



ПРИМЕНЕНИЕ БАЙЕСОВСКОГО ПОДХОДА

В ОЦЕНКЕ РИСКОВ
РЕДКИХ ПРИРОДНЫХ
АВАРИЙ
НА ЖЕЛЕЗНОЙ ДОРОГЕ

Чикир Мария



Уральский Федеральный университет
им.первого Президента
России Б.Н.Ельцина



Уральское отделение
Российской
академии наук



Научно-инженерный центр
«Надежность и ресурс
больших систем и машин»

Серия крупных аварий показывает, что природные аномалии могут приводить к более серьезным последствиям, чем типовые технические аварии. Возрастает вероятность каскадного развития событий.



Дагестан, апрель 2026

Размыв полотна

Северо-Кавказской железной дороги из-за наводнения

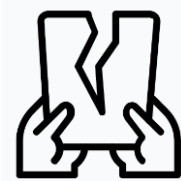
clideo.com 2

Особенности природных рисков:



Неуправляемы

только анализ и количественная оценка



Износ оборудования и железнодорожных путей

Как зависит степень разрушения от первоначального повреждения?
(все методы учитывают "идеальное состояние")



Уязвимость железнодорожного транспорта

к природным рискам по сравнению с другими видами транспорта



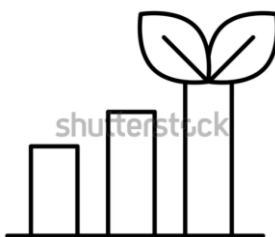
Меры глобального уровня не работают

на местном уровне



Халатность

ошибки персонала
и пренебрежение своими обязанностями



Доступность данных

- сложность климатических наблюдений
- отсутствие четкого алгоритма подачи сообщений об авариях
- противоречивость или неточность данных
- намеренное искажение информации с целью выйти сухим из воды

Почему байесовский подход?

графические модели, сочетающие теорию вероятностей и теорию графов



Байесовский подход сочетает частотные данные со знанием предметной области

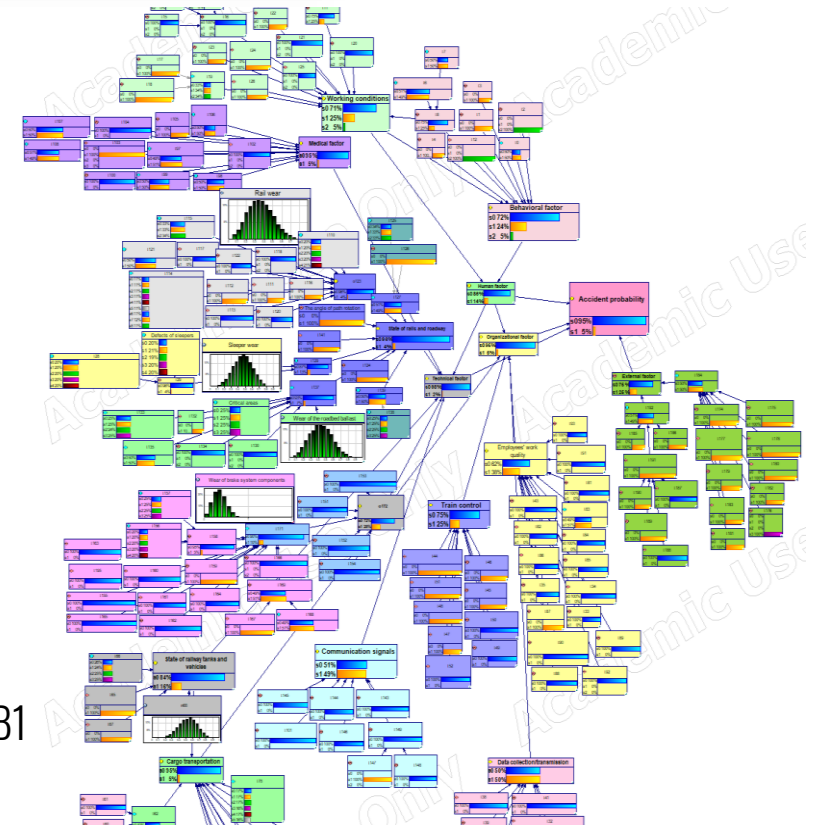
Числовые значения вероятностей могут извлекаться:

- баз данных
- на основе экспертного мнения
- определяются комбинацией вышеперечисленного

Теорема Байеса:

$$P_A(B_1) = \frac{P(B_1) \cdot P_{B_1}(A)}{P(A)} = \frac{P(B_1) \cdot P_{B_1}(A)}{P(B_1) \cdot P_{B_1}(A) + P(B_2) \cdot P_{B_2}(A) + \dots + P(B_n) \cdot P_{B_n}(A)},$$

- $P(B_1)$ – априорная вероятность события B_1
 - $P_A(B_1)$ – апостериорная вероятность B_1 (пересчитанная из-за " вновь открывшихся обстоятельств ")
- A fragment of the Bayesian model

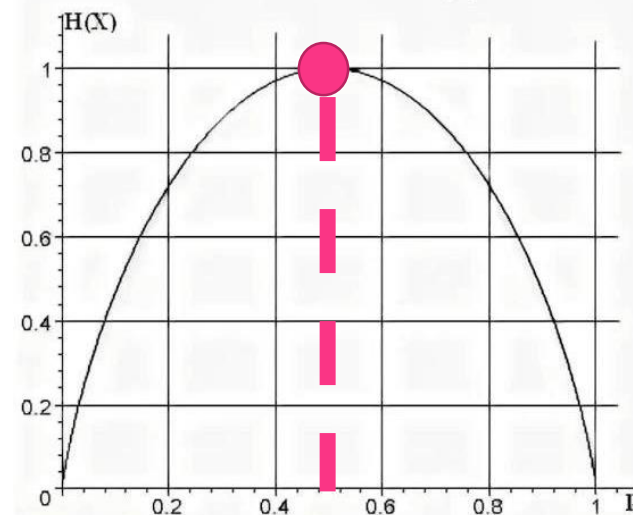


Роль вероятности в оценке неопределенности

БАЙЕСОВСКИЙ МЕТОД СТОХАСТИЧЕСКИЙ СЛУЧАЙНЫЙ ВЕРОЯТНОСТНЫЙ

- вероятность – количественная мера уровня неопределенности параметра

- Теорема Байеса – механизм правильного обновления вероятностей при поступлении новых данных в расчете на уменьшение существующей неопределенности



$$H(\alpha) = -\sum_{i=1}^N P_i \log P_i$$

Exploratory Data Analysis

- Доступность данных

статистика зарубежных аварий 1975-2022

- Предварительная обработка данных и отбор признаков

в среде программирования Python

- Библиотеки Python

Numpy, Scipy, Pandas, Matplotlib, Seaborn, PyMC3

- Пропуски данных

замена средним, наиболее частотным или внутригрупповым значением может привести к потере исходных связей между переменными.

библиотека `missongno`

- Анализ основных свойств данных

Категориальные ранговые и номинальные переменные

Визуальное сравнение качественных признаков

Количественные переменные

Анализ целевой переменной

Фрагмент набора данных по железнодорожным авариям

столбцы соответствуют будущим узлам сети

Report Year	Accident Month	Time	Accident Type	Hazmat Cars	Persons Evacuated	Temperature	Visibility	Weather Condition	Track Type	Maximum Speed
2017.0	6.0	2:14 PM	Derailment	0.0	149	65.0	Day	Clear	Yard	10.0
2017.0	6.0	2:14 PM	Derailment	0.0	0	65.0	Day	Clear	Yard	10.0
1981.0	4.0	7:20 AM	Side collision	0.0	0	28.0	Day	Snow	Yard	4.0
2007.0	1.0	7:10 AM	Derailment	0.0	0	56.0	Day	Cloudy	Industry	4.0
2017.0	10.0	3:55 AM	Hwy-rail crossing	7.0	0	66.0	Dark	Clear	Siding	0.0
2017.0	10.0	6:00 AM	Derailment	0.0	0	60.0	Dawn	Rain	Yard	3.0
2017.0	12.0	7:57 PM	Obstruction	0.0	0	62.0	Dark	Cloudy	Main	80.0
2016.0	12.0	6:30 AM	Other (describe in narrative)	0.0	0	16.0	Dark	Clear	Main	60.0
2017.0	3.0	8:45 AM	Derailment	23.0	0	44.0	Day	Rain	Industry	10.0
2017.0	4.0	10:40 AM	Derailment	0.0	0	38.0	Day	Clear	Yard	10.0
2006.0	10.0	5:10 PM	Derailment	0.0	0	70.0	Day	Clear	Yard	5.0
2010.0	6.0	10:00 PM	Derailment	1.0	0	55.0	Dark	Cloudy	Yard	6.0
2013.0	5.0	6:15 PM	Derailment	0.0	0	48.0	Day	Cloudy	Yard	7.0
2013.0	11.0	1:20 AM	Raking collision	0.0	0	0.0	Dark	Clear	Yard	3.0
2006.0	3.0	3:30 AM	Derailment	22.0	0	31.0	Dark	Clear	Yard	5.0
2006.0	3.0	9:55 AM	Derailment	3.0	0	45.0	Day	Cloudy	Yard	10.0
2014.0	3.0	6:00 PM	Derailment	0.0	0	40.0	Dusk	Cloudy	Industry	4.0
2011.0	4.0	6:15 AM	Derailment	0.0	0	42.0	Dawn	Rain	Yard	5.0
2007.0	7.0	10:25 AM	Derailment	0.0	0	75.0	Day	Clear	Siding	1.0
2006.0	10.0	4:50 AM	Derailment	4.0	0	40.0	Dark	Rain	Yard	10.0
2010.0	7.0	6:05 PM	Side collision	0.0	0	85.0	Day	Clear	Yard	6.0
2010.0	5.0	11:22 AM	Hwy-rail crossing	0.0	0	82.0	Day	Cloudy	Main	38.0
2010.0	5.0	5:45 AM	Other impacts	0.0	0	65.0	Dark	Clear	Yard	7.0
2012.0	5.0	6:00 AM	Other impacts	0.0	0	42.0	Dawn	Clear	Yard	6.0
2006.0	11.0	11:18 AM	Hwy-rail crossing	1.0	0	75.0	Day	Clear	Main	58.0
2016.0	8.0	5:30 AM	Other impacts	0.0	0	70.0	Dawn	Clear	Yard	16.0
2016.0	8.0	5:30 AM	Other impacts	1.0	0	70.0	Dawn	Clear	Yard	16.0

записи соответствуют различным значениям этих переменных

импорт в ПК GeNIe

Анализ целевой переменной «Общая стоимость ущерба»



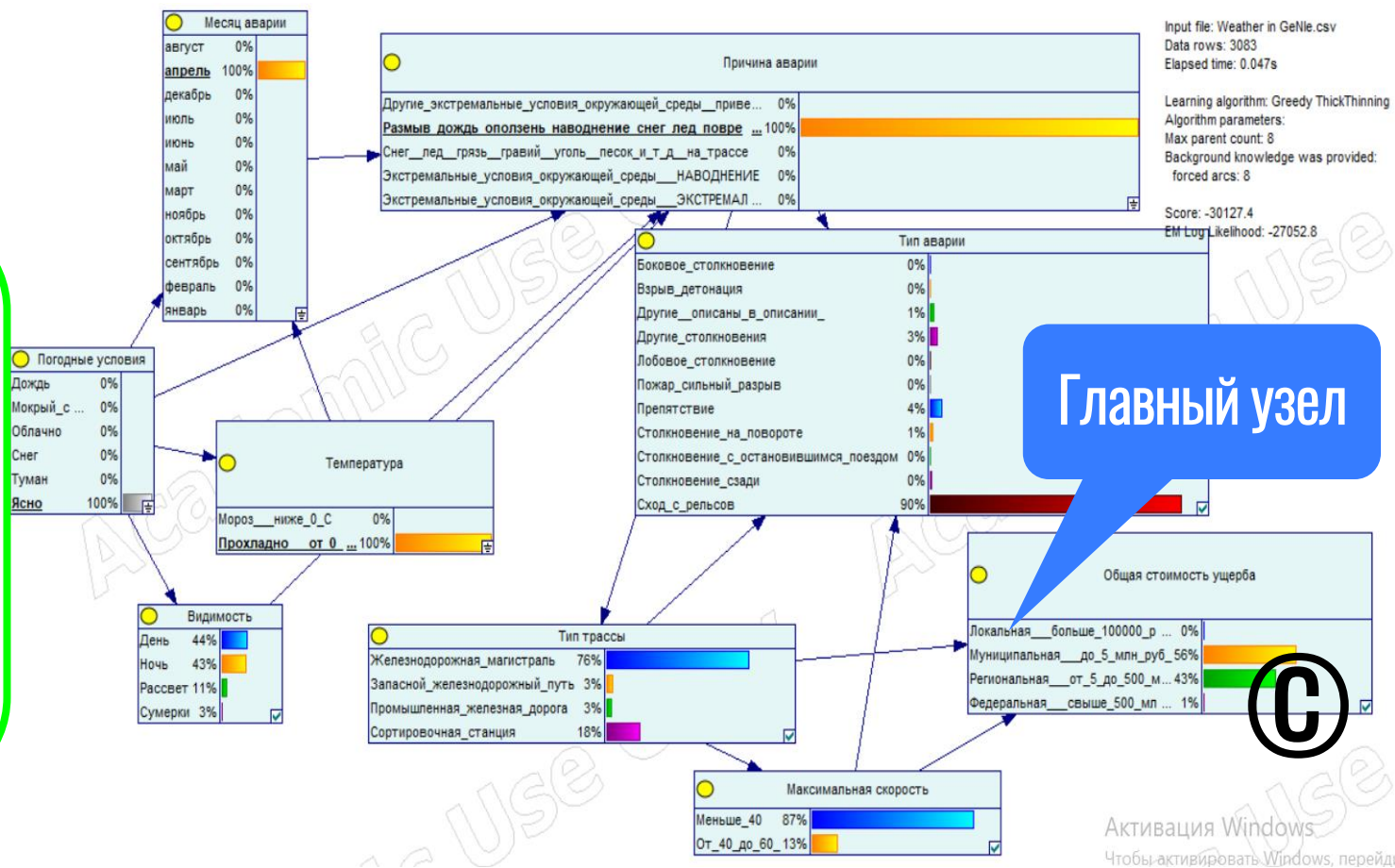
Фрагменты распределения переменной «Общая стоимость ущерба»

- Критерии статистической значимости
Естественные выбросы не являются ошибочными или неверно зарегистрированными
Удаление приведет к неверной интерпретации результатов расчета и потере ценной информации
- Правосторонняя асимметрия
среднее значение прижато к началу координат
- Потенциальные выбросы
145 экстремальных значений из 3083
(4,7 % от общего числа значений)
- Семейство лептокуртических распределений
- Тяжелый хвост
и принцип большого скачка Н.Талеба
«серые и черные лебеди»

Структурное обучение. Первая пробная модель

Сильный перекоз в распределении целевой переменной «Общая стоимость ущерба» приводит к разной точности прогнозов для состояний

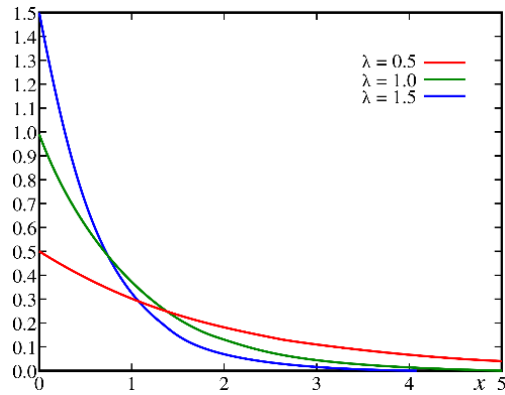
Причина – малое количество зарегистрированных случаев для определенных классов (редкие события)



Байесовская сеть для расчета ущерба природной аварии на железной дороге

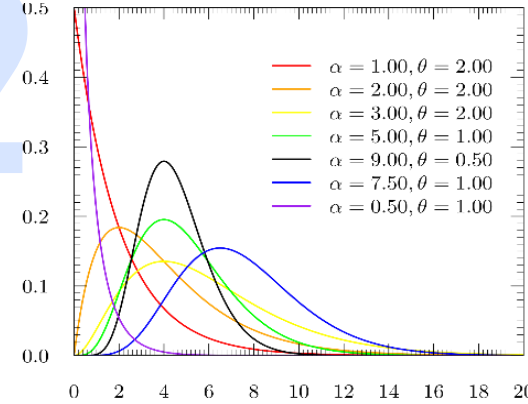
Подбор распределений целевой переменной

1



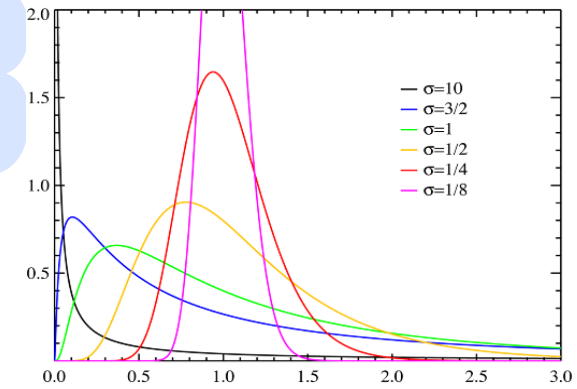
Экспоненциальное распределение

2



Гамма-распределение

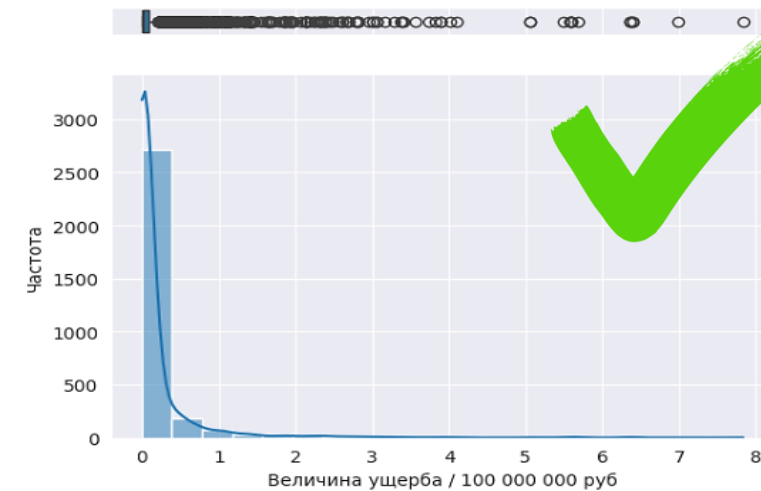
3



Логнормальное распределение

Статистические метрики распределений

Тип распределения	Mean (среднее)	Standard Deviation (среднестатистическое отклонение)	Skewness (скошенность)	Kurtosis (эксцесс)
Логнормальное	0.143	0.618	93.325	164591.228
Экспоненциальное	0.194	0.194	2.000	6.000
Парето	inf	inf	NaN	NaN
Гамма	0.194	0.319	3.283	16.171



Фактическое распределение целевой переменной

Первичная оценка соответствия данным

Количественная оценка

Подбор параметров распределений
на основе метода максимального правдоподобия
(Maximum Likelihood Estimation, MLE)



Расчет информационных критериев:

AIC – информационный критерий Акаике
BIC – байесовский информационный критерий
Тест Колмогорова-Смирнова
MAE – средняя абсолютная ошибка квантилей
 R^2 квантилей

```
#AIC-критерий
def calculate_aic(log_likelihood, num_params):
    return 2 * num_params - 2 * log_likelihood

#BIC-критерий
def calculate_bic(log_likelihood, num_params, num_samples):
    return num_params * np.log(num_samples) - 2 * log_likelihood

#Тест Колмогорова-Смирнова
def perform_ks_test(data, distribution, args, loc, scale):
    # Аргументы, передаваемые в kstest.cdf, должны быть параметрами формы, местоположением и масштабом.

    try:
        #случаи для одно- и двухпараметрических распределений
        #т.к. некоторые распределения содержат или только параметры формы (shape),
        #или только параметры положения/масштаба (loc/scale).
        if args is None or (isinstance(args, (tuple, list)) and len(args) == 0):
            # Распределения без параметра shape (ЭКСПОНЕНЦИАЛЬНОЕ)
            statistic, p_value = stats.kstest(data, lambda x: distribution.cdf(x, loc=loc, scale=scale))
        elif isinstance(args, (tuple, list)):
            # Распределения с параметрами shape (ЛОГНОРМАЛЬНОЕ, ГАММА, ПАРЕТО)
            statistic, p_value = stats.kstest(data, lambda x: distribution.cdf(x, *args, loc=loc, scale=scale))
```

Пользовательский функции для расчета
информационных критериев точности в Python

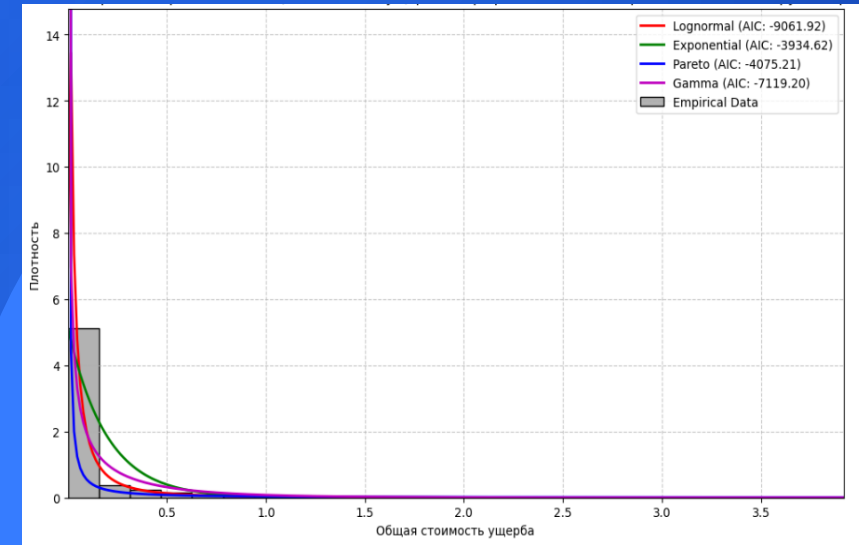
Тип распределения	AIC	BIC	KS Test p-value	Quantile MAE	R ²
Логнормальное	- 9061.919	- 9049.849	0.0000	0.062	8.076 e-01
Экспоненциальное	- 7119.199	- 7107.129	0.0000	0.082	8.026 e-01
Парето	- 4075.213	- 4063.143	0.0000	535190.347	-1.606 e+14
Гамма	- 3934.624	- 3928.589	0.0000	0.137	5.454 e-01

Показатели точности подбора распределений

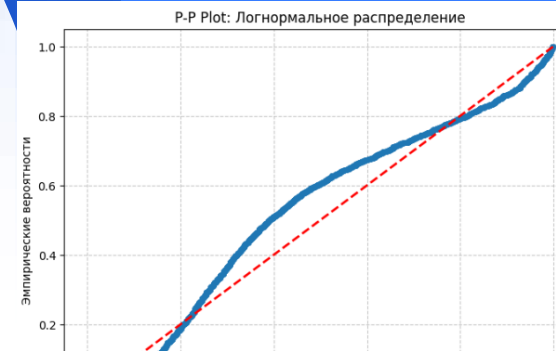
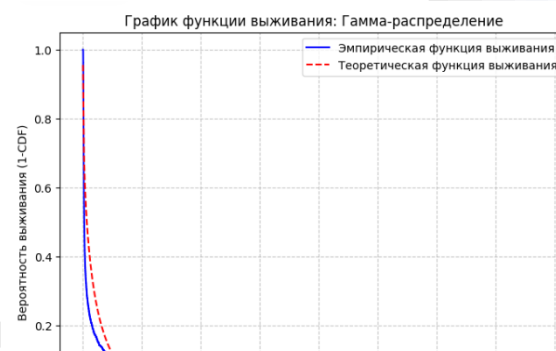
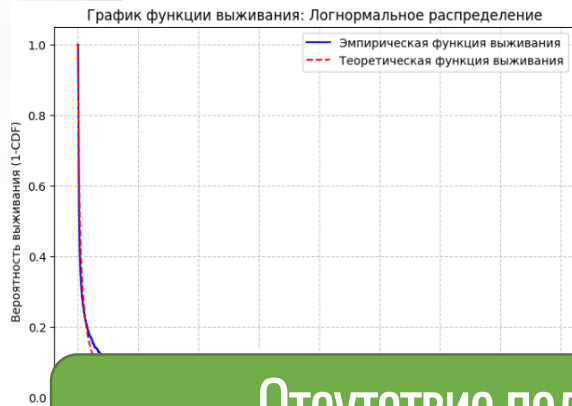
Первичная оценка соответствия данным

Качественная оценка

Графики плотности распределений
Q-Q plot – графики для сравнения квантилей
P-P plot – графики для сравнения
эмпирической функции распределения с теоретической
Survival Function Plot – график функции выживания



Графики функций плотностей вероятности



Отсутствие полного соответствия данным ни для одного из традиционных распределений
Наибольшее соответствие наблюдается для гамма- и логнормального распределений

Байесовский анализ

Математическое описание модели в байесовском анализе состоит из описания функций правдоподобия и априорной вероятности

Подбор априорной вероятности

СЛАБОИНФОРМАТИВНЫЙ
АПРИОР

Подбор правдоподобия

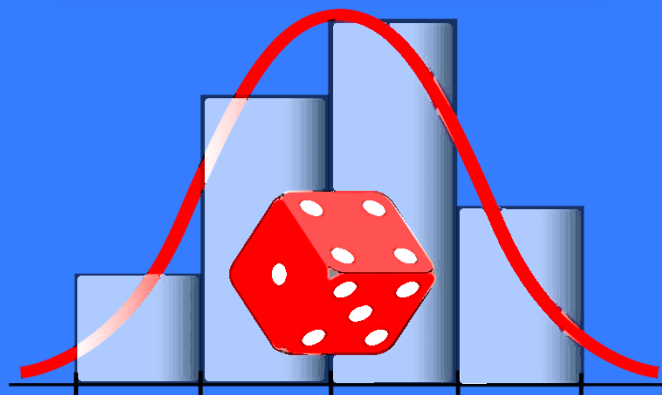
СОПРЯЖЕННОСТЬ РАСПРЕДЕЛЕНИЙ
АПРИОРНОЙ И УСЛОВНОЙ
ВЕРоятности

Выбор оценки апостериорной вероятности

Markov Chain Monte Carlo

- наименьшее влияние на анализ
- в случае малого количества данных или их высокой неопределённости
- целенаправленная минимизация влияния предшествующих знаний во избежание перегрузки модели избыточной информацией

- выражает правдоподобность данных с учетом принятых параметров априорного распределения



- относится к методам численного приближения
- семейство алгоритмов, использующих случайные выборки для расчета апостериорного распределения
- использование сложных апостериоров
- более точные приближения

Модель gamma_bayesian_model

Нотация модели

$$\begin{aligned}\alpha &\sim |N(0, \sigma_\alpha)| \\ \beta &\sim |N(0, \sigma_\beta)| \\ y &\sim \text{Gamma}(\alpha, \beta)\end{aligned}$$

```
alpha_prior_std = 0.371
beta_prior_std = 0.524

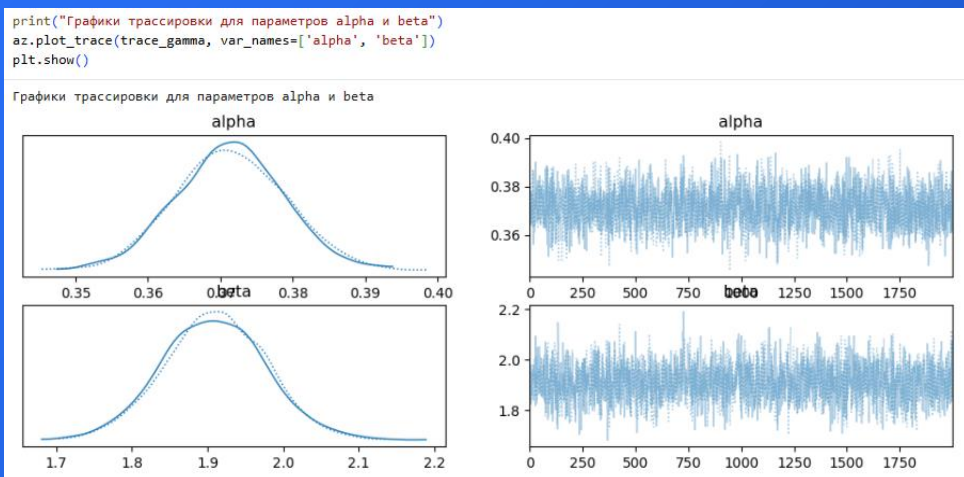
with pm.Model() as gamma_bayesian_model:
    # Априорные вероятности для параметров гамма-распределения
    alpha = pm.HalfNormal('alpha', sigma=100)
    beta = pm.HalfNormal('beta', sigma=100)

    # Правдоподобие должно быть гамма-распределением, т.к. только сопряженное гамма-распределение порождает апостериорное гамма
    likelihood = pm.Gamma('likelihood', alpha= alpha, beta= beta, observed=damage_costs)

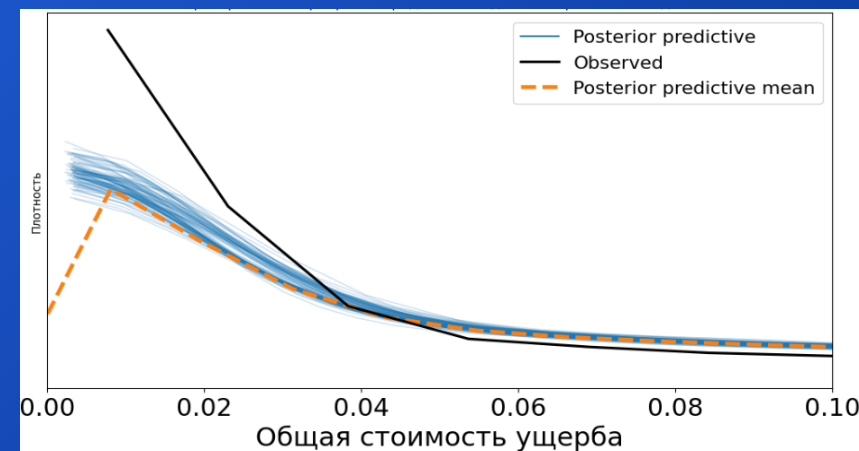
    # Выборка из апостериорного распределения
    trace_gamma = pm.sample(2000, tune=1000, return_inferencedata=True, random_seed=42, log_likelihood=True)
```

Progress	Draw	Divergences	Step size	Grad evals	Speed	Elapsed	Remaining
=====	3000	0	1.144	5	1560.78 draws/s	0:00:01	0:00:00
=====	3000	0	1.107	7	780.82 draws/s	0:00:03	0:00:00

Процесс создания модели и запуска MCMC



Графики трассировки для параметров μ и σ

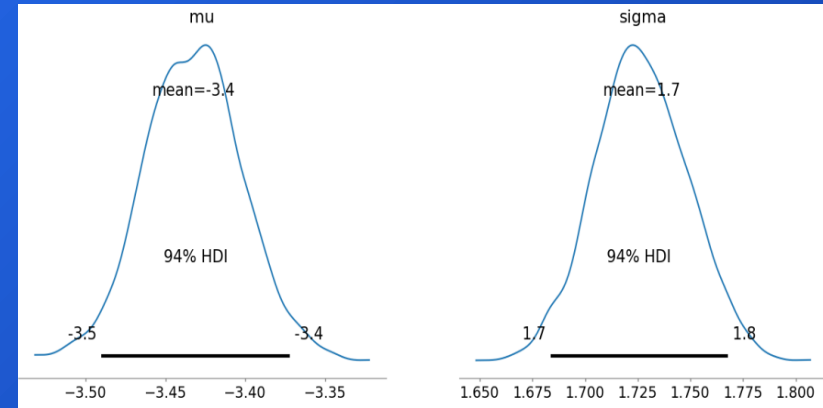


Проверка прогнозируемого апостериорного распределения

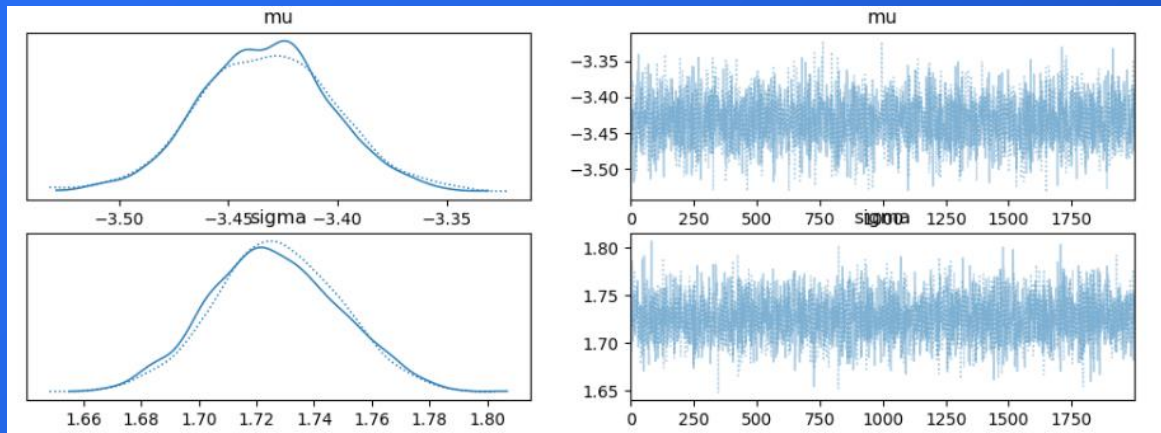
Модель lognormal_bayesian_model

Нотация модели

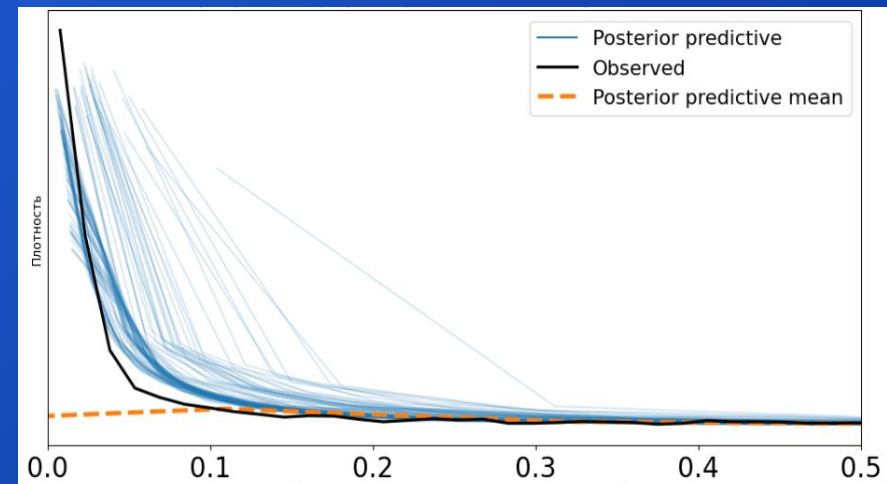
$$\begin{aligned}\mu &\sim N(\mu_{ln}, \sigma_{ln})| \\ \sigma &\sim |N(0, \tau^2)| \\ y &\sim \text{Lognormal}(\mu, \sigma)\end{aligned}$$



Проверка апостериорного распределения



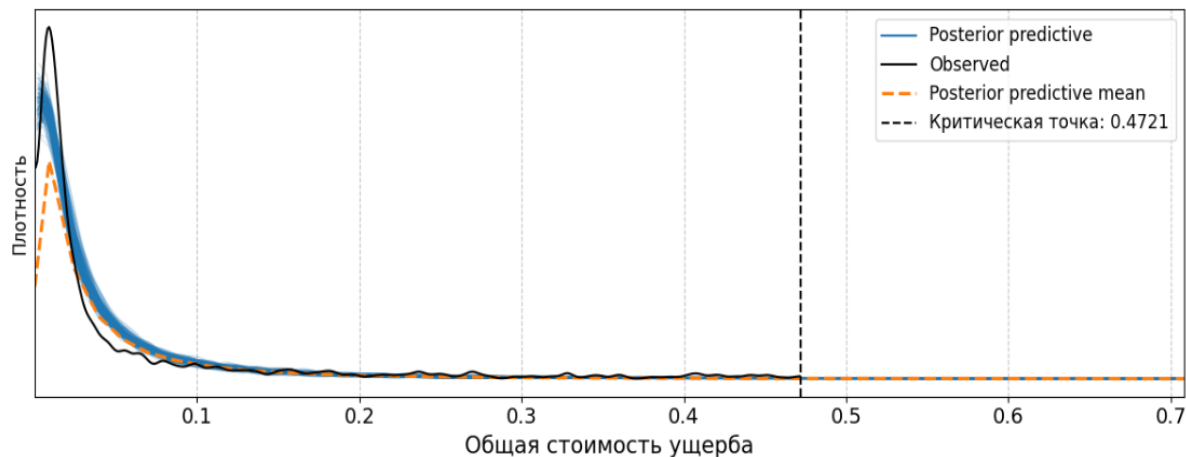
Графики трассировки для параметров μ и σ



Проверка прогнозируемого апостериорного распределения

Смешанная двухкомпонентная логнормальной-модель

Проверка апостериорного распределения в целом



Нотация модели

$$\mu_{centr} \sim |N(0, \sigma_{\mu c})|$$

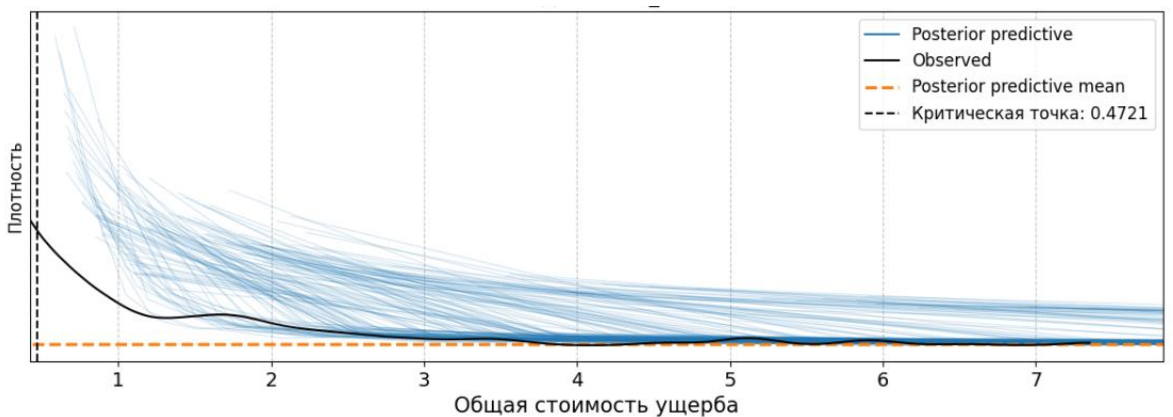
$$\sigma_{centr} \sim |N(0, \sigma_{\sigma c})|$$

$$\mu_{tail} \sim |N(0, \sigma_{\mu t})|$$

$$\sigma_{tail} \sim |N(0, \sigma_{\sigma t})|$$

$$y \sim \text{Lognormal}(\mu_{centr}, \sigma_{centr}, \mu_{tail}, \sigma_{tail})$$

Проверка апостериорного распределения для хвостовой части



Проверка апостериорного распределения для центральной части

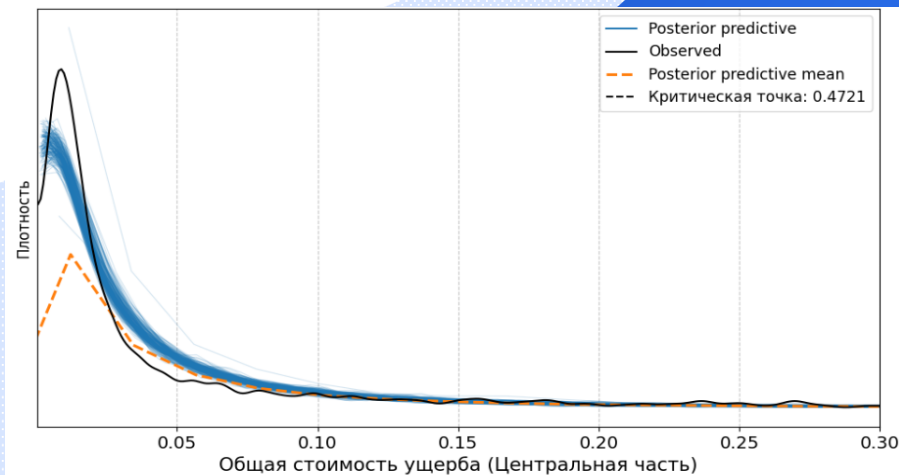


Схема исследования редких природных аварий на железной дороге



Исходные данные

- 1) Набор данных по отказам и предаварийным ситуациям на железной дороге (min количество записей – 100);
- 2) Фильтрация записей по природно-климатическому признаку;
- 3) Входные переменные (зависимые) – совокупность природно-климатических характеристик, а также ряд технических и организационных факторов.



Подготовка данных

- 1) Анализ характера природных аварий, предварительная оценка масштаба аварий, анализ влияния загрузки вагонов на аварийность;
- 2) Оценка полноты данных, закономерностей между пропущенными значениями;
- 3) Исследование экстремальных значений и выбросов зависимых и независимых переменных.



Исследовательский анализ данных (Exploratory Data Analysis)

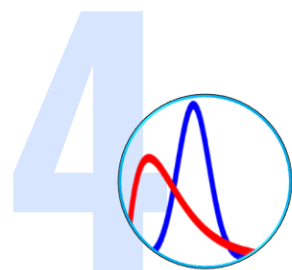
- 1) Анализ основных свойств данных. Проверка на дубликаты, противоречивые записи и логические ошибки;
- 2) Исследование типов переменных (количественные, категориальные – номинальные, ранговые), оценка мультимодальности, кодирование переменных;

Схема исследования редких природных аварий на железной дороге



Исследовательский анализ данных (Exploratory Data Analysis)

- 3) Визуальное сравнение природных аварий в зависимости от типа железнодорожного участка и вида происшествия;
- 4) Масштабирование переменных, определение целевой переменной «Общая стоимость ущерба»;
- 5) Проверка автокорреляции, корреляционный анализ. Отбор наиболее значимых признаков;
- 6) Создание пробных сетей Байеса, анализ гетероскедастичности для целевой переменной;
- 7) Заключение о вероятном виде распределения переменной «Общая стоимость ущерба».



Байесовское моделирование

- 1) Подбор распределений для целевой переменной с учетом асимметричности и тяжелохвостости;
- 2) Создание байесовских моделей: определение априорного распределения, функции правдоподобия и способа расчёта априорной вероятности;
- 3) Построение двухкомпонентной байесовской модели для описания тяжелохвостых распределений ущерба: выбор порогового значения, моделирование центральной и хвостовой части;
- 4) Количественное и качественное сравнение моделей, выводы о возможности прогнозирования ущерба природных аварий на железной дороге.
- 5) **Выходная переменная (независимая)** – функция плотности распределения ущерба природной аварии на железной дороге с учётом аномальных ЧС.

Контакты

Научно-инженерный центр
«Надежность и ресурс больших систем и машин»
Уральского отделения Российской академии наук



sec@sec.uran.ru

mary-catchikir2012@yandex.ru



+7 (343) 374-16-82

Чикир Мария
аспирант
младший научный сотрудник

Спасибо за внимание!

