Классификация категориальных рядов на основе модели негативно биномиального распределения с приложением в неврологии

Комарова Елена Сергеевна, гр. 522

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

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

Санкт-Петербург
2014г.

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.

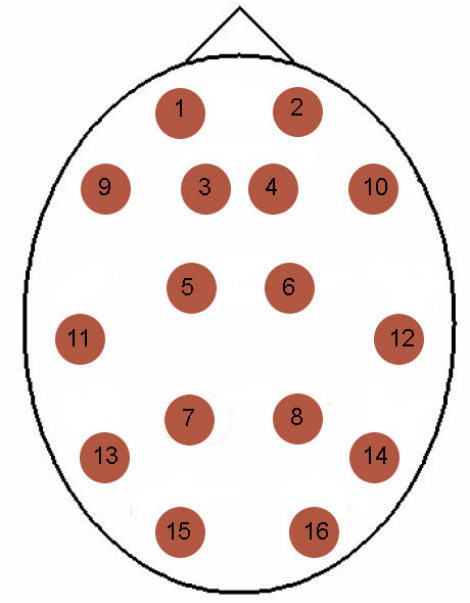


Рис. 1: Расположение датчиков на голове

Рис. 2: Пример отображения отклонений на ЭЭГ

Цель работы и задачи

Цель работы: разработка методов классификации заболеваний по ЭЭГ и идентификации квантов электрофизиологической активности мозга.

Задачи:

1. методы классификации (LDA, QDA, RDA);
2. два способа описания ЭЭГ: амплитудно–частотный и лингвистический;
3. оценка параметров модели негативно-биномиального распределения встречаемости отдельных участков ЭЭГ;
4. критерий согласия с негативно-биномиальным распределением в случае оценки параметров по выборке;
5. зависимость оценок параметров негативно-биномиального распределения от длины ЭЭГ.

Методы классификации популяции

По G группам распределены n человек, π_g — доля группы g, при этом $\sum\limits_{g=1}^{G} \pi_g = 1$. Группа g имеет распределение $\mathrm{N}(\mu_g, \Sigma_g)$, $g = 1, \ldots, G$.
Пусть $x_{1,1}, \ldots, x_{1,n_1}$ из группы 1, $x_{2,1}, \ldots, x_{2,n_2}$ из группы 2 и т. д., где $n_1 + \ldots + n_G = n$. Тогда

$$x_{g,i} \in \left(\tilde{g} = \underset{g'}{\operatorname{argmin}} \left[\frac{1}{2} (x_{g,i} - \hat{\mu}_{g'})^T \hat{\Sigma}_{g'}^{-1} (x_{g,i} - \hat{\mu}_{g'}) - \ln(\pi_{g'}) \right] \right),$$

где $\hat{\mu}_{g'} = \bar{x}_{g'} = \dfrac{1}{n_{g'}} \sum\limits_{i=1}^{n_{g'}} x_{g',i}$, а оценка $\Sigma_{g'}$ зависит от метода.

1. **Линейный дискриминантный анализ (LDA)**:
   $\hat{\Sigma}_{g'} = \hat{\Sigma} = \sum\limits_{g=1}^{G} \pi_g S_g$, где $S_g = \frac{1}{n_g} \sum\limits_{i=1}^{n_g} (x_{g,i} - \hat{\mu}_g)(x_{g,i} - \hat{\mu}_g)^T$.
2. **Квадратичный дискриминантный анализ (QDA)**:
   $\hat{\Sigma}_{g'} = S_{g'}$.
3. **Регуляризованный дискриминантный анализ (RDA)** [Guo Y., Hastie T., Tibshirani R., 2007]:
   $\hat{\Sigma}_{g'} = R_{g'} = (1 - \gamma)\hat{\Sigma}(\lambda) + \gamma d I$, $\hat{\Sigma}(\lambda) = (1 - \lambda)\hat{\Sigma}_{g'} + \lambda \hat{\Sigma}$,
   где $0 \leqslant \gamma, \lambda \leqslant 1$, d — среднее диагональных элементов $\hat{\Sigma}(\lambda)$.

Амплитудно–частотный метод описания ЭЭГ

Пик определяется следующим образом: предыдущая и следующая точки должны быть меньше текущей ($\max$) или, наоборот, предыдущая и следующая точки — больше текущей ($\min$), тогда

- **Частота1** $= 254/($количество чисел между соседними $\max + 1)$;
- **Частота2** $= 254/($количество чисел между соседними $\min + 1)$;
- **Амплитуда1** $=$ разность между значениями $\max$ и последующего $\min$;
- **Амплитуда2** $=$ разность между значениями $\min$ и последующего $\max$.

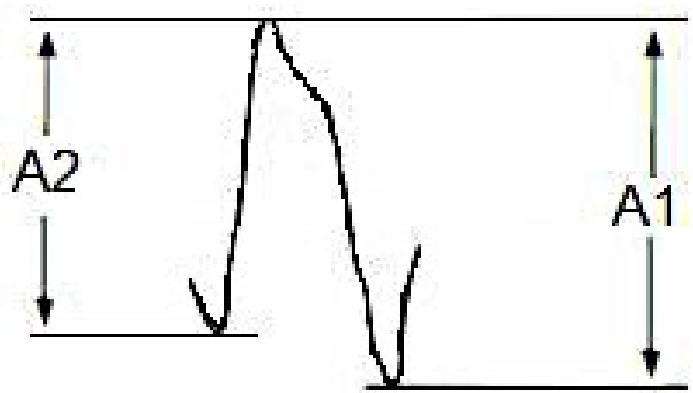


Рис. 3: Амплитуда1 (A1) и Амплитуда2 (A2)

Результаты дисперсионного анализа для зависимых выборок (ANOVA Repeated Measures)

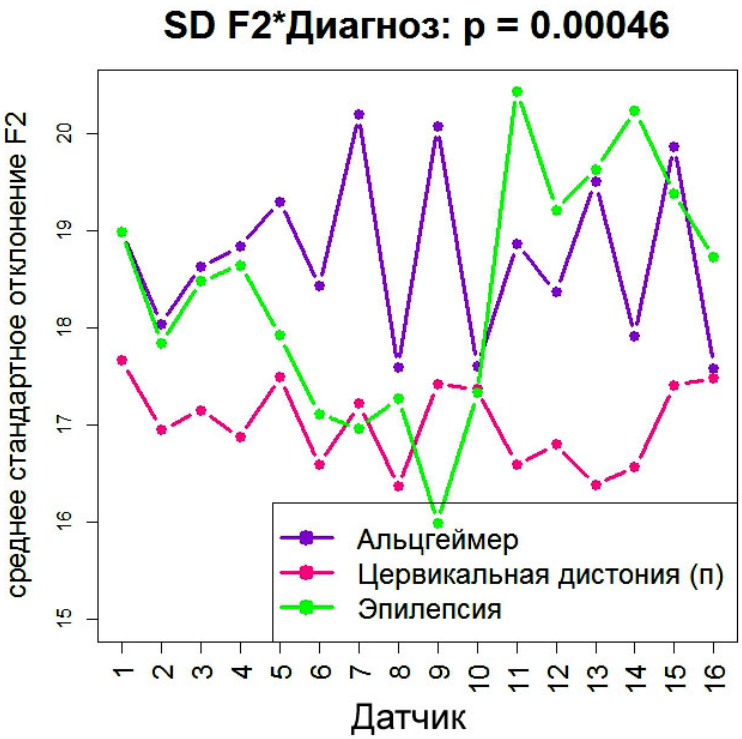


Рис. 4: Стандартное отклонение частоты1 для всех датчиков с учетом заболеваний

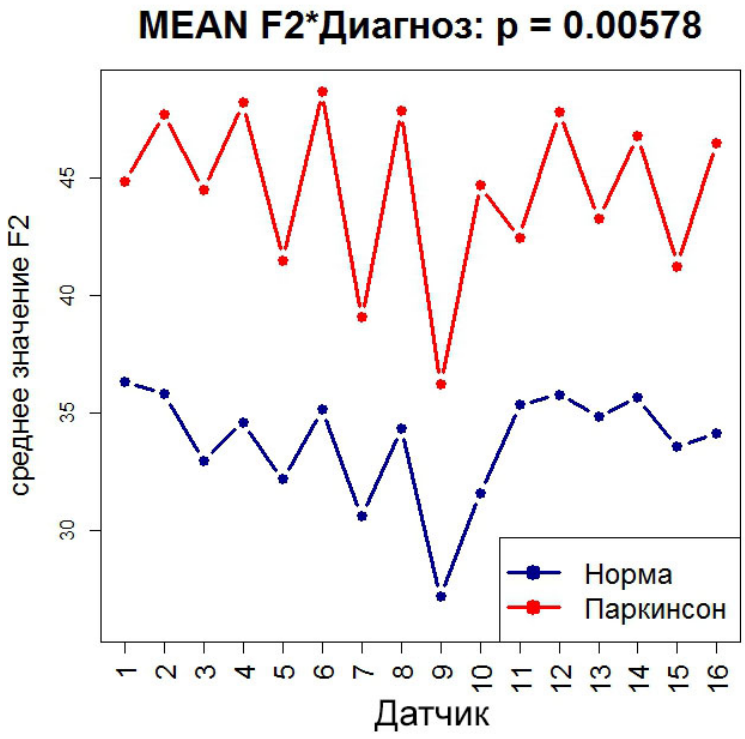


Рис. 5: Средние значения частоты2 для всех датчиков с учетом заболеваний

Результаты классификации на основе амплитуд и частот

Индивиды: $n_1 = 20$ (здоровые), $n_2 = 24$ (болезнь Паркинсона);
Признаки: стандартное отклонение, среднее частоты2 и стандартное отклонение амплитуды2 всех датчиков (всего 48 признаков);
Метод: RDA с параметрами $\gamma = 0.05$, $\lambda = 0.2$.

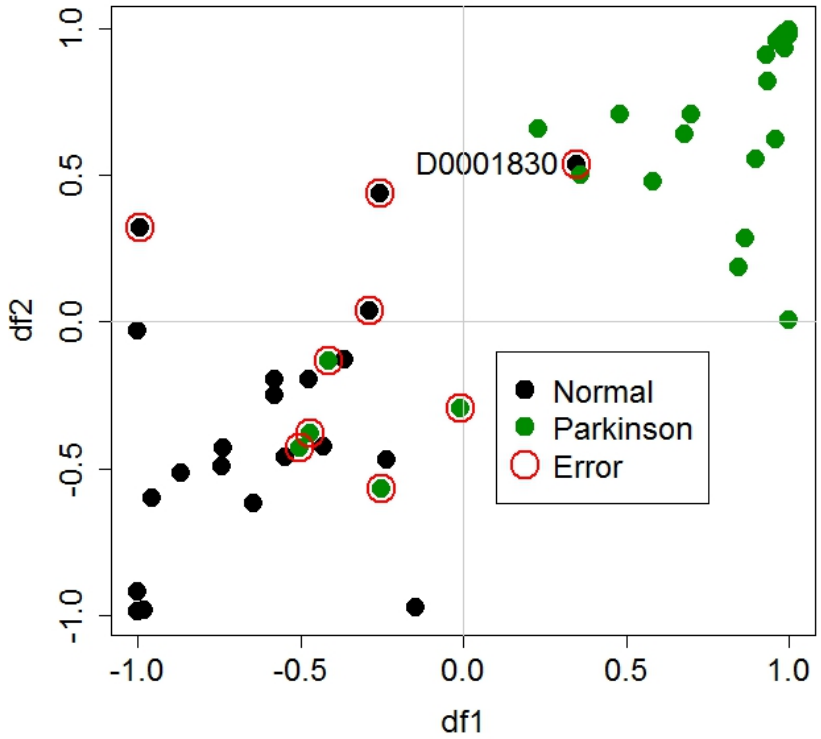
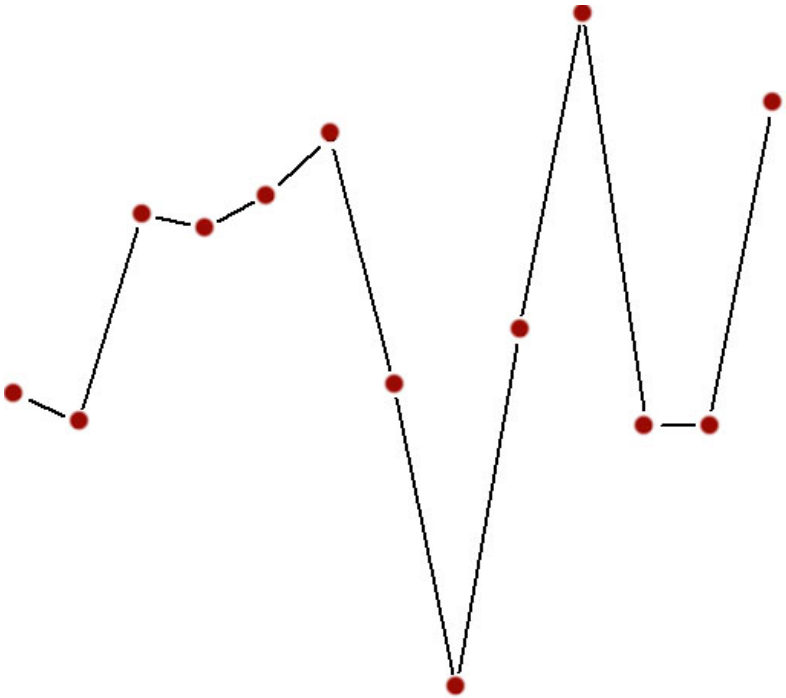


Рис. 6: Двухмерная диаграмма апостериорных вероятностей отнесения к группе больных, построенных по первым 24 признакам (df1) и по 24 остальным (df2)

Представление ЭЭГ в виде категориального ряда

Обозначение: a — средняя по модулю величину скачка последовательности ЭЭГ, тогда

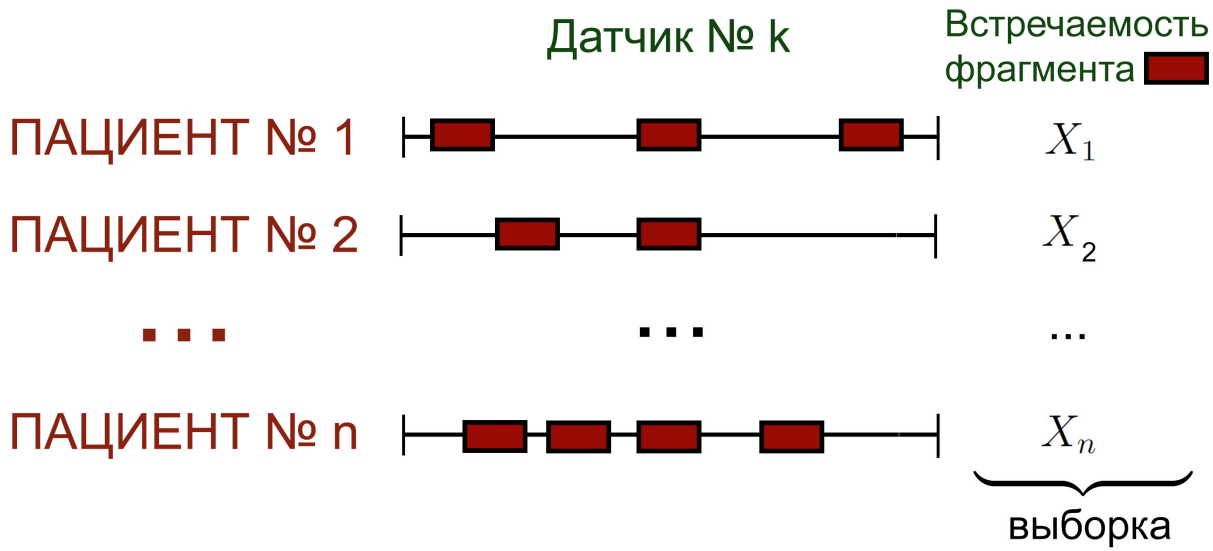
- если величина скачка от 0 до a, то приращение заменяем на “up” (сокращенно “u”);
- если от $(-a)$ до 0, то на “down” (“d”);
- если $> a$, то на “UP” (“U”);
- если $< (-a)$, то на “DOWN” (“D”).



“d U d u u D D U U D d U”

Построение выборки по данным ЭЭГ

Выбираем пациента i $(i = 1, \ldots, n = 58)$.
Фиксируем датчик k $(k = 1, \ldots, 16)$.
Переводим числовую последовательность в категориальный ряд.



В качестве фрагментов берем всевозможные последовательности длины 5, составленные из алфавита {“u”, “d”, “U”, “D”}.

Модель негативно-биномиального распределения (НБР)

Модель НБР: $Y \backsim \mathrm{NB}(p, k)$

$$\mathrm{P}(Y = j) = \frac{\Gamma(k+j)}{\Gamma(j+1)\Gamma(k)} p^k (1-p)^j,\ j = 0, 1, \ldots$$

Гамма-пуассоновская схема:

$$\int_0^\infty P(j|\lambda(1-p))\gamma(\lambda p|k)d(\lambda p) = C_{k+j-1}^{k-1} p^k (1-p)^j.$$

Интерпретация параметров НБР в паразитологии [Барт, 2003]:

- Y — число выживших личинок;
- p — вероятность гибели личинки;
- k — число погибших личинок.

Интерпретация параметров НБР в лингвистике [Alexeyeva, Sotov, 2013]:

- Y — число словоупотреблений в тексте;
- p — вероятность потери слова;
- k — число потерянных слов.

Способы оценивания параметров НБР

- **Метод моментов** по выборке $X_1, ..., X_n$:

$$\widehat{p} = \frac{\overline{X}}{S^2}, \quad \widehat{k} = \frac{\overline{X}^2}{S^2 - \overline{X}},$$

где $\overline{X}$ — выборочное среднее, а S^2 — выборочная дисперсия.

- **Оценка максимального правдоподобия** для параметров p и k — это решение системы:

$$\begin{cases} \sum\limits_{i=1}^{n} \Big(\ln(\hat{p}) + \psi(\hat{k} + X_i) - \psi(\hat{k}) \Big) = 0, \\ \hat{p} = \frac{\hat{k}}{\hat{k} + \bar{X}}, \end{cases}$$

где $\psi(x) = \ln'(\Gamma(x))$.

Критерий Крамера-фон Мизеса и его применимость

Обозначения: $\xi \backsim \mathrm{NB}(p, k)$ [Choulakian, Lockhart, Stephens, 1994]

- N — объем выборки, $p_i = \mathsf{P}(\xi = i)$, s_i — эмпирическая частота выпадения числа i, $Np_i = t_i$ — теоретическая частота выпадения числа i;
- $S_j = \sum\limits_{i=0}^{j} s_i$, $T_j = \sum\limits_{i=0}^{j} t_i$, $H_j = T_j/N$, $\bar{Z} = \sum\limits_{j=1}^{k} Z_j p_j$, $Z_j = S_j - T_j$.

Статистики критерия
для дискретного распределения

$$W^2 = N^{-1} \sum_{j=0}^{k} Z_j^2 p_j,$$

$$U^2 = N^{-1} \sum_{j=0}^{k} (Z_j^2 - \bar{Z})^2 p_j,$$

$$A^2 = N^{-1} \sum_{j=0}^{k} \frac{Z_j^2 p_j}{H_j(1 - H_j)}.$$

Применимость

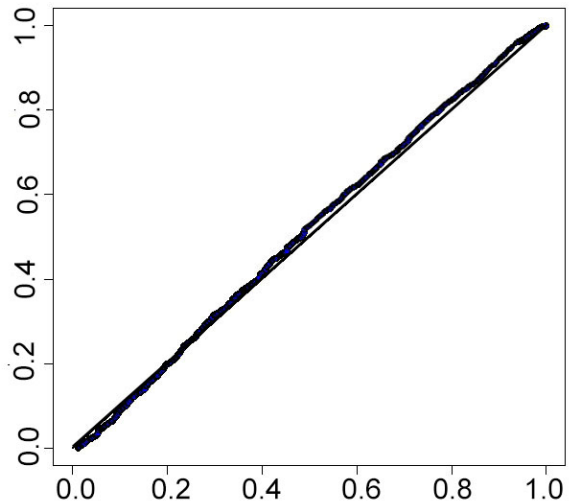


Рис. 7: Эмпирическая функция распределения $p - value$, выборка получена с помощью процедуры bootstrap

Результаты классификации на основе встречаемости “слов”

Индивиды: $n_1 = 20$ (здоровые), $n_2 = 24$ (болезнь Паркинсона);
Признаки: встречаемость “слов” (всего их 97);
Метод: RDA с параметрами $\gamma = 0.05$, $\lambda = 0.2$.

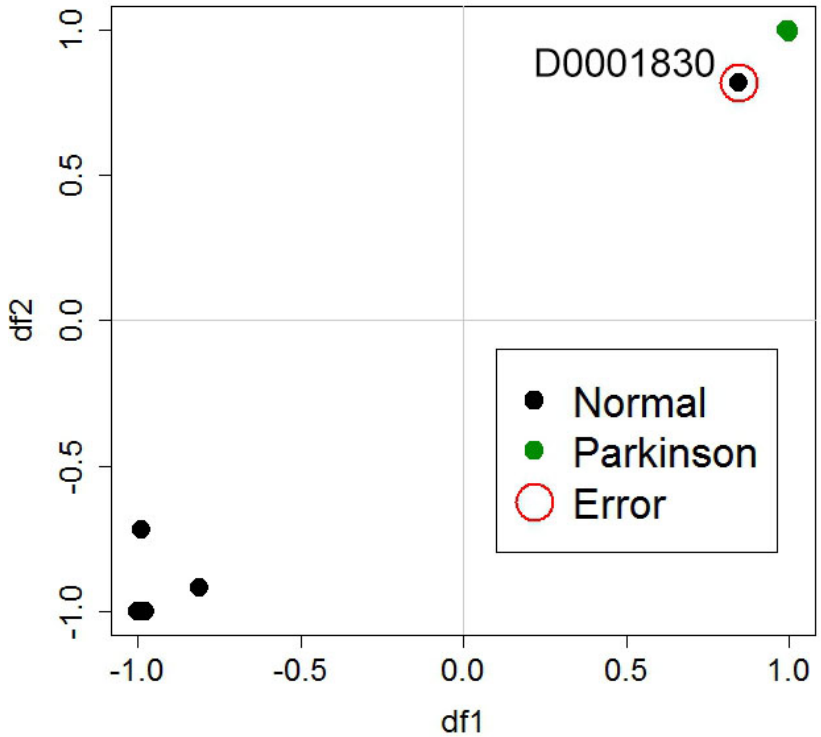


Рис. 8: Двухмерная диаграмма апостериорных вероятностей отнесения к группе больных, построенных по первым 48 признакам (df1) и по 49 остальным (df2)

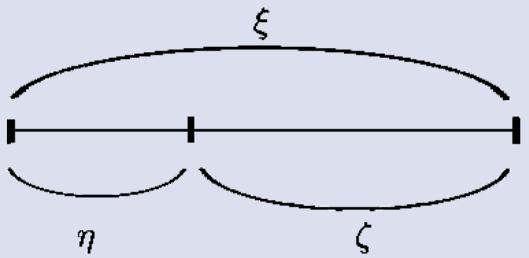
Проблема идентификации кванта э/ф активности мозга

Утверждение

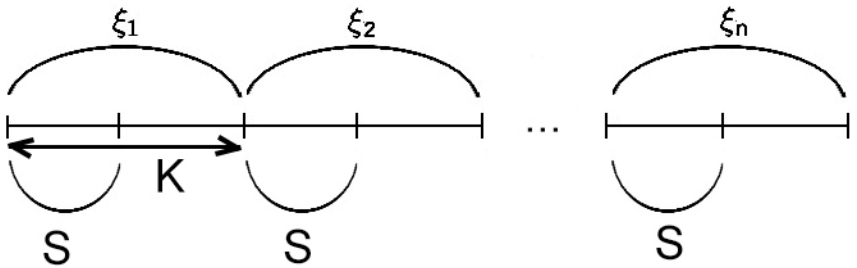
Пусть $\xi \sim \mathrm{NB}(p, r)$ *и* $\xi = \eta + \zeta$, *где* $\eta = \sum_{j=1}^{\xi} \tau_j(\tilde{p})$,

$$\tau_j(\tilde{p}) = \begin{cases} 1 \text{ с вероятностью } \tilde{p}, \\ 0 \text{ с вероятностью } 1 - \tilde{p}, \end{cases}$$

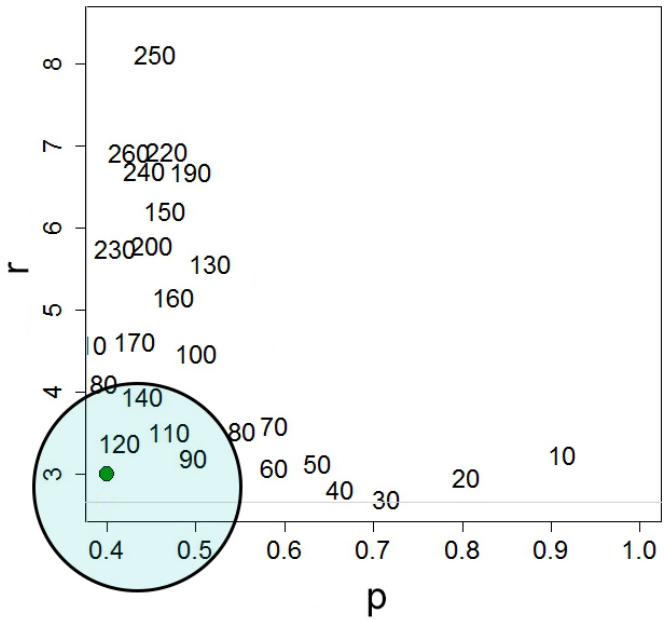
Тогда $\eta \sim \mathrm{NB}\left(\frac{p}{1-(1-p)(1-\tilde{p})}, r\right)$.



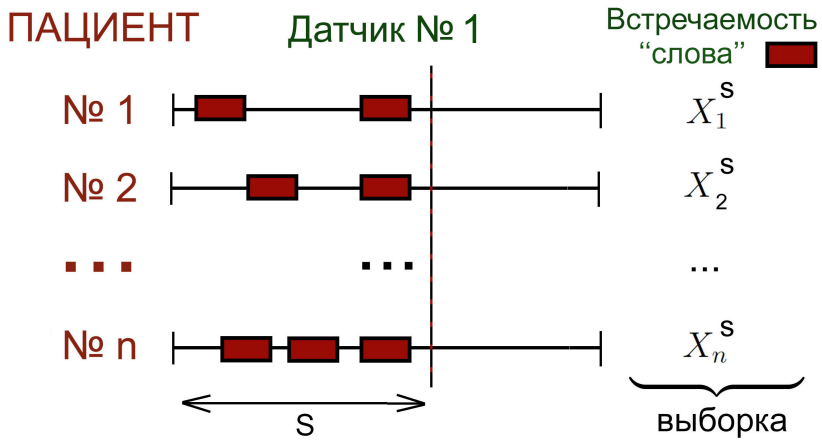
Пусть $\xi_1, \ldots, \xi_n \sim \mathrm{NB}(p, r)$,
$p = 0.4$, $r = 3$,
длина кванта $K = 120$.



Меняем $S = 10 : 260$ с шагом $m = 10$.



Зависимость оценок параметров НБР от длины ЭЭГ



Оценки НБР в зависимости от длины ЭЭГ

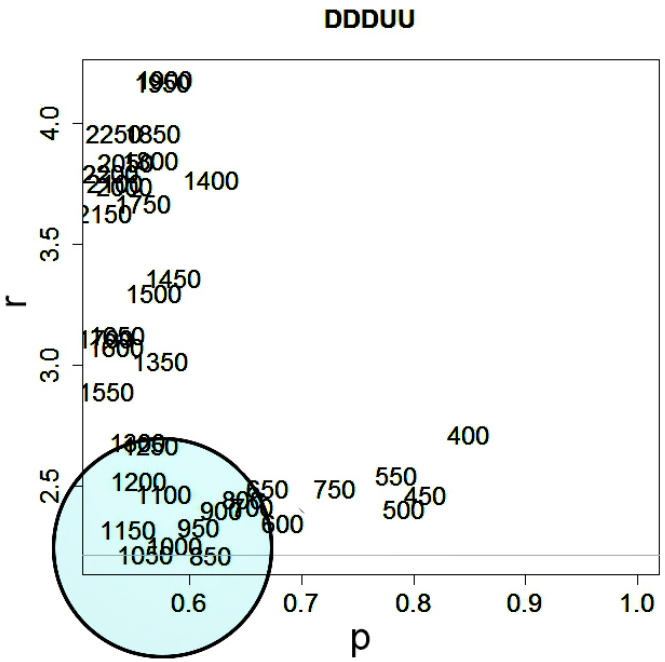


Рис. 9: $S = 400 : 2250$, шаг $m = 50$

Распределение свертки и смеси двух НБР с разными параметрами

Утверждение (Распределение свертки двух НБР)

Пусть $\eta \sim \mathrm{NB}(p, r)$, $\eta = \eta_1 + \eta_2$, *где* $\eta_1 \sim \mathrm{NB}(p_1, r_1)$ *и* $\eta_2 \sim \mathrm{NB}(p_2, r_2)$, η_1 *и* η_2 *— независимые.*
Тогда $\mathrm{NB}(prob, size)$ *— аппроксимирующее распределение для распределения* η, *где* $prob = \frac{A}{B}$, $size = \frac{A^2}{B-A}$,

$$A = \frac{r_1(1-p_1)}{p_1} + \frac{r_2(1-p_2)}{p_2},\ B = \frac{r_1(1-p_1)}{p_1^2} + \frac{r_2(1-p_2)}{p_2^2}.$$

Утверждение (Распределение смеси двух НБР)

Пусть ξ *— смесь двух случайных величин* $\xi_1 \sim \mathrm{NB}(p_1, r_1)$ *и* $\xi_2 \sim \mathrm{NB}(p_2, r_2)$ *с весами* α *и* $1-\alpha$; $A = \mathrm{E}\,\xi_1 = \frac{r_1(1-p_1)}{p_1}$, $B = \mathrm{E}\,\xi_2 = \frac{r_2(1-p_2)}{p_2}$.
Тогда $\mathrm{NB}(prob, size)$ *— аппроксимирующее распределение для распределения* ξ, *где*

$$prob = \frac{\alpha A + (1-\alpha)B}{\frac{\alpha A}{p_1} + \frac{(1-\alpha)B}{p_2} + \alpha(1-\alpha)(A-B)^2},\ size = \frac{p}{1-p}(\alpha A + (1-\alpha)B).$$

Результаты

1. идентификация амплитуд и частот с классификацией заболеваний (ошибка классификации составляет 20%);
2. метод преобразования ЭЭГ в категориальный ряд;
3. выделение наиболее информативных для классификации заболеваний фрагментов в соответствии с моделью НБР (ошибка классификации составляет 2%);
4. проверка адекватности модели НБР с ОМП и проверкой согласия по bootstrap процедуре модифицированного критерия Крамера фон Мизеса для дискретных распределений;
5. соотношения между параметрами суммы и смеси НБР с аппроксимацией НБР;
6. подтверждение гипотезы о квантованности электрофизиологической активности мозга.