Модель негативного бинома в статистическом анализе последовательностей событий

Рожненко Людмила Валерьевна, гр. 422

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

Научный руководитель: к.ф.–м.н., доцент Н.П. Алексеева
Рецензент: к.ф.–м.н., доцент Т.М. Товстик

Санкт-Петербург
2017г.

Цель работы и задачи

Цель работы:

- изучение модели процесса восстановления с негативно биномиальным распределением: интервалов между событиями и соответствующего целочисленного процесса.

Задачи:

- нахождение законов распределения целочисленного процесса для различных межинтервальных распределений (Бернулли, геометрическое, негативно–биномиальное);
- оценка параметров негативно–биномиального распределениями;
- применение критерия согласия дискретного распределения при неизвестных параметрах.

Описание данных

1. **Данные**: электроэнцефалограммы (ЭЭГ) пациентов с болезнями Альцгеймера, Паркинсона, с правосторонней цервикальной дистонией, эпилепсией и здоровые (всего 58 пациентов).
2. **Структура данных для одного пациента**: числовая таблица из 16 столбцов, каждый столбец отвечает одному датчику, и 5080 строк, что составляет 20 секунд записи ЭЭГ.

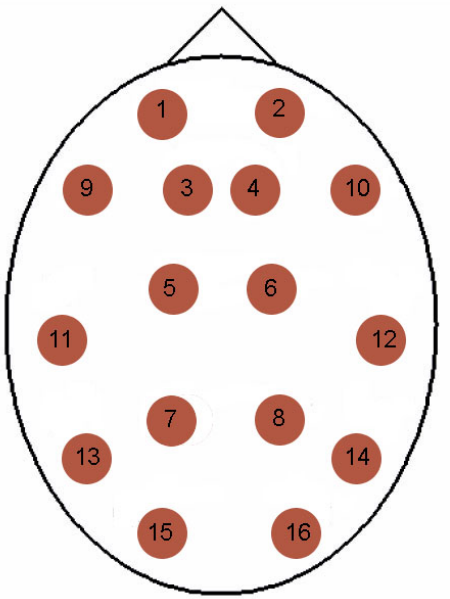


Рис. 1: Расположение датчиков на голове

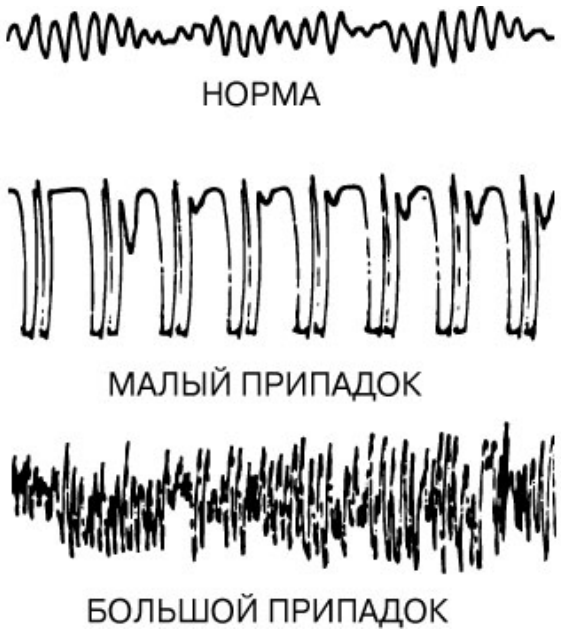


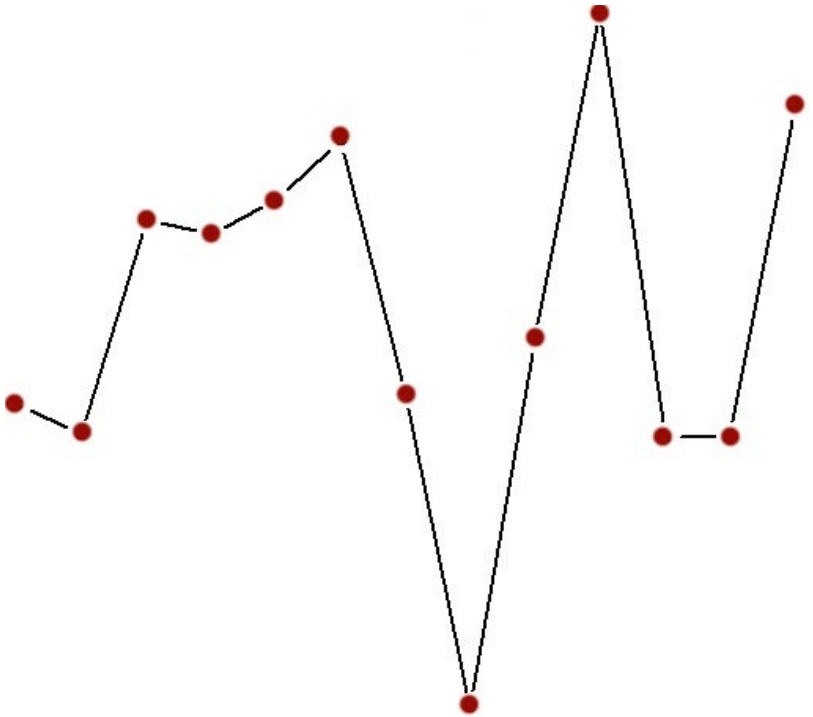
Рис. 2: Пример отображения отклонений на ЭЭГ

Представление ЭЭГ в виде категориального ряда

Обозначение: a — средняя по модулю величина скачка ЭЭГ ($a > 0$).

Таблица: Перевод последовательности ЭЭГ в категориальный ряд

Последовательность	Величина скачка x	Слово
возрастает	$x \in (0; a)$	«u» (up)
	$x \geq a$	«U» (Up)
убывает	$x \in (-a; 0)$	«d» (down)
	$x \leq -a$	«D» (Down)



“d U d u u D D U U D d U”

Многомерная выборка для целочисленного процесса в ЭЭГ

- Датчик k $(k = 1, \ldots, K)$, $K = 16$.
- Пациент i $(i = 1, \ldots, N)$, $N = 58$.
- Фрагмент категориального ряда (uUud, uddu, udDd).

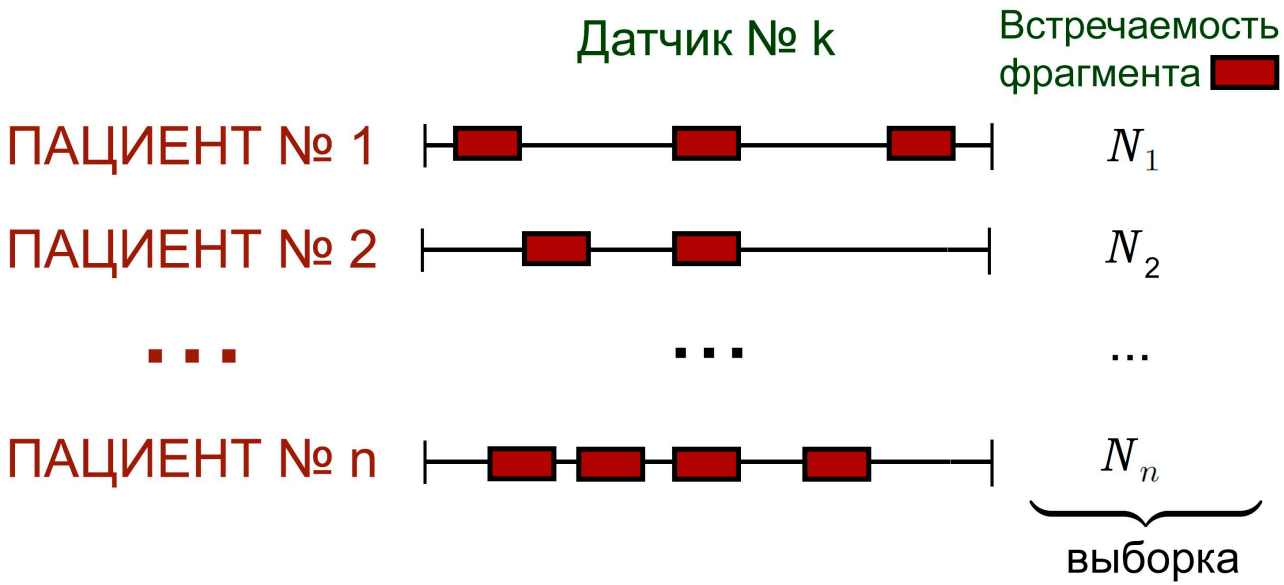
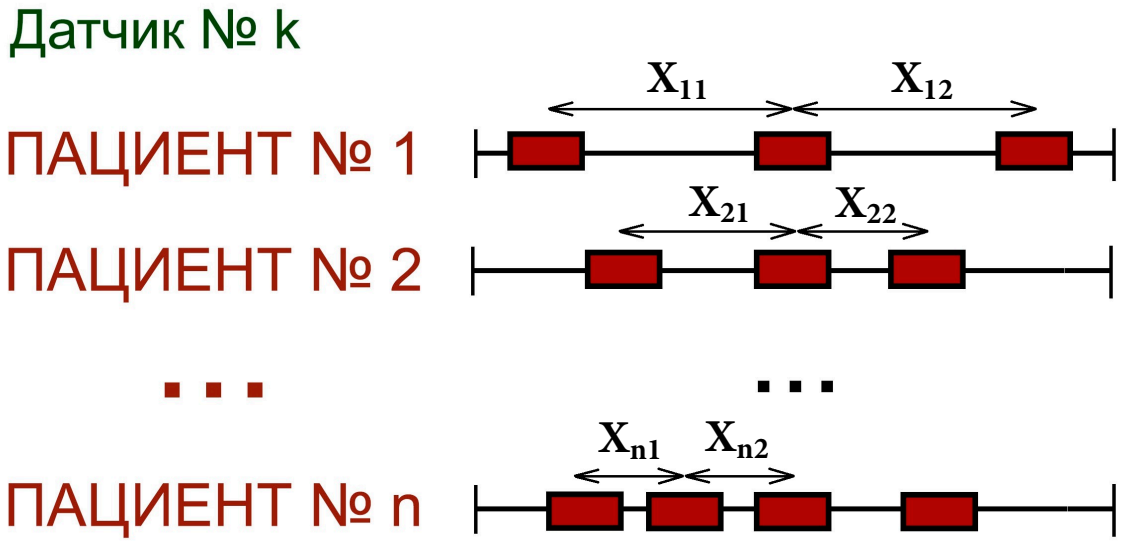


Рис.: Построение выборки встречаемости шаблона

- Датчик k $(k = 1, \ldots, K)$, $K = 16$.
- Фрагмент категориального ряда, встречаемость которого подчиняется негативно биномиальному распределению.



- Выборка $Y_i = \frac{1}{n} \sum_{j=1}^{n} X_{ji}$ усредненных по n индивидам интервалов между событиями, $i = 1, \ldots, m$, где $m = \min_j(N_j)$, $j = 1, \ldots, n$.

Модель НБР: $\xi \backsim NB(r,p)$,

$$P(\xi = k) = \frac{\Gamma(r+k)}{\Gamma(k+1)\Gamma(r)} p^r (1-p)^k,\ k = 0, 1, \ldots$$

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt,\ \ z \in \mathbb{C} :\ \mathrm{Re}(z) > 0.$$

Логарифм функции правдоподобия:

$$\log(L(r,p)) = (nr)\log(p) + (n\bar{X})\log(1-p) + \\ + \sum_{i=1}^{n} \left[\log(\Gamma(r+X_i)) - \log(\Gamma(r)) - \log(\Gamma(X_i+1))\right].$$

Оценки параметров p и r — решение системы:

$$\begin{cases} \frac{\partial log(L(r,p))}{\partial r} = 0, \\ \frac{\partial log(L(r,p))}{\partial p} = 0; \end{cases} \quad \begin{cases} \sum_{i=1}^{n} (\ln(p) + \phi(r+X_i) - \phi(r)) = 0, \\ p = \frac{r}{r+\bar{X}}; \end{cases}$$

где $\phi(x) = \ln'(\Gamma(x))$.

Модификация критерия согласия Колмогорова-Смирнова для дискретных распределений [Conover, 1972]

Обозначения: $S_n(x)$ – эмпирическая ф.р., $H(x)$ – теоретическая ф.р.

Статистики критерия:

$$D = \sup_x |H(x) - S_n(x)|, \qquad P(D \geq d) = P(D^+ \geq d) + P(D^- \geq d),$$

$$D^- = \sup_x (H(x) - S_n(x)), \qquad P(D^- \geq d^-) = \sum_{j=0}^{n(1-d^-)} \binom{n}{j} c_j^{n-j} b_j,$$

$$D^+ = \sup_x (S_n(x) - H(x));$$

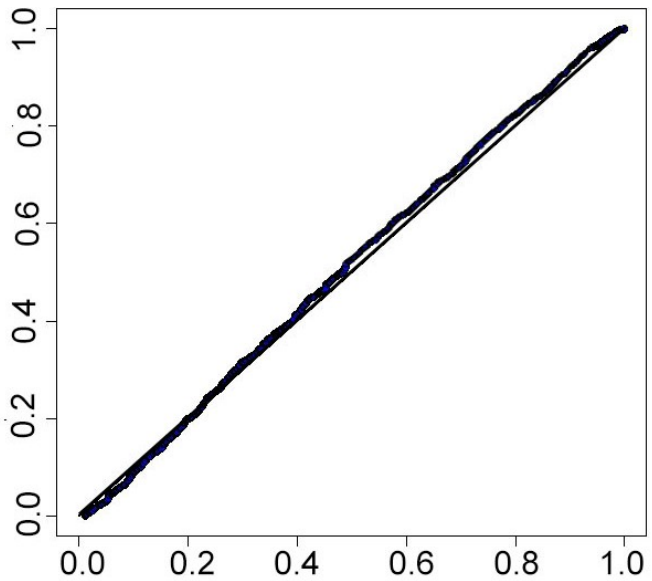
- $d,\ d^-$ — наблюдаемые значения $D,\ D^-$ соответственно;

$$b_k = 1 - \sum_{j=0}^{k-1} \binom{k}{j} c_j^{k-j} b_j, \quad k \geq 1, \quad \text{для всех } k : c_k > 0;$$

- На графике $H(x)$ строим прямую $f = d^- + j/n$.
- Если f пересекает $H(x)$ в скачке функции, то $c_k = 1 - h$, а если в основании скачка, то $c_k = 1 - y$;
- h – ордината вершины скачка; y – ордината основания.

Обозначения: $\mathcal{F} = F(\cdot, \theta)$ – параметрическое семейство ф. р., $H_0 : F_0 \in \mathcal{F}$, F_0 – распределение с параметром $\hat{\theta} = (\hat{k}, \hat{p})$.

1. По выборке $X_1, \ldots, X_n$ вычисляем оценку $\hat{\theta}$;
2. находим статистику критерия Колмогорова–Смирнова t и p_value;
3. генерируем выборку $X_1^*, \ldots, X_n^*$, имеющую распределение $F(\cdot, \hat{\theta})$;
4. по выборке $X_1^*, \ldots, X_n^*$ вычисляем оценку $\hat{\theta}^*$ оценки $\hat{\theta}$;
5. находим статистику критерия t^* и p_value;
6. N раз повторяем 3 – 5, упорядочиваем по возрастанию $t_1^* \leq \cdots \leq t_N^*$, берем $\lceil N(1-\alpha) \rceil$ по порядку элемент: $x_{1-\alpha}$;
7. гипотеза H_0 отвергается, если $t > x_{1-\alpha}$.



Эмпирическая ф. р. p_value, выборка получена с помощью bootstrap–метода

Проверка процедуры оценки параметров на модельных данных

- Промоделируем N раз целочисленный процесс до момента времени $T = 1270$.
- Получим выборку для с.в. τ — количество произошедших событий.
- Построим эмпирическую ф. р. τ и ф. р. $NB(r, p)$ с параметрами, оцененными по выборке процесса встречаемости шаблона.

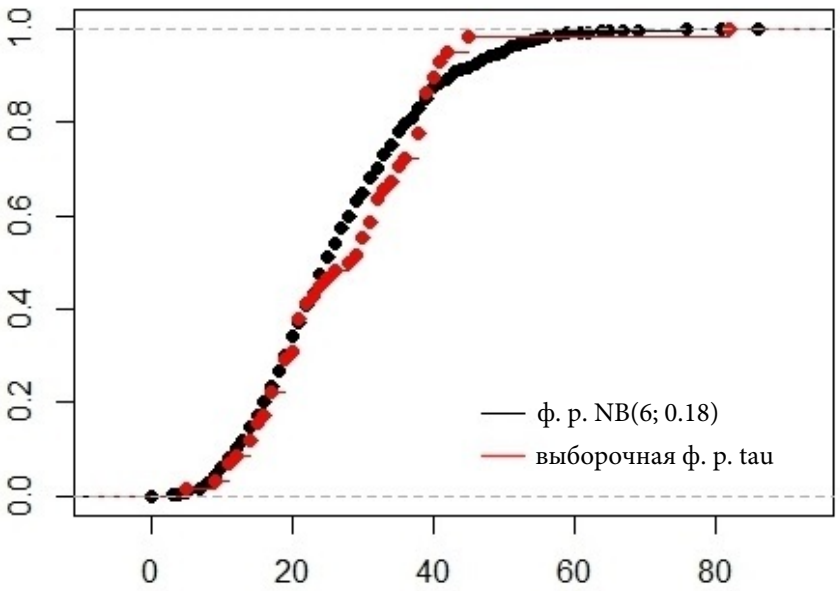


Рис.: Интервалы: $NB(2; 0.005)$

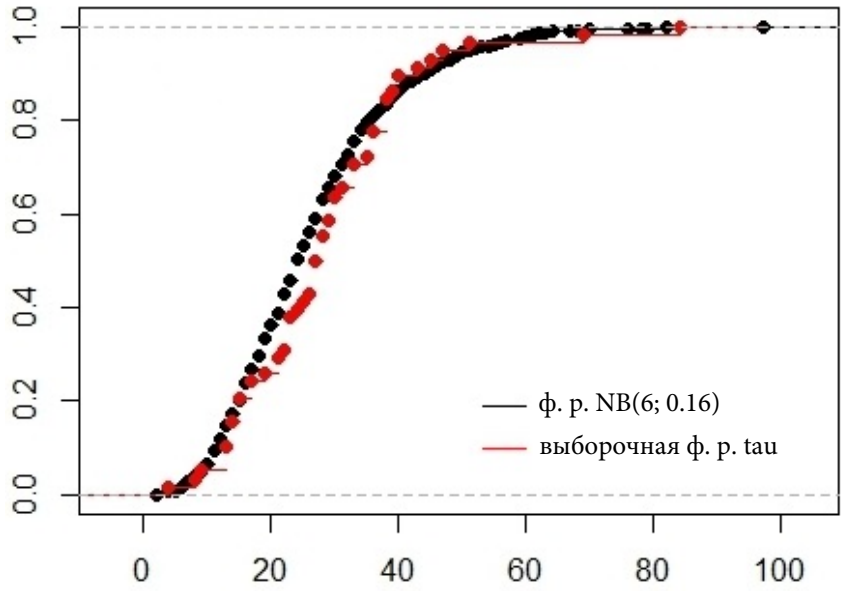


Рис.: Интервалы: $NB(2; 0.0042)$

Результаты анализа ЭЭГ

Результаты для шаблонов, наиболее согласуемых с НБР.

- $\xi \backsim NB(r, p)$

Таблица: Результаты для 12 фрагментов ЭЭГ

	встречаемость шаблона			**интервалы времени между событиями**		
Шаблон	*p*	*r*	*p_value*	*p*	*r*	*p_value*
uUud	0.13	4	0.6	0.004	1	0.65
uddu	0.09	3.7	0.8	0.004	0.8	0.67
udDd	0.18	6	0.5	0.005	2	0.45
Dduu	0.07	3.2	0.6	0.0046	0.79	0.53
duUu	0.16	5.6	0.65	0.0042	1.7	0.54
dDdu	0.17	6	0.68	0.0048	1.8	0.73
dddu	0.06	2.7	0.9	0.0047	0.77	0.52
uUuu	0.3	7	0.72	0.0053	1.6	0.84
dduU	0.08	2.9	0.67	0.0032	0.72	0.5
uudD	0.09	3.4	0.52	0.0045	0.87	0.53
uddd	0.077	3.1	0.86	0.0038	0.79	0.78
uddU	0.2	3.3	0.63	0.002	0.56	0.4

- В строках, выделенных желтым цветом, параметр r распределения интервалов примерно равен 2, в остальных — 1.

Две точки зрения на стационарные точечные процессы

1. Целочисленный процесс N_T — число событий в интервале $[0; T]$.
2. Последовательность с. в. $\{X_i\}$ — интервалы между событиями.

Обозначим $F_i(t) = P(X_1 + \cdots + X_i \le t),\ i = 0, 1, \ldots$

Утверждение [Cox, Lewis, 1966]

$$P(N_T = k) = F_k(T) - F_{k+1}(T)$$

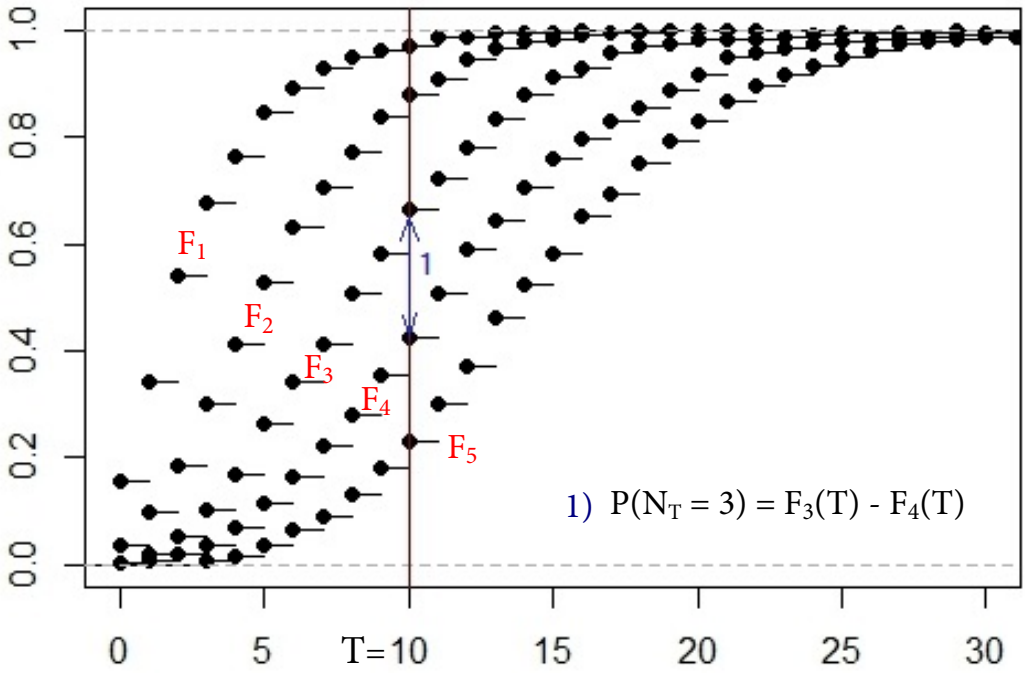


Рис.: Иллюстрация соотношения; распределение интервалов $NB(2; 0.4)$

- С. в. $\xi_1, \ldots, \xi_t, \ldots$ — интервалы времени между событиями.
- Требуется найти распределение целочисленного процесса N_T.

$$P\{N_T = k\} = P\left\{\max\{t : \sum_{j=1}^{t} \xi_j \le T\} = k\right\}$$

- С.в. ξ_j принимают только целые значения, тогда верно

Утверждение

$$P\{N_T = k\} = \sum_{i=0}^{\lfloor T \rfloor} P\left\{\sum_{j=1}^{k} \xi_j = i\right\} P\{\xi_{k+1} > T - i\}$$

- Если $\xi_j \sim Ber(p)$, то распределение процесса N_T является сдвинутым отрицательно–биномиальным: $N_T = \eta + \lfloor T \rfloor$, где $\eta \sim NB(\lfloor T \rfloor + 1, p)$.

- Интервалы $\xi_j \sim NB(r, p)$.

Утверждение

$$P\{N_T = k\} = p^{rk} q^{\lfloor T \rfloor + 1} \sum_{i=0}^{\lfloor T \rfloor} \frac{\Gamma(rk+i)}{\Gamma(i+1)\Gamma(rk)} \left(1 + \sum_{s=1}^{r-1} C^s_{\lfloor T \rfloor - i + s} \cdot p^s\right)$$

- Если $\xi_j \sim Geom(p)$, то распределение целочисленного процесса негативно–биномиальное $NB(\lfloor T \rfloor + 1, q)$.
- $\xi_j \sim NB(2, p)$, тогда выражение для вероятностей процесса N_T:

$$P\{N_T = k\} = p^{2k} q^{\lfloor T \rfloor + 1} \cdot \frac{\Gamma(2k + \lfloor T \rfloor + 1) \cdot (2k(p+1) + \lfloor T \rfloor p + p + 1)}{\Gamma(2k+2)\Gamma(\lfloor T \rfloor + 1)}$$

- $\xi_j \sim NB(2,p)$; τ — накопленное число событий в интервале $[0;T]$.
- Обозначим $\lfloor T \rfloor + 1 = l,\ l > 1$.

Моменты целочисленного распределения при межинтервальном $NB(2,p)$:

- $\mathsf{E}\tau = \dfrac{lp}{2(1-p)} + \dfrac{1}{4}\left(\dfrac{1-p}{1+p}\right)^l - \dfrac{1}{4}$;
- $\mathsf{D}\tau = \dfrac{1}{16} - \dfrac{1}{16}\left(\dfrac{1-p}{1+p}\right)^{2l} + \dfrac{pl}{2(1-p)^2}\left(\dfrac{1}{2} - pl - \left(\dfrac{1-p}{1+p}\right)^{l+1}\right)$.

Распределение случайной величины η *перерассеянное*, если $\mathsf{D}\eta > \mathsf{E}\eta$.

- $\eta \sim Bin(n,p) \Longrightarrow \mathsf{D}\eta < \mathsf{E}\eta$, распределение недорассеянное;
- $\eta \sim NB(r,p) \Longrightarrow \mathsf{D}\eta > \mathsf{E}\eta$, распределение перерассеянное.

Утверждение

Распределение τ перерассеянное при $p \in (0;a)$,

$$a = \frac{-2l - 5 + 2\sqrt{11l^2 - 5l}}{8l^2 - 8l - 5} < 1 \ \ \forall l.$$

Результаты анализа ЭЭГ. Аппроксимация негативным биномом

Таблица: Результаты анализа ЭЭГ. Распределение интервалов времени $NB(2, p)$

	встречаемость шаблона			**интервалы времени между событиями**		
Шаблон	*p*	*r*	*p_value*	*p*	*r*	*p_value*
udDd	0.18	6	0.5	0.005	2	0.45
duUu	0.16	5.6	0.65	0.0042	1.7	0.54
dDdu	0.17	6	0.68	0.0048	1.8	0.73
uUuu	0.3	7	0.72	0.0053	1.6	0.84

- $T = 1270 \Longrightarrow l = \lfloor T \rfloor + 1 = 1271.$
- Распределение числа событий перерассеянное, если $p \in \left(0; \dfrac{-2l - 5 + 2\sqrt{11l^2 - 5l}}{8l^2 - 8l - 5}\right) \approx (0; 0.007).$
- При анализе ЭЭГ получили значения параметра $p \in (0; 0.007)$ для распределения интервалов $\Longrightarrow$ распределение целочисленного процесса N_T перерассеянное.

Результаты

- Построены оценки максимального правдоподобия параметров НБР.
- Применены модификация критерия согласия Колмогорова-Смирнова для дискретных распределений и параметрический bootstrap-метод.
- Показано, что при бернуллиевских и геометрически распределенных интервалах имеет место НБР целочисленный процесс.
- Для $NB(2, p)$ интервалов найдены соотношения между параметрами, при которых целочисленное распределение перерассеянное.