Санкт-Петербургский государственный университет

Мандрикова Анастасия Андреевна

Выпускная квалификационная работа

Модель двумерного гамма-распределения с медико-биологическими приложениями

Уровень образования: магистратура

Направление 01.04.02 «Прикладная математика и информатика»

Основная образовательная программа ВМ.5751.2020 «Математическое моделирование, программирование и искусственный интеллект»

Научный руководитель:

Доцент, кафедра статистического моделирования

к. ф.-м. н., доцент Н. П. Алексеева

Рецензент:

Биостатистик,

ООО «Парексель Интернэшнл»

Е. С. Комарова

Санкт-Петербург

2022

Saint Petersburg State University
Applied Mathematics and Computer Science
Statistical Modelling

Mandrikova Anastasia Andreevna

Graduation Project

Two-dimensional model of gamma distribution with medical biologic applications

Scientific Supervisor:
Associate Professor, Department of
Statistical Modelling N. P. Alexeyeva

Reviewer:
Biostatistics, «Parexel International»
E. S. Komarova

Saint Petersburg
2022

Оглавление

Введение

Медико-биологические системы обладают свойством изменчивости. При анализе подобных систем интерес может представлять изучение динамики развития биологических процессов. Сами наблюдаемые данные обычно характеризуются наличием большого числа признаков. Если эти признаки являются категориальными, то их обработка невозможна с помощью линейных статистических методов.

В данной работе рассмотрены методы исследования систем с большим числом факторов с точки зрения сбалансированности. Первый такой инструмент — симптомно-синдромальный анализ, второй — псевдорандомизация. Второй метод получил широкое распространение в медицинских исследованиях благодаря простоте своего применения. Рассмотрению и применению этих методов на практике посвящена глава 2.

Вопросу поиска модели, наиболее точно удовлетворяющей особенностям рассматриваемой системы, посвящены остальные три главы. Базовая модель одномерного гамма-распределения, ее свойства, способы получения оценок параметров и проверки гипотезы о подчинении теоретическому закону распределения приведены в главе 1. В ситуации, когда данные могут быть описаны несколькими моделями сразу имеет место неопределенность и синонимия. Так и в случае согласия с гамма-распределением может наблюдаться согласие с целым семейством степенных гамма-распределений. Проблема поиска наилучшего представителя этого семейства рассматривается в главе 3.

Модель двумерного гамма распределения, рассмотренная в главе 4, была предложена Н. П. Алексеевой. Эта модель может быть полезна для изучения динамики процесса, который отражают изучаемые данные. Так же в главе 4 приведено применение модели к реальным данным на примере показателей состояния организма наркоманов, проходящих реабилитацию. Для пациентов известна динамика показателей, отвечающих за состояние организма и тяги к наркотическим средствам. Пациентов так же можно разделить на три группы в зависимости от получаемого лекарственного препарата. Подробное описание изучаемых признаков находится в разделе 4.3. С помощью модели исследовались значимость и направленность изменений показателей состояния здоровья, как в зависимости от времени измерения, так и от вида лечения или лекарственного препарата, которое получает пациент.

Глава 1

Одномерная модель распределения

1.1. Модель гамма-распределения

Определение 1. *Пусть распределение случайной величины X задаётся плотностью вероятности, имеющей вид*

$$\gamma(x,\alpha,\beta) = \begin{cases} x^{\alpha-1}\frac{e^{-x/\beta}}{\beta^{\alpha}\,\Gamma(\alpha)}, & x \geq 0 \\ 0, & x < 0 \end{cases},$$

где $\Gamma(\alpha) = \int\limits_0^\infty x^{\,\alpha-1}e^{-x}\,dx$ *— гамма-функция. Тогда говорят, что случайная величина X имеет гамма-распределение с параметрами β — масштаба и α — формы [1].*

Обозначение: $X \sim \gamma(x,\alpha,\beta)$.

Свойства гамма-распределения: Пусть $\xi \sim \gamma(x,\alpha,\beta)$, тогда

- $\mathbb{E}\xi = \alpha\beta$
- $\mathbb{D}\xi = \alpha\beta^2$
- $\operatorname{cov}(\xi, \ln\xi) = \beta$

Покажем последнее свойство: $\operatorname{cov}(\xi, \ln\xi) = \mathbb{E}(\xi \cdot \ln\xi) - \mathbb{E}\xi \cdot \mathbb{E}\ln\xi$.

$$\mathbb{E}\ln\xi = \int\limits_0^\infty \ln(x)\cdot\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}dx = \int\limits_0^\infty \ln(\beta z)\cdot\frac{z^{\alpha-1}e^{-z}}{\Gamma(\alpha)}dz =$$

$$= \ln(\beta)\int\limits_0^\infty \frac{z^{\alpha-1}e^{-z}}{\Gamma(\alpha)}dz + \int\limits_0^\infty \ln(z)\cdot\frac{z^{\alpha-1}e^{-z}}{\Gamma(\alpha)}dz$$

Здесь была сделана замена $x/\beta = z$, получим $\mathbb{E}\ln\xi = \ln(\beta) + \psi(\alpha)$, где $\psi(\alpha)$ — дигамма функция. Используя ту же замену, вычислим:

$$\mathbb{E}(\xi\cdot\ln\xi) = \int\limits_0^\infty \ln(x)\cdot\frac{x^{\alpha}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}dx = \int\limits_0^\infty \ln(\beta z)\cdot\frac{z^{\alpha}e^{-z}}{\Gamma(\alpha)}\beta dz =$$

$$= \beta\ln(\beta)\int\limits_0^\infty \frac{z^{\alpha}e^{-z}}{\Gamma(\alpha)}dz + \beta\int\limits_0^\infty \ln(z)\cdot\frac{z^{\alpha}e^{-z}}{\Gamma(\alpha)}dz$$

Получаем $\beta\alpha\ln(\beta) + \beta\alpha\psi(\alpha+1)$. Подставим все в формулу для $\operatorname{cov}(\xi,\ln\xi) = \beta\alpha\ln(\beta) + \beta\alpha\psi(\alpha+1) - \beta\alpha\cdot(\ln(\beta)+\psi(\alpha)) = \beta$. Так как $\psi(\alpha+1) = \psi(\alpha) + \alpha^{-1}$.

1.2. Оценка параметров гамма-распределения с помощью метода максимального правдоподобия (ММП)

Пусть $f(x, \theta)$ — генеральная плотность распределения, зависящая от параметра θ, $\mathbf{x} = (x_1, \ldots, x_n)$ — случайная выборка наблюдений.

Определение 2. *Для фиксированной реализации выборки* $\mathbf{x}$ *функция*

$$L(x, \theta) = f(x_1, \theta) \cdot \ldots \cdot f(x_n, \theta)$$

называется функцией правдоподобия.

Рассмотрим функцию правдоподобия для выборки из гамма-распределения:

$$L = \prod_{i=1}^{n} \frac{x_i^{\alpha-1} e^{\frac{-x_i}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)} = \frac{\prod\limits_{i=1}^{n} x_i^{\alpha-1} e^{\frac{-x_i}{\beta}}}{\beta^{n\alpha}\Gamma^n(\alpha)} = (\prod_{i=1}^{n} x_i)^{\alpha-1} e^{-\frac{1}{\beta}\sum\limits_{i=1}^{n} x_i} \beta^{-n\alpha}\Gamma^{-n}(\alpha).$$

В данном случае удобнее перейти к логарифму функции правдоподобия:

$$\ln L = (\alpha - 1)\sum_{i=1}^{n} \ln x_i - \frac{1}{\beta}\sum_{i=1}^{n} x_i - n\alpha \ln \beta - n \ln \Gamma(\alpha).$$

Найдем максимум функции:

$$\frac{\partial \ln L}{\partial \alpha} = \sum_{i=1}^{n} \ln x_i - n \ln \beta - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = 0,$$

$$\frac{\partial \ln L}{\partial \beta} = \beta^{-2}\sum_{i=1}^{n} x_i - \frac{n\alpha}{\beta} = 0.$$

Тогда оценки метода максимума правдоподобия получаются из уравнений:

$$\begin{cases} \frac{1}{n}\sum\limits_{i=1}^{n} \ln x_i - \ln \bar{x} + \ln \hat{\alpha} - \Gamma'(\hat{\alpha})/\Gamma(\hat{\alpha}) = 0, \\ \hat{\beta} = \bar{x}/\hat{\alpha}, \end{cases} \tag{1.1}$$

где $\Gamma'(\alpha)/\Gamma(\alpha) = \psi(\alpha)$ — дигамма функция.

1.3. Построение доверительных интервалов

Для построения доверительного интервала по методу МП нужно вычислить информант второго рода. Так как оцениваются два параметра распределения, информант будет матрицей следующего вида:

$$I = n \begin{bmatrix} \psi'(\alpha) & \frac{1}{\beta} \\ \frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{bmatrix}.$$

Обратим матрицу, после этого на ее главной диагонали будут стоять дисперсии соответствующих параметров, а элементы на побочной диагонали соответствовать ковариации.

$$I^{-1} = \frac{\beta^2}{n(\psi'(\alpha)\alpha - 1)} \begin{bmatrix} \frac{\alpha}{\beta^2} & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \psi'(\alpha) \end{bmatrix}.$$

Обратная матрица получена по формуле $A^{-1} = \frac{1}{\det A} A_*^{\mathrm{T}}$, где A_*^{T} — транспонированная матрица алгебраических дополнений. Так как оценки, полученные по методу максимального правдоподобия, являются асимптотически нормальными, построим асимптотические доверительные интервалы вида:

$$P(\hat{\theta} - u_{1-\alpha/2} \cdot \frac{\sigma(\hat{\theta})}{\sqrt{n}} < \theta < \hat{\theta} + u_{1-\alpha/2} \cdot \frac{\sigma(\hat{\theta})}{\sqrt{n}}) = 1 - \alpha,$$

где $u_{1-\alpha/2}$ — квантиль стандартного нормального распределения, n — размер выборки. Примем уровень значимости α, тогда получим доверительный интервал, в который случайная величина θ попадает с вероятностью $1 - \alpha$. Дисперсии оценок параметров распределения являются элементами $\mathrm{diag}\,(I^{-1})$. Извлекая корень из элементов, получим стандартное отклонение $\sigma(\hat{\theta})$, участвующее в построении доверительных интервалов.

1.4. Критерий хи-квадрат Пирсона для проверки согласия эмпирического распределения с теоретическим

Рассмотрим вопрос согласия эмпирического распределения с теоретическим. Для проверки был выбран критерий Пирсона. Формулируется нулевая гипотеза H_0: случайная величина ξ имеет генеральную функцию распределения $F(x)$. Множество S значений случайной величины разбивается на r непересекающихся интервалов $S_1, ..., S_r$. По-другому нулевую гипотезу можно сформулировать как равенство вероятностей p_i, отвечающих выборке $x_1, x_2, \ldots, x_n$, и p_i^0 — теоретических вероятностей попадания случайной величины в интервал S_i:

$$H_0 : p_i = p_i^0.$$

Вычисляются следующие значения:

$$n = \sum_{i=1}^{r} v_i, \qquad p_i^0 = P\{\xi \in S_i\} > 0, \qquad \sum_{i=1}^{r} p_i^0 = 1,$$

где v_i — эмпирические частоты. Для проверки гипотезы воспользуемся теоремой:

Теорема 1 (Критерий Пирсона)**.** *При справедливости H_0 статистика*

$$\chi^2 = \sum_{i=1}^{r} \frac{(v_i - np_i^0)^2}{np_i^0} \xrightarrow[n\to\infty]{d} \chi^2(r-1). \qquad [2]$$

Такая форма критерия используется для проверки гипотез в случае известных параметров. В случае, когда параметры оцениваются по выборке методом максимума правдоподобия (смотри раздел 1.2), предельное распределение имеет число степеней свободы уменьшенное на количество оцениваемых параметров, эта поправка была предложена Фишером. Так как рассматриваемое распределение зависит от двух параметров, число степеней свободы предельного распределения хи-квадрат $df = r - 1 - 2$, где r — количество интервалов, на которые разбивается множество значений случайной величины.

Алгоритм метода bootstrap

Пусть $\mathcal{F} = F(\cdot, \theta)$ — параметрическое семейство функций распределения (в данном случае рассматриваются функции гамма-распределения с параметрами $\theta = (\alpha, \beta)$). Имеется выборка $X_1, \ldots, X_n$. Нужно проверить является ли она выборкой распределения F_0 с параметрами $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ (оценка параметров находится по формуле (1.1)), где $F_0 \in \mathcal{F}$. Приведем параметрический тест проверки гипотезы $H_0 : F_0 \in \mathcal{F}$. Шаги алгоритма, рассмотрение которого на примере другого распределения приведено в [3]:

1. По имеющейся выборке $X_1, \ldots, X_n$ вычисляется значение оценки $\hat{\theta}_0$.
2. Используя оценку $\hat{\theta}_0$ вычисляется тестовая статистика критерия χ^2 (используем рассмотренный выше критерий Пирсона), обозначим найденное значение за χ_0^2.
3. Используя оценку $\hat{\theta}_0$ строится параметрический bootstrap следующим образом:
   а. Генерируется случайная выборка $X_1^*, \ldots, X_n^*$ из распределения $F(\cdot, \hat{\theta}_0)$.
   б. На основании значений $X_1^*, \ldots, X_n^*$ вычисляется оценка $\hat{\theta}_1$ как на шаге 1.
   в. На основании оценки $\hat{\theta}_1$ вычисляется тестовая статистика χ_1^2 как на шаге 2.
4. Шаг 3 повторяется, чтобы получить m значений тестовой статистики, $\chi_1^2, \ldots, \chi_m^2$.
5. Полученные значения упорядочиваются $\chi_{(1)}^2 \leq \chi_{(2)}^2 \leq \ldots \leq \chi_{(m)}^2$. Критическое значение $\chi_\alpha^2(n)$ аппроксимируется $\chi_\alpha^2\left((1-\alpha)m\right)$.
6. Нулевая гипотеза H_0 отклоняется, если $\chi_0^2 > \chi_\alpha^2(n)$.

На рисунке 1.1 представлено сравнение эмпирической функции распределения $p-value$ и функции равномерного распределения, при $m = 10000$. Алгоритм был применен к смоделированной выборке объема 100 из гамма-распределения с параметрами $(5, 2.5)$.

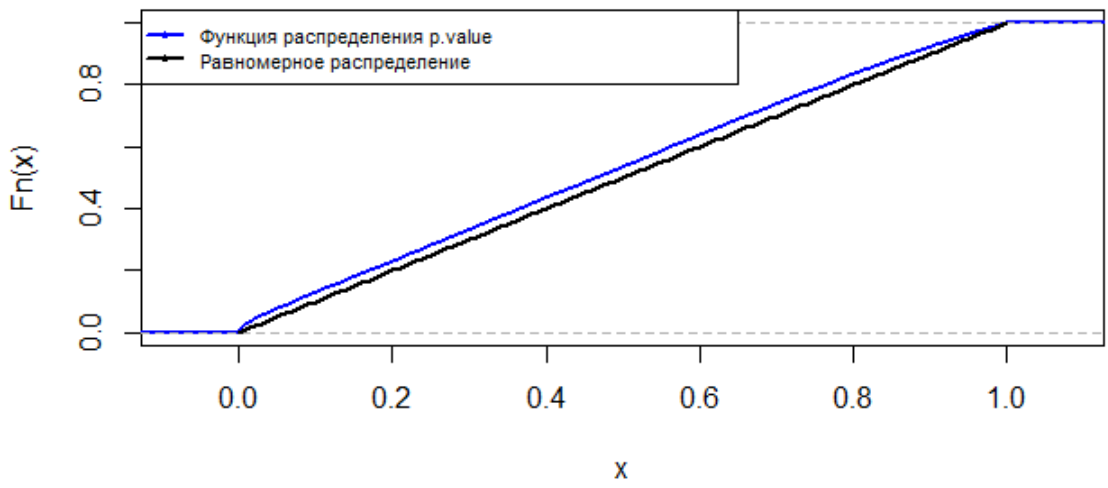

Рис. 1.1. График функции распределения $p - value$

Видим, что распределение p.value приближается к равномерному. После применения процедуры bootstrap критерий становится менее радикальным.

Глава 2

Методы получения однородных выборок

2.1. Расстояние Кульбака-Лейблера

Определение 3. *Расстояние Кульбака-Лейблера для абсолютно непрерывных распределений F_1, F_2 задается формулой*

$$D_{KL}(F_1||F_2) = \int\limits_X f_1(x) \ln \frac{f_1(x)}{f_2(x)} dx,$$

где $f_1(x)$, $f_2(x)$ — функции плотности распределений F_1 и F_2 соответственно, определённые на интервале $X \subseteq R^k$ [4].

Свойства расстояния Кульбака-Лейблера:

1. Несимметрично, симметричная версия: $\frac{1}{2}\left(D_{KL}(F_1||F_2) + D_{KL}(F_2||F_1)\right)$
2. Неотрицательно
3. Равно нулю только если $F_1 = F_2$ с вероятностью 1.

Рассмотрим детальнее симметричную вариацию расстояния Кульбака-Лейблера.

$$D_{KL} = \frac{1}{2}\left(\int\limits_X f_1(x) \ln \frac{f_1(x)}{f_2(x)} dx - \int\limits_X f_2(x) \ln \frac{f_1(x)}{f_2(x)} dx\right) = \frac{1}{2}\int\limits_X (f_1(x) - f_2(x)) \ln \frac{f_1(x)}{f_2(x)} dx \tag{2.1}$$

И выведем формулу расстояния между двумя распределениями гамма с параметрами соответственно равными $\alpha_1, \beta_1, \alpha_2, \beta_2$:

$$D_{KL} = \frac{1}{2}\int\limits_0^\infty \left(x^{\alpha_1 - 1}\frac{e^{-x/\beta_1}}{\beta_1^{\alpha_1}\,\Gamma(\alpha_1)} - x^{\alpha_2 - 1}\frac{e^{-x/\beta_2}}{\beta_2^{\alpha_2}\,\Gamma(\alpha_2)}\right) \ln\left(\frac{x^{\alpha_1 - \alpha_2} e^{-x/\beta_1 + x/\beta_2}\beta_2^{\alpha_2}\,\Gamma(\alpha_2)}{\beta_1^{\alpha_1}\,\Gamma(\alpha_1)}\right) dx.$$

Рассмотрим два интеграла соответствующие двум слагаемым и сделаем в них замены $z = x/\beta_1$ и $z = x/\beta_2$. Первый интеграл (пока опустим множитель 1/2):

$$\int\limits_0^\infty \left((z\beta_1)^{\alpha_1 - 1}\frac{\beta_1 e^{-z}}{\beta_1^{\alpha_1}\,\Gamma(\alpha_1)}\right) \ln\left(\frac{(z\beta_1)^{\alpha_1 - \alpha_2}\exp\left(-z + z\frac{\beta_1}{\beta_2}\right)\beta_2^{\alpha_2}\,\Gamma(\alpha_2)}{\beta_1^{\alpha_1}\,\Gamma(\alpha_1)}\right) dz =$$

$$= \int\limits_0^\infty \left(z^{\alpha_1 - 1}\frac{e^{-z}}{\Gamma(\alpha_1)}\right)\left((\alpha_1 - \alpha_2)\ln(z) + z\left(\frac{\beta_1}{\beta_2} - 1\right) + \ln\left(\frac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)}\left(\frac{\beta_2}{\beta_1}\right)^{\alpha_2}\right)\right) dz.$$

Подынтегральное выражение можно разбить на три логарифма, которые вычисляются аналогично третьему свойству гамма-распределения из пункта 1.1, что дает:

$$(\alpha_1 - \alpha_2)\psi(\alpha_1) + \left(\frac{\beta_1}{\beta_2} - 1\right)\frac{\Gamma(\alpha_1+1)}{\Gamma(\alpha_1)} + \ln\left(\frac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)}\left(\frac{\beta_2}{\beta_1}\right)^{\alpha_2}\right).$$

Перейдем ко второму интегралу, котором была проведена замена $z = x/\beta_2$:

$$\int_0^\infty \left((z\beta_2)^{\alpha_2-1}\frac{\beta_2 e^{-z}}{\beta_2^{\alpha_2}\,\Gamma(\alpha_2)}\right)\ln\left(\frac{(z\beta_2)^{\alpha_1-\alpha_2}\exp\left(z\left(1-\frac{\beta_2}{\beta_1}\right)\right)\beta_2^{\alpha_2}\,\Gamma(\alpha_2)}{\beta_1^{\alpha_1}\,\Gamma(\alpha_1)}\right)dz =$$

$$= \int_0^\infty \left(z^{\alpha_2-1}\frac{e^{-z}}{\Gamma(\alpha_2)}\right)\left((\alpha_1-\alpha_2)\ln(z) + z\left(1-\frac{\beta_2}{\beta_1}\right) + \ln\left(\frac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)}\left(\frac{\beta_2}{\beta_1}\right)^{\alpha_1}\right)\right)dz.$$

Аналогично получаем

$$(\alpha_1 - \alpha_2)\psi(\alpha_2) + \left(1 - \frac{\beta_2}{\beta_1}\right)\frac{\Gamma(\alpha_2+1)}{\Gamma(\alpha_2)} + \ln\left(\frac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)}\left(\frac{\beta_2}{\beta_1}\right)^{\alpha_1}\right).$$

Окончательно перейдем к формуле самого расстояния:

$$D_{KL} = \frac{1}{2}\left[(\alpha_1-\alpha_2)\left(\psi(\alpha_1) - \psi(\alpha_2) + \ln\left(\frac{\beta_2}{\beta_1}\right)\right) + (\beta_1-\beta_2)\left(\frac{\alpha_1}{\beta_2} - \frac{\alpha_2}{\beta_1}\right)\right]. \tag{2.2}$$

Заметим, что формула симметрична, что и ожидалось. Действительно, если поменять аргументы местами, то в финальной формуле поменяются местами параметры распределений. А в каждом слагаемом оба сомножителя изменят знак.

2.2. Информационные метрики

Определение 4. *Для случайной величины ξ из дискретного распределения с вероятностями p_i, $i = 1, \dots n$ определим согласно формуле Шеннона информационную энтропию*

$$H(\xi) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}.$$

Аналогом энтропии Шеннона является дифференциальная энтропия.

Определение 5. *Дифференциальная энтропия для непрерывной случайной величины с плотностью $f_i(x)$ задается формулой*

$$H_{ii} = -\int_{-\infty}^{+\infty} \ln f_i(x) f_i(x) dx. \tag{2.3}$$

Смешанную дифференциальную энтропию зададим формулой

$$H_{ij} = -\int\limits_{-\infty}^{+\infty} \ln f_j(x) f_i(x) dx. \tag{2.4}$$

Заметим, что через введенные функционалы, мы можем получить симметричное задание расстояния Кульбака-Лейблера:

$$D_{KL} = \frac{1}{2}\int\limits_{-\infty}^{+\infty} (f_i(x) - f_j(x)) \ln \frac{f_i(x)}{f_j(x)} dx =$$

$$= \frac{1}{2}\int\limits_{-\infty}^{+\infty} \left[f_i(x) \ln f_i(x) - f_j(x) \ln f_i(x) - f_i(x) \ln f_j(x) + f_j(x) \ln f_j(x))\right] dx =$$

$$= \frac{1}{2}(H_{ij} - H_{ii} + H_{ji} - H_{jj}).$$

2.3. Симптомно-синдромальный подход к анализу данных

Определения данного раздела опираются на определения из монографии [5].

Определение 6. *Пусть* $X = (X_1, \ldots, X_m)^T$ *случайный вектор с компонентами принимающими только значения* 0 *или* 1. *Обозначим* $k-$*подмножество из* m *чисел через*

$$\tau = (t_1, \ldots, t_k) \subseteq (1, 2, \ldots, m)$$

и зададим вектор-строку A_τ *с компонентами*

$$a_j = \begin{cases} 1, \ \textit{если} \ j \in \tau, \\ 0, \ \textit{иначе.} \end{cases}$$

Линейная комбинация $X_\tau = A_\tau X (\mod 2)$ *называется симптомом ранга* k.

Определение 7. *Рассмотрим* $\tau = (t_1, \ldots, t_k) \subseteq (1, 2, \ldots, m)$ *(одно из* $2^m - 1$ *подмножеств, кроме пустого), где* $k \leq m$. *Обозначим* $X^\tau = \prod_{i=t_1}^{t_k} X_i$, *где* $X_{t_1}, ..., X_{t_k}$ *дихотомические переменные со значениями* 0, 1. *Будем называть* X^τ *супер-симптомом.*

Определение 8. *Пусть имеется* $k > 0$ *симптомов* $X_1, \ldots, X_k$. *Совокупность* $2^k - 1$ *симптомов вида*

$$\beta_1 X_1 + \cdots + \beta_k X_k \pmod 2,$$

где коэффициенты $\beta_i \in \mathbb{F}_2$ *не равны нулю одновременно называется синдромом* S_k *порядка* k.

Заметим, что единичный симптом можно рассматривать как синдром первого порядка $S_1 = S_1(X_\tau)$. Основываясь на статье [6] опишем рекуррентное построение синдромов ранга k из синдромов ранга $k-1$:

$$\begin{gathered} S_1 = S_1(X_{t_1}), \\ S_2 = S_2(X_{t_1},\, X_{t_2}) = (X_{t_1},\, X_{t_2},\, X_{t_1} + X_{t_2} \pmod 2), \\ \vdots \\ S_i = S_i(X_{t_1}, \ldots, X_{t_i}) = (S_{i-1},\, X_{t_i},\, S_{i-1} + X_{t_i} \pmod 2), \\ \text{где} X_{t_i} \notin S_{i-1},\, i > 2. \end{gathered} \tag{2.5}$$

Базовыми элементами назовем те элементы синдрома, которые расположены на местах с номерами 2^i. Так как некоторые синдромы ранга k различаются только порядком элементов, можно обобщить импульсную последовательность (2.5) (импульсными последовательностями в алгебре называют последовательности чисел, построенные в конечном поле в результате рекуррентных соотношений типа Фибоначчи), используя любую операцию над полем $\mathbb{F}_2$. Определим специальную импульсную последовательность из элементов $(X_{t_1}, \ldots, X_{t_k})$, $Xt_k \in S_k$, используя операцию $(*)$:

$$\begin{gathered} M_1 = M_1(X_{t_1}|*), \\ M_2 = M_2(X_{t_1},\, X_{t_2}|*) = (X_{t_1},\, X_{t_2},\, X_{t_1} * X_{t_2}), \\ \vdots \\ M_i = M_i(X_{t_1}, \ldots, X_{t_i}|*) = (M_{i-1},\, M_{t_i},\, M_{i-1} * X_{t_i}), \\ \text{где} X_{t_i} \notin M_{i-1},\, i > 1. \end{gathered} \tag{2.6}$$

Видно совпадение $M_k(X|+) = S_k(X),\ X = (X_{t_1}, \ldots, X_{t_k})$. Рассмотрим в качестве базовых элементов синдрома $M_k(\cdot|\odot)$, где $(\odot)$ умножение над полем $\mathbb{F}_2$. Таким образом, легко можно построить суперсиндром $S_K(M_k(X_{t_1}, \ldots, X_{t_k}|\odot))$, где $K = 2^k - 1$, который включается в себя все $2^K - 1$ многочлены над полем $\mathbb{F}_2$ со свободным членом, равным 0.

Допустим, что наблюдаются $X_1, \ldots, X_m$ категориальные признаки с двумя значениями 0 и 1. Будем рассматривать всевозможные комбинации $X_{t_1},\, X_{t_2},\, X_{t_3}$ и вычислять всевозможные элементы из $S_7(M_3(X_{t_1},\, X_{t_2},\, X_{t_3}))$, где $t_1,\, t_2,\, t_3 \in \{1, 2, \ldots, m\}$. Для признаков $a,\, b,\, c \in \{X_1, \ldots, X_m\}$ построение всех возможных $(2^7 - 1 = 127)$ суперсимптомов на языке Matlab приведено в листинге 2.1.

Листинг 2.1. Алгоритм построения суперсиндрома ранга 7

```
syms a b c;
M3 = [a, b, a*b, c, a*c, b*c, a*b*c];
S = [M3(1)];
for i=2:7
    n = length(S);
    S_new = sym(zeros(1, 2*n+1));
    S_new(1:n) = S;
    S_new(n+1) = M3(i);
    S_new((n+2):(2*n+1)) = S + M3(i);
    S = S_new;
end
```

Среди элементов из S необходимо выбрать наиболее информативные. С этой целью для каждого элемента из S рассчитывается энтропия по формуле из определения 4. По значению рассматриваемого суперсимптома наблюдаемые количественные признаки разбиваются на две подвыборки. Для каждой подвыборки проверяется гипотеза о принадлежности гамма-распределению по критерию Пирсона (раздел 1.4) и вычисляется расстояние Кульбака-Лейблера (формула 2.1) между распределениями, если обе выборки удовлетворяют гамма. Параметры распределения определяются по формуле 1.1.

2.3.1. Результаты

Анализируются данные наркоманов, проходящих реабилитацию. Для категориальных признаков:

- Год — продолжительность последнего периода добровольного воздержания от героина составила больше года,
- Возд — последнее воздержание от героина закончилось больше года назад,
- Лечение — ранее лечился от злоупотребления наркотиками,
- Кокаин — обнаружены следы кокаина в анализах,
- Бенз — обнаружены следы бензодиазепина в анализах,
- Амф — обнаружены следы амфетамина в анализах,

- Кон — обнаружена конопля в анализах,
- В — диагноз гепатит В,
- Вич — вич-положительный,
- Галлюц — использует галлюциногены,
- Сед — использует седативные препараты,

получены результаты, представленные в таблице 2.1.

Таблица 2.1. Результаты анализа значимых суперсимптомов

Суперсимптом	Энтропия	$p.v_0$	$p.v_1$	D_{KL}
фермент АЛТ				
Сед+Возд·Амф+Возд·Сед·Амф	0.6422	0.0701	0.1573	0.2446
Амф+Возд·Кон+Кон·Амф	0.6653	0.0503	0.0843	0.1997
В+Возд·Амф+Амф·В	0.7084	0.0798	0.1673	0.1787
фермент АСТ				
Галлюц·Вич+Галлюц·В	0.5197	0.6503	0.4149	0.2818
Галлюц·Год+Галлюц·Вич	0.5574	0.7527	0.3178	0.2263
Лечение+Амф+Бенз·Амф	0.6343	0.1759	0.668	0.213
Возд·Кон+Кон·В	0.5389	0.4841	0.3331	0.1983
Кокаин+Галлюц·Вич	0.5929	0.7012	0.332	0.1979

В таблице представлена энтропия Шеннона, рассчитанная согласно определению 4, для каждого суперсимптома, D_{KL} — расстояние Кульбака-Лейблера между двумя гамма-распределениями, определяемое формулой 2.2, $p.v_{0,1}$ — p.value критерия Хи-квадрат (раздел 1.4) для подвыборки из значений фермента, которая соответствует нулевому или единичному значению суперсимптома. В качестве значимых были выбраны суперсимптомы, приводящие к большому значению расстояния Кульбака-Лейблера, и имеющие достаточно большое значение энтропии. При этом так же учитывалась необходимость интерпретации суперсимптомов, поэтому были выбраны более простые комбинации категориальных признаков.

Далее рассмотрим расслоение выборок ферментов АСТ и АЛТ. Будем сравнивать подвыборки, относящиеся к разным значениям суперсимптомов, путем рассмотрения

интервальных оценок параметров гамма-распределения. Соответствующие результаты приведены в таблицах 2.2–2.3. Нижней индекс у параметра распределения относится к значению, принимаемого суперсимптомом. Необходимо чтобы параметры были однородны по всем слоям соответствующих выборок.

Таблица 2.2. Оценки параметров формы распределения для расслоения выборок

Суперсимптом	α_0	α_1
фермент АЛТ		
Сед+Возд·Амф+Возд·Сед·Амф	3.296 ± 0.035	2.935 ± 0.157
Амф+Возд·Кон+Кон·Амф	3.471 ± 0.037	2.231 ± 0.111
В+Возд·Амф+Амф·В	3.258 ± 0.036	2.953 ± 0.134
фермент АСТ		
Галлюц·Вич+Галлюц·В	3.557 ± 0.036	8.812 ± 0.685
Галлюц·Год+Галлюц·Вич	3.539 ± 0.036	8.103 ± 0.564
Лечение+Амф+Бенз·Амф	5.824 ± 0.327	3.699 ± 0.039
Возд·Кон+Кон·В	3.756 ± 0.038	5.47 ± 0.398
Кокаин+Галлюц·Вич	3.499 ± 0.036	7.975 ± 0.504

Таблица 2.3. Оценки параметров масштаба распределения для расслоения выборок

Суперсимптом	β_0	β_1
фермент АЛТ		
Сед+Возд·Амф+Возд·Сед·Амф	0.323 ± 0.004	0.535 ± 0.011
Амф+Возд·Кон+Кон·Амф	0.311 ± 0.004	0.661 ± 0.008
В+Возд·Амф+Амф·В	0.327 ± 0.004	0.504 ± 0.01
фермент АСТ		
Галлюц·Вич+Галлюц·В	0.158 ± 0.002	0.072 ± 0.028
Галлюц·Год+Галлюц·Вич	0.159 ± 0.002	0.077 ± 0.024
Лечение+Амф+Бенз·Амф	0.121 ± 0.007	0.147 ± 0.001
Возд·Кон+Кон·В	0.147 ± 0.002	0.131 ± 0.012
Кокаин+Галлюц·Вич	0.162 ± 0.002	0.074 ± 0.023

2.4. Псевдорандомизация как статистический метод устранения систематических различий сравниваемых групп

Определение 9. *РКИ — форма научного эксперимента, используемая для отслеживания факторов, не находящихся под экспериментальным контролем.*

Одним из ключевых преимуществ рандомизированных экспериментов для оценки причинных эффектов является то, что экспериментальная и контрольная группы гарантированно отличаются друг от друга только случайным образом по всем признакам, как наблюдаемым, так и ненаблюдаемым. Так как РКИ имеют существенные ограничения в использовании, к числу которых относится невозможность организовать процедуру рандомизации, а также неэтичность исследований, направленных на изучение лечебного воздействия, наиболее часто встречающимся инструментом оценки связи между фактором и исходом в медицине являются обсервационные исследования [7].

Кохрен определил обсервационное исследование как эмпирическое исследование, в котором «цель состоит в выяснении причинно-следственных связей ... в условиях, в которых невозможно использовать контролируемый эксперимент» [8]. Согласно этому определению, обсервационное исследование имеет ту же цель, что и рандомизированный эксперимент. Тем не менее, обсервационное исследование отличается от рандомизированного эксперимента использованием рандомизации для распределения индивидов по лечебной и контрольной группам.

В структуре эксперимента возможно воздействие на субъект или его отсутствие. Каждый индивид i имеет пару потенциальных исходов: $R_i(0)$ и $R_i(1)$ — результаты при отсутствии воздействия и лечении соответственно. Однако каждый индивид либо испытывает воздействие, либо нет. Пусть Z — индикаторная переменная, обозначающая воздействие на индивида ($Z = 0$ для индивида из контрольной группы, который не получает лечение; $Z = 1$ для индивида из группы воздействия, который получает лечение). Таким образом, для каждого субъекта наблюдается только один исход:

$$R_i = Z_i R_i(1) + (1 - Z_i) R_i(0).$$

Для каждого индивида эффект воздействия равен $R_i(1) - R_i(0)$. Средний эффект воздействия (сокращенно ATE от average treatment effect) определяется как

$$\mathrm{E}\left[R_i(1) - R_i(0)\right].$$

Родственным показателем среднего эффекта воздействия является средний эффект воздействия для индивидов, подвергшихся лечению (сокращенно ATT от the average treatment effect for the treated)

$$\mathrm{ATT} = \mathrm{E}\left[R_i(1) - R_i(0)|Z = 1\right].$$

В РКИ эти две меры эффективности лечения совпадают, потому что из-за рандомизации популяция, получающая лечение, в среднем не будет систематически отличаться от популяции в целом. В рандомизированном эксперименте эффект воздействия оценивают прямым сравнением двух групп

$$\widehat{\mathrm{ATE}} = \mathrm{E}\left[R_i(1)\right] - \mathrm{E}\left[R_i(0)\right].$$

В обсервационных исследованиях субъекты, получавшие лечение, часто систематически отличаются от субъектов, не получавших лечения, в общем случае

$$\mathrm{E}\left[R(1)|Z = 1\right] \neq \mathrm{E}\left[R(1)\right]$$

(и аналогично для контрольной группы.) Следовательно, объективная оценка среднего эффекта воздействия не может быть получена путем прямого сравнения результатов между двумя группами.

Определим «соответствие» в широком смысле как любой метод, направленный на уравновешивание распределения признаков в экспериментальной и контрольной группах. Цель сопоставления — уменьшить погрешность в оценке эффекта воздействия. Метод «Псевдорандомизации» (PSM сокращенно от propensity score matching), разработанный P. R. Rosenbaum и D. B. Rubin в 1983 году [9], позволяет рассчитать меру склонности — предрасположенность к воздействию с учетом наблюдаемых признаков.

Определение 10. *Условная вероятность отнесения к лечению, учитывая признаки $e(X) = P(Z = 1|X)$ называется мерой склонности (propensity score – PS), то есть предрасположенностью к воздействию лечения с учетом наблюдаемых признаков. Здесь предполагаем*

$$P(z_1, z_2, \ldots, z_n|x_1, x_2, \ldots, x_n) = \prod_{i=1}^{N} e(x_i)^{z_i}(1 - e(x_i))^{1-z_i}.$$

Особенность метода PSM заключается в том, что он позволяет свести широкий набор характеристик каждого наблюдения к единому вариационному ряду значений PS.

Есть два ключевых свойства мер склонности. Во-первых, меры склонности являются уравновешивающими мерами: при каждом значении меры склонности распределение признаков X, определяющих меру склонности, одинаково в экспериментальной и контрольной группах.

Определение 11. *Уравновешивающая мера (Balancing score)* $b(X)$ *— функция от наблюдаемых признаков* X*, такая что условное распределение* X *при заданном* $b(X)$ *одинаково для обработанных* $(Z = 1)$ *и контрольных* $(Z = 0)$ *единиц:* $Z \perp X \mid b(X)$*, где* $\perp$ *обозначает статистическую независимость. Самая тривиальная функция* $b(X) = X$*.*

Таким образом, группировка индивидов с одинаковыми показателями предрасположенности повторяет рандомизированный эксперимент в отношении наблюдаемых признаков.

Определение 12. *Назначенное лечение можно назвать полностью игнорируемым, если выполняются следующие два условия:*

$$(r_1, r_0) \perp Z|X \quad и \quad 0 < P(Z = 1|X) < 1 \; \forall \, X.$$

Первое условие говорит о том, что назначение лечения не зависит от потенциальных исходов, зависящих от наблюдаемых исходных признаков. Второе условие гласит, что каждый субъект имеет ненулевую вероятность получить какое-либо лечение. Во-вторых, в статье [9] продемонстрировано, что, если назначением лечения можно пренебречь с учетом наблюдаемых признаков, то назначение лечения также игнорируется с учетом меры склонности. Это оправдывает сопоставление на основе меры склонности, а не на полном многомерном наборе признаков. И значит сопоставление по значению меры склонности позволяет получить объективные оценки средних эффектов лечения. Приведем основные теоремы псевдорандомизации, доказанные в статье [9].

Теорема 2. *Назначение лечения и наблюдаемые признаки условно независимы с учетом меры склонности, т.е.* $X \perp Z|e(X)$*.*

Теорема 3. *Пусть* $b(X)$ *— функция от* X*. Тогда* $b(X)$ *является уравновешивающей мерой* $X \perp Z \mid b(X)$*, тогда и только тогда, когда* $b(X)$ *лучше, чем* $e(X)$ *в том смысле, что* $e(X) = f\{b(X)\}$ *для некоторой функции* f*.*

Теорема 4. *Если назначение лечения совершенно игнорируется с учетом X, тогда оно также совершенно игнорируется при любой уравновешивающей мере $b(X)$, иначе:*

$$(r_1, r_0) \perp Z|X \quad и \quad 0 < P(Z = 1|X) < 1 \text{ для всех } X$$

подразумевает

$$(r_1, r_0) \perp Z|b(X) \quad и \quad 0 < P(Z = 1|b(X)) < 1 \text{ для всех } b(X).$$

Теорема 5. *Предположим, что назначение лечения совершенно игнорируется, а $b(X)$ — уравновешивающая мера. Тогда ожидаемая разница в наблюдаемых исходах при $b(X)$ равна среднему эффекту воздействия при $b(X)$, то есть*

$$\mathrm{E}\{r_1|b(x), Z = 1\} - \mathrm{E}\{r_0|b(x), Z = 0\} = \mathrm{E}\{r_1 - r_0|b(x)\}.$$

На практике значение меры склонности чаще всего оценивается с использованием модели логистической регрессии. В этой модели группа воздействия является зависимой переменной, а наблюдаемые признаки — независимыми переменными.

$$\mathbb{P}\{Z = 1 \mid X\} = \frac{1}{1 + e^{-\theta^T X}},$$

где θ — вектор-столбец коэффициентов регрессии. На основе подобранной регрессионной модели получается прогнозируемая вероятность, попасть в группу воздействия, которая является оценкой меры склонности.

2.5. Применение псевдорандомизации

В качестве контрольной и экспериментальной группы будем рассматривать пациентов, которые принимали плацебо ($Z = 0$) и лекарственное средство ($Z = 1$). Были выбраны следующие признаки, по которым нужно добиться сбалансированности групп: пол (gender), возраст (age), уровень образования (educat), работает в текущий момент (curwor), принятие алкоголя (asid3), длительность последнего периода отказа от наркотиков (asid15), сколько месяцев назад закончился последний период отказа от наркотиков (asid16), сколько раз происходил отказ от наркотиков (asid20), стимулянты (tlfbst), галлюциногены (tlfbha), седативные (tlfbse), содержание кокаина в моче (dtcoc), содержание конопли в моче (dtcan), содержание бензодиазепина в моче (dtbenz), содержание

амфетамина в моче (dtamph), вич (hiv), гепатит B (hepb), гепатит C (hepc). Сопоставление проводилось несколькими методами до получения приемлемых результатов.

В [10] предлагается несколько мер выполнения баланса: 1) Стандартизированная разница средних $\frac{\overline{X}_1 - \overline{X}_0}{\sigma_1}$; 2) Отношение дисперсии оценки склонности в экспериментальной и контрольной группах $\frac{\sigma_1^2}{\sigma_0^2}$. Абсолютные стандартизированные различия средних должны быть менее 0.25, а отношение дисперсий должны быть в пределах от 0.5 до 2.

Для начало был выполнен наиболее простой методов сопоставления — сопоставление ближайших соседей $1:1$. Исходная выборка содержала 204 индивида, получавших препарат, и 102 индивида, получавших плацебо. Таким образом было составлено 102 пары, имеющие близкие значения мер склонности. В результате отброшено 102 индивида из группы лечения и получены довольно большие значения отношения дисперсий по признакам: содержание конопли 5.734, содержание бензодиазепина 7.086, содержание амфетамина 9.045. Чтобы не отбрасывать треть исходной выборки можно сопоставлять в отношении 2:1, и выбирать оптимальные по расстоянию пары. Этот способ не приводит к балансу по тем же признакам: содержание конопли 3.501, содержание бензодиазепина 4.022, содержание амфетамина 5.229). Тогда введем ограничение на максимальную абсолютную разницу мер склонности равное 0.25. В итоговую сопоставленную выборку вошли 99 контрольных элементов и 156 элементов из группы воздействия. Абсолютное значение стандартизированной разницы средних не превышает 0.1 по всем признакам. Отношение дисперсий для всех признаков лежит в пределах 0.9–1.23, кроме содержания бензодиазепина 0.43.

Если требуется сохранить как можно большую часть исходной выборки можно использовать метод полного сопоставления. Полное соответствие — это особый тип подклассификация, который оптимальным образом формирует подклассы. Цель подклассификации состоит в том, чтобы сформировать подклассы, в каждом из которых распределение (а не точные значения) ковариат для экспериментальной и контрольной групп будет максимально схожим. Существуют различные схемы подклассификации, включая схему, основанную на мере расстояния, такой как мера склонности. Полное сопоставление выбирает оптимальный наборов подклассов, где каждый подкласс содержит по крайней мере одного индивида из группы лечения и по крайней мере одного контрольного индивида (и в каждом наборе соответствий может быть несколько представителей любой группы). Полное соответствие является оптимальным с точки зрения

минимизации средневзвешенного значения расчетной меры расстояния между каждым лечащимся субъектом и каждым контрольным субъектом в каждом подклассе. Полное сопоставление может оценивать и ATE и ATT, в отличие от сопоставления ближайшего соседа, которое приводит только к оценке ATT. Полное сопоставление является оптимальным с точки зрения минимизации среднего значения расстояний между каждым индивидом, подвергшимся воздействию, и каждым контрольным индивидом в каждом наборе. Абсолютное значение стандартизированной разницы средних не превышает 0.1 по всем признакам. Отношение дисперсий для всех признаков лежит в пределах 0.84–1.55.

С большим числом признаков может быть трудно тщательно изучать числовые характеристики баланса; графическая диагностика может быть полезна для быстрой оценки баланса признаков. Первым шагом является изучение распределения показателей склонности в исходной и согласованной группах, распределения представлены на рисунках Б.1–Б.2. Для непрерывных признаков также можем исследовать графики квантиль-квантиль (QQ), на которых сравниваются эмпирические распределения каждой переменной в экспериментальной и контрольной группах. Графики QQ сравнивают квантили переменной в экспериментальной группе с соответствующими квантилями в контрольной группе. Если две группы имеют одинаковое эмпирическое распределение, все точки лежат на диагональной линии. Графики квантиль-квантиль представлены на рисунках Б.3–Б.5 и Б.6–Б.8 для сопоставления ближайшего соседа, с ограниченным максимальным расстоянием и полного сопоставления соответственно.

Наконец, график стандартизированных разностей средних дает нам краткую информацию об улучшении баланса для отдельных признаков. В некоторых ситуациях стандартизованная разница средних значений нескольких признаков будет увеличиваться. Это может быть особенно верно для признаков, которые находились в равновесии до сопоставления, поскольку они не будут сильно влиять на модель оценки склонности. Сравнение плотностей мер склонности до сопоставления и после представлено на рисунках Б.9 и Б.10 для сопоставления ближайшего соседа, с ограниченным максимальным расстоянием и полного сопоставления соответственно.

Таким же образом можем рассмотреть в качестве контрольной и экспериментальной группы пациентов, которые принимали налтрексон ($Z = 0$) и налтрексон-имплант ($Z = 1$). Добиваться сбалансированности групп будем по тем же признакам. Примене-

ние метода полного сопоставления приводит к абсолютным значениям стандартизированной разницы средних, не превышающих 0.1 по всем признакам. Отношение дисперсий для всех признаков лежит в пределах 0.5–1.6. Распределения показателей склонности в исходной и согласованной группах представлены на рисунке Б.11. Графики квантиль-квантиль представлены на рисунках Б.12–Б.14. Наконец, сравнение плотностей мер склонности до сопоставления и после представлено на рисунке Б.15.

Глава 3

Степенная модель

3.1. Модель степенного гамма-распределения

Одной из особенностью гамма-распределения является то, что оно может быть аппроксимировано степенным гамма-распределением с достаточно широким спектром параметров [11].

Определение 13. *Пусть случайная величина ξ имеет гамма-распределение $\gamma(\alpha, \beta)$ с параметром масштаба β и параметром формы α. Тогда случайная величина $X = \xi^{\frac{1}{\kappa}}$, $\kappa > 0$ имеет степенное гамма-распределение $\gamma_p(\alpha, \beta, \kappa)$ с плотностью*

$$\begin{cases} \frac{\kappa}{\beta^\alpha \Gamma(\alpha)} x^{\kappa\alpha-1} e^{-\frac{x^\kappa}{\beta}}, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

Свойства степенного гамма-распределения: Пусть $\xi \sim \gamma_p(x, \alpha, \beta, \kappa)$, тогда

- $\mathbb{E}\xi = \frac{\Gamma(\alpha+1/\kappa)}{\Gamma(\alpha)} \beta^{1/\kappa}$
- $\mathbb{D}\xi = \frac{\Gamma(\alpha+2/\kappa)\Gamma(\alpha) - \Gamma^2(\alpha+1/\kappa)}{\Gamma^2(\alpha)} \beta^{2/\kappa}$

3.1.1. Синонимичные степенные гамма-распределения

Определение 14. *Распределение с плотностью $f_2(x)$ синонимично с односторонним уровнем синонимии δ_1 распределение с плотностью $f_1(x)$, если $H_{12} - H_{11} < \delta_1$.*

Для гамма-распределения $\gamma(\alpha, \beta)$ (которое совпадает с распределением $\gamma_p(\alpha, \beta, 1)$) в качестве номинативного выбирается синонимичное степенное гамма-распределение с наименьшей смешанной дифференциальной энтропией H_{12} [5]. Рассмотрим H_{12}, где за $f_1(x)$ возьмем плотность распределения $\gamma(\alpha_1, \beta_1)$, а за $f_2(x)$ плотность распределения $\gamma_p(\alpha_2, \beta_2, \kappa_2)$. Получаем:

$$H_{12} = -\int_0^{+\infty} \frac{x^{\alpha_1-1} e^{-x/\beta_1}}{\beta_1^{\alpha_1} \Gamma(\alpha_1)} \cdot \ln\left(\frac{\kappa_2}{\beta_2^{\alpha_2} \Gamma(\alpha_2)} x^{\kappa_2\alpha_2-1} e^{-\frac{x^{\kappa_2}}{\beta_2}} \right) dx.$$

Аналогично рассуждениям проделанным ранее, с помощью той же замены $z = x/\beta_1$ получим окончательно:

$$H_{12} = -\ln \frac{\kappa_2 \beta_1^{\kappa_2 \alpha_2 - 1}}{\beta_2^{\alpha_2} \Gamma(\alpha_2)} - (\kappa_2 \alpha_2 - 1)\psi(\alpha_1) + \frac{\beta_1^{\kappa_2} \Gamma(\alpha_1 + \kappa_2)}{\beta_2 \Gamma(\alpha_1)} \tag{3.1}$$

Чтобы найти аргументы минимизирующие H_{12} продифференцируем 3.1 по α_2, β_2, κ_2.

$$\frac{\partial H_{12}}{\partial \alpha_2} = -\kappa_2 \ln \beta_1 + \ln \beta_2 + \psi(\alpha_2) - \kappa_2 \psi(\alpha_1) = \ln \beta_2 + \psi(\alpha_2) - \kappa_2(\psi(\alpha_1) + \ln \beta_1) = 0, \tag{3.2}$$

$$\frac{\partial H_{12}}{\partial \beta_2} = \frac{\alpha_2}{\beta_2} - \frac{\beta_1^{\kappa_2} \Gamma(\alpha_1 + \kappa_2)}{\beta_2^2 \Gamma(\alpha_1)} = 0, \tag{3.3}$$

$$\frac{\partial H_{12}}{\partial \kappa_2} = -\left(\frac{1}{\kappa_2} + \alpha_2(\ln \beta_1 + \psi(\alpha_1))\right) + \frac{\beta_1^{\kappa_2} \Gamma(\alpha_1 + \kappa_2)(\psi(\alpha_1 + \kappa_2) + \ln \beta_1)}{\beta_2 \Gamma(\alpha_1)} = 0. \tag{3.4}$$

Из уравнения 3.2 выразим κ_2, из уравнения 3.3 — β_2 и упрощая уравнение 3.4 получим α_2. В результате:

$$\begin{cases} \alpha_2 = \left(\kappa_2(\psi(\alpha_1 + \kappa_2) - \psi(\alpha_1))\right)^{-1}, \\ \beta_2 = \frac{\beta_1^{\kappa_2} \Gamma(\alpha_1 + \kappa_2)}{\alpha_2 \Gamma(\alpha_1)}, \\ \kappa_2 = \frac{\ln \beta_1 + \psi(\alpha_1)}{\ln \beta_2 + \psi(\alpha_2)}. \end{cases} \tag{3.5}$$

При $\kappa_2 = \kappa_1 = 1$ решение будет тривиальным.

Таким образом, имея гамма-распределение можем перейти к номинативному степенному гамма-распределению. Для этого при фиксированном κ_2 вычислим параметры масштаба и формы α_2, β_2 степенного распределения согласно системе 3.5. Параметр κ_2 затем находится путем минимизации энтропии степенного распределения. Дифференциальная энтропия согласно 2.3 задается формулой:

$$H_{22} = -\int_0^{+\infty} \frac{\kappa_2 x^{\kappa_2 \alpha_2 - 1} e^{-x^{\kappa_2}/\beta_2}}{\beta_2^{\alpha_2} \Gamma(\alpha_2)} \ln \left[\frac{\kappa_2 x^{\kappa_2 \alpha_2 - 1} e^{-x^{\kappa_2}/\beta_2}}{\beta_2^{\alpha_2} \Gamma(\alpha_2)}\right] dx.$$

Сделаем замену аналогичную предыдущим: $z = \frac{x^{\kappa_2}}{\beta_2}$, $dx = \frac{\beta_2^{1/\kappa_2} z^{1/\kappa_2 - 1}}{\kappa_2}$. После замены и преобразований получим интеграл:

$$H_{22} = -\int_0^{+\infty} \frac{z^{\alpha_2 - 1} e^{-z}}{\Gamma(\alpha_2)} \cdot \left[\ln \kappa_2 - \frac{\ln \beta_2}{\kappa_2} + (\alpha_2 - 1/\kappa_2) \ln z - \ln \Gamma(\alpha_2) - z\right] dz$$

$$H_{22} = -\ln \kappa_2 + \frac{\ln \beta_2}{\kappa_2} - (\alpha_2 - 1/\kappa_2)\psi(\alpha_2) + \ln \Gamma(\alpha_2) + \alpha_2$$

$$H_{22} = -\ln\kappa_2 + \frac{\ln\beta_2 + \psi(\alpha_2)}{\kappa_2} - \alpha_2\psi(\alpha_2) + \ln\Gamma(\alpha_2) + \alpha_2. \tag{3.6}$$

При значениях параметров $\alpha_1 = 5,\ \beta_1 = 0.3$ было получено передельное значения параметра κ_2 равное 0.75. При этом получаются следующие значения информационных метрик $H_{22} = 0.949,\ H_{12} = 0.95$.

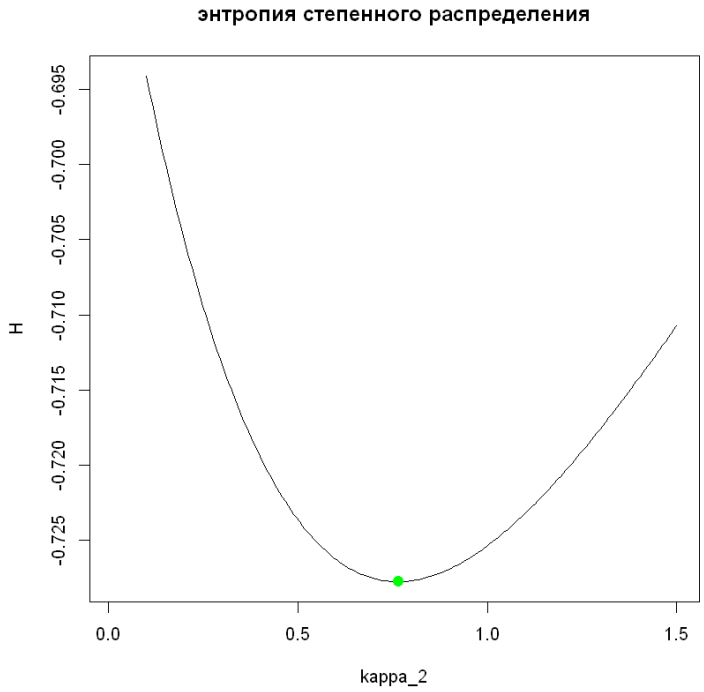

Рис. 3.1. Зависимость дифференциальной энтропии степенного распределения от параметра κ_2

3.2. Теорема Кульбака-Санова

Теорема 6 (Кульбака-Санова [12])**.** *Если $f_2(x)$ — произвольная, а $f_1(x)$ — фиксированная плотность распределения из семейства вероятностных мер с доминирующей мерой, $Y = T(x)$ измеримая статистика, что существуют $\Theta = \int T(x) f_2(x) d\lambda(x)$ и $M_1(\tau) = \int f_1(x) e^{\tau T(x)} d\lambda(x)$ для τ из некоторого интервала, то*

$$I(2:1) \geq \Theta\tau - \log M_1(\tau) = I(*:1),\ \Theta = \frac{d}{d\tau}\log M_1(\tau)$$

с равенством $f_2(x) = f^*(x) = e^{\tau T(x)} f_1(x)/M_1(\tau)[\lambda]$.

Здесь $I(2:1)$ — информация для различия в пользу распределения с плотностью $f_2(x)$. Иначе это несимметричное расстояние Кульбака-Лейблера и, как было определе-

но ранее, оно задается формулой

$$I(2:1) = H_{21} - H_{22} = \int f_2(x) \log \frac{f_2(x)}{f_1(x)} dx.$$

Воспользуемся теоремой 6, чтобы получить номинативного представителя среди синонимичных степенных гамма-распределений. Пусть $f_1(x) = \frac{1}{\beta_1^{\alpha_1}\Gamma(\alpha_1)} x^{\alpha_1 - 1} e^{-x/\beta_1}$, при $x > 0$. Статистика $T(x) = x$. Тогда существует

$$\Theta = \int_0^{+\infty} x f_2(x) dx.$$

$$M_1(\tau) = \int_0^{+\infty} f_1(x) \cdot e^{\tau T(x)} dx = \int_0^{+\infty} \frac{1}{\beta_1^{\alpha_1}\Gamma(\alpha_1)} x^{\alpha_1 - 1} e^{-x\frac{1-\tau\beta_1}{\beta_1}} dx = (1 - \beta_1\tau)^{-\alpha_1}.$$

$$f_2(x) = \frac{e^{\tau x} f_1(x)}{M_1(\tau)} = \frac{e^{\tau x}(1-\beta_1\tau)^{\alpha_1} x^{\alpha_1 - 1} e^{-\frac{x}{\beta_1}}}{\beta_1^{\alpha_1}\Gamma(\alpha_1)} = \frac{x^{\alpha^* - 1} e^{-\frac{x}{\beta^*}}}{\beta^{*\alpha^*}\Gamma(\alpha^*)},$$

где $\alpha^* = \alpha_1,\ \beta^* = \frac{\beta_1}{1-\tau\beta_1},\ \tau = 1/\beta_1 - 1/\beta^*$.

$$\Theta = \frac{d}{d\tau} \log M_1(\tau) = \frac{d}{d\tau} \left(-\alpha_1 \log(1 - \beta_1\tau)\right) = \frac{\alpha_1\beta_1}{1 - \beta_1\tau} = \alpha^*\beta^*.$$

Следовательно, $1/\beta^* = \alpha^*/\Theta = \alpha_1/\Theta$, а $\tau = 1/\beta_1 - \alpha_1/\Theta$. То есть

$$I(*:1) = \Theta\tau - \ln M_1(\tau) = \frac{\Theta}{\beta_1} - \alpha_1 + \alpha_1 \ln\left(1 - \beta_1(1/\beta_1 - \alpha_1/\Theta)\right) = \frac{\Theta}{\beta_1} - \alpha_1 + \alpha_1 \ln\left(\frac{\alpha_1\beta_1}{\Theta}\right).$$

Пусть $f_3(x) = \frac{1}{\beta_2^{\alpha_2}\Gamma(\alpha_2)} x^{\alpha_2 - 1} e^{-\frac{x}{\beta_2}}$, при $x > 0$. Тогда

$$I(3:1) = \ln \frac{\beta_1^{\alpha_1}\Gamma(\alpha_1)}{\beta_2^{\alpha_1}\Gamma(\alpha_2)} + (\alpha_2 - \alpha_1)\psi(\alpha_2) - \alpha_2 + \frac{\beta_2\Gamma(\alpha_2)}{\beta_1\Gamma(\alpha_2)}.$$

Ранее мы определили параметры номинативного представителя исходя из минимизации энтропии H_{12} и H_{22}. Проверим совпадает ли полученный результат с утверждением теоремы. Так как для $\xi \sim \gamma_p(\alpha, \beta, \kappa)$ верно $\xi^\kappa \sim \gamma(\alpha, \beta)$, то можем вычислять различающую информацию для гамма-распределения с параметрами, совпадающими с параметрами степенного распределения. Параметры степенного распределения: $\beta_2 = \frac{\beta_1^{\kappa_2}\Gamma(\alpha_1 + \kappa_2)}{\alpha_2\Gamma(\alpha_1)}$, $\alpha_2 = \left(\kappa_2(\psi(\alpha_1 + \kappa_2) - \psi(\alpha_1))\right)^{-1}$. Проверяем выполнение неравенства:

$$\ln \frac{\beta_1^{\alpha_1}\Gamma(\alpha_1)}{\beta_2^{\alpha_1}\Gamma(\alpha_2)} + (\alpha_2 - \alpha_1)\psi(\alpha_2) - \alpha_2 + \frac{\beta_2\Gamma(\alpha_2)}{\beta_1\Gamma(\alpha_1)} \geq \Theta/\beta_1 - \alpha_1 + \alpha_1 \ln\left(\frac{\alpha_1\beta_1}{\Theta}\right),$$

где Θ — математическое ожидание случайной величины с плотностью $f_3(x)$.

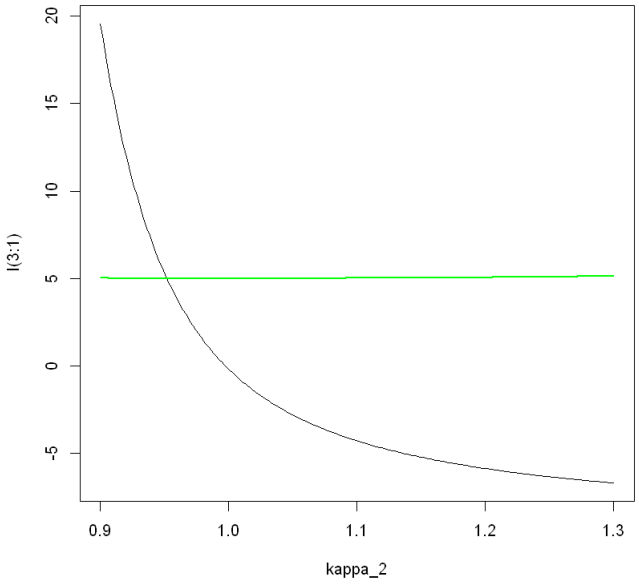

Рис. 3.2. Зависимость различающей информации I(3:1) от параметра κ_2

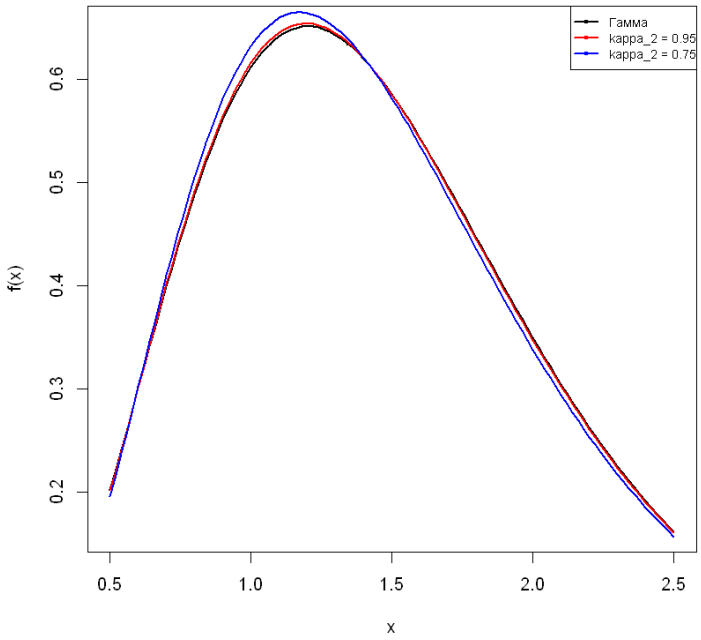

Рис. 3.3. Сравнение плотностей исходного гамма и номинативных степенных распределений

На рисунке 3.2 зеленая линия — предельное значение различающей информации. Таким образом, неравенство выполняется для значения параметра $\kappa_2 \leq 0.95$. Полученное ранее предельное значение $\kappa_2 = 0.75$ удовлетворяет утверждении теоремы, но не доставляет минимум различающей информации. При параметрах исходного гамма-распределения $\alpha_1 = 5$, $\beta_1 = 0.3$, получаем значения $H_{22} = 0.949$ и $H_{12} = 0.95$. На рисунке 3.3 представлено сравнение плотности гамма-распределения и полученных но-

минативных представителей по одному из представленных выше способов. Таким образом, при достаточно большом значении параметра α_1 предельное значение κ_2 приводит к результатам, аналогичным результатам полученным с использованием теоремы Кульбака-Санова.

3.3. Построение доверительных интервалов для параметров степенного распределения

Из свойств гамма-распределения легко получить, что $E\ln(\xi) = \ln\beta + \psi(\alpha)$ и $E(\xi^\kappa) = \beta^\kappa \cdot \frac{\Gamma(\alpha+\kappa)}{\Gamma(\alpha)}$, где $\xi \sim \gamma(\alpha, \beta)$. Таким образом можно переписать формулу для дифференциальной энтропии 3.1 следующим образом:

$$H_{12} = -\ln\kappa_2 + \alpha_2\ln\beta_2 + \ln\Gamma(\alpha_2) - (\kappa_2\alpha_2 - 1)E(\ln X_1) + \frac{E(X_1^{\kappa_2})}{\beta_2}. \tag{3.7}$$

Здесь мы продолжаем рассуждения раздела 3.1.1 и предполагаем, что случайная величина X_1 из гамма-распределения с параметрами α_1, β_1 и имеет плотность $f_1(x)$. Ранее мы дифференцировали функцию H_{12} для получения параметров номинативного распределения, в новых обозначениях удобно получить свойства моментов случайной величины из этого распределения. Напомним, что плотность степенного распределения обозначалась $f_2(x)$.

Утверждение 1. *Параметры номинативного распределения удовлетворяют формуле*

$$EX_1^{\kappa_2} = \alpha_2\beta_2 = EX_2^{\kappa_2},$$

где X_2 случайная величина из номинативного распределения.

Доказательство 1. *Дифференцирование функции H_{12} по параметру β_2 приводит к выражению:*

$$\frac{\alpha_2}{\beta_2} - \frac{EX_1^{\kappa_2}}{\beta_2^2},$$

которое приравнивается к нулю для минимизации дифференциальной энтропии. Следовательно после преобразования получим $EX_1^{\kappa_2} = \alpha_2\beta_2$. Случайная величина из степенного гамма-распределения по определению при возведении в соответствующую степень становится случайной величиной из гамма-распределения с теми же параметрами масштаба и формы. А по свойствам гамма-распределения, верно $\alpha_2\beta_2 = EX_2^{\kappa_2}$.

Утверждение 2. *Параметр масштаба номинативного распределения удовлетворяет формуле*

$$\beta_2 = \operatorname{cov}(X_1^{\kappa_2}, \ln X_1^{\kappa_2})$$

где X_1 случайная величина из гамма-распределения.

Доказательство 2. *Вычислим моменты некоторых случайных величин.*

$$E \ln X_1^{\kappa_2} = \int\limits_0^\infty \ln(x^{\kappa_2}) \cdot \frac{x^{\alpha_1 - 1} e^{-x/\beta_1}}{\beta_1^{\alpha_1} \Gamma(\alpha_1)} dx = \int\limits_0^\infty \kappa_2 \left(\ln \beta_1 + \ln z\right) \cdot \frac{z^{\alpha_1 - 1} e^{-z}}{\Gamma(\alpha_1)} dz =$$

$$= \kappa_2 \left[\ln \beta_1 + \psi(\alpha_1)\right];$$

$$E(X_1^{\kappa_2} \ln X_1^{\kappa_2}) = \int\limits_0^\infty \ln(x^{\kappa_2}) \cdot \frac{x^{\alpha_1 + \kappa_2 - 1} e^{-x/\beta_1}}{\beta_1^{\alpha_1} \Gamma(\alpha_1)} dx = \int\limits_0^\infty \kappa_2 \beta_1^{\kappa_2} \left(\ln \beta_1 + \ln z\right) \cdot \frac{z^{\alpha_1 + \kappa_2 - 1} e^{-z}}{\Gamma(\alpha_1)} dz$$

$$= \frac{\kappa_2 \beta_1^{\kappa_2} \Gamma(\alpha_1 + \kappa_2) \left(\ln \beta_1 + \psi(\alpha_1 + \kappa_2)\right)}{\Gamma(\alpha_1)}.$$

Теперь с помощью полученных математических ожиданий перепишем формулу 3.4 для частной производной H_{12}

$$-\frac{1}{\kappa_2} - \frac{\alpha_2}{\kappa_2} E \ln X_1^{\kappa_2} + \frac{E(X_1^{\kappa_2} \ln X_1^{\kappa_2})}{\kappa_2 \beta_2} = 0$$

$$E(X_1^{\kappa_2} \ln X_1^{\kappa_2}) - \alpha_2 \beta_2 E \ln X_1^{\kappa_2} = \beta_2$$

Воспользуемся утверждением 1 и окончательно получим

$$\beta_2 = E(X_1^{\kappa_2} \ln X_1^{\kappa_2}) - E(X_1^{\kappa_2}) E \ln X_1^{\kappa_2} = \operatorname{cov}(X_1^{\kappa_2}, \ln X_1^{\kappa_2})$$

3.4. Доверительные интервалы по методу максимума правдоподобия для параметров степенного распределения

Для построения доверительного интервала по методу МП нужно вычислить информант второго рода. Выполним необходимые вычисления для случайной величины ξ, имеющей степенное гамма-распределение $\gamma_p(\alpha, \beta, \kappa)$ с плотностью

$$\begin{cases} \frac{\kappa}{\beta^\alpha \Gamma(\alpha)} x^{\kappa\alpha - 1} e^{-\frac{x^\kappa}{\beta}}, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

Функция правдоподобия имеет вид:

$$L(x, \Theta) = \frac{\kappa^n}{\beta^{n\alpha} \Gamma^n(\alpha)} \prod_{i=1}^n x_i^{\kappa\alpha - 1} e^{-x_i^\kappa/\beta},$$

где $\Theta = (\kappa, \alpha, \beta)$ — вектор параметров. Перейдем к логарифму функции правдоподобия

$$\ln L(x, \Theta) = n \ln(\kappa) - n\alpha \ln \beta - n \ln \Gamma(\alpha) + (\kappa\alpha - 1) \sum_{i=1}^{n} \ln x_i - \frac{1}{\beta} \sum_{i=1}^{n} x_i^{\kappa}.$$

Найдем информант первого рода $s(x, \Theta)$ — производную по параметру логарифма функции правдоподобия:

$$\begin{cases} \frac{\partial \ln L}{\partial \kappa} = \frac{n}{\kappa} + \alpha \sum\limits_{i=1}^{n} \ln x_i - \frac{\kappa}{\beta} \sum\limits_{i=1}^{n} x_i^{(\kappa-1)}, \\ \frac{\partial \ln L}{\partial \alpha} = -n \ln \beta - n\psi(\alpha) + \kappa \sum\limits_{i=1}^{n} \ln x_i, \\ \frac{\partial \ln L}{\partial \beta} = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2} \sum\limits_{i=1}^{n} x_i^{\kappa}. \end{cases}$$

Найдем информант второго рода $I(\theta) = -\mathbb{E} \frac{\partial^2}{\partial \theta^2} \ln L(x, \theta) = -\mathbb{E} \frac{\partial}{\partial \theta} s(x, \theta)$. Для этого сначала составим матрицу вторых производных:

$$\begin{pmatrix} -\frac{n}{\kappa^2} - \frac{1}{\beta} \sum\limits_{i=1}^{n} x_i^{(\kappa-1)} - \frac{\kappa(\kappa-1)}{\beta} \sum\limits_{i=1}^{n} x_i^{(\kappa-2)} & \sum\limits_{i=1}^{n} \ln x_i & \frac{\kappa}{\beta^2} \sum\limits_{i=1}^{n} x_i^{(\kappa-1)} \\ \sum\limits_{i=1}^{n} \ln x_i & -n\psi'(\alpha) & -\frac{n}{\beta} \\ \frac{\kappa}{\beta^2} \sum\limits_{i=1}^{n} x_i^{(\kappa-1)} & -\frac{n}{\beta} & \frac{n\alpha}{\beta^2} - \frac{2}{\beta^3} \sum\limits_{i=1}^{n} x_i^{\kappa} \end{pmatrix}.$$

Также найдем необходимые моменты:

- $\mathbb{E}x^k = \alpha\beta$,
- $\mathbb{E}x^{(k-1)} = \frac{\Gamma(\alpha+1-1/\kappa)}{\Gamma(\alpha)} \beta^{1-1/\kappa}$,
- $\mathbb{E}x^{(k-2)} = \frac{\Gamma(\alpha+1-2/\kappa)}{\Gamma(\alpha)} \beta^{1-2/\kappa}$,
- $\mathbb{E} \ln x = \frac{\ln \beta + \psi(\alpha)}{\kappa}$.

Тогда информационная матрица имеет вид:

$$I(\theta) = n \begin{pmatrix} \frac{1}{\kappa^2} + \frac{\beta^{-1/\kappa}\Gamma(\alpha+1-1/\kappa) + \kappa(\kappa-1)\beta^{-2/\kappa}\Gamma(\alpha+1-2/\kappa)}{\Gamma(\alpha)} & -\frac{\ln \beta + \psi(\alpha)}{\kappa} & -\frac{\kappa\Gamma(\alpha+1-1/\kappa)}{\beta^{1/\kappa}\Gamma(\alpha)} \\ -(\ln \beta + \psi(\alpha))/\kappa & \psi'(\alpha) & \beta^{-1} \\ -\frac{\kappa\Gamma(\alpha+1-1/\kappa)}{\beta^{1/\kappa}\Gamma(\alpha)} & \beta^{-1} & \alpha\beta^{-2} \end{pmatrix}.$$

Построим асимптотические доверительные интервалы (с уровнем доверия $1 - \alpha$), основанные на теории нормального распределения, как это было сделано в разделе 1.3:

$$P(\hat{\theta} - u_{1-\alpha/2} \cdot \frac{\sigma(\hat{\theta})}{\sqrt{n}} < \theta < \hat{\theta} + u_{1-\alpha/2} \cdot \frac{\sigma(\hat{\theta})}{\sqrt{n}}) = 1 - \alpha,$$

где $u_{1-\alpha/2}$ — квантиль стандартного нормального распределения, n — размер выборки. Чтобы получить стандартное отклонение параметров $\hat{\sigma}(\theta)$ нужно извлечь корень из диагональных элементов матрицы, обратной к информационной I^{-1}. Для упрощения формул приведем только искомые элементы обратной матрицы:

$$\operatorname{diag}\left(I^{-1}(\theta)\right) = \frac{1}{nD}\begin{bmatrix} \left(\psi'(\alpha)\alpha - 1\right)/\beta^2 \\ \alpha \cdot \frac{\beta^{2/\kappa}\Gamma(\alpha)+\kappa^2\beta^{1/\kappa}\Gamma(\alpha+1-1/\kappa)+\kappa^3(\kappa-1)\Gamma(\alpha+1-2/\kappa)}{\kappa^2\beta^{2+2/\kappa}\Gamma(\alpha)} - \frac{\kappa^2\Gamma^2(\alpha+1-1/\kappa)}{\beta^{2/\kappa}\Gamma^2(\alpha)} \\ \psi'(\alpha) \cdot \frac{\Gamma(\alpha)+\kappa^2\beta^{-1/\kappa}\Gamma(\alpha+1-1/\kappa)+\kappa^3(\kappa-1)\beta^{-2/\kappa}\Gamma(\alpha+1-2/\kappa)}{\kappa^2\Gamma(\alpha)} - \frac{\alpha^2}{\beta^4} \end{bmatrix},$$

где

$$D = \frac{\Gamma(\alpha) + \kappa^2\beta^{-1/\kappa}\Gamma(\alpha+1-1/\kappa) + \kappa^3(\kappa-1)\beta^{-2/\kappa}\Gamma(\alpha+1-2/\kappa)}{\kappa^2\beta^2\Gamma(\alpha)} \cdot \left(\psi'(\alpha)\alpha - 1\right) -$$

$$- \frac{\left(\ln\beta + \psi(\alpha)\right)^2\alpha}{\kappa^2\beta^2} - \frac{2\left(\ln\beta + \psi(\alpha)\right)\Gamma(\alpha+1-1/\kappa)}{\beta^{1+1/\kappa}\Gamma(\alpha)} - \frac{\psi'(\alpha)\kappa^2\Gamma^2(\alpha+1-1/\kappa)}{\beta^{2/\kappa}\Gamma^2(\alpha)}.$$

3.5. Сравнение параметров степенного гамма-распределения для групп

Перейдем к сравнению параметров степенных распределений. Учитывая свойства и связь степенного и обычного гамма-распределений, можно перейти от степенного распределения к гамма-распределению и получить оценки параметров масштаба и формы по методу максимума правдоподобия. В таблицах 3.1–3.2 эти оценки с доверительными интервалами обозначены без знака штрих и получены, как и ранее, с помощью функции fitdistr. Параметр κ_2, используемый при переходе к гамма-распределению, взят из условия минимизации энтропии распределения 3.1.1. Если же нужно получить оценки параметров не прибегая к такому преобразованию, то утверждение 2 позволяет преобразовать случайную величину таким образом, что параметр β_2 будет оценен как среднее. Как построить доверительный интервал для среднего уже известно и он может выглядеть как асимптотический интервал, построенный в разделе 1.3. Однако параметр α_2 придется оценивать исходя из свойств гамма-распределения, также используя среднее значение случайной величины. Оценки, полученные таким образом, в таблицах представлены со знаком штрих. Параметры оценивались по исходной выборке и после выполнения псевдорандомизации (2.4) на преобразованной выборке (проводилось полное сопоставление).

Таблица 3.1. Сравнение параметров формы степенных гамма-распределений

	α_0	α_1	$\alpha_0^{'}$	$\alpha_1^{'}$
фермент АЛТ				
Без балансирования	5.224 ± 0.142	5.179 ± 0.069	6.685 ± 3.701	7.149 ± 2.87
С балансированием	1.705 ± 0.044	2.986 ± 0.039	3.64 ± 1.71	6.686 ± 3.175
фермент АСТ				
Без балансирования	6.42 ± 0.175	6.541 ± 0.088	8.607 ± 8.547	8.886 ± 6.651
С балансированием	1.864 ± 0.048	3.568 ± 0.047	4.054 ± 2.783	8.052 ± 5.869

Таблица 3.2. Сравнение параметров масштаба степенных гамма-распределений

	β_0	β_1	$\beta_0^{'}$	$\beta_1^{'}$
фермент АЛТ				
Без балансирования	0.217 ± 0.006	0.202 ± 0.003	0.169 ± 0.053	0.146 ± 0.042
С балансированием	0.693 ± 0.021	0.357 ± 0.019	0.292 ± 0.118	0.151 ± 0.067
фермент АСТ				
Без балансирования	0.104 ± 0.003	0.096 ± 0.002	0.077 ± 0.019	0.071 ± 0.017
С балансированием	0.369 ± 0.011	0.178 ± 0.011	0.176 ± 0.052	0.086 ± 0.031

Перейдем к сравнению тех же параметров, но в качестве нулевой группы рассмотрим пациентов, принимавших налтрексон, а в качестве первой группы — пациентов, котором был назначен налтрексон-имплант. Результаты сравнения соответствующих параметров представлены в таблицах 3.3–3.4.

Таблица 3.3. Сравнение параметров формы степенных гамма-распределений

	α_0	α_1	$\alpha_0^{'}$	$\alpha_1^{'}$
фермент АЛТ				
Без балансирования	4.279 ± 0.114	6.668 ± 0.179	6.095 ± 3.005	8.816 ± 5.953
С балансированием	1.629 ± 0.041	2.172 ± 0.056	3.457 ± 1.727	4.701 ± 2.799
фермент АСТ				
Без балансирования	5.403 ± 0.145	8.537 ± 0.23	7.473 ± 6.867	14.356 ± 11.374
С балансированием	2.093 ± 0.054	2.431 ± 0.063	4.534 ± 3.685	5.326 ± 4.695

Таблица 3.4. Сравнение параметров масштаба степенных гамма-распределений

	β_0	β_1	β_0'	β_1'
фермент АЛТ				
Без балансирования	0.253 ± 0.007	0.151 ± 0.012	0.178 ± 0.077	0.115 ± 0.035
С балансированием	0.713 ± 0.021	0.475 ± 0.031	0.302 ± 0.166	0.207 ± 0.101
фермент АСТ				
Без балансирования	0.121 ± 0.003	0.071 ± 0.006	0.088 ± 0.032	0.053 ± 0.013
С балансированием	0.319 ± 0.009	0.251 ± 0.012	0.154 ± 0.075	0.124 ± 0.046

Глава 4

Двумерная модель

4.1. Обоснование модели двумерного гамма-распределения

Рассмотрим модель двумерного гамма-распределения. Определение для многомерного случая дается в [13].

Определение 15. *Пусть* $\xi_i \sim G(\alpha_i, \beta_i, \gamma_i), i = \overline{0:k}$ *независимые случайные величины. Положим*

$$Z_i = \frac{\beta_i}{\beta_0}\xi_0 + \xi_i, i = \overline{1:k}.$$

Совместное распределение вектора $Z = (Z_1, \ldots, Z_k)$ *будем называть многомерным гамма-распределением. Обозначение* $X \sim G(\alpha, \beta, \gamma)$ *значит, что* X *имеет плотность*

$$f(x, \alpha, \beta, \gamma) = \frac{(x-\gamma)^{\alpha-1}\exp\left(-(x-\gamma)/\beta\right)}{\beta^\alpha \Gamma(\alpha)},\ x > \gamma, \alpha > 0, \beta > 0,$$

и ноль иначе. Здесь α, β, γ *параметры формы, масштаба и местоположения.*

Определение 16. *Производящей функцией моментов случайной величины* X *называется функция вида:* $\varphi_X(t) = E(e^{tX})$. *С помощью производящей функции могут быть вычислены моменты случайной величины по формуле* $E(X^n) = \frac{d^n}{dt^n}\varphi_X(t)|_{t=0}$.

Некоторые свойства распределения следуют из производящей функции моментов 16.

$$E(e^{t\xi}) = \int_\gamma^\infty \frac{(x-\gamma)^{\alpha-1}\exp\left(-(x-\gamma)/\beta\right)\exp(tx)}{\beta^\alpha\Gamma(\alpha)}dx = \frac{\exp(\gamma t)}{(1-\beta t)^\alpha}.$$

Тогда

$$\varphi_Z(t) = E(e^{t_1 z_1 + \ldots + t_k z_k}) = E(e^{(1/\beta_0)(\beta_1 t_1 + \ldots + \beta_k t_k)\xi_0})E(e^{t_1\xi_1})\ldots E(e^{t_k\xi_k}) =$$
$$= \frac{e^{(\gamma_0/\beta_0)(\beta_1 t_1 + \ldots + \beta_k t_k)}e^{\gamma_1 t_1 + \ldots + \gamma_k t_k}}{[1-(\beta_1 t_1 + \ldots + \beta_k t_k)]^{\alpha_0}(1-\beta_1 t_1)^{\alpha_1}\ldots(1-\beta_k t_k)^{\alpha_k}} =$$
$$= \frac{\exp[(g + (\gamma_0/\beta_0)b)^{\mathrm{T}}t]}{(1-b^{\mathrm{T}}t)^{\alpha_0}\prod_{i=1}^k(1-\beta_i t_i)^{\alpha_i}},$$

где $b = (\beta_1, \ldots, \beta_k)^{\mathrm{T}}$, $g = (\gamma_1, \ldots, \gamma_k)^{\mathrm{T}}$, $t = (t_1, \ldots, t_k)^{\mathrm{T}}$, $|\beta_i t_i| < 1$ для всех i, а также $|b^{\mathrm{T}}t| < 1$. Теперь легко получить следующие свойства непосредственно из определения или из производящей функции моментов:

- $Z_i \sim G(\alpha_0 + \alpha_i, \beta_i, (\gamma_0/\beta_0)\beta_i + \gamma_i)$,
- $E(Z_i) = (\alpha_0 + \alpha_i)\beta_i + (\gamma_0/\beta_0)\beta_i + \gamma_i$,
- $D(Z_i) = (\alpha_0 + \alpha_i){\beta_i}^2$,
- $\mathrm{cov}(Z_i, Z_j) = \alpha_0\beta_i\beta_j, \quad i \neq j$.

Из последнего очевидно, что Z_i и Z_j при $i \neq j$ положительно коррелируют.

Упростим модель для дальнейшего рассмотрения, а именно положим все параметры местоположения γ_i равными нулю и перейдем к величинам $Z_i \sim G(\alpha_0 + \alpha_i, \beta_i)$. Моменты $E({Z_i}^m)$ будут получены из определения, их вычисление требует знания моментов $E({\xi_i}^m)$, которые известны из производящей функции моментов. А именно:

$$\varphi_{\xi_i}(t) = (1 - \beta t)^{-\alpha},$$

$$\frac{d^m \varphi_{\xi_i}(t)}{dt^m} = \alpha_i(\alpha_i + 1)\dots(\alpha_i + m - 1){\beta_i}^m(1 - \beta t)^{-\alpha - m}.$$

Положим $t = 0$, тогда получим

$$E({\xi_i}^m) = \alpha_i(\alpha_i + 1)\dots(\alpha_i + m - 1){\beta_i}^m. \tag{4.1}$$

Перейдем к моментам случайных величин Z_i:

$$E({Z_i}^m) = E\left(\frac{\beta_i}{\beta_0}\xi_0 + \xi_i\right)^m = E\left(\sum_{r=0}^{m} C_m^r \left(\frac{\beta_i}{\beta_0}\right)^r {\xi_0}^r {\xi_i}^{m-r}\right) =$$

$$= \sum_{r=0}^{m} C_m^r \left(\frac{\beta_i}{\beta_0}\right)^r E(\xi_0)^r E(\xi_i)^{m-r} =$$

$$= \sum_{r=0}^{m} C_m^r \left(\frac{\beta_i}{\beta_0}\right)^r \alpha_0(\alpha_0 + 1)\dots(\alpha_0 + r - 1){\beta_0}^r \cdot \alpha_i(\alpha_i + 1)\dots(\alpha_i + m - r - 1){\beta_i}^{m-r} =$$

$$= \sum_{r=0}^{m} C_m^r \alpha_0(\alpha_0 + 1)\dots(\alpha_0 + r - 1)\alpha_i(\alpha_i + 1)\dots(\alpha_i + m - r - 1){\beta_i}^m.$$

С помощью той же процедуры, что и выше, можно напрямую получить смешанные моменты:

$$E({Z_i}^m {Z_j}^n) = E\left(\left(\frac{\beta_i}{\beta_0}\xi_0 + \xi_i\right)^m \left(\frac{\beta_j}{\beta_0}\xi_0 + \xi_j\right)^n\right) =$$

$$= E\left(\sum_{r=0}^{m} C_m^r \left(\frac{\beta_i}{\beta_0}\right)^r {\xi_0}^r {\xi_i}^{m-r} \sum_{s=0}^{n} C_n^s \left(\frac{\beta_j}{\beta_0}\right)^s {\xi_0}^s {\xi_j}^{n-s}\right) =$$

$$= \sum_{r=0}^{m}\sum_{s=0}^{n} C_m^r C_n^s \left(\frac{\beta_i}{\beta_0}\right)^r \left(\frac{\beta_j}{\beta_0}\right)^s (E{\xi_0}^{r+s})(E{\xi_i}^{m-r})(E{\xi_j}^{n-s}) =$$

$$= \sum_{r=0}^{m} \sum_{s=0}^{n} C_m^r C_n^s \left(\frac{\beta_i}{\beta_0}\right)^r \left(\frac{\beta_j}{\beta_0}\right)^s {M_0}^{(r+s)} {M_i}^{(m-r)} {M_j}^{(n-s)},$$

где ${M_i}^{(m)}$ для всех $i = 1, \ldots, k$ получены по формуле (4.1).

Если положить параметры масштаба равными единице и рассматривать случай двумерного распределения, то получим модель, рассматриваемую в работе [5] (Обозначения параметров сохраняют обозначения, используемые в работе). Далее рассматриваем частный случай двумерного гамма-распределения, представленный в этой работе:

$$\begin{cases} y_1 = x_0 + x_1, \\ y_2 = x_0 + x_2, \end{cases} \tag{4.2}$$

где x_0, x_1, x_2 — независимые гамма-распределенные случайные величины с параметрами формы $\lambda_i, i = 0, 1, 2$ и масштаба 1.

Теорема 7. *Пусть x_0, x_1, x_2 — независимые гамма-распределенные случайные величины с параметрами формы λ_i, $i = 0, 1, 2$ и масштаба 1. Тогда $y_i = x_0 + x_i$ имеет гамма-распределение с параметрами $\lambda_0 + \lambda_i$ и 1.*

Определим связь одномерного и двумерного гамма-распределения. Ранее были представлены моменты случайных гамма величин. В нашем случае из формулы (4.1) следует, что $E({x_i}^2) = \lambda_i(\lambda_i + 1)$, а при $i \neq j$ $E(x_i x_j) = \lambda_i \lambda_j$, так как случайные величины независимы. Отсюда

$$E(y_1 y_2) = E(x_0 + x_1)(x_0 + x_2) = E({x_0}^2 + x_0(x_1 + x_2) + x_1 x_2) =$$

$$= \lambda_0(\lambda_0 + 1) + \lambda_0(\lambda_1 + \lambda_2) + \lambda_1\lambda_2 = \lambda_0(\lambda_0 + \lambda_1 + \lambda_2 + 1) + \lambda_1\lambda_2,$$

$$\operatorname{cov}(y_1, y_2) = E(y_1 y_2) - Ey_1 Ey_2 =$$

$$= \lambda_0(\lambda_0 + \lambda_1 + \lambda_2 + 1) + \lambda_1\lambda_2 - (\lambda_0 + \lambda_1)(\lambda_0 + \lambda_2) = \lambda_0.$$

Из последнего получим $\rho(y_1, y_2) = \frac{\lambda_0}{\sqrt{(\lambda_0+\lambda_1)(\lambda_0+\lambda_2)}}$.

Оценим параметры двумерного гамма-распределения по методу моментов. Пусть есть две гамма-распределенные случайные величины с параметром масштаба, равным 1, и с параметрами формы $\lambda_0+\lambda_1$ и $\lambda_0+\lambda_2$. Если $\bar{y}_1, \bar{y}_2$ — выборочные средние, а m_{11} — второй выборочный смешанный центральный момент, то оценки по методу моментов представляются в виде: $\hat{\lambda}_0 = m_{11}$, $\hat{\lambda}_1 = \bar{y}_1 - m_{11}$ и $\hat{\lambda}_2 = \bar{y}_2 - m_{11}$ [5].

Примем за $\Lambda_i = \lambda_0 + \lambda_i$, где $i = 1, 2$. И пусть имеются две гамма-распределенные случайные величины Y_1, Y_2 с параметрами формы, равными Λ_1, Λ_2 соответственно, единичными параметрами масштаба и с коэффициентом корреляции равным ρ.

Теорема 8. *Параметры двумерного гамма-распределения могут быть получены следующим образом:*

$$\lambda_0 = \rho\sqrt{\Lambda_1\Lambda_2},\ \lambda_1 = \Lambda_1 - \rho\sqrt{\Lambda_1\Lambda_2},\ \lambda_2 = \Lambda_2 - \rho\sqrt{\Lambda_1\Lambda_2}.\ [5]$$

Параметры из теоремы 8 называются параметрами экстенсивности.

4.2. Вывод плотности двумерного гамма-распределения

Чтобы перейти к методу максимального правдоподобия нужно получить формулу для плотности двумерного гамма-распределения. Рассмотрим совместную плотность трех независимых гамма-распределенных случайных величин (x_0, x_1, x_2 подходящих условиям предыдущего пункта):

$$f(x_0, x_1, x_2) = \frac{1}{\Gamma(\lambda_0)\Gamma(\lambda_1)\Gamma(\lambda_2)} x_0^{\lambda_0-1} x_1^{\lambda_1-1} x_2^{\lambda_2-1} e^{-(x_0+x_1+x_2)}, x_0, x_1, x_2 > 0.$$

Сделаем замену переменных

$$\begin{cases} u = x_0 + x_1 \\ v = x_0 + x_2 \\ t = x_0 \end{cases} \Leftrightarrow \begin{cases} x_1 = u - t \\ x_2 = v - t \\ x_0 = t \end{cases}.$$

Якобиан этого преобразования

$$|J| = \begin{vmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{vmatrix} = 1.$$

Таким образом, подставляя новые переменные, обозначая $T = \min(u, v)$ и интегрируя по переменной t, получаем

$$f_2(u, v) = \frac{1}{\Gamma(\lambda_0)\Gamma(\lambda_1)\Gamma(\lambda_2)} \int_0^T t^{\lambda_0-1}(u-t)^{\lambda_1-1}(v-t)^{\lambda_2-1}e^{-(u+v-t)}dt.$$

Пусть сначала $u < v$, тогда $T = u$. Разложим экспоненту в ряд Тейлора $e^t = \sum\limits_0^\infty \frac{t^n}{n!}$. Тогда

$$f_2(u,v) = \frac{1}{\Gamma(\lambda_0)\Gamma(\lambda_1)\Gamma(\lambda_2)} \int\limits_0^u t^{\lambda_0-1}(u-t)^{\lambda_1-1}(v-t)^{\lambda_2-1} \sum_{n=0}^\infty \frac{t^n}{n!} e^{-(u+v)} dt.$$

Вынесем из-под знака интеграла все независящее от переменной t и по свойствам интеграла поменяем местами знаки суммирования и интегрирования

$$f_2(u,v) = \frac{1}{\Gamma(\lambda_0)\Gamma(\lambda_1)\Gamma(\lambda_2)} e^{-(u+v)} \sum_{n=0}^\infty \frac{1}{n!} \int\limits_0^u t^{n+\lambda_0-1}(u-t)^{\lambda_1-1}(v-t)^{\lambda_2-1} dt =$$

$$= C \sum_{n=0}^\infty \frac{1}{n!} \int\limits_0^u t^{n+\lambda_0-1}(u-t)^{\lambda_1-1}(v-t)^{\lambda_2-1} dt,\ C = \frac{e^{-(u+v)}}{\Gamma(\lambda_0)\Gamma(\lambda_1)\Gamma(\lambda_2)}.$$

Сделаем замену $t = su$

$$C \sum_{n=0}^\infty \frac{1}{n!} \int\limits_0^1 (su)^{n+\lambda_0-1}(u-su)^{\lambda_1-1}(v-su)^{\lambda_2-1} u ds =$$

$$= C \sum_{n=0}^\infty \frac{u^{n+\lambda_0+\lambda_1-1} v^{\lambda_2-1}}{n!} \int\limits_0^1 s^{n+\lambda_0-1}(1-s)^{\lambda_1-1} \left(1 - s\frac{u}{v}\right)^{\lambda_2-1} ds.$$

Заметим, что при $v < u$ рассуждения будут проводиться аналогично: последняя замена $t = sv$ и изменятся соответствующие показатели степеней. Таким образом, приходим к гипергеометрической функции и получаем выражение для плотности двумерного гамма-распределения:

$$f_2(u,v) = \begin{cases} C \sum\limits_{n=0}^\infty \frac{u^{n+\lambda_0+\lambda_1-1} v^{\lambda_2-1}}{n!} \cdot \frac{\Gamma(n+\lambda_0)\Gamma(\lambda_1)}{\Gamma(n+\lambda_0+\lambda_1)} \cdot {}_2F_1(1-\lambda_2, n+\lambda_0, n+\lambda_0+\lambda_1, \frac{u}{v}), u < v, \\ C \sum\limits_{n=0}^\infty \frac{u^{\lambda_1-1} v^{n+\lambda_0+\lambda_2-1}}{n!} \cdot \frac{\Gamma(n+\lambda_0)\Gamma(\lambda_2)}{\Gamma(n+\lambda_0+\lambda_2)} \cdot {}_2F_1(1-\lambda_1, n+\lambda_0, n+\lambda_0+\lambda_2, \frac{v}{u}), v < u. \end{cases}$$

Определение 17. *Интегральное представление для гипергеометрической функции Гаусса при $Re(c) > Re(b) > 0$:*

$${}_2F_1(a,b,c,z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int\limits_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a} dt,$$

где $\Gamma(x)$ — гамма-функция. Гипергеометрическая функция определяется внутри круга $|z| < 1$ как сумма гипергеометрического ряда, а при $|z| > 1$ — как её аналитическое продолжение. Ряд, через который определяется гипергеометрическая функция, расходится, если $a + b - c \geq 1$.

Условие сходимости $a + b - c < 1$ гипергеометрического ряда для функции ${}_2F_1(a,b,c,z)$ в обоих случаях приобретает вид $\lambda_1 + \lambda_2 > 0$.

4.3. Исследование изменения состояния пациентов, проходящих лечение от наркомании

Имеются данные 306 пациентов, и измерения показателей здоровья в нескольких временных точках. Будем рассматривать изменение в зависимости от группы, в которой проходит лечение пациент, следующих показателей состояния:

- Аспартатаминотрансфераза (далее АСТ) - клеточный фермент содержится в клетках сердца, печени и почек. Аспартатаминотрансфераза синтезируется внутриклеточно, и в норме лишь небольшая часть этого фермента попадает в кровь. При повреждении миокарда, печени в результате цитолиза (разрушения клеток) этот фермент попадает в кровь, что выявляется лабораторными методами. При циррозе печени с цитолитическим синдромом уровень АСТ чаще повышен, но в поздних стадиях цирроза уровни трансаминаз редко бывают повышены.

- Аланинаминотрансфераза (далее АЛТ) относится к классу ферментов-трансфераз, которые участвуют в переносе определенных химических молекул, катализируя тем самым биохимические процессы. На основании высокого уровня АЛТ выставляется лабораторный диагноз цитолитического синдрома, который характеризует поражение печени с разрушением целостности ее мембраны и попаданием в кровь внутриклеточных ферментов. При этом АЛТ является более специфичным для диагностики печеночных патологий, чем АСТ. Для мониторинга состояния пациента с патологией гепатобилиарного тракта более целесообразно определять уровень АЛТ в динамике лечения.

  Повышение АСТ, превышающее повышение АЛТ, характерно для повреждения сердечной мышцы; если же показатель АЛТ выше, чем АСТ, то это, как правило, свидетельствует о разрушении клеток печени.

Рассматриваются две группы пациентов. Пациенты каждой группы получали соответственно: плацебо и налтрексон. Плацебо — вещество без явных лечебных свойств, используемое для имитации лекарственного средства; налтрексон — антагонист опиоидных рецепторов. Обозначения: χ^2 — значение статистики критерия Пирсона, χ^2_b — значение статистики, вычисленное по выборкам bootstrap при $m = 1000$, $\alpha = 0.05$. Алгоритм описан в разделе 1.4.

4.3.1. Анализ изменения содержания фермента АЛТ в группах, принимавших плацебо и налтрексон

Для фермента АЛТ наблюдается согласие с гамма-распределением только в средней точке обследования для каждой группы пациентов. А именно для точки 7: $\chi_1^2 = 1.605 < \chi_{b1}^2 = 3.278$, $\chi_2^2 = 3.678 < \chi_{b2}^2 = 8.59$. После проведения взвешивания для рассматриваемого фермента согласие с гамма-распределением наблюдается в средней и конечной точках обследования для каждой группы пациентов. А именно для точки 7: $\chi_1^2 = 1.14 < \chi_{b1}^2 = 8.373$, $\chi_2^2 = 0.059 < \chi_{b2}^2 = 8.525$. Для точки 13: $\chi_1^2 = 2.827 < \chi_{b1}^2 = 3.727$, $\chi_2^2 = 3.837 < \chi_{b2}^2 = 7.623$.

Перейдем к модели, описанной в разделах 4.1 и 4.2. Рассмотрим параметры интенсивности, которые получаются по формуле $\mathrm{cov}(y_i, \ln(y_i))$. Исходя из модели и свойств гамма-распределения, обозначенных в определении 1, интенсивность равна параметру масштаба распределения. Отсюда могут быть получены доверительные интервалы для параметров, они представлены в таблице 4.1.

Таблица 4.1. Сравнение доверительных интервалов параметров интенсивности фермента АЛТ

АЛТ		
Лекарственный препарат	Точка 7	Точка 13
Плацебо	1.105 ± 0.643	1.239 ± 1.213
Налтрексон	1.031 ± 0.292	1.057 ± 0.482
После взвешивания		
Плацебо	1.189 ± 1.078	1.23 ± 1.025
Налтрексон	1.142 ± 0.828	1.102 ± 0.68

После сравнения интервалов из таблицы 4.1 различий параметров интенсивности ни для одной группы обнаружено не было, уровень значимости ($p \leq 0.05$). Перейдем к рассмотрению параметров экстенсивности из теоремы 8. В таблице 4.2 приведено сравнение параметров, полученных по методам максимального правдоподобия и моментов.

Таблица 4.2. Сравнение оценок параметров распределения, полученных методами моментов и максимума правдоподобия

АЛТ						
Параметр	λ_0		λ_1		λ_2	
Лечебная группа	1	2	1	2	1	2
Оценка ММ	0.227	2.01	1.889	1.309	1.319	3.266
ОМП	1.44	1.29	1.72	1.91	0.988	2.018
После взвешивания						
Оценка ММ	0.469	2.393	0.937	0.052	0.687	1.509
ОМП	1.479	0.961	2.12	1.19	0.687	1.312

Для получения оценок параметров распределения методом максимума правдоподобия был использован пакет bbmle и соответствующая функция mle2. Также эта функция дает возможность построения доверительных интервалов для полученных оценок. В разделе 4.2 было приведено построение плотности двумерного гамма-распределения, основанное на модели из [5]. Зная вид плотности построим функцию правдоподобия для использования функции mle2.

Таблица 4.3. Сравнение доверительных интервалов параметров экстенсивности фермента АЛТ

АЛТ			
Лекарственный препарат	λ_0	λ_1	λ_2
Плацебо	(0.00; 3.086)	(0.308; 4.517)	(0.182; 3.633)
Налтрексон	(0.00; 3.149)	(0.453; 4.682)	(0.557; 4.628)
После взвешивания			
Плацебо	(0.00; 2.53)	(1.11; 4.589)	(0.00; 2.314)
Налтрексон	(0.00; 2.343)	(0.227; 3.256)	(0.311; 3.309)

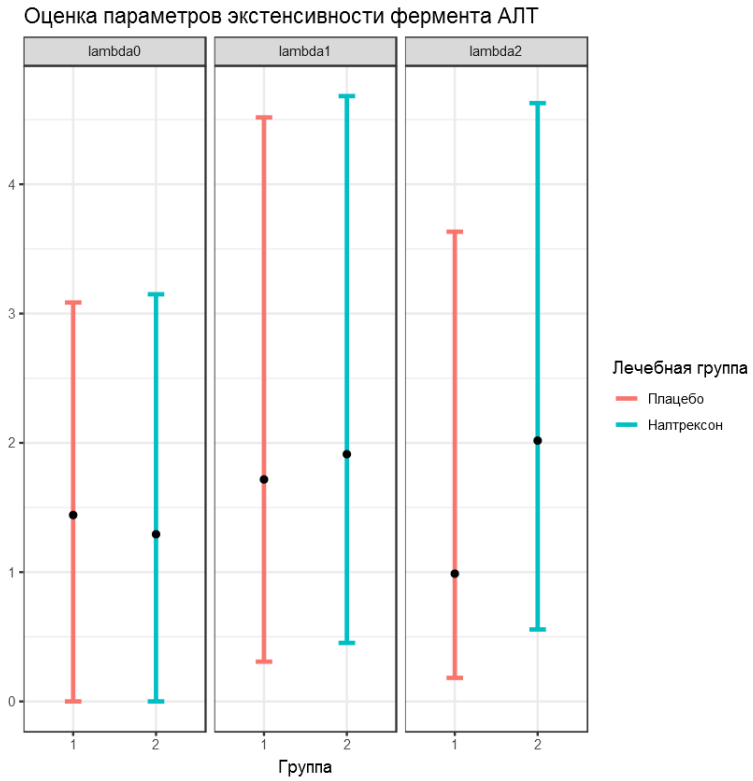

Рис. 4.1. Сравнение доверительных интервалов параметров экстенсивности фермента АЛТ

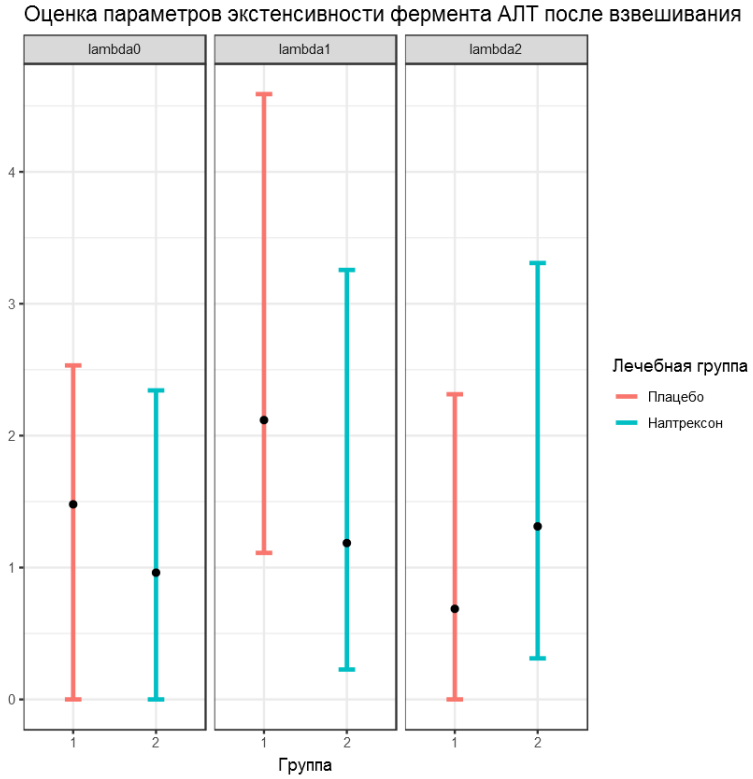

Рис. 4.2. Сравнение доверительных интервалов параметров экстенсивности фермента АЛТ после взвешивания

На уровне значимости 0.05 различие между параметрами экстенсивности не удается найти ни в одной точке, смотри таблицу 4.3 и графическое представление доверительных интервалов на рисунках 4.1–4.2.

4.3.2. Анализ изменения содержания фермента АСТ в группах, принимавших плацебо и налтрексон

Перейдем к рассмотрению фермента АСТ. Для этого фермента наблюдается согласие с гамма-распределением в начальной и средней точках обследования для каждой группы пациентов. А именно для точки 0: $\chi_1^2 = 1.892 < \chi_{b1}^2 = 8.181$, $\chi_2^2 = 2.524 < \chi_{b2}^2 = 10.426$. Для точки 7: $\chi_1^2 = 0.582 < \chi_{b1}^2 = 3.607$, $\chi_2^2 = 6.39 < \chi_{b2}^2 = 8.376$. После проведения взвешивания согласие с гамма-распределением также наблюдается в тех же точках. А именно для точки 0:
$\chi_1^2 = 9.751 < \chi_{b1}^2 = 12.033$, $\chi_2^2 = 10.081 < \chi_{b2}^2 = 12.778$. Для точки 7: $\chi_1^2 = 4.756 < \chi_{b1}^2 = 8.32$, $\chi_2^2 = 6.995 < \chi_{b2}^2 = 9.867$. Сравним доверительные интервалы для параметров интенсивности, результаты представлены в таблице 4.4.

Таблица 4.4. Сравнение доверительных интервалов параметров интенсивности фермента АСТ

АСТ		
Лечебный препарат	Точка 0	Точка 7
Плацебо	1.119 ± 0.796	1.096 ± 0.676
Налтрексон	0.993 ± 0.324	1.034 ± 0.397
После взвешивания		
Плацебо	1.194 ± 1.048	1.143 ± 0.891
Налтрексон	1.03 ± 0.447	1.091 ± 0.649

После сравнения интервалов из таблицы 4.4 различий параметров интенсивности ни для одной группы обнаружено не было, уровень значимости ($p \leq 0.05$).

Перейдем к сравнению параметров экстенсивности. Сравнение оценок, полученных по методам моментов и максимума правдоподобия, представлено в таблице 4.5.

Таблица 4.5. Сравнение оценок параметров распределения, полученных методами моментов и максимума правдоподобия

АСТ						
Параметр	λ_0		λ_1		λ_2	
Лечебная группа	1	2	1	2	1	2
Оценка ММ	1.688	2.446	1.99	4.287	0.611	2.942
ОМП	4.86	3.23	0.538	2.09	0.31	2.14
После взвешивания						
Оценка ММ	1.551	2.986	0.466	2.043	0.042	0.925
ОМП	2.59	2.25	2.88	1.23	0.611	1.347

На уровне значимости 0.2 найдено значимое различие интервальных оценок параметров распределения λ_2 выборок, к которым не было применено взвешивание. Все доверительные интервалы для параметров экстенсивности представлены в таблице 4.6 и на рисунках 4.3–4.4.

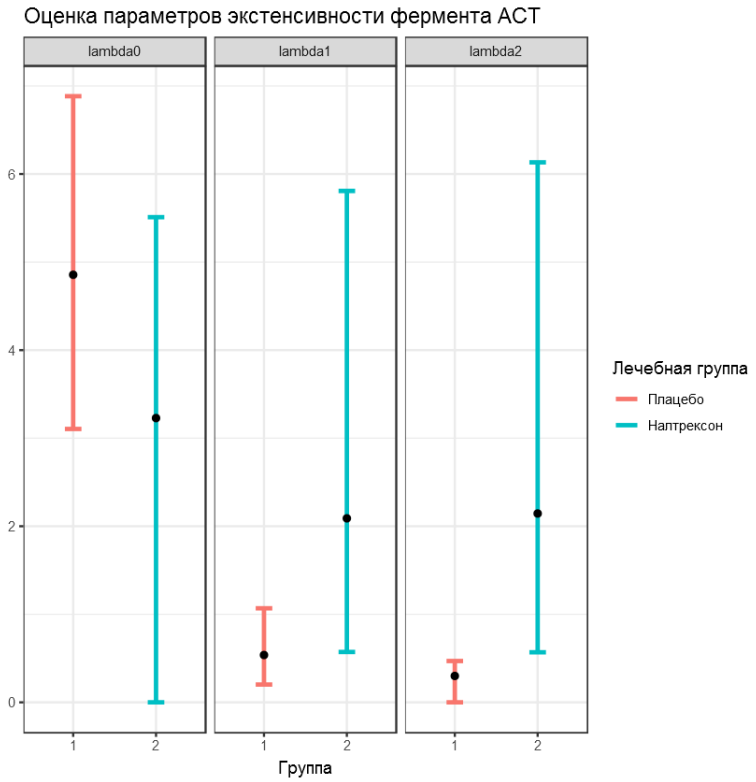

Рис. 4.3. Сравнение доверительных интервалов параметров экстенсивности фермента АСТ

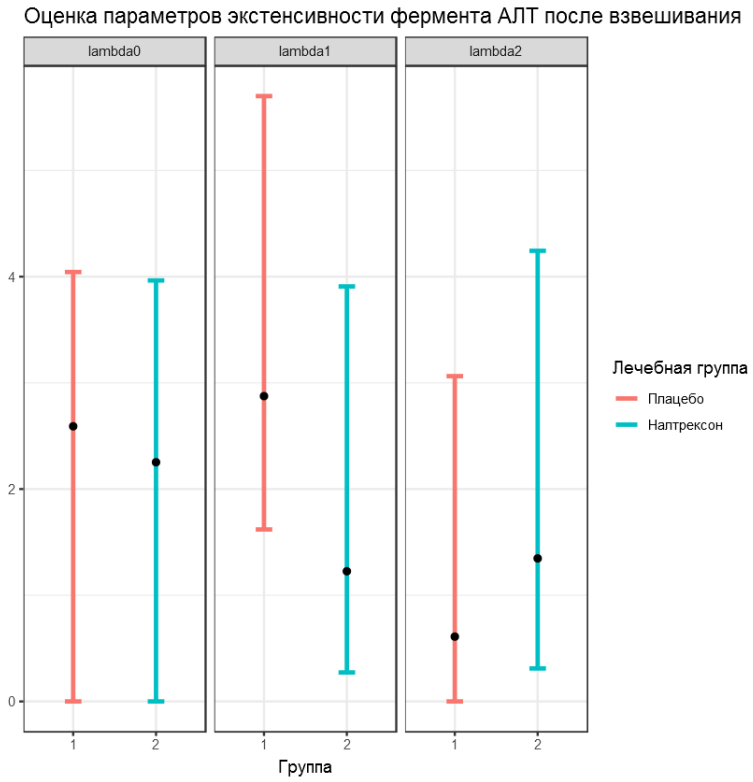

Рис. 4.4. Сравнение доверительных интервалов параметров экстенсивности фермента АСТ после взвешивания

Таблица 4.6. Сравнение доверительных интервалов параметров экстенсивности фермента АСТ

АСТ			
Лекарственный препарат	λ_0	λ_1	λ_2
Плацебо	(3.105; 6.884)	(0.202; 1.068)	(0.00; 0.469)
Налтрексон	(0.00; 5.51)	(0.572; 5.808)	(0.568; 6.133)
После взвешивания			
Плацебо	(0.00; 4.043)	(1.619; 5.7)	(0.00; 3.063)
Налтрексон	(0.00; 3.964)	(0.273; 3.908)	(0.311; 4.244)

4.3.3. Анализ изменения содержания фермента АЛТ в группах, которым был назначен налтрексон и налтрексон-имплант

Рассмотрим другое разделение пациентов на две группы. Пациенты первой группы получали налтрексон, а пациенты второй группы налтрексон-имплант. Налтрексон-имплант вводится подкожно и исключает возможность того, что пациент не примет

необходимое лекарство. Определим снова с помощью алгоритма bootstrap в каких точках лечения для изучаемых ферментов наблюдается согласие с гамма-распределением.

Для фермента АЛТ наблюдается согласие с гамма-распределением в начальной и средней точках обследования для каждой группы пациентов. А именно для точки 0: $\chi_1^2 = 7.717 < \chi_{b1}^2 = 8.818$, $\chi_2^2 = 1.511 < \chi_{b2}^2 = 8.058$. Для точки 7: $\chi_1^2 = 1.488 < \chi_{b1}^2 = 4.447$, $\chi_2^2 = 4.184 < \chi_{b2}^2 = 7.489$. После проведения взвешивания не наблюдается согласие для двух точек и каждой группы с гамма-распределением. Поэтому оценки на взвешенных выборках не будет рассматриваться.

Вернемся к рассмотрению оценок параметров интенсивности (таблица 4.7), экстенсивности (таблица 4.8) и доверительных интервалов для оценок, полученных по методу максимума правдоподобия. Интервальные оценки параметров представлены в таблице 4.7 и на рисунке 4.5.

Таблица 4.7. Сравнение доверительных интервалов параметров интенсивности фермента АЛТ

АЛТ		
Лекарственный препарат	Точка 0	Точка 7
Налтрексон	1.081 ± 0.783	1.018 ± 0.369
Налтрексон-имплант	1.051 ± 0.44	1.219 ± 1.061

Таблица 4.8. Сравнение оценок параметров распределения, полученных методами моментов и максимума правдоподобия

АЛТ						
Параметр	λ_0		λ_1		λ_2	
Лечебная группа	1	2	1	2	1	2
Оценка ММ	1.529	0.707	0.925	3.123	0.777	1.389
ОМП	1.484	0.707	1.09	1.76	0.777	2.13

Таблица 4.9. Сравнение доверительных интервалов параметров экстенсивности фермента АЛТ

АЛТ			
Лекарственный препарат	λ_0	λ_1	λ_2
Налтрексон	(0.446; 2.539)	(0.447; 2.026)	(0.00; 1.078)
Налтрексон-имплант	(0.00; 1.644)	(0.499; 2.947)	(0.641; 3.432)

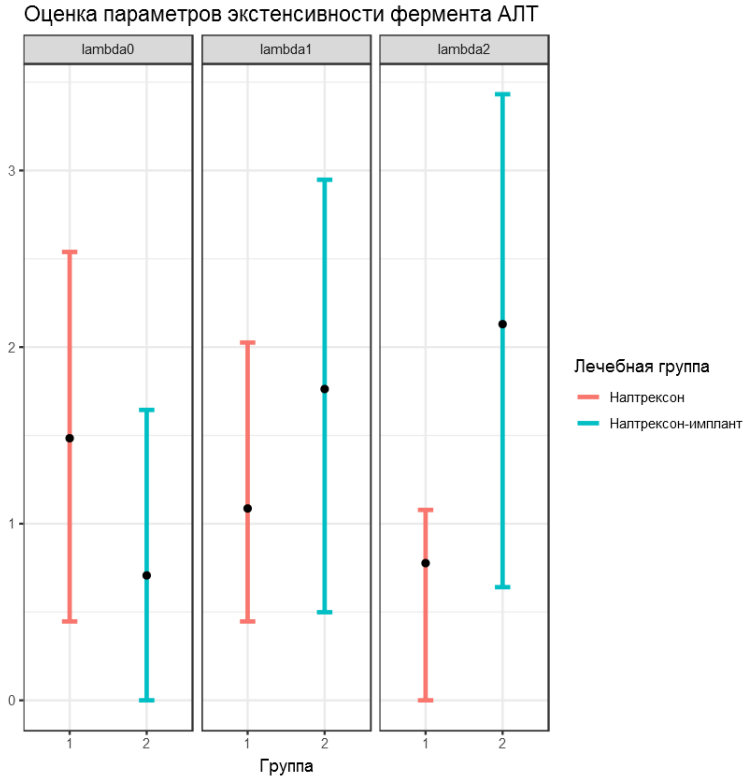

Рис. 4.5. Сравнение доверительных интервалов параметров экстенсивности фермента АЛТ

Для получения оценок методом максимума правдоподобия объем выборок был искусственно увеличен в три раза путем дублирования каждого индивида. Даже при этом объеме на уровне значимости 0.2 значимого различия обнаружить не удалось.

4.3.4. Анализ изменения содержания фермента АСТ в группах, которым был назначен налтрексон и налтрексон-имплант

Для фермента АСТ наблюдается согласие с гамма-распределением в начальной и средней точках обследования для каждой группы пациентов. А именно для точки 0: $\chi^2_1 = 3.583 < \chi^2_{b1} = 8.703$, $\chi^2_2 = 1.035 < \chi^2_{b2} = 9.411$. Для точки 7: $\chi^2_1 = 0.721 < \chi^2_{b1} = 4.989$, $\chi^2_2 = 2.589 < \chi^2_{b2} = 7.508$. После проведения взвешивания не наблюдается

согласие для двух точек и каждой группы с гамма-распределением. Поэтому оценки на взвешенных выборках не будет рассматриваться.

Снова рассмотрим оценки параметров интенсивности (таблица 4.10), экстенсивности (таблица 4.11) и доверительных интервалов для оценок, полученных по методу максимума правдоподобия. Интервальные оценки параметров представлены в таблице 4.10 и на рисунке 4.6.

Таблица 4.10. Сравнение доверительных интервалов параметров интенсивности фермента АСТ

АСТ		
Лекарственный препарат	Точка 0	Точка 7
Налтрексон	1.072 ± 0.792	1.031 ± 0.545
Налтрексон-имплант	1.011 ± 0.263	1.059 ± 0.374

Таблица 4.11. Сравнение оценок параметров распределения, полученных методами моментов и максимума правдоподобия

АСТ						
Параметр	λ_0		λ_1		λ_2	
Лечебная группа	1	2	1	2	1	2
Оценка ММ	2.126	2.608	2.373	2.774	3.185	1.847
ОМП	4.286	2.767	2.08	1.675	0.214	1.7

Таблица 4.12. Сравнение доверительных интервалов параметров экстенсивности фермента АСТ

АСТ			
Лекарственный препарат	λ_0	λ_1	λ_2
Налтрексон	(2.407; 6.186)	(1.087; 3.817)	(0.00; 1.768)
Налтрексон-имплант	(0.00; 4.766)	(0.428; 4.765)	(0.439; 5.021)

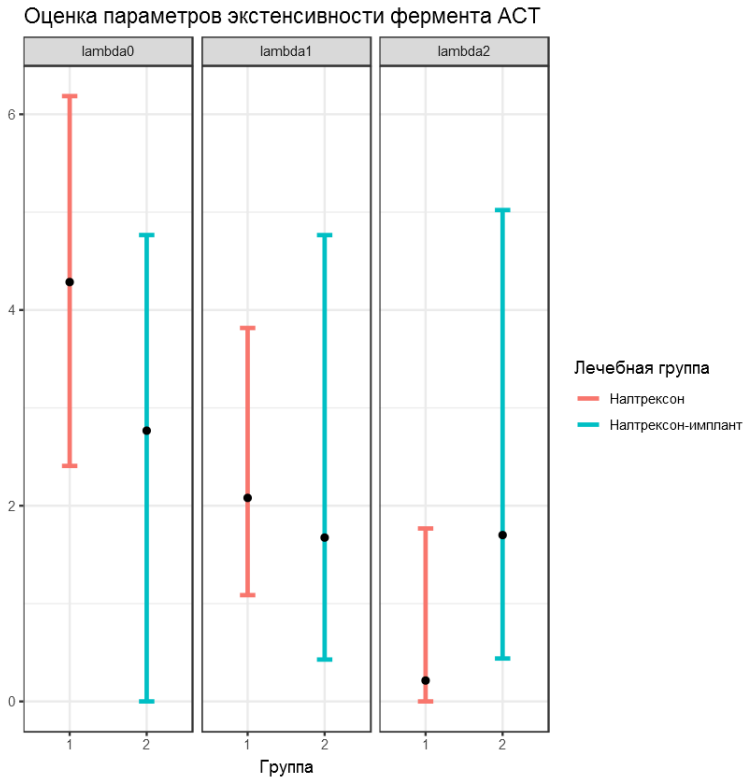

Рис. 4.6. Сравнение доверительных интервалов параметров экстенсивности фермента АСТ

Как и в разделе 4.3.3 для получения оценок методом максимума правдоподобия объем выборок был увеличен в три раза. Снова на уровне значимости 0.2 значимого различия оценок параметров экстенивности не было обнаружено.

Заключение

В работе были рассмотрены три модели: одномерное гамма-распределение, степенное гамма-распределение и двумерное гамма-распределение. Приведены способы получения оценок параметров распределения. В частности для двумерного распределения построена плотность распределения, которая затем используется для получение оценок максимального правдоподобия.

Практическая часть работы представляет собой изучение изменения содержания ферментов печени АЛТ и АСТ пациентов, проходивших лечении от наркотической зависимости. Для корректного сравнения разных групп пациентов (принимавших разный препарат) необходимо исключить влияние сопутствующих факторов. С помощью симптомно-синдромального анализа можно убедиться, что группы однородны по представленным признакам. А с помощью метода псевдорандомизации можно добиться сбалансированности групп. Эти методы были применены до основного анализа динамики содержания ферментов.

Основной анализ включал: проверку согласия с гамма-распределением с помощью критерия Пирсона (p.value $>$ 0.5) и boostrap выборок, получение оценок параметров распределения по методам моментов и максимума правдоподобия, построение доверительных интервалов для оценок параметров двумерного распределения. После сравнения интервальных оценок получено различие на уровне значимости 0.2 для параметра экстенсивности фермента АСТ для пациентов, получавших плацебо и препарат налтрексон.

Список литературы

1. Севастьянов Б. А. Курс теории вероятностей и математической статистики. — Издательство «Наука», 1982. — 256 с.
2. Бородин А. Н. Элементы теории вероятностей и математической статистики. — Издательство «Лань», 1999. — 224 с.
3. González E. G., Villaseñor-Alva J. A. Panteleeva .O V., Huerta H. V. On testing the log-gamma distribution hypothesis by bootstrap // Computational Statistics. — 2013. — Vol. 28, no. 6. — P. 2761–2776.
4. Кульбак С. Теория информации и статистика. — Наука. Гл. ред. физ.-мат. лит., 1967.
5. Алексеева Н. П. Анализ медико-биологических систем. Реципрокность, эргодичность, синонимия. — Издательство С.-Петербурского университета, 2012. — 184 с.
6. Alexeyeva N. P. Al-Juboori F. S. Skurat E. P. Symptom analysis of multidimensional categorical data with applications // Periodicals of Engineering and Natural Sciences (PEN). — 2020. — Vol. 8, no. 3. — P. 1517–1524.
7. Гржибовский А. М. и др. Псевдорандомизация (propensity score matching) как современный статистический метод устранения систематических различий сравниваемых групп при анализе количественных исходов в обсервационных исследованиях // Экология человека. — 2016. — № 7. — С. 51–60.
8. Cochran W. G., Chambers S. P. The planning of observational studies of human populations // Journal of the Royal Statistical Society. Series A (General). — 1965. — Vol. 128, no. 2. — P. 234–266.
9. Rosenbaum P. R., Rubin D. B. The central role of the propensity score in observational studies for causal effects // Biometrika. — 1983. — Vol. 70, no. 1. — P. 41–55.
10. Austin P. C. An introduction to propensity score methods for reducing the effects of confounding in observational studies // Multivariate behavioral research. — 2011. — Vol. 46, no. 3. — P. 399–424.
11. Alexeyeva N. P. Alexeyev A. O. Synonymy of power gammadistributions in the statistical model of muscles // Simulation. Proceedings of the 5th St. Petersburg Workshop on Simulation. Edited by Ermakov S. M., Melas V. B. and Pepelyshev A. N. — 2005. — P. 39–43.
12. Thomas M. Joy A. T. Elements of Information Theory. — Hoboken, N.J. : Wiley-

Interscience, 2006. — 774 p.

13. Mathai A. M., Moschopoulos P. G. On a multivariate gamma // Journal of Multivariate Analysis. — 1991. — Vol. 39, no. 1. — P. 135–153.

Приложение А

Сравнение гистограмм и плотностей

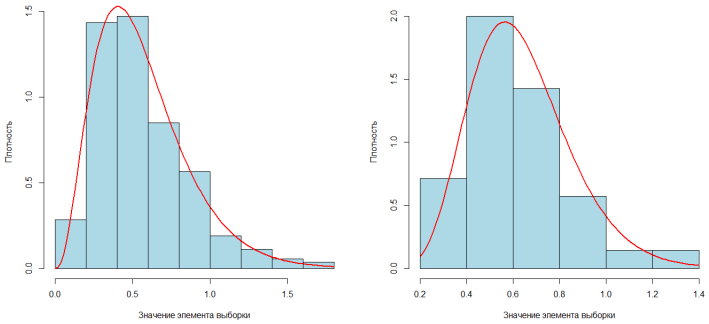

Рис. А.1. Сравнение гистограммы фермента АСТ и плотности гамма-распределения в группах, соответствующих суперсимптому Галлюц·Вич+Галлюц·В

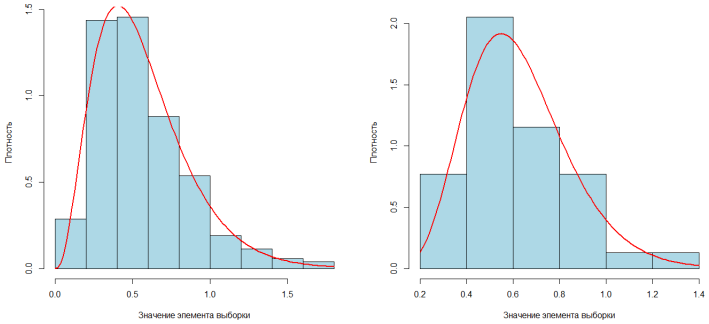

Рис. А.2. Сравнение гистограммы фермента АСТ и плотности гамма-распределения в группах, соответствующих суперсимптому Галлюц·Год+Галлюц·Вич

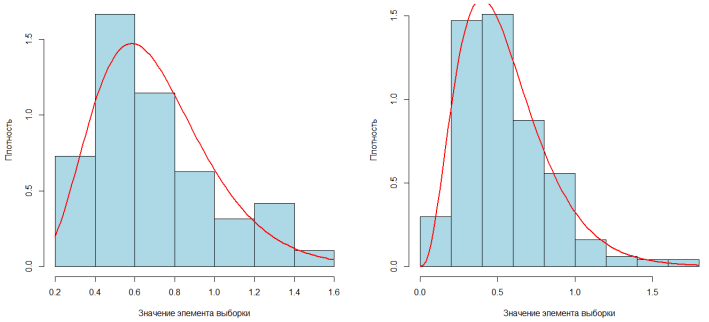

Рис. А.3. Сравнение гистограммы фермента АСТ и плотности гамма-распределения в группах, соответствующих суперсимптому Лечение+Амф+Бенз·Амф

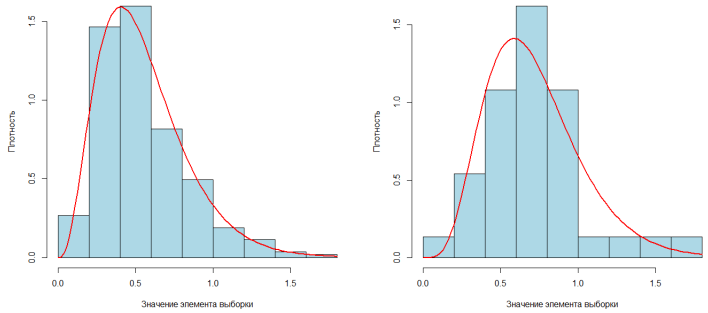

Рис. А.4. Сравнение гистограммы фермента АСТ и плотности гамма-распределения в группах, соответствующих суперсимптому Возд·Кон+Кон

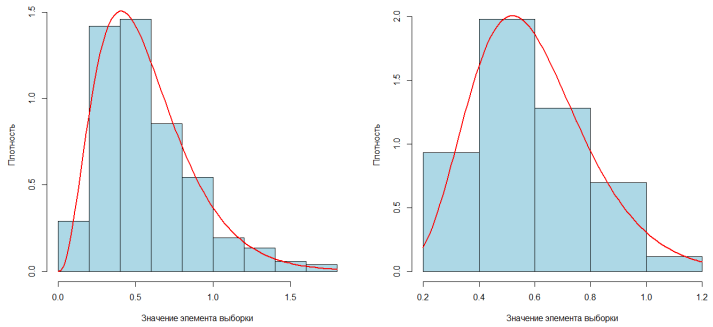

Рис. А.5. Сравнение гистограммы фермента АСТ и плотности гамма-распределения в группах, соответствующих суперсимптому Кокаин+Галлюц·Вич

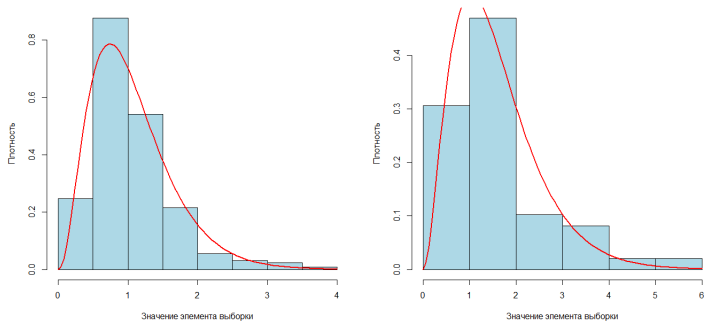

Рис. А.6. Сравнение гистограммы фермента АЛТ и плотности гамма-распределения в группах, соответствующих суперсимптому Сед+Возд·Амф+Возд·Сед·Амф

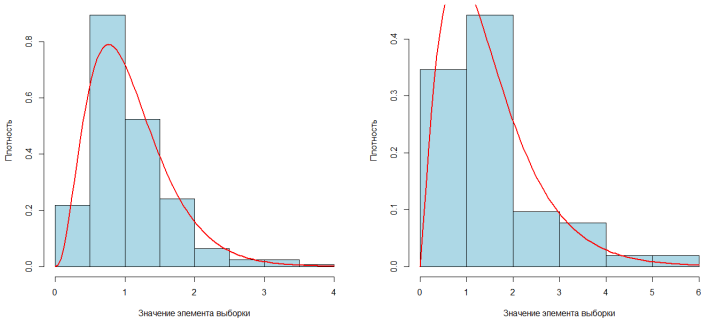

Рис. А.7. Сравнение гистограммы фермента АЛТ и плотности гамма-распределения в группах, соответствующих суперсимптому Амф+Возд·Кон+Кон·Амф

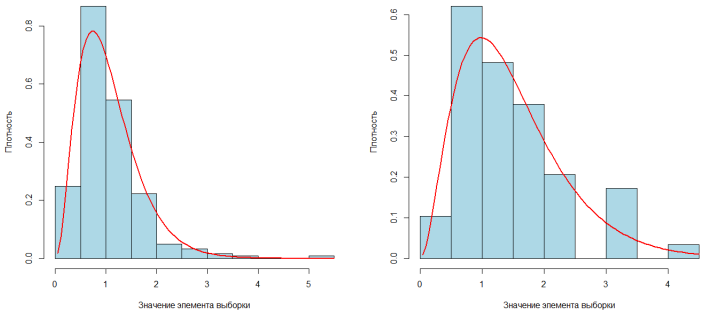

Рис. А.8. Сравнение гистограммы фермента АЛТ и плотности гамма-распределения в группах, соответствующих суперсимптому В+Возд·Амф+Амф·В

Приложение Б

Графическая диагностика баланса

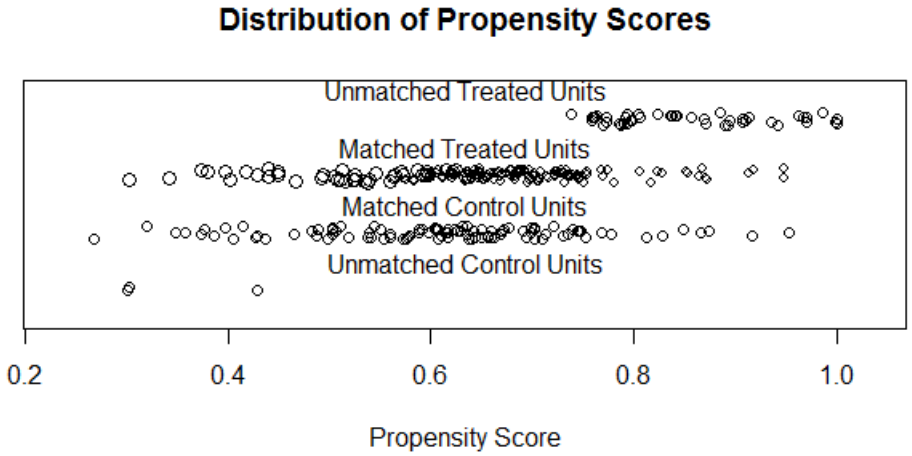

Рис. Б.1. Плотность меры склонности. Сопоставление по методу ближайшего соседа

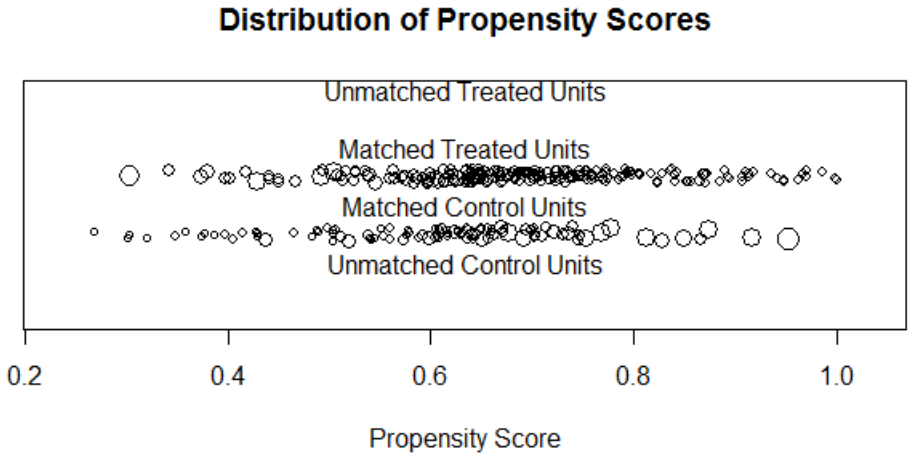

Рис. Б.2. Плотность меры склонности. Полное сопоставление

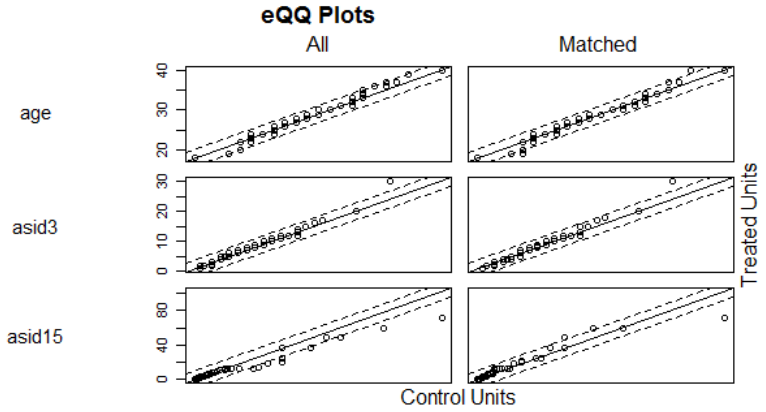

Рис. Б.3. График квантиль–квантиль сопоставления ближайшего соседа (признаки: возраст, принятие алкоголя, длительность последнего периода отказа от наркотиков)

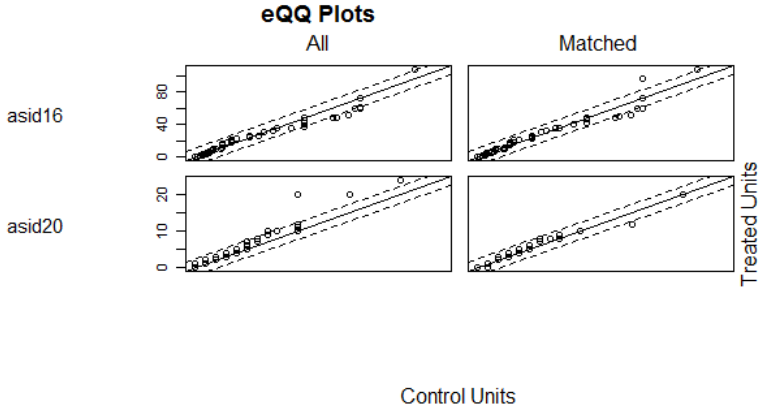

Рис. Б.4. График квантиль–квантиль сопоставления ближайшего соседа (признаки: сколько месяцев назад закончился последний период отказа от наркотиков, сколько раз происходил отказ от наркотиков)

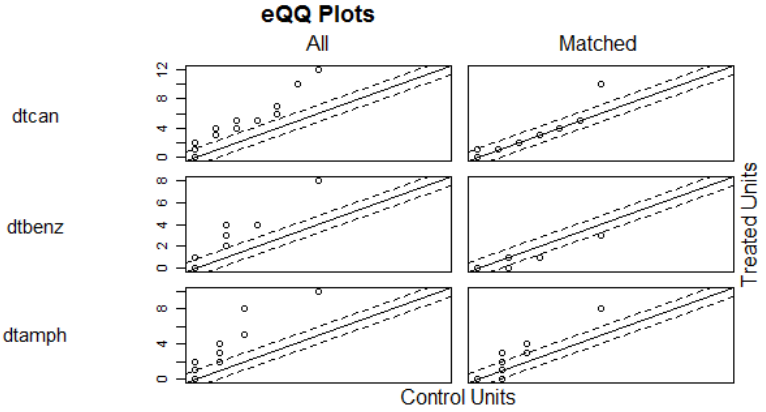

Рис. Б.5. График квантиль–квантиль сопоставления ближайшего соседа (признаки: содержание конопли в моче, содержание бензодиазепина в моче, содержание амфетамина в моче)

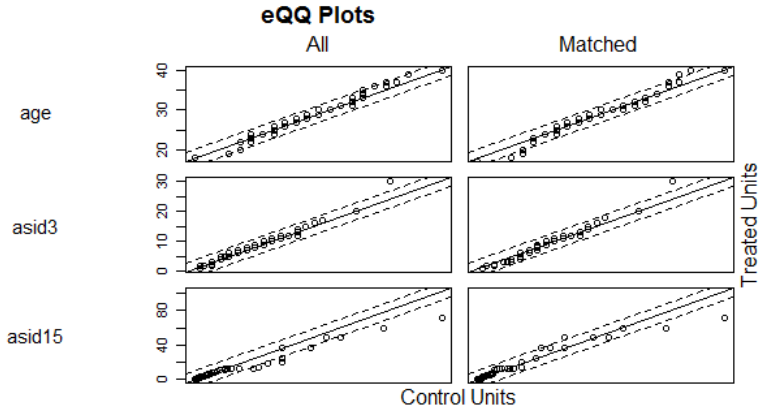

Рис. Б.6. График квантиль–квантиль полного сопоставления (признаки: возраст, принятие алкоголя, длительность последнего периода отказа от наркотиков)

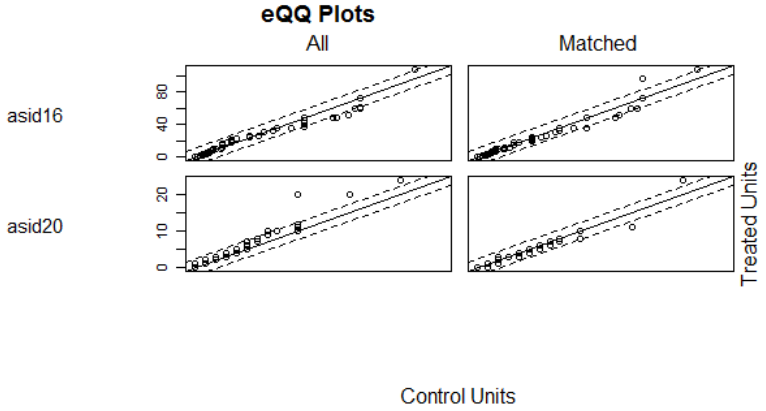

Рис. Б.7. График квантиль–квантиль полного сопоставления (признаки: сколько месяцев назад закончился последний период отказа от наркотиков, сколько раз происходил отказ от наркотиков)

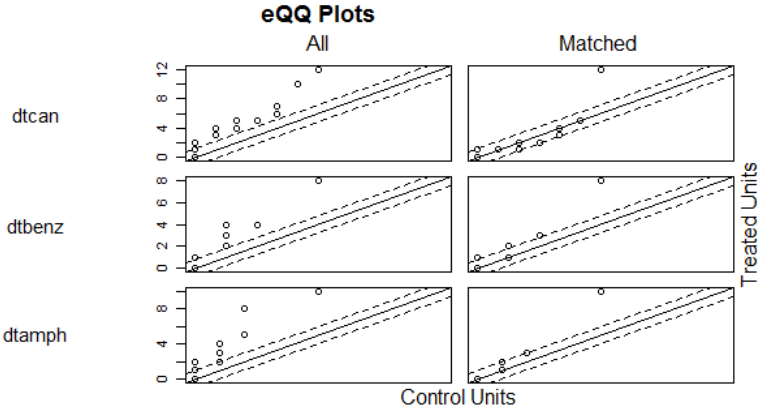

Рис. Б.8. График квантиль–квантиль полного сопоставления (признаки: содержание конопли в моче, содержание бензодиазепина в моче, содержание амфетамина в моче)

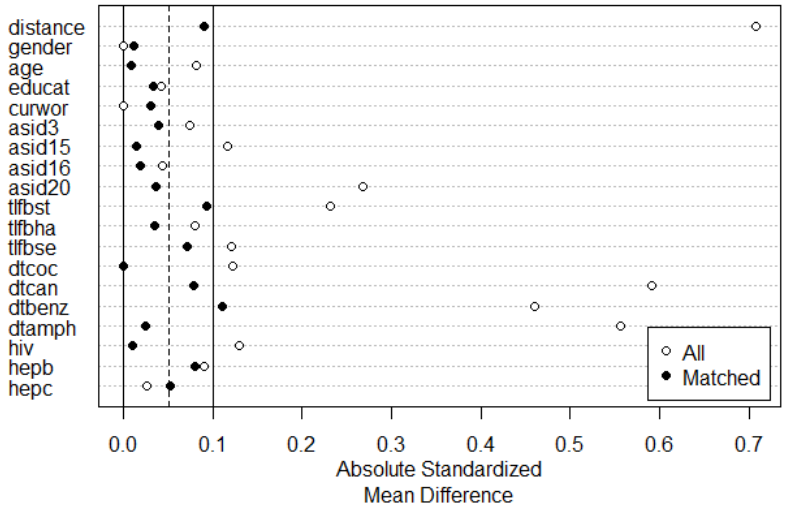

Рис. Б.9. График стандартизированных разностей средних. Сопоставление ближайшего соседа

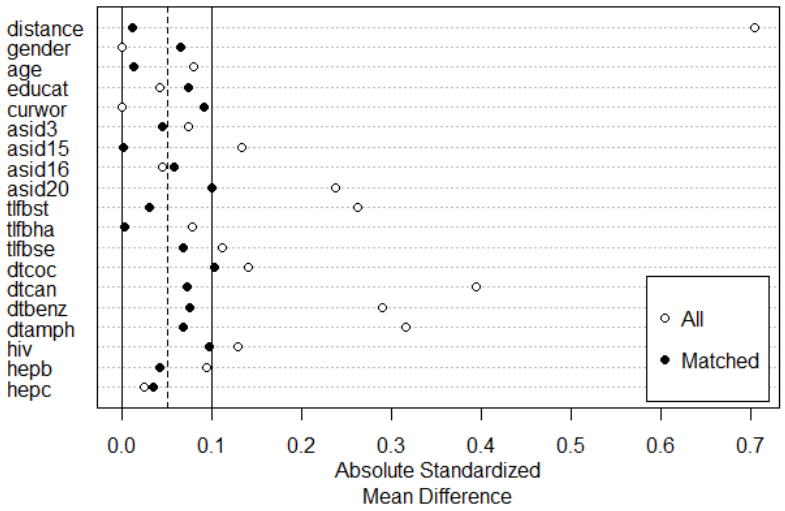

Рис. Б.10. График стандартизированных разностей средних. Полное сопоставление

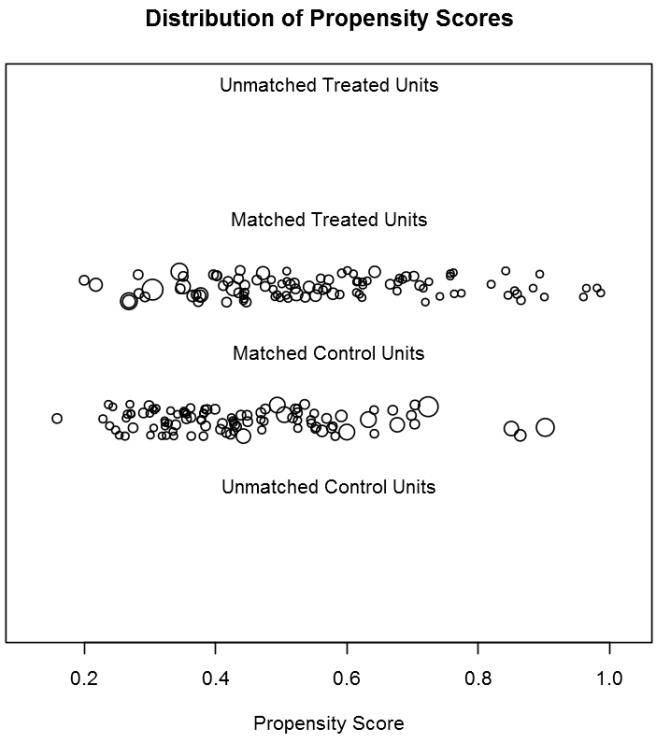

Рис. Б.11. Плотность меры склонности. Полное сопоставление. Для пациентов получавших лечение препаратом.

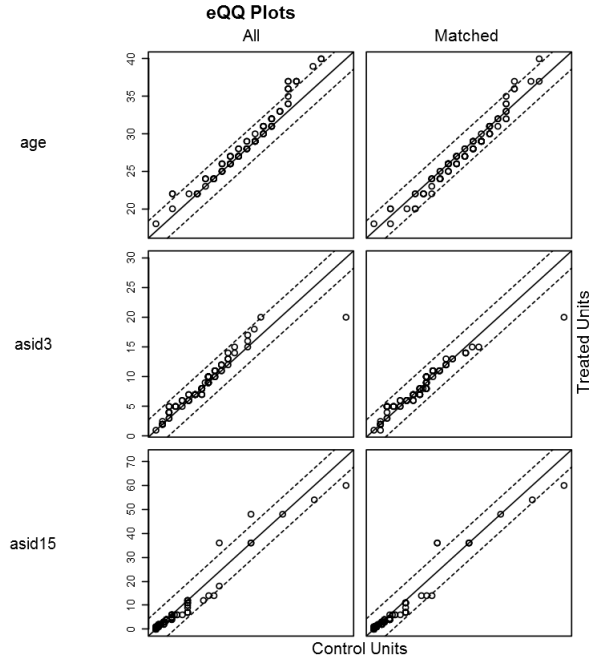

Рис. Б.12. График квантиль–квантиль полного сопоставления для пациентов получавших лечение препаратом (признаки: возраст, принятие алкоголя, длительность последнего периода отказа от наркотиков)

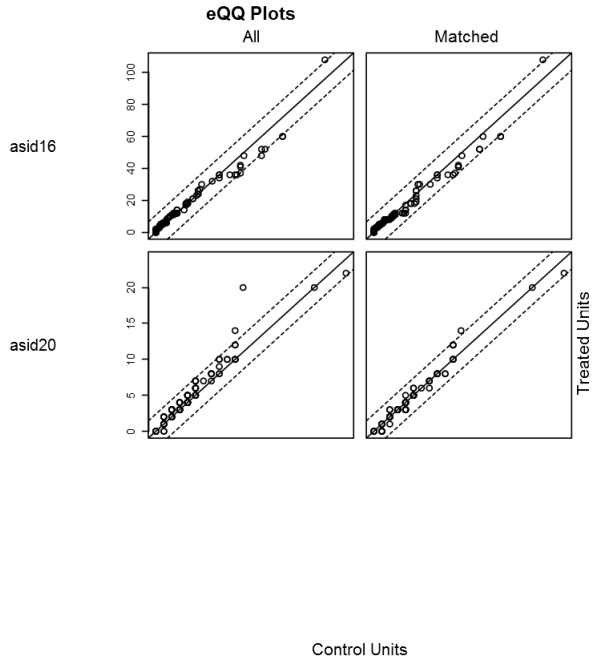

Рис. Б.13. График квантиль–квантиль полного сопоставления для пациентов получавших лечение препаратом (признаки: сколько месяцев назад закончился последний период отказа от наркотиков, сколько раз происходил отказ от наркотиков)

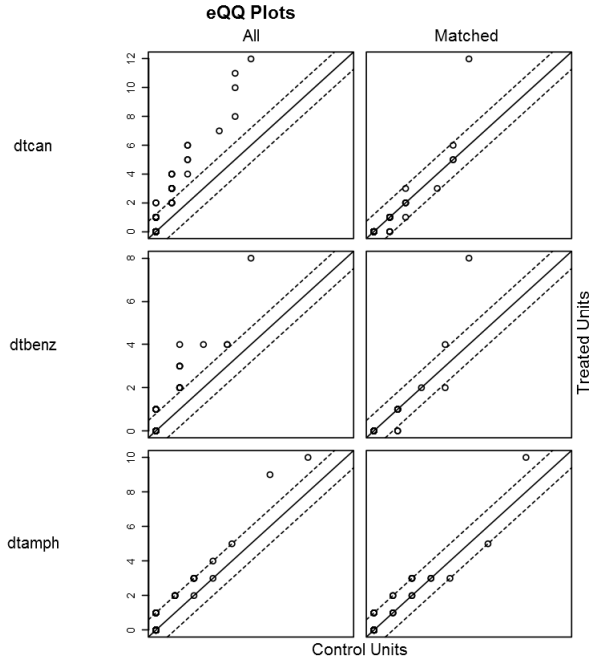

Рис. Б.14. График квантиль–квантиль полного сопоставления для пациентов получавших лечение препаратом (признаки: содержание конопли в моче, содержание бензодиазепина в моче, содержание амфетамина в моче)

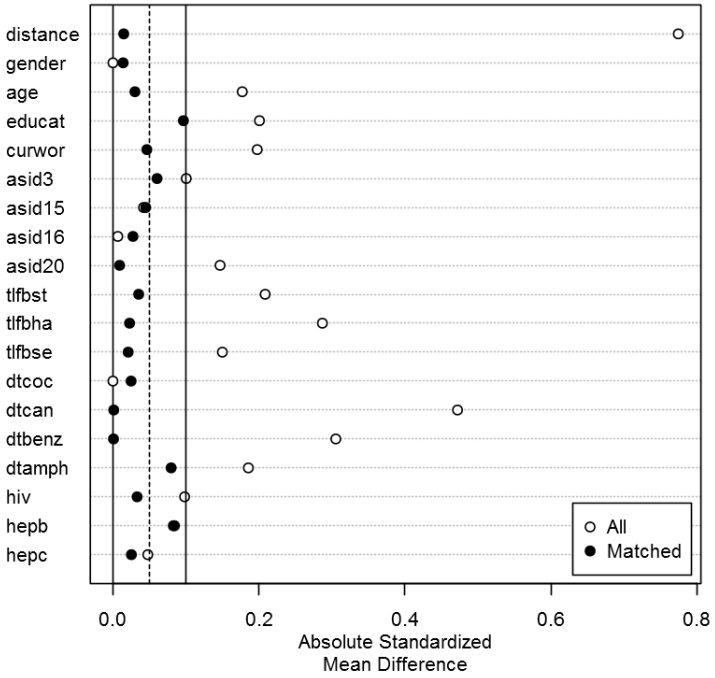

Рис. Б.15. График стандартизированных разностей средних. Полное сопоставление. Для пациентов получавших лечение препаратом.