Статистические модели на основе гамма-распределения в анализе медико-психологических данных

Чебакова Майя Владимировна, гр.21.Б04-мм
Научный руководитель: кандидат физико-математических наук, доцент Алексеева Нина Петровна
Рецензент: научный сотрудник ПСПбГМУ им. акад. И.П. Павлова, Белякова Людмила Анатольевна

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

2025

Введение: цель и задачи

Рассматривается большой набор медико-психологических данных, включающий значения индекса массы тела (ИМТ), результаты психологических опросников и дополнительную информацию о респондентах.

Цель работы: применить статистические модели на основе гамма-распределения для анализа влияния факторов на параметры распределения ИМТ.

Задачи:

- Проверить согласие с гамма-распределением
- Проверить однородность параметров гамма-распределения между группами в зависимости от значений факторов.
- Расширить подход с помощью применения специальной регрессионной модели с несимметричными остатками.

Общая информация о данных

Данные: 27,770 наблюдений по 38 переменным (данные получены сотрудниками ПСПбГМУ им. акад. И.П. Павлова).

Переменные:

- 3 опросника на РПП, алекситимию, перфекционизм, депрессию, тревожность, манию.
- ИМТ (индекс массы тела, вес/рост^2(кг/м^2)), возраст, рост, вес, пол, дата тестирования.
- Город, регион, округ.

Фрагмент данных (4 из 38 переменных):

ИМТ	**Возраст**	**DEBQ (опросник)**	**EDE (опросник)**
25.86	55	11.00	3.07
42.21	28	7.86	2.57
34.06	50	8.75	1.68
30.12	45	6.10	1.96
41.00	50	12.10	4.07

Проверка согласия с гамма-распределением

- Рассматриваются значения ИМТ у женщин.
- Для проверки согласия распределения ИМТ с гамма, взято **20 случайных выборок объемом** $n = 200$**.**

Для каждой выборки из 20:

- Параметры гамма-распределения λ, β оценены методом максимального правдоподобия.
- Проверена гипотеза H_0: ИМТ согласуется с гамма-распределением.
- Критерий проверки: хи-квадрат Пирсона, уровень значимости $\alpha = 0.05$.

H_0 **отвергается** в 12 из 20 выборок.

Согласие с гамма-распределением в подгруппах

Стратификация ИМТ:

- **Пищевое поведение:** 5 категорий в зависимости от значения по Голландскому опроснику пищевого поведения (DEBQ).
- **Возраст:** 4 категории (18–25, 26–40, 41–59, 60–65 лет).

Для каждой категории взято **20 выборок объемом** $n = 200$ и проверяется аналогичная гипотеза H_0, $\alpha = 0.05$.

Число выборок, где H_0 отвергнута:

DEBQ	< 7	7–8	8–9	9–10	> 10
H_0 отвергнута	14	16	13	18	14

Возраст	18–25	26–40	41–59	60–65
H_0 отвергнута	15	16	13	16

Линейная регрессия и анализ остатков

- Модель:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2),$$

где y_i — ИМТ, x_i — возраст, ε_i — остаток.

- По тесту Шапиро-Уилка p-value < 0.05 во всех выборках ($\alpha = 0.05$).

- Сдвинутые остатки каждой модели:

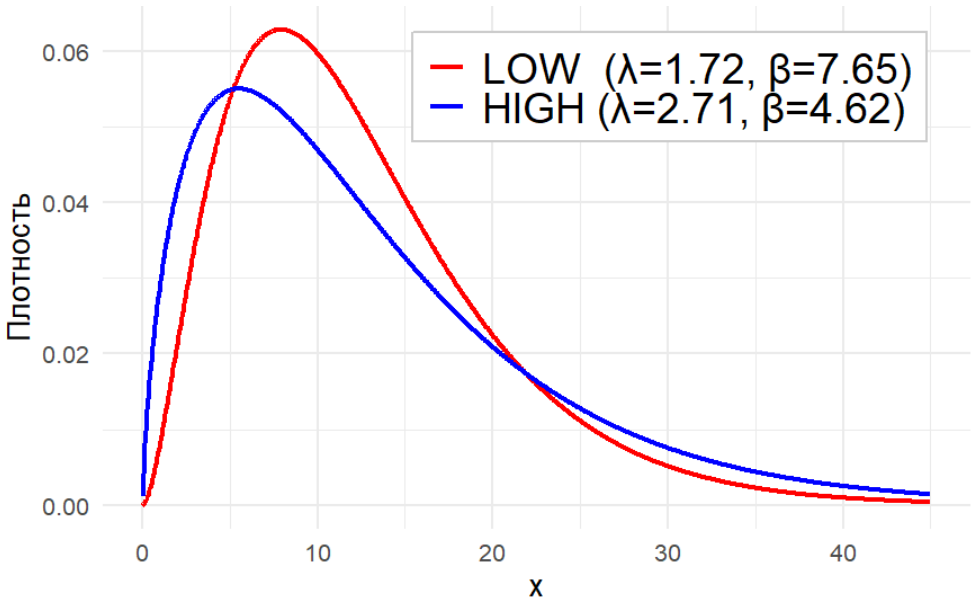
$$\varepsilon_i^{\text{shift}} = \varepsilon_i + |\min_j \varepsilon_j| + 10^{-4}, \quad i, j \in \{1, \ldots, 200\}.$$

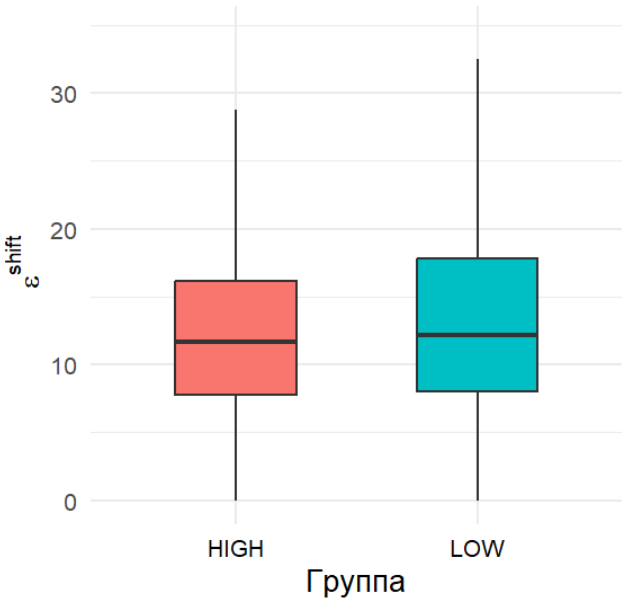
В 17 из 20 выборок $\varepsilon^{\text{shift}}$ **согласуются** с гамма-распределением (критерий Пирсона, $\alpha = 0.05$).

Деление на группы по риску РПП

$\varepsilon^{\text{shift}}$ делятся на 2 группы:

- **LOW** – DEBQ ≤ 7, $n = 446$, p-value = 0.07 (критерий Пирсона, $\alpha = 0.05$).
- **HIGH** – DEBQ > 7, $n = 3554$, p-value = 0.2.





- p-value = 0.08 (критерий Стьюдента), p-value = 0.06 (Колмогорова-Смирнова), $\alpha = 0.05$.

Проверка однородности параметров

- **Параметр масштаба β.**
  Для $X \sim \Gamma(\lambda, \beta)$ верно $\operatorname{cov}(X, \ln X) = \beta$.
  Гипотеза H_0 проверяется по выборочным ковариациям с использованием критерия Стьюдента.
- **Параметр формы λ.**
  Выборки нормируются: $X^* = \dfrac{X}{\widehat{\beta}} \sim \Gamma(\lambda, 1)$.
  Сравниваются выборочные средние используя тот же критерий.

Результаты проверки гипотез для групп HIGH и LOW ($\alpha = 0.05$):

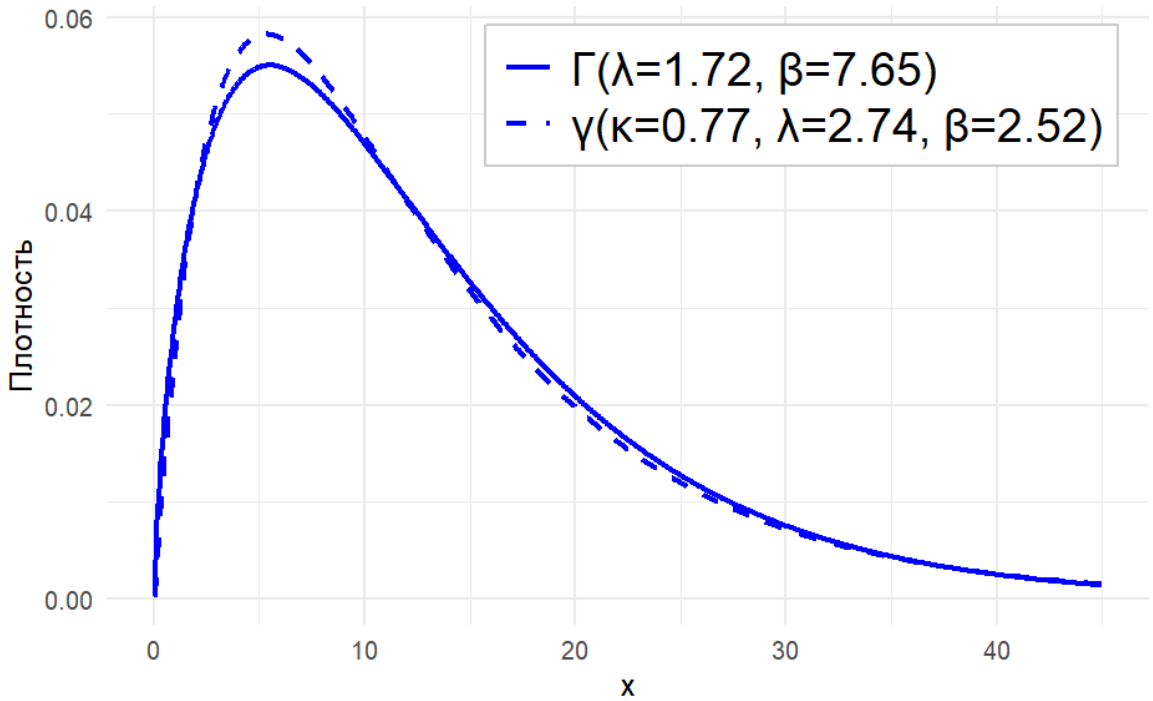
- $\beta_{LOW} = 7.65$, $\beta_{HIGH} = 4.62$, p-value $= 0.12$.
- $\lambda_{LOW} = 1.72$, $\lambda_{HIGH} = 2.71$, p-value $< 2.2 \cdot 10^{-16}$.

Степенное гамма-распределение

Степенное гамма-распределение

Распределение случайной величины $\xi^{1/\kappa}$, где $\xi \sim \Gamma(\lambda, \beta)$, с плотностью

$$\gamma(x|\kappa, \lambda, \beta) = \frac{\kappa}{\beta^{\lambda}\Gamma(\lambda)} x^{\kappa\lambda-1} e^{-x^{\kappa}/\beta} \quad (x, \lambda, \beta, \kappa > 0).$$



Синонимичные распределения

Синонимичные распределения

Распределение P_j **синонимично** распределению P_i с заданным уровнем синонимии δ^*, если $I(i:j) < \delta^*$, где:

$$I(i:j) = H_{ij} - H_{ii}, \quad -\text{средняя информация},$$

$$H_{ij} = -\int_X f_i(x) \log f_j(x)\, dx, \quad -\text{дифференциальная энтропия}.$$

Предложение 1 (Н. П. Алексеева, 2012 [1])

Если известны параметры степенного гамма-распределения $(\kappa_1, \beta_1, \lambda_1)$, *то, фиксируя значения* κ_2, *параметры* β_2, λ_2 *синонимичного степенного гамма-распределения находятся из системы:*

$$\lambda_2 = \left(\theta\left(\psi(\lambda_1+\theta) - \psi(\lambda_1)\right)\right)^{-1}, \quad \alpha_2 = \lambda_2 \alpha_1^{\theta} \frac{\Gamma(\lambda_1)}{\Gamma(\lambda_1+\theta)}, \text{ где}$$

$$\theta = \frac{\kappa_2}{\kappa_1}, \quad \psi-\textit{дигамма-функция}, \ \alpha = 1/\beta.$$

Информация I будет минимальна при данном κ, $\delta^* = I$.

Номинативное распределение

Номинативное распределение

При фиксированном уровне синонимии, **номинативным** называется синонимичное распределение с минимальной собственной энтропией H_{jj}.

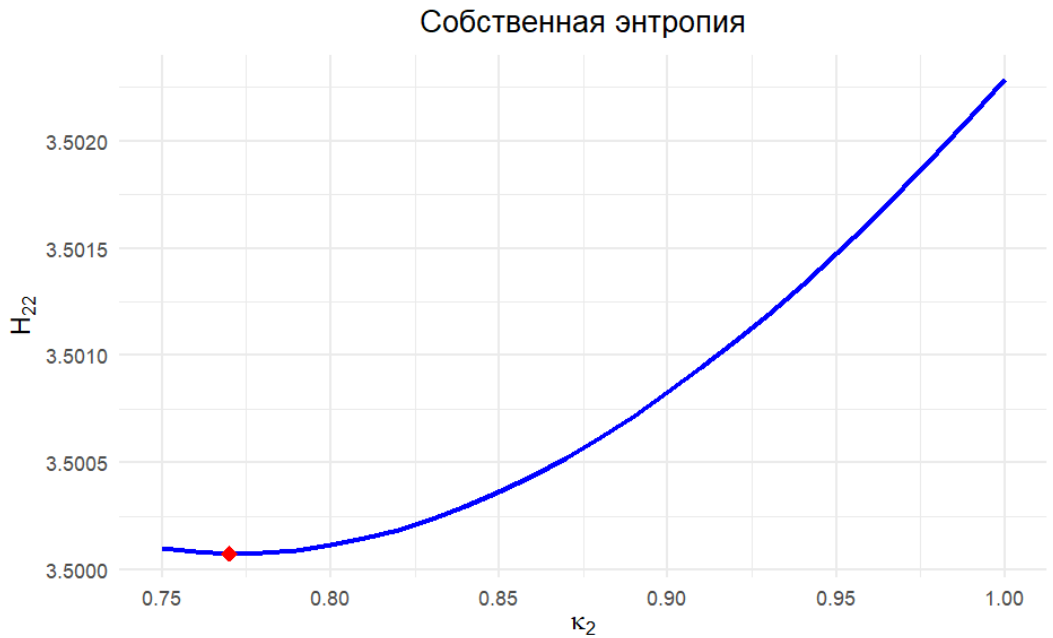
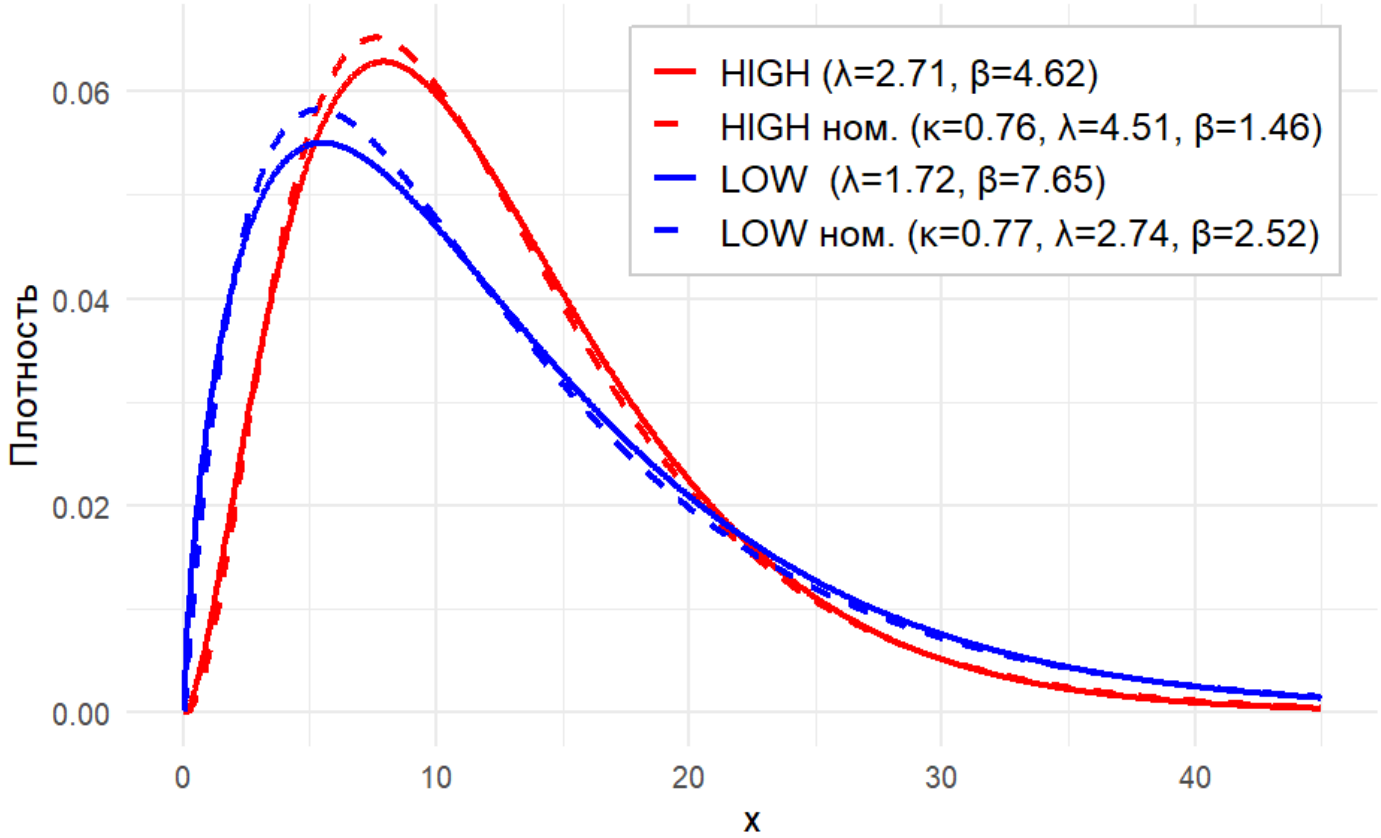


Рис. 1: График зависимости собственной энтропии от κ для синонимичных распределений группы LOW

Номинативные распределения групп LOW и HIGH



Проверка гипотез об однородности параметров номинативных распределений ($\alpha = 0.05$):

- β: p-value $= 0.05$ (p-value $= 0.12$ для групп HIGH и LOW).
- λ: p-value $< 2.2 \cdot 10^{-16}$, сохраняются **значимые** различия.

Информационное расстояние

Симметризованное информационное расстояние между распределениями

$$J(i,j) = I(i:j) + I(j:i) = H_{ij} + H_{ji} - H_{ii} - H_{jj}.$$

Для сравнения:

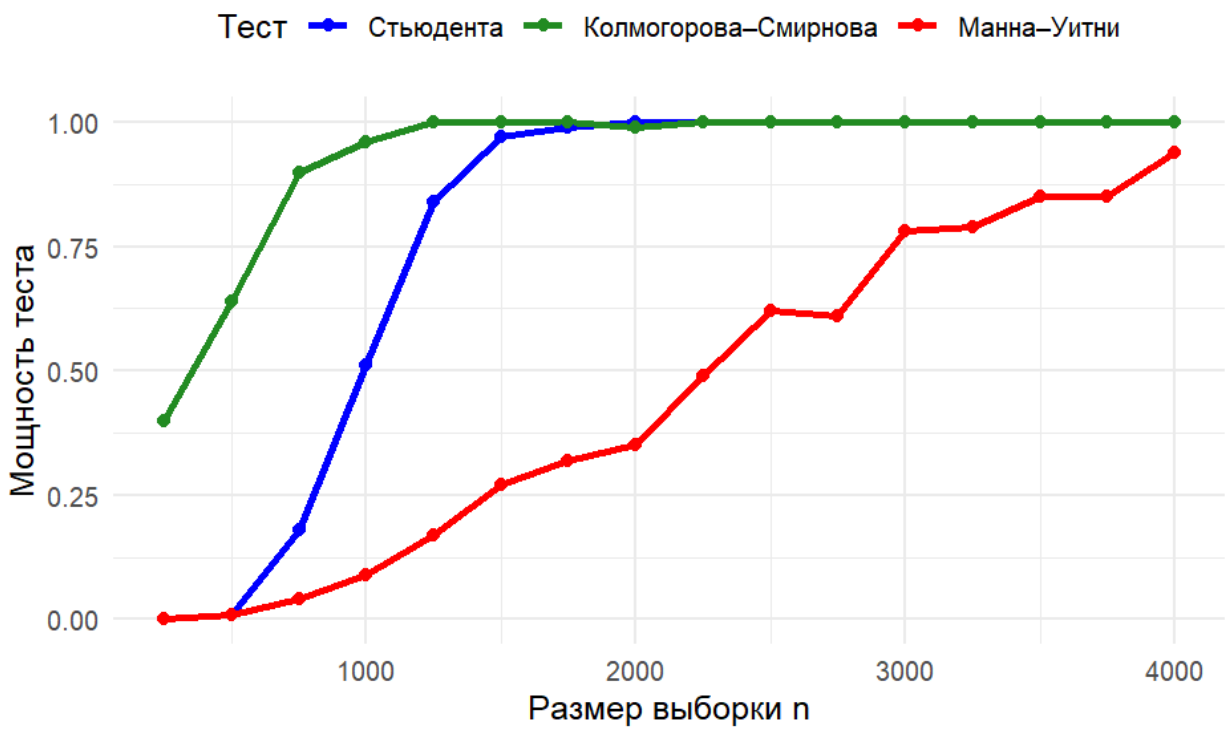
- $J(\text{LOW}, \text{LOW номинативное}) = 0.0043.$
- $J(\text{HIGH}, \text{HIGH номинативное}) = 0.0032.$
- $J(\text{LOW}, \text{HIGH}) = 0.1265.$

Расстояние между распределением и его номинативным

Рис. 2: Информационное расстояние J между распределением и его номинативным в зависимости от объема выборки. Параметры распределения - параметры группы LOW ($\lambda = 1.72$, $\beta = 7.65$).

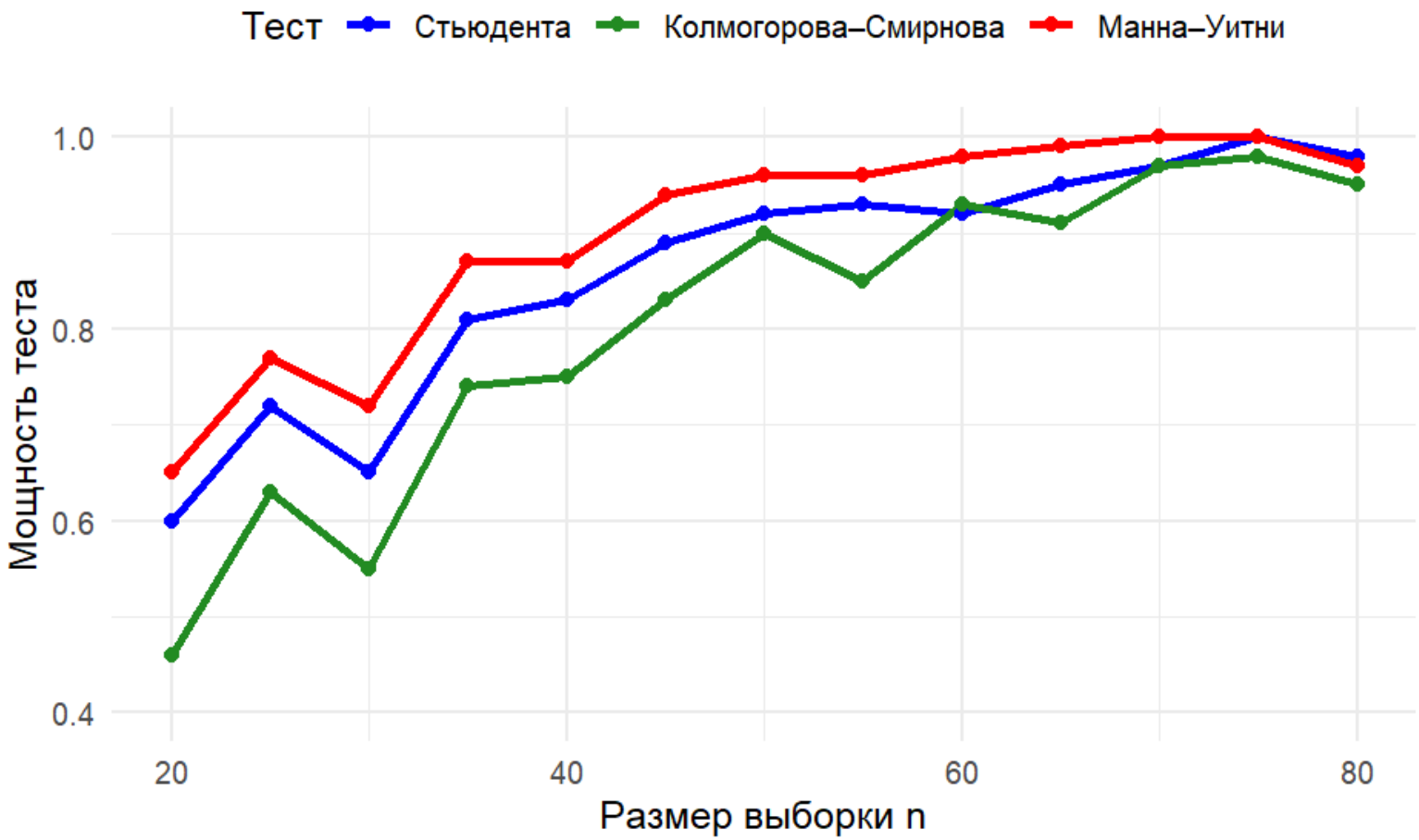
Чувствительность тестов к различиям в параметре β

- Для 100 генераций с параметрами как в группах LOW и HIGH для каждого n.
- Мощность ≥ 0.84 достигается при:
  - $n \geq 1250$ для критерия Стьюдента;
  - $n \geq 750$ для Колмогорова–Смирнова;
  - $n \geq 3500$ для Манна–Уитни.



Чувствительность тестов к различиям в параметре λ

- Мощность $\geq$ достигается уже при малых n:
  - $n \geq 35$ для Стьюдента и Манна–Уитни;
  - $n \geq 45$ для Колмогорова–Смирнова.



Регрессия с гамма-распределённой ошибкой

Модель:

$$y_i = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon_i, \quad \varepsilon_i \sim \Gamma(\lambda, \beta)$$

- y_i – ИМТ.
- x_i – TAS (алекситимия), EDE и DEBQ (опросники по РПП).
- 20 выборок объемом $n = 200$.
- Для каждой переменной x_i в проверялась однородность параметров гамма-распределения остатков между полной моделью с 4 предикторами и укороченной моделью (без этой переменной, с 3 предикторами).

Влияние переменных на параметры распределения остатков

Таблица 1: Количество выборок (из 20 по 200 человек), в которых исключение переменной привело к значимому различию в параметрах гамма-распределения остатков ($\alpha = 0.05$, критерий Стьюдента).

Переменная	λ **(форма)**	β **(масштаб)**
TAS (алекситимия)	1	0
EDE (РПП)	9	0
DEBQ (РПП)	1	0
Возраст	**18**	0

Заключение: интерпретация результатов

- Различие между группами LOW (норма) и HIGH (риск РПП):
  - параметр формы λ: большая неоднородность в группе HIGH $=>$ большее количество факторов, влияющих на ИМТ (психологических, поведенческих, генетических).
- Модель с гамма-распределенными остатками:
  - Возраст влияет на форму распределения остатков (параметр формы λ), что отражает возрастную динамику в физиологии и пищевом поведении.

Заключение: результаты работы

- Предложен подход к сравнению групп с помощью проверки однородности параметров гамма-распределения.
- Использованы синонимичные распределения для расширения подхода.
- Проведён анализ мощности статистических критериев.
- Установлена устойчивость приближения: расстояние J между распределением и номинативной моделью стабильно мало при любом объёме выборки.
- Построена регрессионная модель с гамма-распределённой ошибкой; проанализирован вклад переменных в структуру остатков.

Список литературы

 Алексеева Н.П. Анализ медико-биологических систем. Реципрокность, эргодичность, синонимия. — Российская Федерация : Издательство Санкт-Петербургского университета, 2012. — ISBN: 978-5-288-05347-4.