Исследование распределения k-меров в геноме

Елена Николаевна Картышева, гр. 15.Б04-мм

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

Научный руководитель: к. ф.-м. н., доцент Н. П. Алексеева
Рецензент: биостатистик «Парексель», Е. С. Комарова

2019 г.

Отрицательное биномиальное распределение

- Гамма-пуассовская модель [Bart, 2003]

$$P(\xi = j) = \int_0^\infty P(j|\lambda q)\gamma(\lambda p|r,1)d(\lambda p) = \beta_-(j|r,p),$$

где $P(j|\lambda q) = \frac{\lambda^j}{j!}q^j e^{-\lambda q}$, $\gamma(\lambda p|r,1) = \frac{\lambda^{r-1}}{\Gamma(r)}p^{r-1}e^{-\lambda p}$.

- Модель отрицательного биномиального распределения (ОБР) в лингвистике [Alexeyeva, Sotov, 2013]
- Структура данных: текст, разделенный на N глав.
- $\xi \sim NB(p,r)$ — количество вхождений слова A в главу.
  - p – вероятность неупотребления слова, выражает намерение рассказчика.
  - r – число потерянных слов, можно рассматривать как меру «обычности» слова.

Пример на тексте

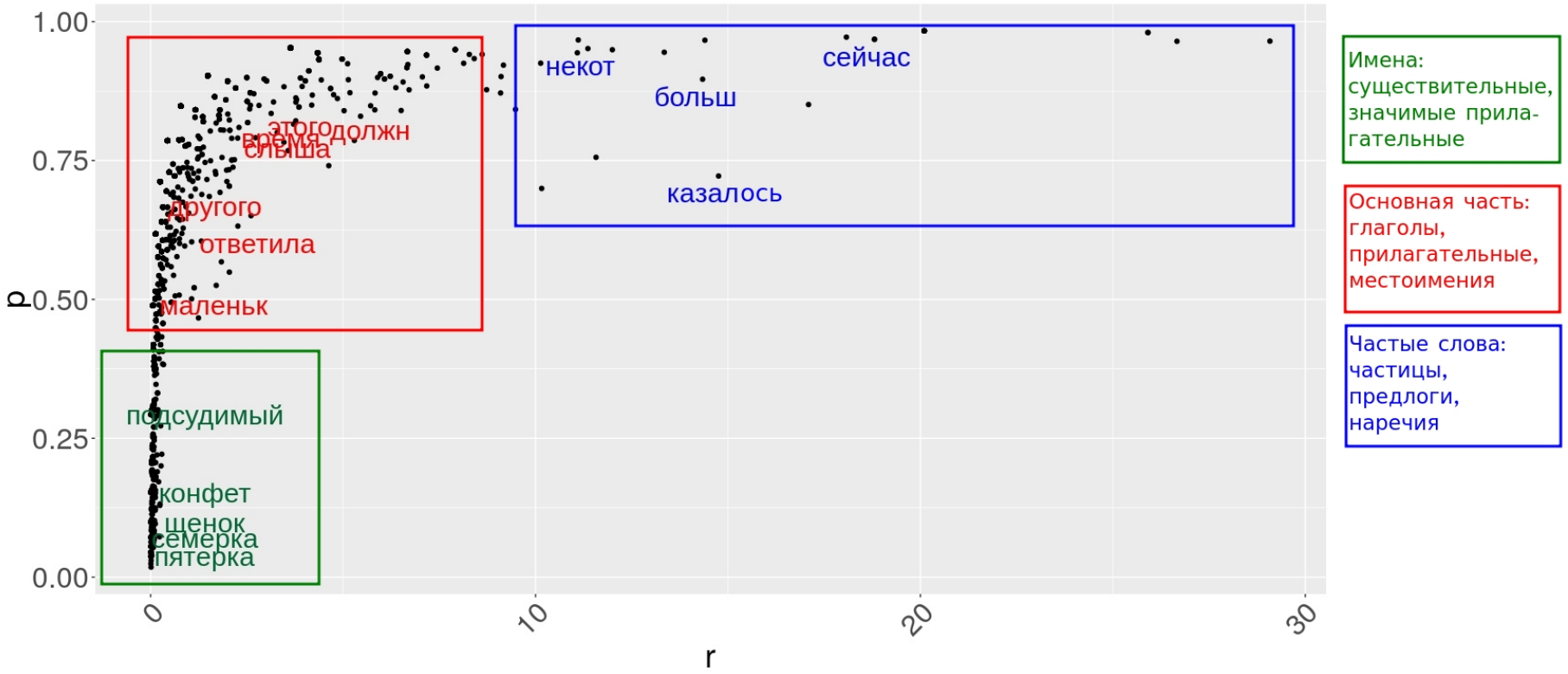


Рис. 1: Двумерная диаграмма оценок параметров для текста «Алисы в Стране Чудес»

Постановка задачи

Основные термины

- **Геном** — строка над четырёхбуквенным алфавитом $\{A, T, G, C\}$.
- k-**мер** — подстрока генома длины k.

Идея

Геном можно рассматривать как текст, искусственно разделенный на N глав, и вычислить встречаемость k-меров по главам.

Задачи

- Проверка согласия встречаемости k-меров с ОБР.
- Применение оценок параметров словоупотребления для дифференциации геномов.

Исследуемые критерии согласия

Проверяем гипотезу $H_0 : F_0 \in NB$, где F_0 — эмпирическая функция распределения.

Для исследования модели применим следующие критерии:

1. Колмогорова–Смирнова для дискретных распределений [Arnold, Emerson, 2011].
2. Хи-квадрат.
3. Крамера–фон Мизеса для дискретных распределений.

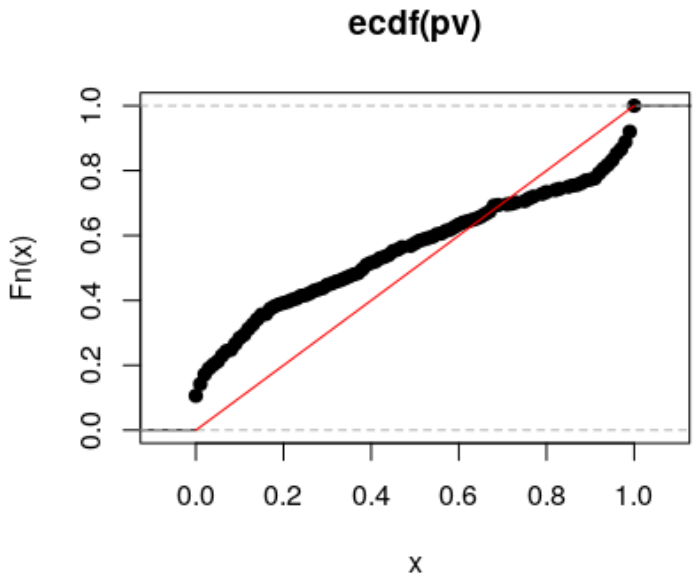
- Известно, что если H_0 верна, и тест применимый, то *p-value* имеет равномерное распределение на $[0, 1]$.
- Критерии Колмогорова–Смирнова и Крамера–фон Мизеса предполагают проверку простой гипотезы.
- Поскольку вместо точных значений параметров используются оценки параметров, необходимо исправить p-value с помощью бутстрепа.

Бутстреп

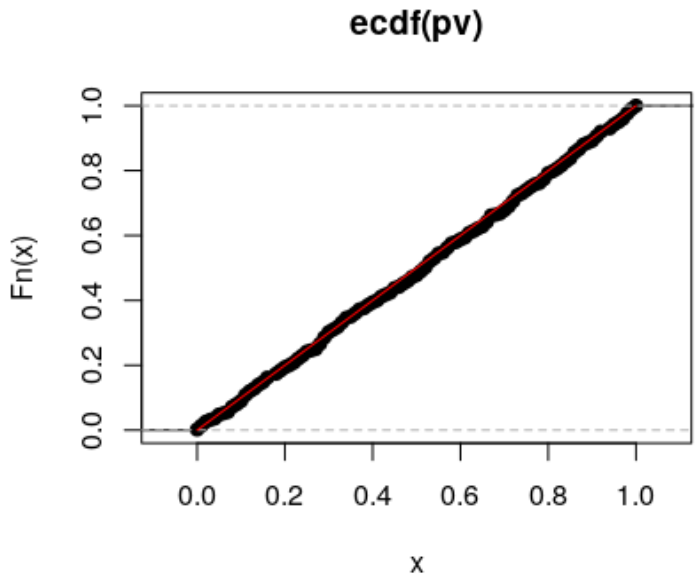
Бутстреп — компьютерный метод исследования распределения статистик вероятностных распределений, основанный на множественной генерации выборок на основе имеющейся выборки.

1. По выборке X вычисляется оценка $\hat{\theta}$.
2. Вычисляется статистика критерия t.
3. Генерируется выборка X^*, имеющая распределение $NB(\hat{\theta})$.
4. По выборке X^* вычисляется оценка $\hat{\theta^*}$ и статистика t^* .
5. Шаги 3–4 повторяем N раз, полученные статистики упорядочиваем по возрастанию $t_1^* \leq \ldots \leq t_N^*$ и берем выборочную квантиль уровня $1 - \alpha$ (обозначим $T_{1-\alpha}$).
6. Гипотеза H_0 отвергается, если $t > T_{1-\alpha}$.

Применимость критериев



(a) Тест Колмогорова–Смирнова



(b) Тест Крамера–фон Мизеса

Рис. 2: Распределение p-value для критериев, предназначенных для дискретных распределений, с применением бутстрепа.

Результаты на тестовых данных

Данные: геномы дрожжей S.cerevisiae и S.paradoxus.

Таблица 1: Процент k-меров, распознанных как слова для различных значений k

k	*S. cerevisiae*	*S. paradoxus*	Alice in Wonderland
5	2.8%	4.2%	7.6%
6	19.9%	19.1%	5.4%
7	36%	34.7%	3.6%
8	36.5%	34%	2.6%
9	18.5%	16.3%	1.9%

Задача сравнения геномов

Цель: Сравнить M геномов на сходства и различия.

- Геномы сопоставляются выравниванием.

Выравнивание — размещение двух или более (в таком случае выравнивание будет множественным) последовательностей друг над другом таким образом, чтобы легко было увидеть сходные участки в этих последовательностях.

```
EEELTKPRLLWALYFNMRDALSSG-
---VEKPRILYALYFNMRD--SSDE
```

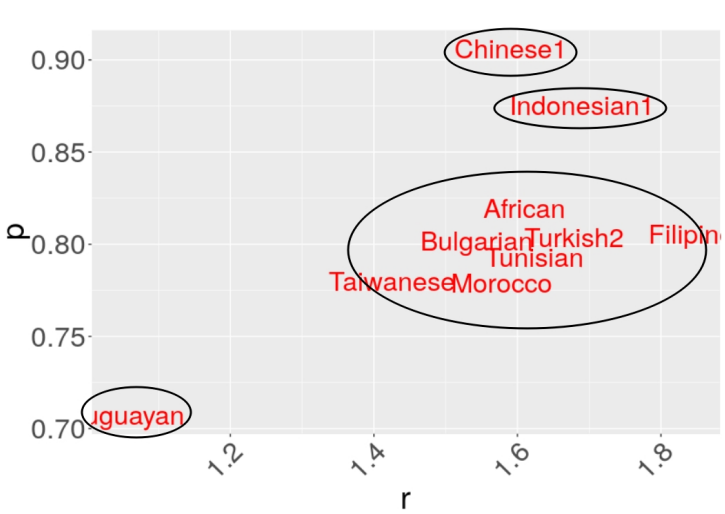
Рис. 3: Пример выравнивания

- Расстояния между наблюдениями вычисляются по метрике Кимура [Kimura, 1980] (чем больше несовпадений в выравнивании, тем дальше последовательности).
- Строится матрица попарных расстояний.

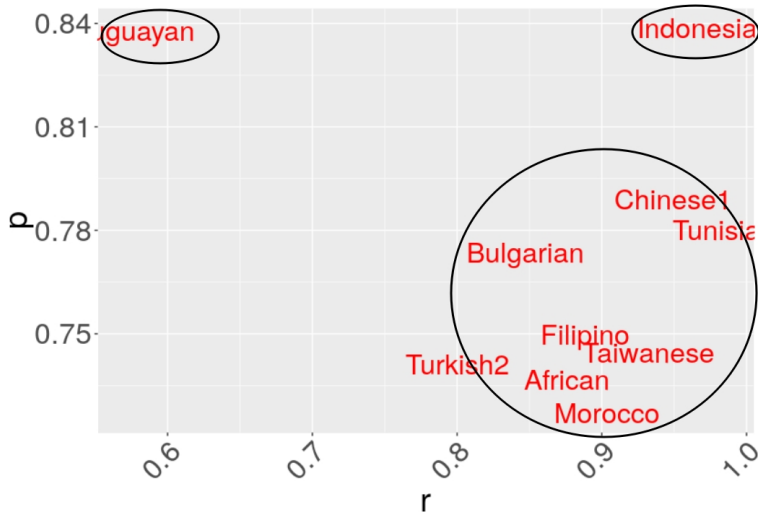
ОБР для сравнений наблюдений генома

Данные: 45 геномов людей различных национальностей.

- Для каждого наблюдения выделили k-меры, подчиняющиеся ОБР (брали $k = 5$).
- Оставили слова, общие для всех геномов (всего слов выделилось 74), обозначим это множество за $\mathcal{L}$.



(a) Для AAGAC



(b) Для TTCCC

Рис. 4: Двумерные диаграммы оценок параметров

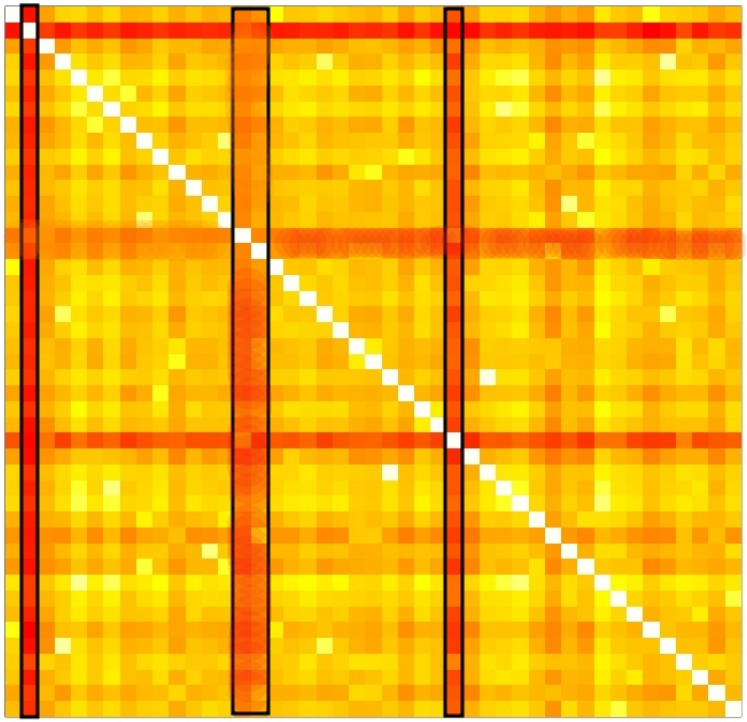
Матрица расстояний по оценкам параметров 5-меров

- Каждое наблюдение представлено вектором $x = (p_1, r_1, \ldots, p_s, r_s)$, где p и r — параметры слов.
- Метрика Канберра: $dist(x, y) = \sum_i \frac{|x_i - y_i|}{|x_i| + |y_i|}$ [Hill-Burns, 2017].
- Вычислим матрицу попарных расстояний по метрике Канберра.

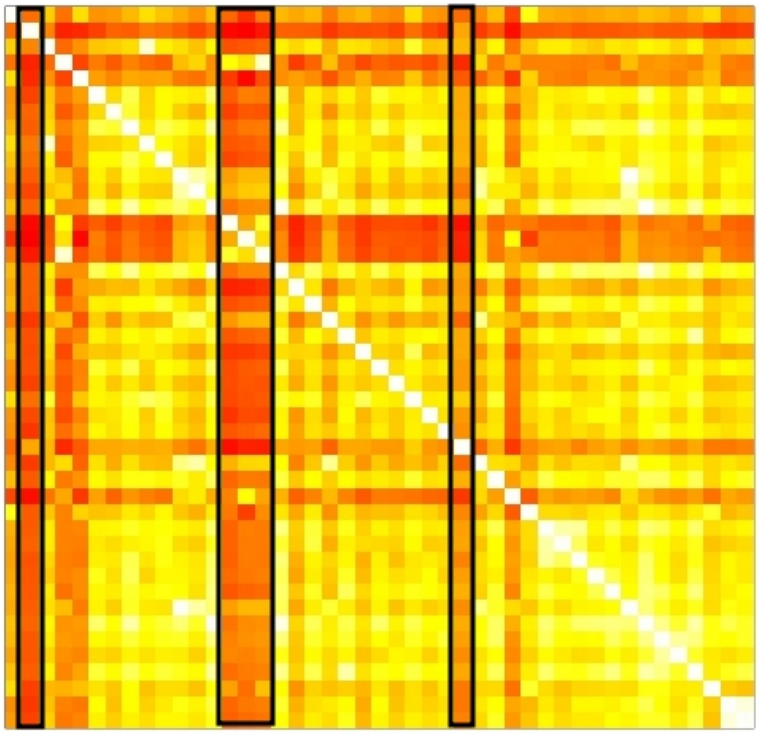
Визуализация

Heatmap — нормированная матрица попарных расстояний, где числа представлены цветом. Белый цвет соответствует нулю, чем краснее, тем значение данного элемента матрицы больше.

Результаты сравнения сходства геномов



(a) Матрица попарных расстояний, построенная с помощью выравнивания



(b) Матрица попарных расстояний, построенная по метрике Канберра

Видно сходство: наблюдения-выбросы в матрице (a) (столбцы красного цвета) являются выбросами и в матрице (b).

Проверка частоты ошибок в словах

- Выберем пару наблюдений.
- За ошибку в выравнивании будем считать несовпадение символов или пропуск символа в одной из последовательностей. Проверим все такие места.

Рис. 5: Пример ошибки в слове AGTCT

- Для каждой ошибки проверим, произошла ли она в слове из множества $\mathcal{L}$.
- Считаем долю ошибок, которые пришлись на слова:

$$\frac{\text{количество ошибок в словах из } \mathcal{L}}{\text{количество всех ошибок}}.$$

В результате, доля ошибок в словах

- для пары далеких друг от друга наблюдений — 60%.
- для пары близких — 18%.

Результаты

- При помощи моделирования показано, что для проверки гипотез согласия критерии Крамера–фон Мизеса и хи-квадрат более предпочтительны, чем модификация критерия Колмогорова–Смирнова для дискретных распределений.
- Создано программное обеспечение для выявления k-меров в геномах, подчиняющихся ОБР.
- Показано, что матрица расстояний по оценкам параметров словоупотребления 5-меров лишь частично соответствует матрице расстояний по метрике Кимура, однако для некоторых геномов удалось выявить 5-меры, параметры словоупотребления которых существенно отличают эти геномы от остальных.