Варианты обобщения геометрического распределения

Есипенко Евгений Вячеславович, 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., мл. науч. сотр. Ананьевская П.В.

Санкт-Петербург
2014г.

Цель работы

Анализ четырех моделей обобщения геометрического распределения, констукция которых основана на операции частичного обращения по Барту А. Г.

- Исследование обобщенного геометрического распределения $\beta_{\alpha}^{-}(p)$ [Барт, 1987], оценка параметров, свойства оценок, свойства распределения.
- Вывод законов обобщенных распределений с различной формой параметризации частично обратных к реализациям испытаний Бернулли, оценка параметров.
- Сравнение моделей при $\alpha = \frac{2}{2k+1}$, $k \in N$ с приложением в лингвистике и использованием критерия Cramer-von Mises(CVM) и процедуры bootstrap.

Обобщение геометрического распределения (GGD) по Барту А. Г.

Крайние частично обратные функции к числу успехов $\xi(n)$ из n испытаний Бернулли с вероятностью успеха p:

$$\xi_0^-(k) = \min\{n : \xi(n) \geqslant k\} \quad \text{левая},$$
$$\xi_1^-(k) = \max\{n : \xi(n) \leqslant k\} \quad \text{правая}.$$

Утверждение (Барт, 2003)

1. $\xi_0^-(0) = 0$ *с вероятностью 1.*
2. $\xi_1^-(0)$ *имеет геометрическое распределение.*
3. $\eta = \lfloor \alpha \xi_1^-(0) \rfloor$ *имеет обобщенное геометрическое распределение* $\beta_\alpha^-(p)$ *при* $0 < p \leqslant 1$, $0 < \alpha \leqslant 1$

$$\mathrm{P}(\eta = j) = (1-p)^{\lceil \frac{j}{\alpha} \rceil} - (1-p)^{\lceil \frac{j+1}{\alpha} \rceil}, \; j = 0, 1, 2, \ldots$$

Обобщение по Алексеевой Н. П.

Изменение правила усечения числа неудач до первого успеха приводит к разным видам обобщения.

Утверждение (Алексеева, 2012)

$$\eta_0 = \lceil \alpha \xi_1^-(0) \rceil \sim {}_\alpha\beta^-(p)$$
$$\mathrm{P}(\eta_0 = j) = (1-p)^{\lfloor \frac{j-1}{\alpha} \rfloor + 1} - (1-p)^{\lfloor \frac{j}{\alpha} \rfloor + 1}, j = 1, 2, \ldots$$
$$\mathrm{P}(\eta_0 = 0) = p,$$

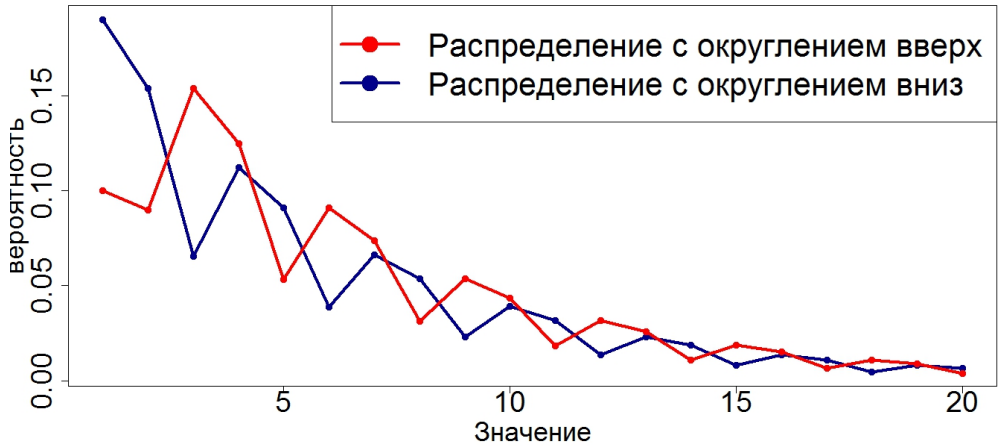

Рис. 1 : Вероятности распределений, $p = 0.1$, $\alpha = 0.6$

Варианты обобщения, введённые в данной работе

Утверждение (Законы распределений)

1. $\eta_1 = [\alpha\xi_1^-(0)] \sim \xi_R^-(p,\alpha)$, *считаем* $[0.5] = 1$,
   $\mathsf{P}(\eta_1 = 0) = 1 - (1-p)^{\lceil \frac{1}{2\alpha} \rceil}$
   $\mathsf{P}(\eta_1 = j) = (1-p)^{\lceil \frac{j-0.5}{\alpha} \rceil} - (1-p)^{\lceil \frac{j+0.5}{\alpha} \rceil}, j \geqslant 1$
2. $\eta_2 = \sum\limits_{i=0}^{\tau} \zeta_i$, $\tau \sim$ *Geom*(p), $\zeta_i \sim$ *Bern*(α),
   $\eta_2 \sim$ *Geom*$(\frac{p}{1-(1-p)(1-\alpha)})$

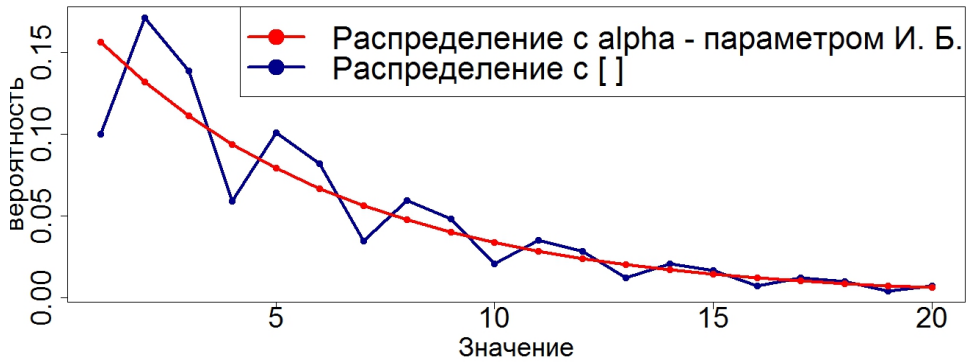

Рис. 2 : Вероятности распределений, $p = 0.1$, $\alpha = 0.6$

ОМП для параметров $\beta_\alpha^-(p)$ при $\alpha = \frac{2}{2k+1}$, $k \in N$

Пусть $q = 1 - p$, n_z — число элементов выборки, равных z $(z = 0, 1, \ldots)$.

Если k известно, то $\hat{q} = \left(1 - \frac{2C}{A}\right)^{\frac{1}{k}}$, иначе ОМП имеют вид

$$\begin{cases} \hat{q} = \dfrac{A^2}{(A+2B)(A-2C)} \\ \hat{k} = \log_{\hat{q}}\left(1 - \dfrac{2C}{A}\right) \end{cases}$$

$$A = \sum_z n_z z + \sum_{\text{нечетн. z}} n_z$$

$$B = \sum_{\text{четн. z}} n_z$$

$$C = \sum_{\text{нечетн. z}} n_z$$

Результаты оценивания параметров

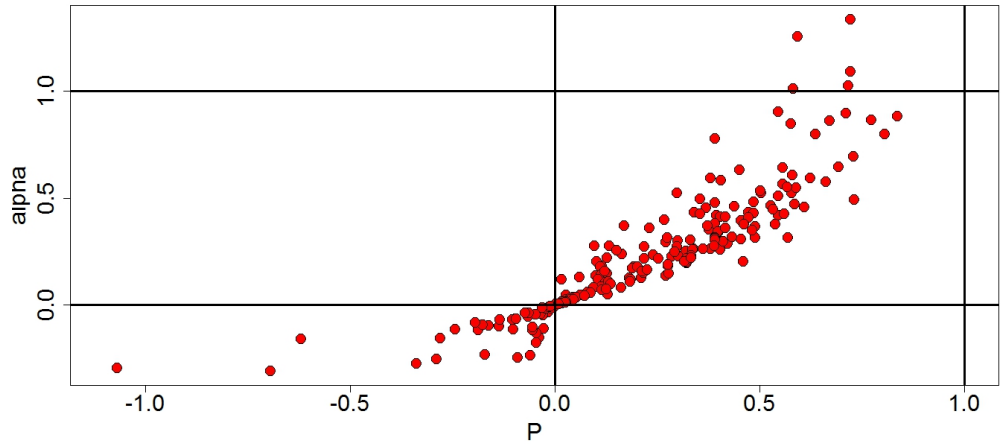

Рис. 3 : ОМП

Утверждение (Свойства оценок)

1. *Оценки* $\hat{p}$, $\hat{\alpha}$ *состоятельны при* $\alpha = \frac{2}{2t+1}$, *где* $t \in N$.
2. *Пусть* $\alpha = \frac{2}{2t+1}$, $t \in N$, $\hat{p} = \hat{p}_n$ *и* $\hat{\alpha} = \hat{\alpha}_n$, *где* n *— объем выборки. Тогда при* $n \to \infty$ *оценки* $\hat{p}_n > 0$ *и* $\hat{\alpha}_n > 0$.
3. *Если* $(A - \frac{2BC}{B-C})(B - C)(A - 2C) < 0$, *то* $\hat{p} < 0$, $\hat{\alpha} < 0$.

Расширение области определения параметров $\beta_\alpha^-(p)$

Пусть I и J — интервалы; $\beta^-(I \times J) \stackrel{\text{def}}{=} \{\beta_\alpha^-(p)\}_{p \in I, \alpha \in J}$.
Доказаны следующие свойства:

Утверждение (Расширение области определения параметров)

1. *между* $\beta^-([0,1) \times [0,+\infty))$ *и* $\beta^-((-\infty,0] \times (-\infty,0])$ *существует биекция, полностью сохраняющая значения вероятностей, она задается следующим образом:* $\beta_\alpha^-(p) \sim \beta_{-\alpha}^-(\frac{-p}{1-p})$
2. *пусть* $\xi \sim \beta_\alpha^-(p)$, *пусть* $\{k_i\}_{i=0,1,\ldots}$ *такая что:*

$$k_i = j \text{ такое, что } \mathrm{P}(\xi = j) > 0$$
$$\text{и } \#\{z : z < j \text{ и } \mathrm{P}(\xi = z) > 0\} = i,$$

тогда при $|\alpha| > 1$

$$\mathrm{P}(\xi = k_i) = p(1-p)^i.$$

Распределение β_α^- при $\alpha = 1/k, k \in N$

Известно, что для распределения β_α^- при $\alpha = 1/k, k \in N$, выполняется: $\beta_\alpha^-(p) \sim Geom(1-(1-p)^k)$.
Отсюда видно, что при $\alpha = 1$, $\beta_\alpha^- \sim Geom(p)$.
Исходя из этого свойства, были найдены пары значений параметров, задающие совпадающие распределения:

Утверждение (Классы совпадающих распределений)

Пусть зафиксировано p_g, $p_g \in [0,1]$, тогда все пары точек (p, α), удовлетворяющие соотношениям:

1. $\alpha = 1/k, k \in N$
2. $p = 1 - (1-p_g)^\alpha$

при различных k будут задавать одно и то же распределение $\beta_\alpha^-(p) \sim Geom(p_g)$

Другие свойства распределения β_α^-

О распределении β_α^- были доказаны следующие утверждения:

Утверждение (Переодичнось вероятностей и неоднозначность выбора распределения)

- *Пусть* $\xi \sim \beta_\alpha^-(p)$, $\alpha = \frac{m}{r}$, *где* $m, r \in N$; $m < r$; $r = cm + z$; $c, z \in Z$; $z < m$. *Тогда*

$$\mathsf{P}(\xi = k_i + nm) = \mathsf{P}(\xi = k_i)(q^{cm+z})^n,$$

где $k_i = 0, 1, \ldots, m-1; n \in N$.

- *Пусть* $\alpha = \frac{2}{2t+1}$, $t \in N$. *Тогда для фиксированного* k *и* $\forall$ $n < k$, $\forall\ 0 < l < \frac{2}{(2t+1)(k-1+2t)}$:

$$\mathsf{P}(\xi = n) = \mathsf{P}(\eta = n),$$

где $\xi \sim \beta_\alpha^-(p)$, $\eta \sim \beta_{\alpha+l}^-(p)$.

Оценка параметров ${}_{\alpha}\beta^{-}$ и ${}_{\alpha}\beta^{-}$ при $\alpha = \frac{2}{2k+1}$, $k \in N$

ОМП ${}_{\alpha}\beta^{-}$ была получена в виде решения системы уравнений:

$$\begin{cases} \dfrac{n_0}{x-y} + \dfrac{A+B+D}{x} - \dfrac{C}{1-x} = 0 \\ \dfrac{n_0 x}{y(y-x)} + \dfrac{A+B-C-D}{y} - \dfrac{D}{1-y} = 0, \end{cases}$$

$$A = \frac{1}{2} \sum_{\text{четн. z}} n_z z \qquad B = \frac{1}{2} \sum_{\text{нечетн. z}} n_z (z-1)$$

$$C = \sum_{\text{четн. z}} n_z \qquad D = \sum_{\text{нечетн. z}} n_z$$

где n_z — число элементов выборки, равных z $(z = 0, 1, \ldots)$, $x = \hat{q}^{\hat{k}+1}$, $y = \hat{q}^{\hat{k}}$.

Для распределения β_R^- ОМП были найдены в виде решения одной из двух систем, лучшего по принципу правдоподобия, чем решение второй системы.

Критерий CVM

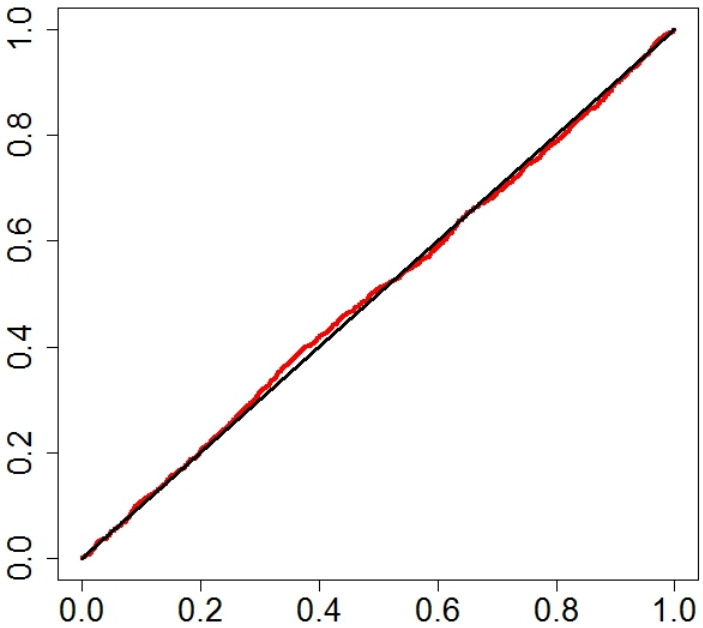

Рис. 4 : Функция распределения $p - value$, верная гипотеза, $\beta^{-}_{0.6}(0.1)$

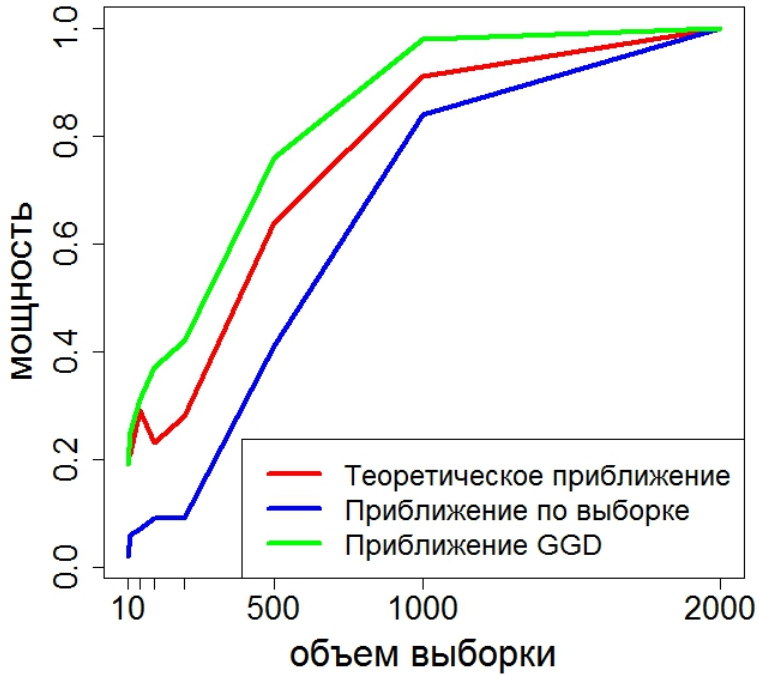

Рис. 5 : Мощность, $H_0 : \beta^{-}_{0.6}(0.1)$, $H_1 : Geom(p)$, уровень значимости $= 0.2$

Мощность CVM

Проблема различимости обобщённых распределений.

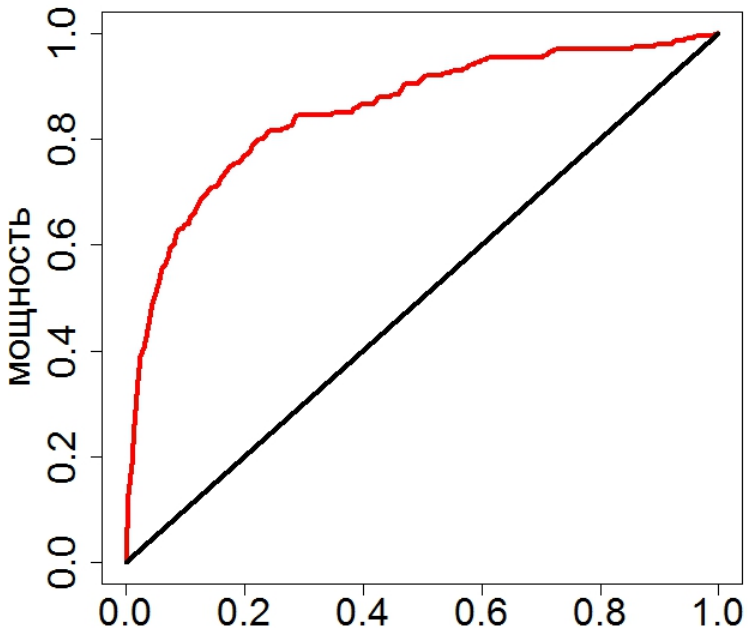

Рис. 6 : Мощность,
$H_0 : \beta^-_{0.6}(0.1)$,
$H_1 : \beta^-_{0.61}(0.11)$, $n = 100$

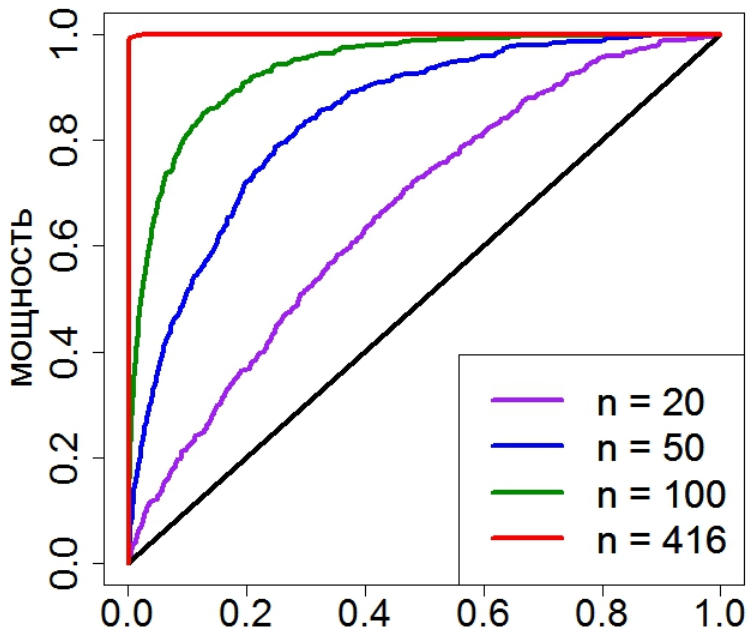

Рис. 7 : Мощность,
$H_0 : \beta^-_{0.6}(0.1)$,
$H_1 : {}_{0.6}\beta^-(0.1)$

Параметрический bootstrap тест

Алгоритм вычисления модифицированного значения $p - value$.
Для выборки выполняются следующие действия:

- получаются оценки параметров $\hat{\theta}$
- проверяется гипотеза согласия для параметров $\hat{\theta}$ (с помощью критерия CVM), получено $p - value$ P
- с оцененными параметрами $\hat{\theta}$ моделируется 100 выборок
- для них оцениваются параметры $\tilde{\theta}$ и проверяется гипотеза согласия для параметров $\tilde{\theta}$, получаются $p - value$ P^*
- строится эмпирическая функция распределения P^*
- по P и функции распределения P^* для выборки получается модифицированное значение $p - value$

Мощность и применимость bootstrap

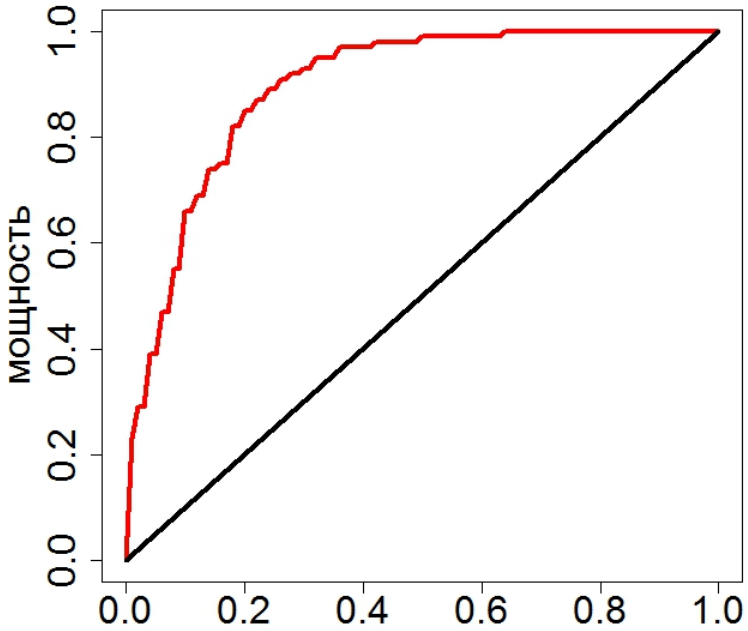

Рис. 8 : Мощность,
$H_0 : \beta^{-}_{0.6}(0.1)$,
$H_1 : Geom(p)$, $n = 416$

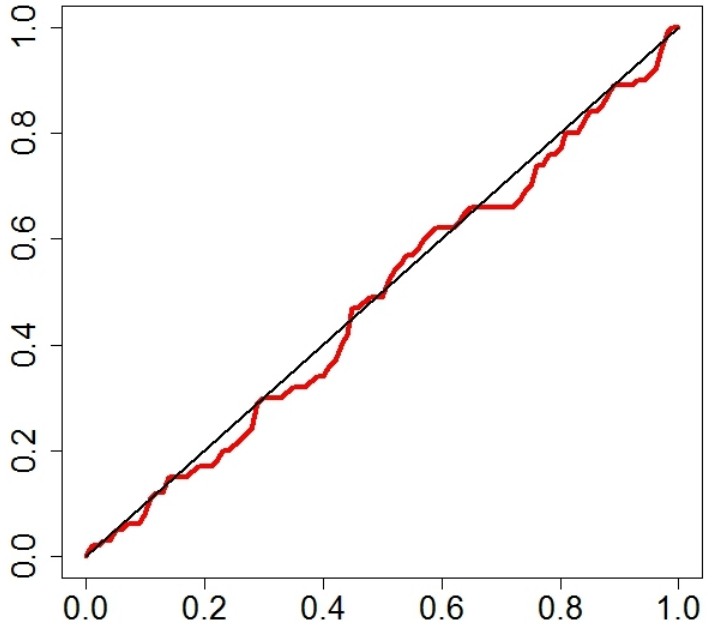

Рис. 9 : Функция распределения $p - value$, верная гипотеза, $\beta^{-}_{0.6}(0.1)$

Результаты применения описанных алгоритмов и критериев к лигвистическим данным

Таблица 1 : Результаты проверки гипотезы

Распределение	β_{α}^{-}	${}_{\alpha}\beta^{-}$	β_{R}^{-}	Geom
Число не отверженных гипотез из 200, CVM	145	125	62	113
Число не отверженных гипотез из 200, CVM с bootstrap	108	60	109	68

После применения процедуры bootstrap были сосчитаны модифицированные значения $p - value$, и число слов, подходящих под все распределения кроме β_{R}^{-}, уменьшилось. В итоге модели с наибольшим числом удовлетворяющих им слов - β_{α}^{-} и β_{R}^{-}. Проверка гипотез выполнялась для уровня значимости 0.2.