- 26562: дрейф->гребля->отряхивание->гребля->дрейф
- 31616: плавание->карабкание->гребля->карабкание->гребля



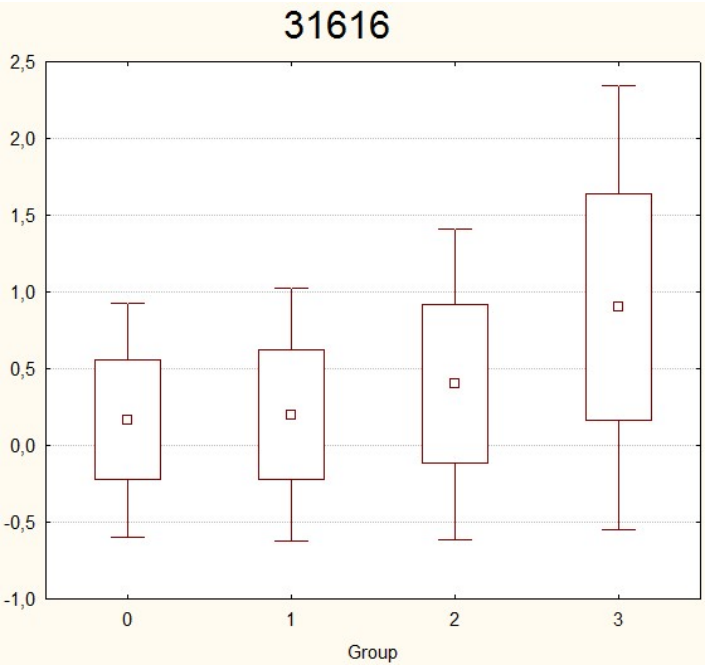


Рис.: Диаграммы размаха значимых для дискриминации переменных

Критерий Тьюки

	X26562	X31616
5 мг/кг - физраствор	0.11667	0.99883
10 мг/кг - физраствор	0.01099	0.73108
20 мг/кг - физраствор	0.04885	0.01247
10 мг/кг - 5 мг/кг	0.77477	0.83112
20 мг/кг - 5 мг/кг	0.98068	0.02545
20 мг/кг - 10 мг/кг	0.93932	0.16509

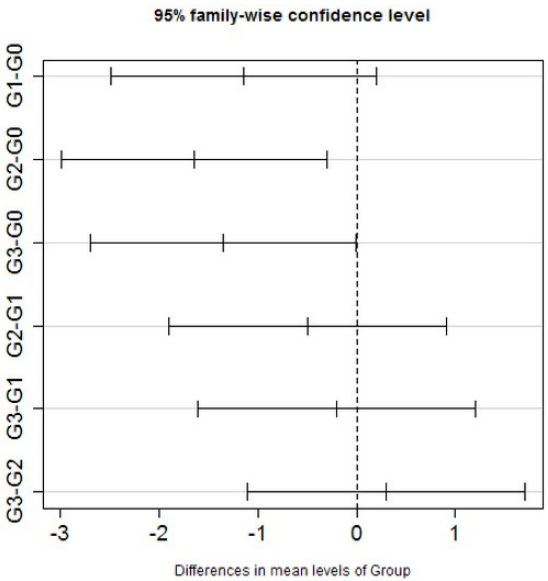


Рис.: 26562

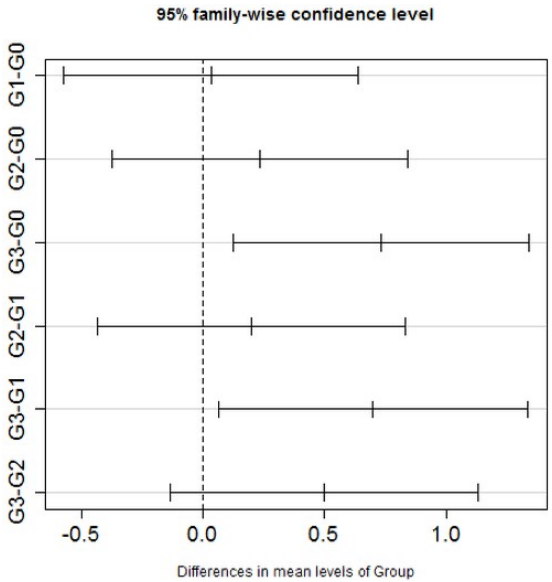


Рис.: 31616

Основные результаты:

- Разработано ПО на C#, вычисляющее встречаемость всевозможных фрагментов заданной длины в категориальном ряду каждой особи;
- Разработаны на R методы анализа однородности категориальных рядов;
- Проведен анализ полученных результатов на примере поведения животных из разных экспериментальных групп.