Модель двумерного гамма-распределения с медико-биологическими приложениями

Мандрикова Анастасия Андреевна, гр. 622

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Научный руководитель: к.ф.-м.н., доцент Н. П. Алексеева
Рецензент: биостатистик Е. С. Комарова

Санкт-Петербург, 2022г.

Постановка задачи

Изучение **безопасности и эффективности** лечения с применением моделей гамма распределения на примере реальных данных из медицинской практики.

- Проверка необходимости устранения **влияния сопутствующих факторов** на зависимую переменную
- Получение оценок и построение доверительных интервалов для параметров **одномерного и двумерного гамма распределения**
- Проверка гипотез о **равенстве параметров** распределения двух и более выборок

Структура данных

Повторные данные 306 пациентов, которые проходили лечение от наркомании (НМИЦ ПН им В. М. Бехтерева и ПСПбГМУ).

Рассматриваются:

- Группа лечения, в которой пациент получает:
  - Плацебо
  - Налтрексон
  - Налтрексон-имплант
- Категориальные признаки, отражающие: социальный статус участника, предыдущий опыт лечения и характер зависимости на момент начала программы.
- Показатели: Ферменты АЛТ и АСТ, уровень депрессии по шкале BDI.

Псевдорандомизация как метод устранения систематических различий сравниваемых групп

Пусть Z — **индикаторная переменная**, обозначающая принадлежность индивида к группе воздействия; X — **наблюдаемые ковариаты**; Y — **наблюдаемая зависимая переменная**.

Определение: мера склонности (Rosenbaum, P. R., Rubin, D. B., 1983)

$e(x) \overset{\text{def}}{=} \mathbb{P}\{Z = 1 \mid X = x\}$, здесь предполагаем $\mathbb{P}(z_1, z_2, \ldots, z_n | x_1, x_2, \ldots, x_n) = \prod\limits_{i=1}^{N} e(x_i)^{z_i}(1 - e(x_i))^{1-z_i}$.

Оценки мер склонности получаются из **модели логистической регрессии** $\mathbb{P}\{Z = 1 \mid X\} = \frac{1}{1+e^{-\theta^T X}}$. Обозначим индивидов из группы контроля индексами i, из группы воздействия — j. Находим веса w для индивидов из условия:

$$\frac{1}{m_1 m_2} \sum_{\substack{1 \le i \le m_1, \\ 1 \le j \le m_2}} (w_i e(x_i) - w_j e(x_j)) \underset{w}{\to} \min .$$

Новая выборка wY.

Симптомно-синдромальный подход

Метод позволяет найти комбинацию факторов, расслоение по которой исходной выборки, приводит к однородным подгруппам наиболее отличающимся друг от друга.

Определение

Линейная комбинация $X_\tau = A_\tau X \pmod 2$, где $X = (X_1, \dots, X_m)^\mathrm{T}$ случайный вектор с компонентами 0 или 1, $\tau = (t_1, \dots, t_k) \subseteq (1, 2, \dots, m)$ и вектор строка A_τ :

$$a_j = \begin{cases} 1, & \text{если } j \in \tau, \\ 0, & \text{иначе;} \end{cases}$$

называется **симптомом ранга** k.

Имея $k > 0$ симптомов $X_1, \dots, X_k$, можем определить некоторую совокупность.

Симптомно-синдромальный подход

Определение

Синдромом S_k **порядка** k называется совокупность $2^k - 1$ симптомов $\beta_0 X_0 + \dots + \beta_k X_k \pmod 2$, где коэффициенты $\beta_i \in \mathbb{F}_2$ не равны нулю одновременно.

Заметим, что единичный симптом можно рассматривать как синдром первого порядка $S_1 = S_1(X_\tau)$.

Рекуррентное построение синдромов:

$$S_i = S_i(X_{t_1}, \dots, X_{t_i}) = (S_{i-1},\, X_{t_i},\, S_{i-1} + X_{t_i} \pmod 2),$$

где $X_{t_i} \notin S_{i-1}$.

Построение суперсиндромов

Последовательность обобщается с использованием любой операции над полем $\mathbb{F}_2$:

$$M_i = M_i(X_{t_1}, \ldots, X_{t_i}|*) = (M_{i-1},\, M_{t_i},\, M_{i-1} * X_{t_i}),$$

где $X_{t_i} \notin M_{i-1}$. **Синдром** $S_k(X)$ **совпадает с** $M_k(X|+)$.

Определение

Рассмотрим в качестве базовых элементов синдрома $M_k(\cdot|\odot)$, где $(\odot)$ умножение над полем $\mathbb{F}_2$. Таким образом, строится **суперсиндром** $S_K(M_k(X_{t_1}, \ldots, X_{t_k}|\odot))$ **порядка** $K = 2^k - 1$.

Из элементов суперсиндрома отбираются элементы по которым осуществляется расслоение выборки на наиболее отличающиеся по параметрам модели подвыборки.

Модель одномерного гамма распределение

Функция плотности гамма распределения

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

β – параметр масштаба, α – параметр формы.

Пусть $\{x_i\}_{i=1}^{n}$ реализация с.в. $\xi\sim\gamma(x,\alpha,\beta)$, обозначим $\psi(\alpha)$ дигамма функцию $\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$.

Оценки максимума правдоподобия (МП)

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

Модель двумерного гамма распределения

Mathai, 1991

Пусть $\xi_i \sim \gamma(\alpha_i, \beta_i), i = 0, 1, 2$ взаимно независимые.
Определим случайные величины η_i следующим образом:

$$\begin{cases} \eta_1 = \frac{\beta_1}{\beta_0}\xi_0 + \xi_1, \\ \eta_2 = \frac{\beta_2}{\beta_0}\xi_0 + \xi_2. \end{cases}$$

Вектор (η_1, η_2) имеет двумерное гамма-распределение.

Оценки метода моментов

$$\beta_i = \frac{\mu_2^i}{\bar{\eta}_i}, \ \alpha_0 = \frac{m_{11}\bar{\eta}_1\bar{\eta}_2}{\mu_2^1\mu_2^2}, \ \alpha_i = \frac{(\bar{\eta}_i)^2}{\mu_2^i} - \frac{m_{11}\bar{\eta}_1\bar{\eta}_2}{\mu_2^1\mu_2^2},$$

где $i = 1, 2$; $\bar{\eta}_i$ — среднее η_i; μ_2^i — второй центральный момент;
m_{11} — смешанный центральный момент.

Частный случай двумерного гамма распределения

Параметры интенсивности β_i могут быть оценены по одномерным выборкам, в результате можно перейти к упрощенной модели.

Оценки параметров двумерного гамма-распределения по методу моментов (ММ)

Пусть x_i, где $i = 0, 1, 2$ — независимые гамма-распределенные случайные величины, с параметрами формы λ_i и масштаба $\beta_i = 1$. Тогда $y_i = x_0 + x_i$ имеет гамма распределение с параметрами Λ_i и 1. Параметры распределения (y_1, y_2) выражаются через параметры одномерного распределения:

$$\begin{cases} \lambda_0 = \rho\sqrt{\Lambda_1 \Lambda_2}, \\ \lambda_1 = \Lambda_1 - \rho\sqrt{\Lambda_1 \Lambda_2}, \\ \lambda_2 = \Lambda_2 - \rho\sqrt{\Lambda_1 \Lambda_2}. \end{cases}$$

[Алексеева, Щербакова, 2007]

Здесь $\rho = \mathrm{cor}(y_1, y_2)$.

Оценки метода МП параметров двумерного гамма распределения

Логарифмическая функция правдоподобия:

$$L(\mathbf{u},\ \mathbf{v}|\lambda_0, \lambda_1, \lambda_2) = \sum_{i=1}^{n} \ln \gamma(u_i, v_i|\lambda_0, \lambda_1, \lambda_2).$$

Используя совместную плотность н. с. в. $x_0, x_1, x_2 \sim \gamma(\alpha_i, 1)$, получим вид **плотности двумерного гамма распределения (Алексеева Н. П., 2012):**

$$\gamma(u, v|\lambda_0, \lambda_1, \lambda_2) = S(u, v) \cdot \frac{\exp^{-(u+v)}}{\Gamma(\lambda_0)\Gamma(\lambda_1)\Gamma(\lambda_2)},$$

$$S(u, v) = \begin{cases} \sum\limits_{n=0}^{\infty} \frac{u^{c-1} v^{-a} \Gamma(b)\Gamma(c-b)}{\Gamma(c) \cdot n!} \cdot {}_2F_1(a, b, c, \frac{u}{v}),\ u < v, \\ \sum\limits_{n=0}^{\infty} \frac{u^{-a} v^{c-1} \Gamma(b)\Gamma(c-b)}{\Gamma(c) \cdot n!} \cdot {}_2F_1(a, b, c, \frac{v}{u}),\ v < u. \end{cases}$$

$$(a, b, c) = \begin{cases} (1 - \lambda_2, n + \lambda_0, n + \lambda_0 + \lambda_1),\ u < v, \\ (1 - \lambda_1, n + \lambda_0, n + \lambda_0 + \lambda_2),\ v < u. \end{cases}$$

Проверка согласия с гамма распределением

Для проверки согласия эмпирического распределения с теоретическим одномерным гамма распределением применялся **параметрический bootstrap** с использованием статистики критерия Хи-квадрат Пирсона. Число bootstrap выборок 1000, уровень значимости $\alpha = 0.05$.

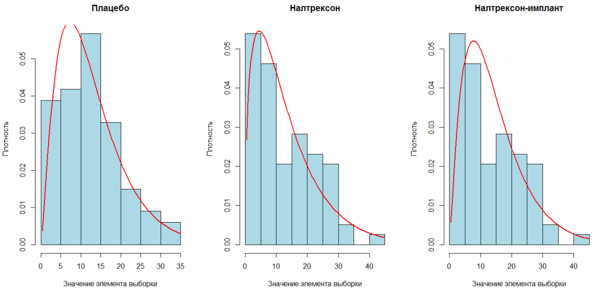


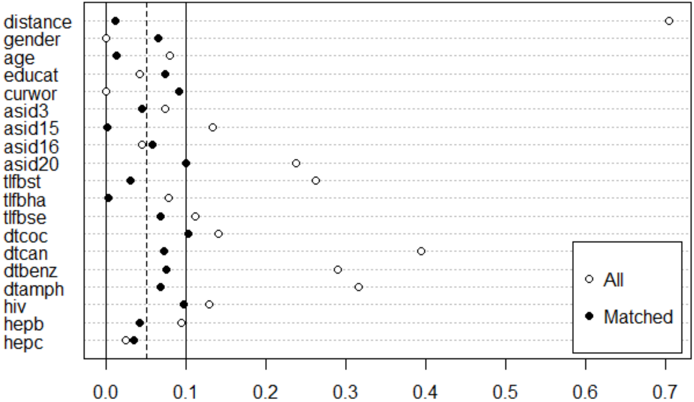
Рис.: Гистограммы значений теста по шкале BDI, полученные во второй опрос пациентов. Значения p-value критерия Хи-квадрат: $\text{p.v}_{\text{Плацебо}} = 0.388$, $\text{p.v}_{\text{Налтрексон}} = 0.175$, $\text{p.v}_{\text{Налтрексон-имплант}} = 0.754$

Получение сбалансированных групп сравнения методом псевдорандомизации

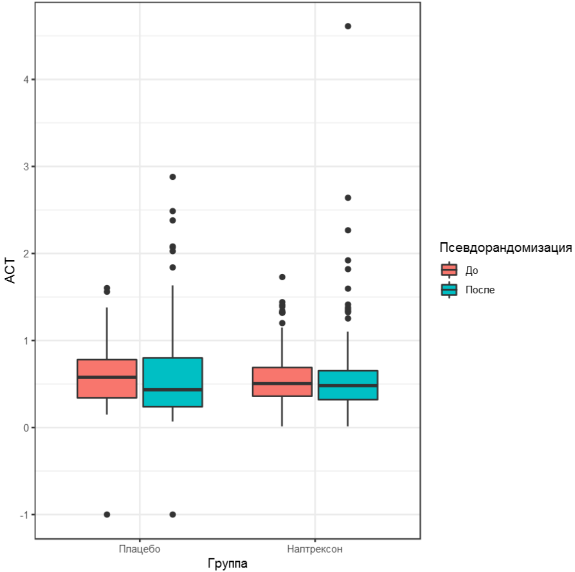
Пусть $\bar{x}_{0,\,1}$ — среднее значение признака в группах контроля и воздействия, σ_1 — стандартное отклонение признака в группе воздействия. **Условие сбалансированности групп:**

$$\left|\frac{\bar{x}_1 - \bar{x}_0}{\sigma_1}\right| < 0.25$$

по всем признакам.

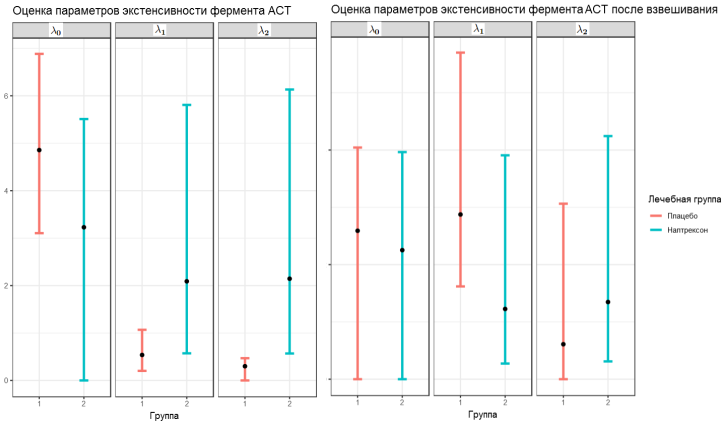


Получение сбалансированных групп сравнения методом псевдорандомизации



Проверка безопасности препарата

С помощью функции mle2 из R-пакета bbmle и функции правдоподобия, получим доверительные интервалы для оценок параметров модели. Уровень значимости равен 0.2.



Применение симптомного подхода

Построим такой суперсимптом, чтобы подгруппы, относящиеся к разному значению суперсимптома, были выборками из гамма распределения и максимально различались в смысле расстояния между распределениями.

Суперсимптом $Z = \bar{A}\bar{B}C$, где

- A — последнее воздержание от героина закончилось больше года назад,
- B — использует седативные препараты,
- C — обнаружены следы бензодиазепина в анализах

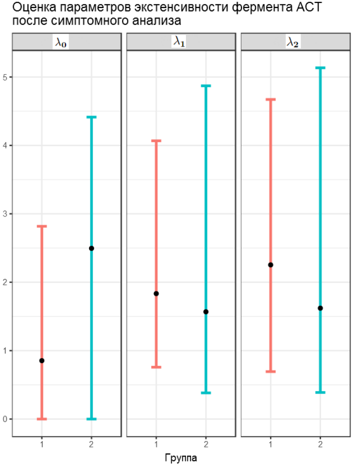
приводит к подгруппам объема 282 и 24 соответственно.

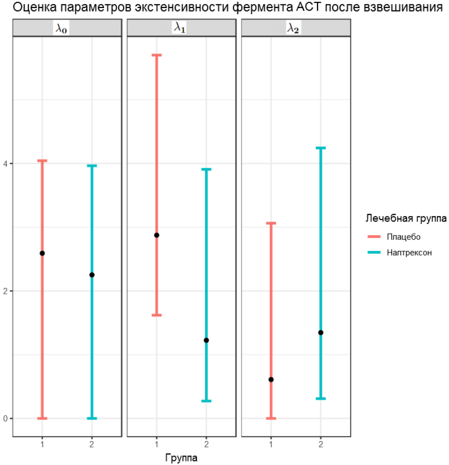
Меньшую группу исключим из исходной выборки как индивидов отличающихся от остальных в смысле значения наблюдаемых признаков.

Проверка безопасности препарата

(сравнение симптомного подхода и псевдорандомизации)

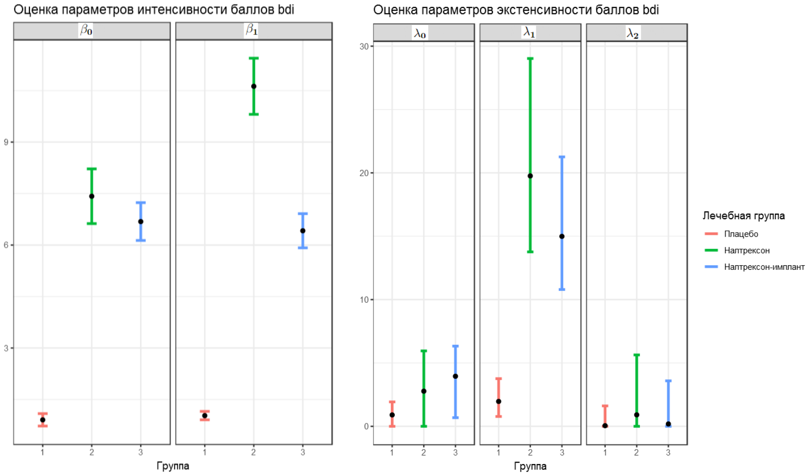
Уровень значимости равен 0.2.





Проверка «эффективности» препарата

Уровень значимости 0.05.



Результаты

- Применены методы устранения влияния сопутствующих факторов: симптомный анализ и псевдорандомизация
- Построены оценки параметров одномерного и двумерного гамма распределений
- Осуществлена проверка однородности параметров распределений посредством сравнения доверительных интервалов для оценок параметров
- *Продемонстрированы безопасность препарата по влиянию на состояние печени и наличие побочного эффекта препарата по снижению уровня депрессии*