Модель двухмерного гамма распределения с приложением в фармакологии

Морозов Никита Денисович, 22.М03-мм

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

Научный руководитель: к. ф.-м. н., доцент Н. П. Алексеева
Рецензент: Скурат Евгения Петровна

Санкт-Петербург
2024г.

Постановка задачи

Цель работы состоит в исследовании возможности применения гамма распределения для кардиологической базы данных.

- Выделение подходящих признаков имеющих согласие с гамма распределением.
- Сравнение оценок параметров у различных признаков, с разбиением их на несколько разных групп данных.
- Построение доверительных интервалов для одномерных и двухмерных оценок параметров гамма распределения.
- Проверка гипотез о согласии, значимости различий оценок параметров. Рассматривалось влияние проведенных операций и этиологии на показатели пациентов.

Структура данных

Данные представляют собой большую кардиологическую базу из 224 признаков и 169 индивидов.

- LA, LVd, LVs – Размеры предсердий и желудочков перед операцией. 2 временные точки.
- RA, RV – Правое предсердие и желудочек.
- EF, EF po – Фракция изгнания левого желудочка. 2 временные точки
- EKK, ISh – Время перфузии и ишемии
- KDO, KSO – Конечный диастолический и систолический объем ЛЖ. В 2 временных точках.
- pAPmax, PMKмах – Градиент давления на МК, АК(аортальный и митральный клапаны) или протезе.

Категориальные признаки: Пол: 1-мужчина, 0-женщина, ИМТ$<$25, возраст$<$35, курение: 1-курит, 0-нет.

Модель одномерного гамма распределение

Функция и плотность гамма распределения

Функция гамма распределения :

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \alpha > 0.$$

Плотность:

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

Где β – масштаб, α – форма.

Модель двумерного гамма распределения

Есть три независимо распределенных гамма величины ξ_1, ξ_2, ξ_3 с параметром масштаба равным 1, и параметрами формы(экстенсивности), равными $\lambda_1, \lambda_2, \lambda_3$. Из них мы можем построить случайные величины

$$\eta_1 = \xi_1 + \xi_2, \eta_2 = \xi_1 + \xi_3.$$

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.
Примем за $\Lambda_i = \lambda_0 + \lambda_i$, где $i = 1, 2$.
Параметры двумерного гамма-распределения могут быть получены следующим образом:

$$\lambda_0 = \rho\sqrt{\Lambda_1\Lambda_2}, \quad \lambda_1 = \Lambda_1 - \rho\sqrt{\Lambda_1\Lambda_2}, \quad \lambda_2 = \Lambda_2 - \rho\sqrt{\Lambda_1\Lambda_2}.$$

[Н.Алексеева, 2012]

Оценки ММП

$$L(x_1,\dots,x_n|\beta,\alpha)=\prod_{i=1}^{n}\frac{x_i{}^{\alpha-1}e^{\frac{-x_i}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}=\frac{1}{\beta^{n\alpha}\Gamma^{n}(\alpha)}\prod_{i=1}^{n}x_i{}^{\alpha-1}e^{\frac{-x_i}{\beta}}=$$

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

$$lnL(x_1,\dots,x_n|\beta,\alpha)=(a-1)\sum_{i=1}^{n}lnx_i-\frac{1}{\beta}\sum_{i=1}^{n}x_i-n\alpha ln\beta-nln\Gamma(\alpha),$$

$$\frac{\partial lnL(x_1,\dots,x_n|\beta,\alpha)}{\partial\alpha}=\sum_{i=1}^{n}lnx_i-nln\beta-n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}=0,$$

$$\Longleftrightarrow\frac{1}{n}\sum_{i=1}^{n}\ln(x_i)+\ln(\hat{\alpha})-\ln(\bar{x})-\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})}=0,$$

$$\frac{\partial lnL(x_1,\dots,x_n|\beta,\alpha)}{\partial\beta}=\frac{n\alpha}{\beta}-\sum_{i=1}^{n}x_i=0\Longleftrightarrow\hat{\beta}=\frac{\hat{\alpha}}{\frac{1}{n}\sum_{i=1}^{n}x_i}\Longleftrightarrow\hat{\beta}=\frac{\hat{\alpha}}{\bar{x}}.$$

Доверительные интервалы

Для построения интервалов используется метод Бутстрап. $\hat{\lambda}_i$ – изначальные оценки полученные по всей выборке. Получаем оценки параметров $\hat{\lambda}_i^*$ используя ММП для 1000 бутстрап выборок. σ_i – стандартное отклонения набора $\hat{\lambda}_i$

Доверительные интервалы для параметров двухмерного гамма распределения

$$P\left(\hat{\lambda}_i - u_{1-\alpha/2}\frac{\sigma_i}{\sqrt{n}} \leq \lambda_i \leq \hat{\lambda}_i + u_{1-\alpha/2}\frac{\sigma_i}{\sqrt{n}}\right) = 1 - \alpha$$

Где $u_{1-\alpha/2}$ — квантиль стандартного нормального распределения, n — размер выборки.

Критерий Манна-Уитни

Гипотеза

$$H_0 : med_1 = med_2.$$

$$H_1 : med_1 \neq med_2.$$

Пусть есть две независимые выборки
$X = \{X_1, X2, ..., X_m\} Y = \{Y_1, Y_2, ..., Y_n\}$ Объединяем выборки в одну:$Z = X \cup Y$ и присваиваем ранги всем значениям в объединенной выборке. Вычисляем статистики U для каждой выборки следующим образом:

$$U_X = R_X - \frac{m(m+1)}{2}$$

$$U_Y = R_Y - \frac{n(n+1)}{2}$$

где R_X и R_Y — суммы рангов для выборок X и Y соответственно, а m и n — размеры выборок. Статистика критерия U определяется как минимум из U_X и U_Y:

Гистограммы гамма распределения

Были отобраны те признаки, которые имеют согласие с гамма распределением. Для проверки согласия использовалась статистика χ^2.

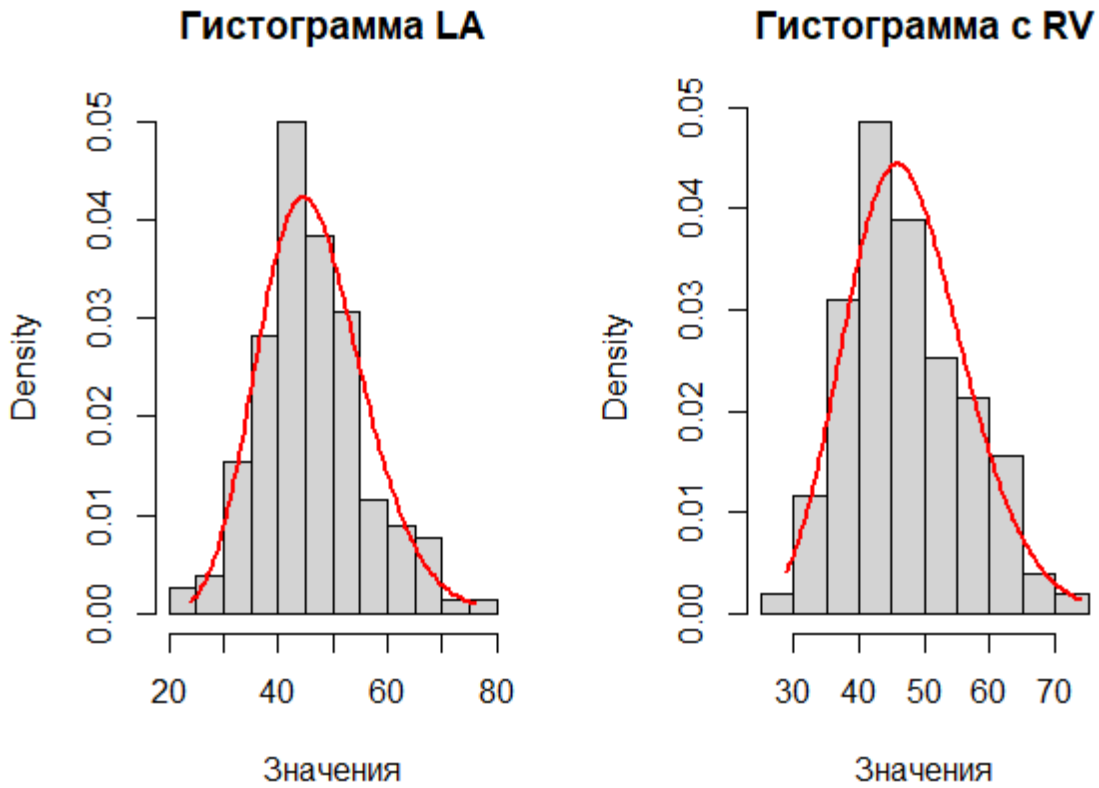

Рис.: Пример гистограмм признаков

Оценка параметров по признакам разбитых на группы

- Рассмотрены группы с значением TRUE и FALSE у категориальных переменных
- Получены оценки параметров и значение p-value для согласия с гамма распределения
- Удалось найти подгруппы у которых по полной выборке нет согласия с гаммой, но в подгруппе есть

Таблица: Пример Оценки сгруппированных параметров RA и p.value

Группа	Форма	Масштаб	P.Value Согласия
Изначальное значение	27	1.7	0.95
Женщина	19.18	2.4	0.9
Мужчина	32.2	1.4	0.7
Лишний вес	61	0.8	0.2
Нормальный вес	24	1.8	0.45

Получены оценки параметров у данных в 2 точках

У нас есть данные, которые представляю две временные точки, с ними были проведены операции оценки параметров и разбиения на группы с последующим получением оценок

Таблица: Пример оценок параметров двумерного гамма распределения в 2 точках

Признак	Λ_1	Λ_2	ρ	λ_0	λ_1	λ_2
LA,LAd po	25	40.5	0.66	21.3	3.7	19.2
LVd,LVd po	37.8	49.6	0.65	28.5	9.2	21
LVs,LVs po	21.01	22.1	0.7	15.2	5.7	6.8
EF,EF po	30.3	23.2	0.4	10.8	19.5	12.4

Создание симптомов

Имеется набор бинарных признаков x_i. Симптомом называется такой новый признак Y, образованный операциями сложения и умножения бинарных признаков x_i над полем F_2. Таким образом, например, имея a, b, c получаем симптомы.
$a \cdot b, a \cdot c, a \cdot b + c, a + c, b + c, a + b + c, ...$
Обозначим в работе переменные:

- Пол – a
- Возраст – b
- Вес – c
- Курение – d

Симптом

Рассмотрим подробнее симптом Вес+Пол·Вес+Возраст·Вес·Пол у признака LVd LVd ро.
Напомню:

- Вес TRUE – ИМТ<25
- Возраст TRUE – Возраст <35
- Пол TRUE – Мужчина
- Курение TRUE – Человек курит

Таблица: Группы в симптоме

Признак	0	1	2	3	4	5	6	7
Вес	0	0	0	0	1	1	1	1
Пол	0	0	1	1	0	0	1	1
Возраст	0	1	0	1	0	1	0	1

	TRUE			FALSE				
Группы	4	6	7	0	1	2	3	5
Вес	1	1	1	0	0	0	0	1
Пол	0	1	1	0	0	1	1	0
Возраст	0	0	1	0	1	0	1	1

В подгруппу TRUE у нас попадают худые люди, либо худые мужчины, либо худые молодые мужчины.
Одна из групп несмотря на небольшой вес(номер 5), попадает в FALSE подвыборку.

Таблица: Оценки параметров двумерной гаммы у симптома.

Группа	Λ_1	Λ_2	ρ	λ_0	λ_1	λ_2	N
TRUE	33	45.8	0.6	24.1	9.7	21.7	58
FALSE	51.74	57.39	0.63	34.33	17.41	23.06	54
Значимость	+	+	0.8	+	+	-	

Группа	βLVd	βLVd po	Med
TRUE	1.86	1.16	63
FALSE	1.34	1.00	69.5
Значимость	+	-	0.02

Доверительные интервалы

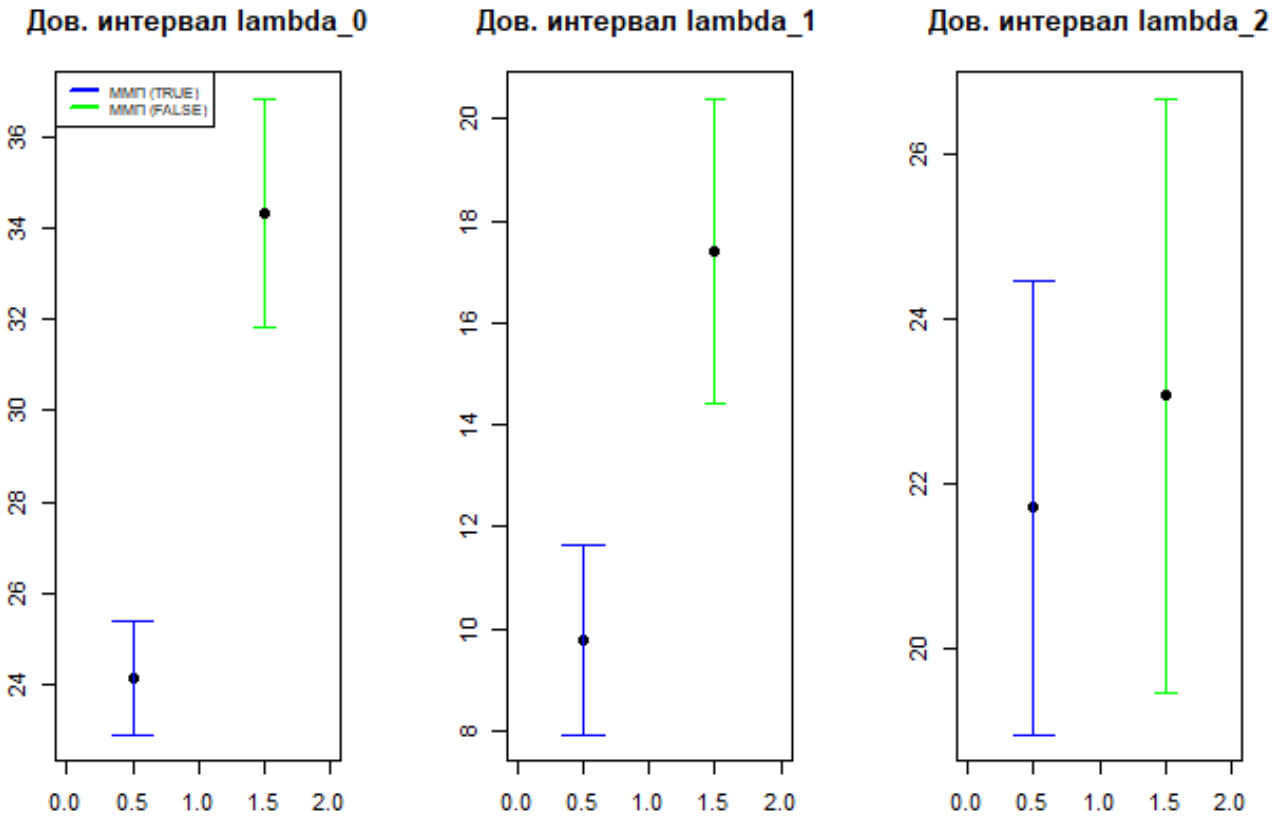

Рис.: Доверительные интервалы двумерной гаммы симптома по весу

Влияние этиологии

В базе данных даны 9 различных категорий этиологии. Для анализа все вторичные этиологии были объеденные в 1 кластер. Оценки этиологии с влиянием веса и возраста (b+bc)·этиология

Таблица: Группы в этиологии с симптомом

Признак	0	1	2	3	4	5	6	7
Пол	0	0	0	0	1	1	1	1
Вес	0	0	1	1	0	0	1	1
Этиология	0	1	0	1	0	1	0	1
N	2	1	18	7	20	31	18	15

	0		1		2		3	
Группы	0	2	5	7	1	3	4	6
Пол	0	0	1	1	0	0	1	1
Вес	0	1	0	1	0	1	0	1
Этиология	0	0	1	1	1	1	0	0
N	24		58		44		11	

0 – Женщины с вторичным, 1 – Мужчины с первичным, 2 – Женщины с первичным, 3 – Мужчины с вторичным

Таблица: Оценки двумерной гаммы Этиологии с симптомом с ограничением

Значение	Λ_1	Λ_2	ρ	λ_0	λ_1	λ_2	N
Мужчины Первичная	56.37	55.94	0.7	38.22	18.14	17.72	58
Женщины Первичная	28.30	47.44	0.5	17.26	10.74	22.34	11
Мужчины Вторичная	32.84	48.15	0.6	32	10.92	35.26	44
Женщины Вторичная	64.57	32.18	0.3	21.87	42.7	43.9	24

Можно заметить как у группы из женщин с вторичной этиологией отличаются параметры λ_1, λ_2 от всех подгрупп, и как мужчины с первичной этиологией отличаются от вторичных групп в целом.

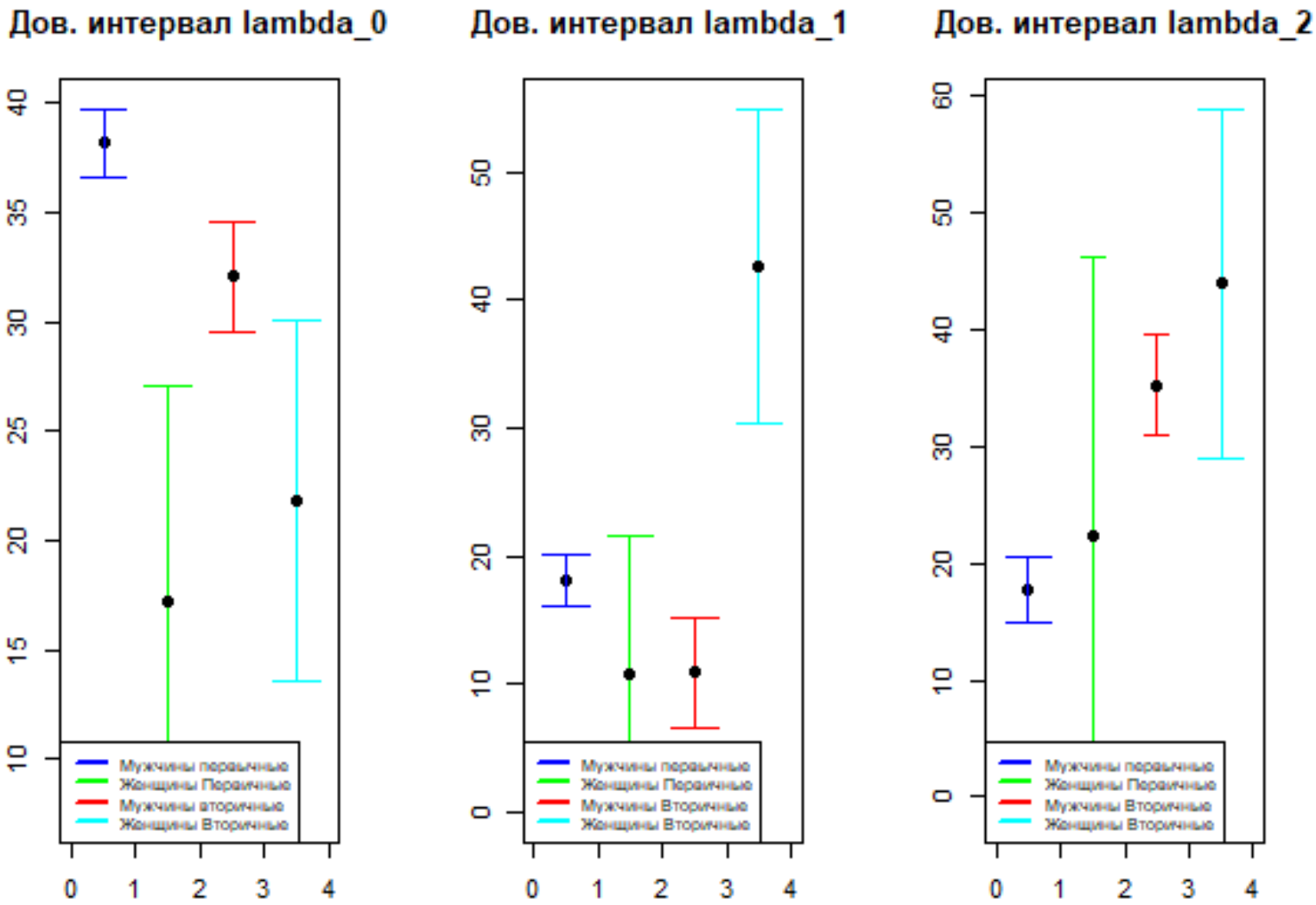

Рис.: Доверительные интервалы двумерной гаммы этиологии

Полученные результаты

- Рассмотрены модели и возможности получения оценок параметров у одномерного и двумерного гамма распределения.
- Модель и методы оценки параметров применены к реальным данным кардиологических наблюдений.
- Выделены признаки имеющие согласия с гамма распределением.
- Построены доверительные интервалы используя метод Бутстрап
- Найдены такие однородные группы, которые имеют различия в характерах процесса по его консервативности.