Новые статистические методы анализа текстов на естественном языке

Артем Павлович Ковалевский

artyom.kovalevskii@gmail.com

Новосибирск, 2022

Лекция 1. Методы, основанные на понятии авторского инварианта

Мы будем строить статистические тесты, позволяющие проверять гипотезы о текстах на естественном языке. Эти тесты работают в рамках очень простых вероятностных моделей текста. Мы будем каждый раз описывать эти вероятностные модели и возможные несоответствия. Приемлемость модели в сочетании с тестом оценивается по степени достижения требуемого результата:

модель $\rightarrow$ тест $\rightarrow$ проверка на текстах

Понятие авторского инварианта введено Фоменко и Фоменко (1984). Это множество слов языка, обладающее следующими свойствами:
1) для разных текстов одного автора относительные частоты появления слов из авторского инварианта примерно одинаковы;
2) для текстов разных авторов эти относительные частоты могут сильно различаться.

Итак, использование авторского инварианта состоит в следующем:
1) выделяется авторский инвариант — множество слов языка;
2) выделяются слова текста, заглавные буквы переводятся в строчные;
3) тексту сопоставляется последовательность индикаторов попадания слов в словарь:
$x_i = 1$, если i-тое слово содержится в инварианте;
$x_i = 0$ иначе.

Поскольку мы будем строить статистические критерии (тесты), поговорим подробнее о том, как они устроены.
Приведем формулировки для двухвыборочного статистического теста, одновыборочный будет его частным случаем.
Пусть $\vec{x}_{n_1} = (x_1, \ldots, x_{n_1})$ и $\vec{y}_{n_2} = (y_1, \ldots, y_{n_2})$ — конечные последовательности чисел.
Статистическая гипотеза предполагает, что эти конечные последовательности являются конечными реализациями бесконечных последовательностей случайных величин $(X_1, X_2, \ldots)$ и $(Y_1, Y_2, \ldots)$ с заданным совместным распределением.
Более подробно, случайные последовательности заданы на некотором вероятностном пространстве $(\Omega, \mathcal{F}, \mathrm{P})$, и для некоторого $\omega \in \Omega$ выполнено:
$x_i = X_i(\omega)$ для всех $i = 1, \ldots, n_1$;
$y_i = Y_i(\omega)$ для всех $i = 1, \ldots, n_2$.

Простыми гипотезами h называются статистические гипотезы, для которых распределения однозначно определены.
Основная гипотеза H и альтернативная гипотеза $\overline{H}$ — это два непересекающихся множества простых гипотез.
Статистикой называется измеримая функция

$$J_{n_1,n_2}(\vec{x}_{n_1}, \vec{y}_{n_2}).$$

Для построения статистического теста будем требовать от статистики следующих свойств:
1. Для любой $h \in H$ имеет место слабая сходимость

$$J_{n_1,n_2}(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}) \Rightarrow J$$

при $n_1 \to \infty$, $n_2 \to \infty$, где случайная величина J имеет известную непрерывную функцию распределения $F_J(x)$.
2. Для любой $h \in \overline{H}$ имеет место сильная сходимость

$$J_{n_1,n_2}(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}) \stackrel{a.s.}{\to} +\infty$$

при $n_1 \to \infty$, $n_2 \to \infty$.

Определим пи-значение (p-value, реально достигаемый уровень значимости)

$$\text{p-value} = \varepsilon^* = 1 - F_J(J_{n_1,n_2}(\vec{x}_{n_1}, \vec{y}_{n_2})).$$

Статистический тест (нерандомизированный статистический критерий согласия) состоит в следующем:

$$\text{гипотеза } H \text{ принимается на уровне } \varepsilon \quad \Longleftrightarrow \quad \varepsilon^* < \varepsilon.$$

Уровень ε выбирается в зависимости от задачи или оптимизируется по результатам исследования.

Разработаем простейший тест однородности и применим его к текстам.
Предполагаем, что $\vec{x}_{n_1}, \vec{y}_{n_2}$ состоят из нулей и единиц.
Согласно основной гипотезе H, случайные величины $X_1, X_2, \ldots$ и $Y_1, Y_2, \ldots$ независимы и имеют распределение Бернулли с одним и тем же параметром p, $0 < p < 1$.
Согласно альтернативной гипотезе $\overline{H}$, случайные величины $X_1, X_2, \ldots$ и $Y_1, Y_2, \ldots$ также независимы, но $X_1, X_2, \ldots$ имеют распределение Бернулли с параметром p_1, $0 < p_1 < 1$, а $Y_1, Y_2, \ldots$ имеют распределение Бернулли с параметром p_2, $0 < p_2 < 1$, причем $p_1 \neq p_2$, за счет чего основная и альтернативная гипотезы не пересекаются.

Построим тест на разности средних $\overline{X} - \overline{Y}$, где $\overline{X} = \sum_{i=1}^{n_1} X_i/n_1$, $\overline{Y} = \sum_{i=1}^{n_2} Y_i/n_2$.
Согласно основной гипотезе,

$$\mathrm{Var}(\overline{X} - \overline{Y}) = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Theorem 1.1

Пусть случайные величины $X_1, X_2, \ldots$ *и* $Y_1, Y_2, \ldots$ *независимы и одинаково распределены с конечной ненулевой дисперсией* σ^2,

$$\overline{X} = \sum_{i=1}^{n_1} X_i / n_1, \quad \overline{Y} = \sum_{i=1}^{n_2} Y_i / n_2.$$

Тогда

$$\frac{\overline{X} - \overline{Y}}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

слабо сходится к стандартному нормальному закону при $n_1 \to \infty$, $n_2 \to \infty$.

Доказательство.
Используем представление

$$\frac{\overline{X} - \overline{Y}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} =: S_{n_1+n_2} = \sum_{i=1}^{n_1+n_2} X_i^0,$$

где

$$X_i^0 = \frac{X_i - \mathrm{E}X_1}{n_1\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad 1 \le i \le n_1,$$

$$X_i^0 = \frac{\mathrm{E}X_1 - Y_{i-n_1}}{n_2\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad n_1 + 1 \le i \le n_1 + n_2.$$

Видим, что $S_{n_1+n_2}$ является суммой независимых случайных величин с нулевым матожиданием и дисперсиями

$$\mathrm{Var}X_i^0 = \mathrm{Var}\left(\frac{X_i - \mathrm{E}X_1}{n_1\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) = \frac{1}{n_1^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}, \quad 1 \le i \le n_1,$$

$$\mathrm{Var}X_i^0 = \mathrm{Var}\left(\frac{\mathrm{E}X_1 - Y_{i-n_1}}{n_2\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) = \frac{1}{n_2^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}, \quad n_1+1 \le i \le n_1+n_2.$$

Поэтому $\mathrm{Var}X_i^0 \to 0$ при $n_1 \to \infty$, $n_2 \to \infty$, и характеристические функции

$$\varphi_{X_i^0}(u) = \mathrm{E}e^{\mathrm{i}uX_i^0}$$

$$= 1+\mathrm{i}u\mathrm{E}X_i^0+\frac{\mathrm{i}^2u^2}{2}\mathrm{E}(X_i^0)^2(1+o(1)) = 1-\frac{u^2(1+o(1))}{2n_1^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)},\ 1 \le i \le n_1,$$

$$\varphi_{X_i^0}(u) = 1 - \frac{u^2(1+o(1))}{2n_2^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}, \quad n_1+1 \le i \le n_1+n_2.$$

В силу независимости случайных величин X_i^0, характеристическая функция

$$\varphi_{S_{n_1+n_2}}(u) = \mathrm{E}e^{\mathrm{i}uS_{n_1+n_2}}$$

$$= \left(1 - \frac{u^2(1+o(1))}{2n_1^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right)^{n_1} \left(1 - \frac{u^2(1+o(1))}{2n_2^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right)^{n_2}.$$

Используя второй замечательный предел, получаем

$$\lim_{n_1,n_2\to\infty} \varphi_{S_{n_1+n_2}}(u) = \lim_{n_1,n_2\to\infty} \exp\left(-\frac{u^2(1+o(1))}{2n_1\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} - \frac{u^2(1+o(1))}{2n_2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right)$$

$$= e^{-u^2/2}.$$

В силу теоремы о непрерывном соответствии, $\frac{\overline{X}-\overline{Y}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ слабо сходится к стандартному нормальному закону. $\square$

Мы будем применять эту теорему к анализу однородности двух текстов. В рамках используемой вероятностной модели предполагается, что слова текстов попадают в инвариант независимо друг от друга.
Основная гипотеза H утверждает, что вероятность попадания в авторский инвариант одинакова для всех слов. Обозначим эту вероятность через p, $0 < p < 1$.
Альтернативная гипотеза $\overline{H}$ утверждает, что эти вероятности разные для разных текстов: $p_1 \neq p_2$.
Обозначим

$$m_1 = \sum_{i=1}^{n_1} x_i = n_1 \overline{x}, \quad m_2 = \sum_{i=1}^{n_2} y_i = n_2 \overline{y}$$

— количества попаданий в авторский инвариант слов из первого и второго текстов.

Согласно теореме 1.1, при верной гипотезе H есть слабая сходимость случайной величины

$$\frac{\overline{X} - \overline{Y}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

к стандартному нормальному закону при $n_1 \to \infty$, $n_2 \to \infty$. Здесь

$$\sigma = \sqrt{\mathrm{Var} X_1} = \sqrt{p(1-p)}.$$

Параметр p неизвестен, поэтому заменим его на оценку

$$p^* = \frac{n_1\overline{X} + n_2\overline{Y}}{n_1 + n_2}.$$

Эта оценка сильно состоятельна согласно УЗБЧ. В силу теоремы об асимптотической нормальности, статистика

$$J_{n_1,n_2}(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}) = \frac{|\overline{X} - \overline{Y}|}{\sqrt{\frac{n_1\overline{X}+n_2\overline{Y}}{n_1+n_2}\left(1 - \frac{n_1\overline{X}+n_2\overline{Y}}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

слабо сходится к модулю стандартного нормального закона $|\mathcal{N}_{0,1}|$ при $n_1 \to \infty$, $n_2 \to \infty$.

Альтернативная гипотеза $\overline{H}$ утверждает, что $p_1 \neq p_2$, причем $0 < p_1 < 1$, $0 < p_2 < 1$.
Если верна $\overline{H}$, то

$$|\overline{X} - \overline{Y}| \to |p_1 - p_2| > 0$$

п.н. согласно УЗБЧ, а знаменатель
$J_{n_1,n_2}(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ сходится к 0 п.н., поэтому

$$J_{n_1,n_2}(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}) \to +\infty \quad \text{п.н.} \tag{1}$$

Exercise 1.1

Предположим, что $p_1 = 0 < p_2 \leq 1$. Всегда ли есть сходимость (1)?

Итак, на этой статистике мы сможем построить статистический тест:

$$\text{p-value} = 1 - F_{|\mathcal{N}_{0,1}|}(J_{n_1,n_2}(\vec{x}_{n_1}, \vec{y}_{n_2}))$$

$$= 2\left(1 - \Phi\left(\frac{|\overline{X} - \overline{Y}|}{\sqrt{\frac{n_1\overline{X}+n_2\overline{Y}}{n_1+n_2}\left(1 - \frac{n_1\overline{X}+n_2\overline{Y}}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}\right)\right)$$

$$= 2\left(1 - \Phi\left(\frac{|\frac{m_1}{n_1} - \frac{m_2}{n_2}|}{\sqrt{\frac{m_1+m_2}{n_1+n_2}\left(1 - \frac{m_1+m_2}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}\right)\right).$$

Здесь $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$ — функция распределения стандартного нормального закона.

Для того, чтобы применять этот стастистический тест к анализу однородности текстов, надо выделить слова из текста и выбрать авторский инвариант.
Слова будем выделять следующей процедурой: все буквы переводим в строчные (lowercase), удаляем символы конца строки, знаки препинания (кроме дефиса и апострофа), удаляем дефисы, окруженные пробелами.

```
line=line.replace('\n',' ')
for sym in  '.,:;«»“”!?—"':
                    line = line.replace(sym, '')
line = line.replace(" – ", " ")
line = line.lower()
ListText.extend(line.split())
```

Теперь надо выбрать множество слов, которое будет использоваться в качестве авторского инварианта.
Результаты при использовании различных авторских инвариантов могут радикально различаться.
Для иллюстрации мы рассмотрим множество, состоящее из двух слов — личного местоимения первого лица и частицы отрицания:
я и *не* в русском языке,
i и *no* в английском,
je и *ne* во французском.

В качестве примера возьмем первые два сонета из книги Les Regrets (1558) французского поэта дю Белле Joachim Du Bellay

I

Je ne veux point fouiller au sein de la nature,
Je ne veux point cercher l'esprit de l'univers,
Je ne veux point sonder les abysmes couvers,
N'y dessigner du ciel la belle architecture.

Je ne peins mes tableaux de si riche peinture,
Et si hauts argumens ne recerche a mes vers :
Mais suivant de ce lieu les accidens divers,
Soit de bien, soit de mal, j'escris a l'adventure.

Je me plains a mes vers, si j'ay quelque regret,
Je me ris avec eux, je leur di mon secret,
Comme estans de mon coeur les plus seurs secretaires.

Aussi ne veux-je tant les peigner et friser,
Et de plus braves noms ne les veux desguiser,
Que de papiers journaux, ou bien de commentaires.

II

Un plus scavant que moy (Paschal) ira songer
Avesques l'Ascrean dessus la double cyme :
Et pour estre de ceux dont on fait plus d'estime,
Dedans l'onde au cheval tout nud s'ira plonger.

Quant a moy, je ne veux, pour un vers allonger,
M'accourcir le cerveau : ni pour polir ma rime,
Me consumer l'esprit d'une soigneuse lime,
Frapper dessus ma table, ou mes ongles ronger.

Aussi veux-je (Paschal) que ce que je compose
Soit une prose en ryme, ou une ryme en prose,
Et ne veux pour cela le laurier meriter.

Et peut estre que tel se pense bien habile,
Qui trouvant de mes vers la ryme si facile,
En vain travaillera, me voulant imiter.

Для первых двух сонетов из Les Regrets дю Белле
$n_1 = 122$, $m_1 = 14$, $n_2 = 114$, $m_2 = 4$
p-value= 0.021218

Следующий пример — песни Владимира Высоцкого

Я не люблю

Я не люблю фатального исхода.
От жизни никогда не устаю.
Я не люблю любое время года,
Когда веселых песен не пою.
Я не люблю открытого цинизма,
В восторженность не верю, и еще,
Когда чужой мои читает письма,
Заглядывая мне через плечо.
Я не люблю, когда наполовину
Или когда прервали разговор.
Я не люблю, когда стреляют в спину,
Я также против выстрелов в упор.
Я ненавижу сплетни в виде версий,
Червей сомненья, почестей иглу,
Или, когда все время против шерсти,
Или, когда железом по стеклу. [...]

Если где-то в чужой, неспокойной ночи

Если где-то в чужой, неспокойной ночи, ночи
Ты споткнулся и ходишь по краю -
Не таись, не молчи, до меня докричи, докричи,
Я твой голос услышу, узнаю.
Может, с пулей в груди ты лежишь в спелой ржи, в спелой ржи?
Потерпи! Я иду, и усталости ноги не чуют.
Мы вернемся туда, где и травы врачуют,
Только - ты не умри, только - кровь удержи.
Если ж конь под тобой - ты домчи, доскачи, доскачи,
Конь дорогу отыщет, буланый,
В те края, где всегда бьют живые ключи, ключи,
И они исцелят твои раны. [...]

Результаты для двух песен Высоцкого:
$n_1 = 176$, $m_1 = 29$, $n_2 = 174$, $m_2 = 9$
p-value= 0.000676

Сравним теперь песню Высоцкого с первым сонетом из Les Regrets дю Белле.
Для этого заменим *я* на *je*, *не* на *ne*:

```
je ne люблю фатального исхода
от жизни никогда ne устаю
je ne люблю любое время года
когда веселых песен ne пою [...]
```

Здесь соответствие весьма хорошее:
$n_1 = 176$, $m_1 = 29$, $n_2 = 122$, $m_2 = 14$
p-value= 0.226935

Требование к авторскому инварианту: чтобы он был инвариантным.
Идея Фоменко и Фоменко: служебные слова.
Служебные слова — это предлоги, союзы и частицы:
1) предлоги: в, на, с, за, к, по, из, у, от, для, во, без, до, о, через, со, при, про, об, ко, над, из-за, из-под, под;
2) союзы: и, что, но, а, да, хотя, когда, чтобы, если, тоже, или, зато, будто;
3) частицы: не, как, же, даже, бы, ли, только, вот, то, ни, лишь, ведь, вон, то-есть, нибудь, уже, либо.

Результаты для этого авторского инварианта и двух песен Высоцкого:
$n_1 = 176$, $m_1 = 52$, $n_2 = 174$, $m_2 = 43$
p-value= 0.309363

Лекция 2. Разладка в текстах

Двигающим мотивом к этой работе послужила процедура написания рефератов студентами в век Интернета: зачастую реферат студента состоит просто в соединении двух или нескольких текстов, найденных с помощью поисковой системы. При такой процедуре «творчества» определить интеллектуальный вклад студента не представляется возможным. Поэтому важен алгоритм, позволяющий быстро выявить наличие разнородных фрагментов текста.

Введем следующие обозначения. Пусть $\{\xi_i^{(1)}\}_{i=1}^{\infty}$ и $\{\xi_i^{(2)}\}_{i=1}^{\infty}$ — две взаимно независимых последовательности независимых одинаково распределенных случайных величин с распределениями $\mathcal{F}_1$ и $\mathcal{F}_2$ соответственно.
Схема серий случайных величин $\{X_i^{(n)}\}_{i=1}^{n}$ задается следующим образом:
$X_i^{(n)} = \xi_i^{(1)}$ при $1 \leq i \leq [nT]$;
$X_i^{(n)} = \xi_i^{(2)}$ при $[nT] + 1 \leq i \leq n$.
Здесь T — неизвестный параметр, $0 < T < 1$.

Нулевая гипотеза H состоит в том, что разладки не произошло, то есть $\mathcal{F}_1 = \mathcal{F}_2$. Ее альтернатива — в том, что не только распределения, но и их математические ожидания различны.
Как при нулевой гипотезе, так и при альтернативе мы будем предполагать, что дисперсии распределений $\mathcal{F}_1$ и $\mathcal{F}_2$ конечны и положительны.
Каждый из рассматриваемых критериев проверки гипотез основан на последовательности статистик $V_1,\ V_2,\ \ldots$. Гипотеза H принимается, если статистика не превосходит некоторого уровня.

Для решения задачи обнаружения разладки строятся статистики, являющиеся функционалами от *эмпирического моста* $Z_n = \{Z_n(t), 0 \le t \le 1\}$ — случайной ломаной, построенной по точкам

$$\left(\frac{k}{n};\ \frac{nS_k - kS_n}{sn\sqrt{n}} \right), k = 0, \ldots, n, \tag{2}$$

где $S_n = \sum_{i=1}^{n} X_i^{(n)} = n\overline{X}$, $s^2 = \overline{X^2} - (\overline{X})^2$.

Доказывается, что для любой простой гипотезы θ_0 из H эмпирический мост C-сходится к стандартному броуновскому мосту W^0, то есть для любого заданного на $C(0;1)$ непрерывного в равномерной метрике функционала g имеет место сходимость по распределению

$$g(Z_n) \overset{\mathsf{P}_{\theta_0}}{\Rightarrow} g(W^0).$$

Также доказывается, что для любой простой гипотезы θ из альтернативы эмпирический мост сходится почти наверное в равномерной метрике на отрезке $[0;\ 1]$ к детерминированному процессу $z_\theta = \{z_\theta(t),\ 0 \le t \le 1\}$, то есть для любого заданного на $C(0;1)$ непрерывного в равномерной метрике функционала J имеет место сходимость почти наверное

$$J(Z_n)/\sqrt{n} \to J(z_\theta)\ \ (\mathrm{P}_\theta) - \text{п.н.}$$

Процесс z_θ — кусочно линейный, с единственным изломом в точке T.

В качестве функционалов рассматриваются различные нормы эмпирического моста, а также нормированные взвешенные суммы случайных величин. Доказано, что нормированные взвешенные суммы с точностью до малого слагаемого (сходящегося п. н. к нулю при выполнении как основной гипотезы, так и альтернативы) представимы в виде интегральных функционалов от эмпирического моста.

В случае, когда последовательность статистик J_n сходится по распределению к случайной величине J с функцией распределения F, для нее определен *приближенный бахадуровский наклон* $c(\theta)$ равенством

$$c(\theta) = 2 \lim_{n\to\infty} (-n^{-1} \ln(1 - F(J_n))) \quad (\mathrm{P}_\theta - \text{п. н.}) \qquad (3)$$

(см. [55], §1.4). Это определение и будет нами использоваться для сравнения критериев: чем больше $c(\theta)$, тем лучше критерий различает гипотезы.

Отметим, что $c(\theta)$ в пределе (при $\theta \to \theta_0$) обратно пропорционально числу испытаний n, необходимому для достижения заданного размера критерия α и заданной мощности β при близких гипотезах $\mathcal{F}_1$ и $\mathcal{F}_2$. Этот подход, изобретенный Э. Питменом, изложен в [57] (глава 7) для статистик, имеющих в пределе нормальное распределение.

Для случая, когда предельное распределение абсолютно непрерывно, но отлично от нормального, модификация подхода Питмена введена Х. С. Виэндом в [115]. Для применения подхода Э. Питмена надо проверить, что статистика является виэндовской, то есть при выполнении альтернативной гипотезы сходится к предельному значению достаточно быстро. Подробное изложение этих результатов можно найти в книге Я. Ю. Никитина [55] (глава 1).

Заданы $\{\xi_i^{(1)}\}_{i=1}^{\infty}$ и $\{\xi_i^{(2)}\}_{i=1}^{\infty}$ — две взаимно независимых последовательности независимых одинаково распределенных случайных величин.
Случайные величины $\xi_i^{(1)}$ имеют распределение $\mathcal{F}_1$ с математическим ожиданием m_1 и дисперсией $0 < {\sigma_1}^2 < \infty$.
Случайные величины $\xi_i^{(2)}$ имеют распределение $\mathcal{F}_2$ с математическим ожиданием m_2 и дисперсией $0 < {\sigma_2}^2 < \infty$.

Схема серий случайных величин $\{X_i^{(n)}\}_{i=1}^n$ задается следующим образом:
$X_i^{(n)} = \xi_i^{(1)}$ при $1 \leq i \leq [nT]$;
$X_i^{(n)} = \xi_i^{(2)}$ при $[nT] + 1 \leq i \leq n$.
Предполагается, что T — неизвестная константа, $0 < T < 1$.

Простыми гипотезами θ здесь являются тройки $\theta = (\mathcal{F}_1,\ \mathcal{F}_2,\ T)$. Будем предполагать, что все множество гипотез $\Theta = \Theta_0 \cup \Theta_1$, где $\Theta_0 = \{\theta :\ \mathcal{F}_1 = \mathcal{F}_2,\ m_1 \neq 0\}$, $\Theta_1 = \{\theta :\ m_1 \neq m_2\}$. Мы будем строить критерии, различающие сложные гипотезы Θ_0 и Θ_1.

Будем использовать обозначения X_i вместо $X_i^{(n)}$ там, где это не приводит к противоречиям.

Remark 2.1

Сформулированные выше условия на случайные величины $X_1^{(n)}, \ldots, X_n^{(n)}$ *близки к необходимым условиям состоятельности критерия проверки гипотез. Если согласно первой гипотезе дисперсия равна 0, это приводит к вырожденной статистической задаче. Случаев бесконечной дисперсии или выполнения равенств* $m_1 = m_2$, $m_1 = 0$ *можно избежать путем подходящих преобразований фазового пространства.*

Критерий принимает альтернативную гипотезу в случае, когда $J_n \geq C$, где J_n — статистика, определяющая критерий.
В дальнейшем рассматриваются различные классы статистик: взвешенные суммы и L_p-нормы.
Введем топологию, порожденную полуметрикой модуля разности математических ожиданий до и после разладки: $||\theta|| = |m_1 - m_2|$. В этом случае $\partial\Theta_0 = \{\theta : \ m_1 = m_2\} = \Theta_0$.

Для сравнения критериев будем вычислять приближенные бахадуровские наклоны, определяемые равенством (3). Отметим, что в случае сходимости почти наверное (как в (3)) говорят о сильных наклонах, а в случае сходимости по вероятности — о слабых. Мы докажем, что для всех рассматриваемых последовательностей статистик есть сходимость почти наверное.

Мы будем говорить, что один критерий *лучше* другого в смысле используемого подхода, если для всех $\theta \in \Theta_1$ из некоторой окрестности $\partial\Theta_0$ значение $c(\theta)$ приближенного бахадуровского наклона (определяемого равенством (3)) для статистики первого критерия больше, чем для второго.

Lemma 2.1

Если для любой $\theta_0 \in \Theta_0$ *имеет место сходимость к одному и тому же распределению* $\mathsf{P}_{\theta_0}\{J_n < t\} \to F(t)$, *а для любой* $\theta \in \Theta_1$ — *сходимость* $J_n/\sqrt{n} \to j(\theta)$ *(*P_θ*-п. н.), причем*

$$\ln(1 - F(t)) \sim -\frac{1}{2}at^2 \tag{4}$$

при $t \to \infty$, *то*

$$c(\theta) = aj^2(\theta)\ (\mathsf{P}_\theta - \text{п. н.}) \tag{5}$$

Доказательство.
Согласно формуле (3) ,

$$c(\theta) = 2 \lim_{n\to\infty}(-n^{-1} \ln(1 - F(J_n)) \ \ (\mathrm{P}_\theta - \text{п. н.})$$

В силу сделанных предположений,

$$c(\theta) = 2 \lim_{n\to\infty}(n^{-1} \frac{1}{2} a J_n^2) = a j^2(\theta) \ \ (\mathrm{P}_\theta - \text{п. н.})$$

Доказательство завершено.

Рассмотрим эмпирический мост $Z_n = \{Z_n(t),\ 0 \le t \le 1\}$ — случайную ломаную, построенную по точкам, определенным формулой (2). Понятие эмпирического моста возникло в [?] при анализе неоднородности энергопотребления в течение суток. По определению,

$$Z_n(t) = \frac{nS_k - kS_n}{sn\sqrt{n}} + \frac{nX_{k+1} - S_n}{s\sqrt{n}}\left(t - \frac{k}{n}\right),$$

$$\frac{k}{n} \le t < \frac{k+1}{n}, \quad k = 0, \ldots, n-1.$$

Lemma 2.2

Для любой $\theta_0 \in \Theta_0$ имеет место C-сходимость эмпирического моста Z_n к броуновскому мосту W^0.

Доказательство.
Представим $Z_n(t)$ в виде

$$Z_n(t) = Z_n^*(t) - tZ_n^*(1),$$

где Z_n^* — случайная ломаная, построенная по точкам

$$\left(\frac{k}{n};\ \frac{S_k - km_1}{s\sqrt{n}}\right), \quad k = 0, \ldots, n.$$

Пусть Z_n^{**} — случайная ломаная, построенная по точкам

$$\left(\frac{k}{n};\ \frac{S_k - km_1}{\sigma_1\sqrt{n}}\right), k = 0, \ldots, n.$$

В силу принципа инвариантности (см. [5], [7]) процесс Z_n^{**} будет C-сходиться к стандартному винеровскому процессу W.

Так как $s \to \sigma_1$ (P_{θ_0}-п.н.), то Z_n^* будет C-сходиться к стандартному винеровскому процессу W.

В силу свойств слабой сходимости Z_n будет C-сходиться к

$$W^0 = \{W^0(t), 0 \leq t \leq 1\},$$

где $W^0(t) = W(t) - tW(1)$.
Доказательство завершено.

Lemma 2.3

Для любой $\theta \in \Theta_1$ процесс $Z_n/\sqrt{n}$ сходится (P_θ-п. н.) в равномерной метрике на отрезке $[0;\ 1]$ к функции z_θ:

$$z_\theta(t) = \begin{cases} \frac{tM_\theta}{T}, & 0 \le t \le T; \\ \frac{(1-t)M_\theta}{1-T}, & T \le t \le 1, \end{cases}$$

$$M_\theta = \frac{(m_1 - m_2)T(1-T)}{\sigma_\theta},$$

$$\sigma_\theta = \sqrt{T({m_1}^2 + {\sigma_1}^2) + (1-T)({m_2}^2 + {\sigma_2}^2) - (Tm_1 + (1-T)m_2)^2}.$$

Доказательство.
Докажем, что

$$\sup_{0\le t\le T} |Z_n(t)/\sqrt{n} - z(t)| \to 0$$

п. н. Так как в условиях теоремы $s \to \sigma_T$ п. н., то достаточно доказать, что

$$\max_{1\le k\le [nT]} \left| \frac{nS_k - kS_n}{n^2} - \frac{k}{n}(m_1 - m_2)(1 - T) \right| \to 0$$

п. н.

Эта сходимость следует из неравенств

$$\max_{1\le k\le [nT]}\left|\frac{nS_k - kS_n}{n^2} - \frac{k}{n}(m_1 - m_2)(1-T)\right|$$

$$\le \max_{1\le k\le [nT]}\left|\frac{S_k - km_1}{n}\right| + \max_{1\le k\le [nT]}\left|\frac{k}{n}\cdot\frac{S_n - m_1 T - m_2(1-T)}{n}\right|$$

$$\le \max_{1\le k\le [nT]}\left|\frac{S_k - km_1}{n}\right| + \left|\frac{S_{[nT]} - m_1 T}{n}\right| + \left|\frac{S_n - S_{[nT]} - m_2(1-T)}{n}\right|.$$

Последние 2 слагаемых сходятся п. н. к нулю в силу усиленного закона больших чисел. Докажем сходимость п. н. к нулю первого слагаемого: пусть $S_n^0 = S_n - nm_1$ — сумма центрированных слагаемых. Сходимость

$$\max_{1\le k\le [nT]} \left|\frac{S_k^0}{n}\right| \to 0$$

п. н. следует из того, что для любого фиксированного $N_0 \ge 1$ имеет место

$$\max_{1\le k\le [nT]} \left|\frac{S_k^0}{n}\right| \le \frac{N_0}{n} \max_{1\le k\le N_0} \left|\frac{S_k^0}{N_0}\right| + \max_{N_0\le k\le [nT]} \left|\frac{S_k^0}{n}\right| \to 0$$

п. н., так как

$$\frac{N_0}{n} \max_{1\le k\le N_0} \left|\frac{S_k^0}{N_0}\right| \to 0$$

п. н. при $n \to \infty$, и

$$\max_{N_0\le k\le [nT]} \left|\frac{S_k^0}{n}\right| \le \sup_{k\ge N_0} \left|\frac{S_k^0}{k}\right| \to 0$$

п. н. при $N_0 \to \infty$ согласно УЗБЧ.

Аналогично

$$\sup_{T \le t \le 1} |Z_n(t)/\sqrt{n} - z(t)| \to 0$$

п. н. Лемма доказана.

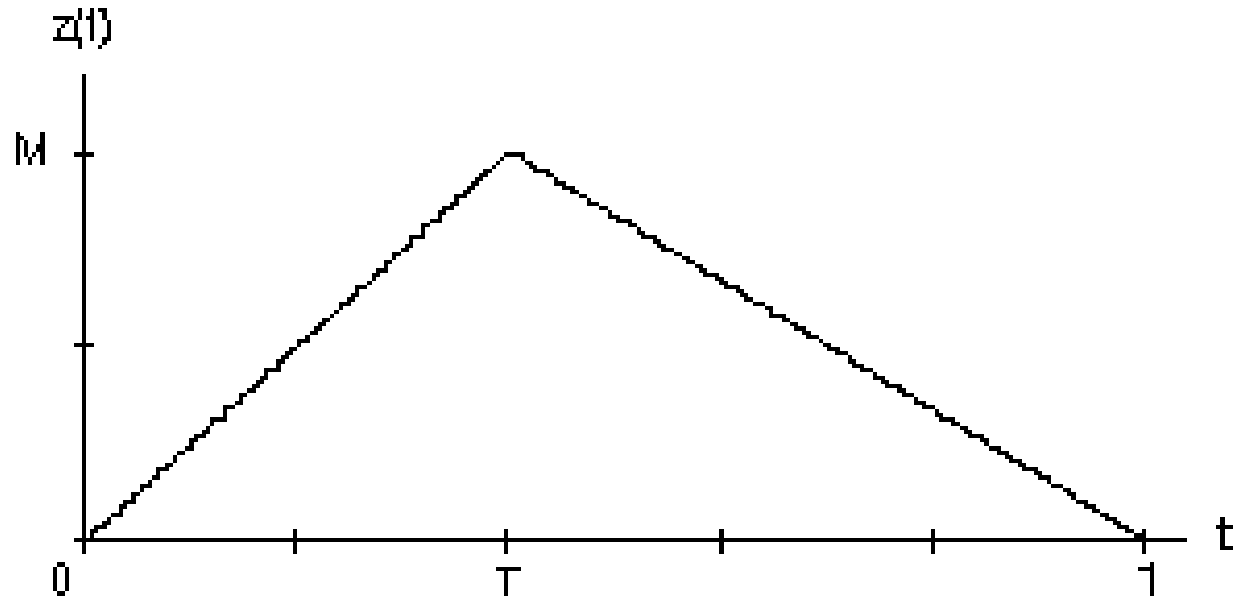

Рис. 2.1. График процесса z_θ. Здесь T — момент разладки, $M = M_\theta$ — экстремальное значение процесса.

Представляется логичным рассматривать различные нормы случайной функции Z_n на отрезке $[0,\ 1]$, а также размах этой функции на отрезке. Обозначим

$$J_n^{(r)} = ||Z_n||_{L_r} = \left(\int_0^1 |Z_n(t)|^r dt \right)^{1/r};$$

$$J_n^{\infty} = ||Z_n||_{L_\infty} = \sup_{t \in [0;\ 1]} |Z_n(t)|;$$

$$J_n^{R} = \sup_{t \in [0;\ 1]} Z_n(t) - \inf_{t \in [0;\ 1]} Z_n(t).$$

Наряду с L_r-нормами эмпирического процесса рассмотрим функционалы вида

$$J_n^{|r|} = \left| \int_0^1 (Z_n(t))^r \, dt \right|^{1/r}. \tag{6}$$

При четном r эти функционалы совпадают с L_r-нормами.

При выполнении первой гипотезы распределение статистики J_n^∞ слабо сходится к распределению Колмогорова, J_n^R — к распределению размаха броуновского моста, а $J_n^{(2)}$ — к распределению корня квадратного из случайной величины, имеющей распределение омега-квадрат.

Отметим, что критерии, основанные на функционалах размаха и L_r-нормы, оказываются при выборе из альтернатив Θ_0 и Θ_1 асимптотически не лучше критериев, основанных на функционалах J_n^{∞} и $J_n^{|r|}$ соответственно:

Theorem 2.1

Статистика J_n^R не лучше статистики J_n^{∞}. Статистика $J_n^{(r)}$ не лучше статистики $J_n^{|r|}$.

Доказательство.
В силу леммы 2.3 для $\theta \in \Theta_1$ имеют место сходимости

$$J_n^R/\sqrt{n} \to |M_\theta|, \qquad J_n^\infty/\sqrt{n} \to |M_\theta|$$

(P_θ-п.н.).
В то же время

$$J_n^R \geq J_n^\infty$$

почти наверное просто по определению, и при выполнении нулевой гипотезы для предельных распределений имеет место неравенство

$$\mathrm{P}\{\sup_{t\in[0;\,1]} W^0(t) - \inf_{t\in[0;\,1]} W^0(t) \geq x\} \geq \mathrm{P}\{\sup_{t\in[0;\,1]} W^0(t) \geq x\}.$$

Согласно определению приближенного бахадуровского наклона (3) статистика J_n^R не лучше статистики J_n^∞.

Аналогично для интегральных функционалов: для $\theta \in \Theta_1$

$$J_n^{(r)}/\sqrt{n} \to \left(\int_0^1 |z_\theta(t)|^r dt\right)^{1/r} = \left|\int_0^1 (z_\theta(t))^r dt\right|^{1/r},$$

$$J_n^{|r|}/\sqrt{n} \to \left|\int_0^1 (z_\theta(t))^r dt\right|^{1/r}$$

(P_θ-п.н.).
В то же время

$$\mathsf{P}\{\left(\int_0^1 |W^0(t)|^r dt\right)^{1/r} \geq x\} \geq \mathsf{P}\{\left|\int_0^1 (W^0(t))^r dt\right|^{1/r} \geq x\}.$$

Доказательство завершено.

Рассмотрим статистику

$$J_n = \left| \frac{\sum_{k=1}^{n} h_{k,n} X_k}{s\sqrt{n}} \right|.$$

Здесь $h_{1,n}, \ldots, h_{k,n}$ — весовые коэффициенты. Они могут выбираться весьма произвольно, в частности, их знаки могут чередоваться, что ухудшает качество статистического критерия. Мы будем предполагать, что весовые коэффициенты заданы следующим регулярным образом:

$$h_{k,n} = g(k/n),$$

где $g(t)$ — функция ограниченной вариации на $[0,\ 1]$. Для введенных с помощью функции g весовых коэффициентов будем использовать обозначение статистики $J_n = J_n(g)$.

Theorem 2.2

Для любой $\theta_0 \in \Theta_0$ *статистика* $J_n(g)$ *сходится по распределению к модулю стандартного нормального закона в том и только том случае, когда выполнены условия*

$$\int_0^1 g(t)dt = 0, \qquad \int_0^1 g^2(t)dt = 1. \tag{7}$$

Доказательство.
Последовательность случайных величин $\{h_{k,n}X_k\}$ удовлетворяет условиям теоремы Линдеберга (см., например, [7], глава 16, параграф 7), поэтому

$$\frac{\sum_{k=1}^{n} h_{k,n}X_k - \mathrm{E}\left(\sum_{k=1}^{n} h_{k,n}X_k\right)}{\sqrt{\mathrm{Var}\left(\sum_{k=1}^{n} h_{k,n}X_k\right)}}$$

сходится по распределению к стандартному нормальному закону.

Заметим, что

$$\frac{1}{n}\mathrm{Var}\left(\sum_{k=1}^{n} h_{k,n} X_k\right) = \sum_{k=1}^{n} \frac{{h_{k,n}}^2}{n} {\sigma_1}^2 \to {\sigma_1}^2 \int_0^1 g^2(t)dt.$$

Так как

$$\mathsf{E}\left(\sum_{k=1}^{n} h_{k,n} X_k\right) = m_1 \sum_{k=1}^{n} g(k/n),$$

и $s \to \sigma_1$ (P_{θ_0}-п.н.), то $J_n(g)$ сходится по распределению к модулю стандартного нормального закона тогда и только тогда, когда

$$\frac{\sum_{k=1}^{n} g(k/n) X_k - m_1 \sum_{k=1}^{n} g(k/n)}{\sigma_1 \sqrt{n}}$$

сходится по распределению к стандартному нормальному закону,

то есть

$$\frac{1}{n{\sigma_1}^2}\mathrm{Var}\left(\sum_{k=1}^{n} h_{k,n}X_k\right) \to 1$$

что равносильно второму из условий теоремы, и

$$\frac{m_1}{\sigma_1\sqrt{n}}\sum_{k=1}^{n} g(k/n) \to 0,$$

что равносильно первому из условий теоремы, так как по теореме о среднем существуют такие $r_k = r_k(n)$, что $r_k \in [k-1;\ k]$ и

$$\int_0^1 g(t)dt = \frac{1}{n}\sum_{k=1}^{n} g(r_k/n),$$

и поэтому

$$\sup_n \left| \sum_{k=1}^{n} g(k/n) - n\int_0^1 g(t)dt \right| \leq$$

$$\leq \sup_n \sum_{k=1}^{n} |g(k/n) - g(r_k/n)| \leq \int_0^1 |dg(t)| < \infty.$$

Теорема доказана.

Theorem 2.3

Пусть g — функция ограниченной вариации, удовлетворяющая условиям (7) теоремы 2.2. Тогда статистику $J_n(g)$ можно представить в виде интеграла Стилтьеса

$$J_n(g) = \left| \int_0^1 Z_n(t) dg(t) + \nu_n \right|,$$

где $\nu_n \to 0$ (P_θ-п.н.) для любой $\theta \in \Theta$.

Доказательство.
По определению

$$J_n(g) = \left| \frac{\sum_{k=1}^{n} g(k/n) X_k}{s\sqrt{n}} \right|,$$

$$\int_0^1 Z_n(t)dg(t) =$$

$$= \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nS_k - kS_n}{sn\sqrt{n}} dg(t) +$$

$$+ \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nX_{k+1} - S_n}{s\sqrt{n}} \left(t - \frac{k}{n} \right) dg(t) =$$

$$= \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \left(\frac{S_k}{s\sqrt{n}} - \frac{k}{n} \frac{S_n}{s\sqrt{n}} \right) dg(t) +$$

$$+ \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nX_{k+1} - S_n}{s\sqrt{n}} \left(t - \frac{k}{n} \right) dg(t) =$$

$$= \frac{1}{s\sqrt{n}} \sum_{k=0}^{n-1} \left(S_k - \frac{k}{n} S_n \right) \left(g\left(\frac{k+1}{n} \right) - g\left(\frac{k}{n} \right) \right) +$$

$$+ \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nX_{k+1} - S_n}{s\sqrt{n}} \left(t - \frac{k}{n} \right) dg(t) =$$

$$= \frac{1}{s\sqrt{n}} \left(\sum_{k=1}^{n} \left(S_{k-1} - \frac{k-1}{n} S_n \right) g\left(\frac{k}{n}\right) - \sum_{k=1}^{n} \left(S_k - \frac{k}{n} S_n \right) g\left(\frac{k}{n}\right) \right) +$$

$$+ \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nX_{k+1} - S_n}{s\sqrt{n}} \left(t - \frac{k}{n} \right) dg(t) =$$

$$= \frac{1}{s\sqrt{n}} \sum_{k=1}^{n} g\left(\frac{k}{n}\right)\left(-X_k + \frac{S_n}{n}\right) +$$

$$+ \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nX_{k+1} - S_n}{s\sqrt{n}} \left(t - \frac{k}{n}\right) dg(t) =$$

$$= \frac{1}{s\sqrt{n}} \left(\frac{S_n}{n} \sum_{k=1}^{n} g\left(\frac{k}{n}\right) - \sum_{k=1}^{n} g\left(\frac{k}{n}\right) X_k \right) +$$

$$+ \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nX_{k+1} - S_n}{s\sqrt{n}} \left(t - \frac{k}{n} \right) dg(t).$$

Итак,

$$\nu_n = -\frac{S_n}{sn}\frac{\sum_{k=1}^n g(\frac{k}{n})}{\sqrt{n}} - \sum_{k=0}^{n-1}\int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nX_{k+1} - S_n}{s\sqrt{n}}\left(t - \frac{k}{n}\right) dg(t).$$

Так как g имеет ограниченную вариацию и $\int_0^1 g(t)\,dt = 0$, то

$$\left|\frac{\sum_{k=1}^n g(\frac{k}{n})}{n}\right| \le \sum_{k=1}^n \left|\frac{g(\frac{k}{n})}{n} - \int_{(k-1)/n}^{k/n} g(t)dt\right|$$

$$\le \sum_{k=1}^n \frac{1}{n}\int_{(k-1)/n}^{k/n} |dg(t)| = \frac{1}{n}\int_0^1 |dg(t)| = O(1/n).$$

Поэтому в силу УЗБЧ получаем, что

$$\frac{S_n}{sn}\frac{\sum_{k=1}^n g(\frac{k}{n})}{\sqrt{n}} \to 0$$

п. н. с ростом n.

Справедлива оценка

$$\left|\sum_{k=0}^{n-1}\int_{\frac{k}{n}}^{\frac{k+1}{n}}\frac{nX_{k+1}-S_n}{s\sqrt{n}}\left(t-\frac{k}{n}\right)dg(t)\right|\leq$$

$$\leq\frac{\max_{k=0}^{n-1}|X_{k+1}|+|S_n|\int_0^1|dg(t)|}{sn\sqrt{n}}\leq$$

$$\leq\frac{\sum_{k=1}^{n}|X_k|\left(1+\int_0^1|dg(t)|\right)}{sn\sqrt{n}}.$$

Правая часть неравенства сходится п. н. к нулю в силу усиленного закона больших чисел.
Теорема доказана.

Corollary 2.1

Для любой $\theta \in \Theta_1$ *имеет место сходимость*

$$\frac{J_n(g)}{\sqrt{n}} \to \left| \frac{m_1 \int_0^T g(t)dt + m_2 \int_T^1 g(t)dt}{\sigma_\theta} \right|$$

(P_θ*-п. н.)*

Доказательство.
Так как $\int_0^1 Z_n(t)dg(t)$ — непрерывный в равномерной метрике функционал от Z_n, лемма 2.3 обосновывает сходимость (P_θ-п. н.) $\int_0^1 Z_n(t)dg(t)/\sqrt{n} \to \int_0^1 z_\theta(t)dg(t)$. Интегрирование по частям дает

$$\int_0^1 z_\theta(t)dg(t) = \frac{m_1 \int_0^T g(t)dt + m_2 \int_T^1 g(t)dt}{\sigma_\theta}.$$

Применяя теорему 2.3, получаем требуемое утверждение.
Следствие доказано.

Напомним, что, согласно введенной терминологии, критерий тем *лучше*, чем больше его приближенный бахадуровский наклон, вычисление которого основано на лемме 2.1.
Для сравнения статистик, введенных в предыдущем параграфе, рассмотрим байесовскую постановку задачи — будем предполагать, что T — случайная величина с известной функцией распределения $F_T(t)$.

Для формулировки следующей теоремы обозначим

$$s_k(u) = \sin(2\pi ku), \quad c_k(u) = \cos(2\pi ku).$$

Theorem 2.4

Если известна функция распределения F_T случайной величины T, то функция

$$-\sqrt{2}\sum_{k=1}^{\infty}\frac{k^{-1}\left(c_k(u)\int_0^1 s_k(u)dF_T(u) + s_k(t)\int_0^1 (1-c_k(u))dF_T(u)\right)}{\sqrt{\sum_{j=1}^{\infty} j^{-2}\left(\left(\int_0^1 s_k(u)dF_T(u)\right)^2 + \left(\int_0^1 (1-c_k(u))dF_T(u)\right)^2\right)}}$$

определяет наилучший критерий в классе критериев, удовлетворяющих условиям теоремы 2.2.

Доказательство.
В силу теоремы 2.2, при $\theta_0 \in \Theta_0$ имеет место сходимость по распределению к стандартному нормальному закону. Поэтому в силу леммы 2.1 требуется найти функцию g, на которой достигается максимум предела последовательности $J_n(g)/\sqrt{n}$.
Согласно следствию 2.1, нужно решить следующую оптимизационную задачу: найти функцию g, максимизирующую выражение

$$\left|(m_1 - m_2)\int_0^1 \left(\int_0^u g(t)dt\right) dF_T(u)\right|,$$

где

$$g(t) = \sum_{k=1}^{\infty}(a_k \cos(2\pi kt) + b_k \sin(2\pi kt))$$

(слагаемое $a_0 = 0$ в силу условия $\int_0^1 g(t)dt = 0$) при ограничении

$$\sum_{k=1}^{\infty}({a_k}^2 + {b_k}^2) = 2$$

(равенство Парсеваля, соответствующее условию $\int_0^1 g^2(t)dt = 1$).
Будем искать функцию $g(t)$ такую, что

$$\int_0^1 \left(\int_0^u g(t)dt \right) dF_T(u) \to \max$$

(функция $-g(t)$ дает симметричное решение).
Для отыскания условного экстремума составляем функцию Лагранжа, дифференцируем по a_k и b_k, приравниваем частные производные к 0 и используем условие нормировки (равенство Парсеваля). Приходим к решению, приведенному в формулировке теоремы.
Теорема доказана.

Corollary 2.2

При известном T функция

$$g_T(t) = \begin{cases} -K_1, & 0 \le t < T; \\ K_2, & T \le t \le 1 \end{cases}$$

определяет наилучший критерий в классе критериев, удовлетворяющих условиям теоремы 2.2. Здесь K_1 и K_2 определяются как решения системы уравнений

$$K_1 T = K_2(1 - T);$$

$$K_1{}^2 T + K_2{}^2(1 - T) = 1,$$

то есть

$$K_1 = \sqrt{T/(1-T)}, \qquad K_2 = \sqrt{(1-T)/T}.$$

Доказательство.

Подставляя в формулировку теоремы 2.4 функцию распределения $F_T(t) = \mathsf{I}\{T < t\}$, получаем разложение в ряд Фурье ступенчатой функции, приведенной в формулировке теоремы.
Доказательство завершено.

Remark 2.2

Важно отметить, что функция $g(t)$ не зависит от значений m_1, m_2. Однако она зависит от значения T, которое в ряде практических задач неизвестно. Следствие 2.2 показывает, что не существует равномерного по T наилучшего критерия. В случае $m_1 = 0$ к асимптотически оптимальному критерию приводит другая функция $g(t)$, зависящая от m_2.

Corollary 2.3

Если T имеет равномерное распределение на $[0;\ 1]$, то функция $g^{(u)}(t) = \sqrt{12}(t - 1/2)$, $t \in [0;\ 1]$, обеспечивает наилучший критерий.

Доказательство.
Подставляя $\int_0^1 \sin(2\pi ku)du = 0$ и $\int_0^1 (1 - \cos(2\pi ku))du = 1$ в утверждение теоремы 2.4, получаем

$$g(t) = -\sum_{k=1}^{\infty} \frac{\sqrt{2}\sin(2\pi kt)}{k\sqrt{\sum_{j=1}^{\infty} j^{-2}}} =$$

$$= -\sum_{k=1}^{\infty} \frac{\sqrt{12}}{k\pi} \sin(2\pi kt),$$

что совпадает с разложением функции $g^{(u)}(t)$ в ряд Фурье на $[0;\ 1]$.
Следствие доказано.

При отсутствии информации о распределении случайной величины наиболее логичным представляется использование либо статистики $J_n(g_{1/2})$ (в качестве предполагаемого значения T выбирается середина интервала), либо статистики $J_n(g^{(u)})$ (предполагается равномерное распределение случайной величины T). В силу теоремы 2.3, изучение этих статистик сводится к анализу частного случая функционалов от Z_n. Такой язык является вполне естественным. Так, статистики $J_n(g^{(u)})$ и $J_n(g_{1/2})$, введенные в следствиях 2.2 (при $T = 1/2$) и 2.3, приводят (с точностью до константы) к функционалам $|Z_n(1/2)|$ и $\left|\int_0^1 Z_n(t)dt\right|$ соответственно. Отметим, что последний функционал — это функционал $J_n^{|1|}$, введенный равенством (6) при $r = 1$.

В результате проведенного исследования (теорема 2.1, следствия 2.2, 2.3) нами отобраны следующие функционалы:
$J_n^{\infty} = \sup_{t\in[0;\ 1]} |Z_n(t)|$ — супремум модуля эмпирического моста;
$J_n^{|r|} = \left|\int_0^1 (Z_n(t))^r dt\right|^{1/r}$, $r = 1,\ 2,\ \ldots,$ — интегральные функционалы, совпадающие при четных r с L_r-нормами, а при нечетных оказывающиеся более предпочтительными в силу теоремы 2.1 (при $r = 1$ функционал совпадает с наилучшим в предположении о равномерности распределения случайной величины T — результат следствия 2.3 и теоремы 2.3);

$J_n(g_{1/2}) = |Z_n(1/2)|$ — оптимальная статистика для случая, когда разладка происходит в середине интервала.
Для этих функционалов найдем приближенные бахадуровские наклоны, определенные формулой (3). В формулировке теоремы использованы обозначения, введенные в лемме 2.3.

Theorem 2.5

Приближенные бахадуровские наклоны для функционалов J_n^{∞}, $J_n^{|r|}$, $J_n(g_{1/2})$ *равны соответственно*

$$c_{\infty}(\theta) = 4{M_\theta}^2, \qquad c_r(\theta) = \frac{2^{2-2/r} B^2\left(\frac{1}{r},\ \frac{1}{2}\right)}{r(r+2)^{1-2/r}(r+1)^{2/r}} {M_\theta}^2,$$

$$c_{g_{1/2}}(\theta) = \left(\min\left\{\frac{1}{T};\ \frac{1}{1-T}\right\}\right)^2 {M_\theta}^2.$$

Здесь $B(\cdot,\cdot)$ *— бета-функция.*

Доказательство.
Согласно лемме 2.1, $c(\theta) = a(j(\theta))^2$ п.н.
Для статистики J_n^∞ получаем

$$a_\infty = -\lim_{x\to\infty} \frac{2}{x^2} \ln \mathrm{P}\{\sup_{t\in[0;\ 1]} |W^0(t)| \geq x\} = -\lim_{x\to\infty} \frac{2}{x^2} \ln(2e^{-2x^2}) = 4,$$

так как $\sup_{t\in[0;\ 1]} |W^0(t)|$ имеет распределение Колмогорова.

Согласно лемме 2.3,

$$j_\infty(\theta) = \sup_{t \in [0;\ 1]} z_\theta(t) = M_\theta.$$

Отсюда $c_\infty(\theta) = 4{M_\theta}^2$.

Для статистики $J_n^{|r|}$ коэффициенты a_r найдены в [103] (см. также [55], §2.6):

$$a_r = \frac{2^{2-2/r} B^2\left(\frac{1}{r},\ \frac{1}{2}\right)}{r(r+2)^{1-2/r}}.$$

Вычислим $j_r(\theta)$:

$$j_r(\theta) = \left(\int_0^1 (z_\theta(t))^r \right)^{1/r} dt$$

$$= \left(\int_0^T \left(\frac{M_\theta}{T} \right)^r t^r dt + \int_0^{1-T} \left(\frac{M_\theta}{1-T} \right)^r t^r dt \right)^{1/r} = \frac{M_\theta}{(r+1)^{1/r}}.$$

Подставляя в выражение для $c_r(\theta)$, получаем соответствующее утверждение теоремы.

Вычислим $a_{g_{1/2}}$. Так как $\operatorname{Var} W^0(1/2) = 1/4$, то

$$a_{g_{1/2}} = -\lim_{x\to\infty} \frac{2}{x^2} \ln \mathrm{P}\{|W^0(1/2)| \geq x\} = 4.$$

Согласно лемме 2.2,

$$j_{g_{1/2}}(\theta) = |z_\theta(1/2)| = \min\left\{\frac{M_\theta}{2T};\ \frac{M_\theta}{2(1-T)}\right\}.$$

Подставляя в выражение для $c_{g_{1/2}}(\theta)$, получаем последнее утверждение теоремы.
Теорема доказана.

Отметим, что наилучшим является критерий, использующий L_∞-норму. Действительно,

$$\min\left\{\frac{1}{T};\ \frac{1}{1-T}\right\} \le 2,$$

а график $c_r(\theta)$ как функции переменной r (при $M_\theta = 1$) изображен на рис. 2.2 — функция строго возрастает.

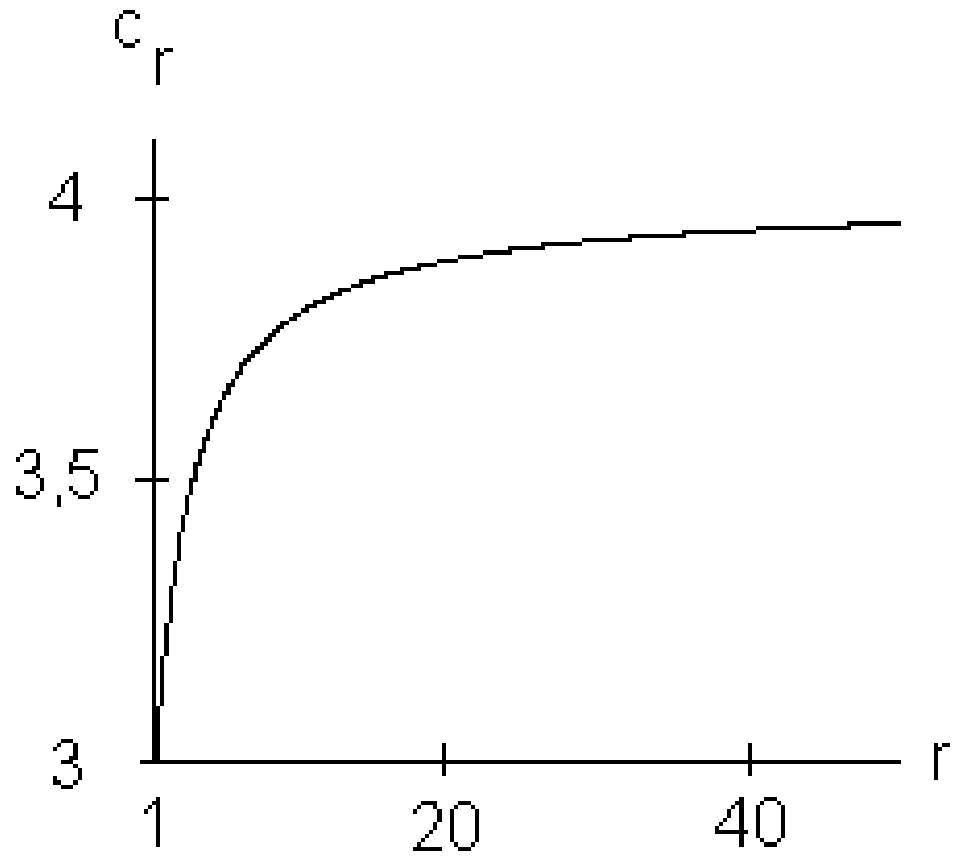

Рис. 2.2. График $c_r(\theta)$ как функции переменной r (при $M_\theta = 1$). Отметим, что $c_1(\theta) = 3M_\theta{}^2$.

Для анализа однородности текста необходимо определить способ, с помощью которого тексту сопоставляется последовательность чисел. Нами была разработана программа, которая сопоставляет тексту последовательность индикаторов вхождения слов текста в некоторый словарь.

Использовалось следующее техническое определение слова: слово русского языка — это последовательность русских букв и соединяющих их дефисов между любыми двумя символами, не являющимися русскими буквами. При этом словом не считается последовательность букв, начинающаяся или заканчивающаяся дефисом. Например, «сигма-алгебра» — это одно слово, «σ-алгебра» — вообще не слово, а «сигма - алгебра» — два слова (дефис, отделенный пробелами с двух сторон, считается за тире).

Важным этапом исследования является выбор словаря. Были испытаны различные варианты словарей. Выяснилось, что наилучшие результаты различения разнородных и однородных текстов получаются при использовании в качестве словаря авторского инварианта, введенного В. П. Фоменко и Т. Г. Фоменко в [67]. Это набор служебных слов, состоящий из 3 частей:

1) предлоги: в, на, с, за, к, по, из, у, от, для, во, без, до, о, через, со, при, про, об, ко, над, из-за, из-под, под;

2) союзы: и, что, но, а, да, хотя, когда, чтобы, если, тоже, или, то есть, зато, будто;

3) частицы: не, как, же, даже, бы, ли, только, вот, то, ни, лишь, ведь, вон, то-есть, нибудь, уже, либо.

В связи с введенным нами техническим определением слова мы исключили из авторского инварианта союз «то есть».

Всего было взято для исследования 25 текстов:

- Илья Ильф и Евгений Петров. Двенадцать стульев 12stul.txt
- Алексей Толстой. Аэлита aelit.txt
- Аркадий и Борис Стругацкие. Трудно быть богом be-god.txt
- Виктор Пелевин. Чапаев и Пустота chapaev.txt
- Михаил Булгаков. Собачье сердце dogheart.txt
- Ник Перумов. Эльфийский клинок (часть 1) elf-klin.txt
- Алексей Толстой. Гиперболоид инженера Гарина giper.txt
- Аркадий и Борис Стругацкие. Град обреченный grad.txt
- Борис Пастернак. Охранная грамота gramota.txt
- Виктор Пелевин. Жизнь насекомых insec.txt
- Михаил Веллер. Все о жизни life.txt
- Владимир Набоков. Лолита lolita.txt
- Владимир Набоков. Защита Лужина lughin.txt
- Михаил Булгаков. Мастер и Маргарита master.txt

- Николай Носов. Незнайка в Солнечном городе nezn-sun.txt
- Николай Носов. Незнайка на Луне nezn-moon.txt
- Иван Ефремов. На краю Ойкумены ojkum.txt
- Виктор Пелевин. Поколение "П"pel-g.txt
- Иван Ефремов. Туманность Андромеды tuman.txt
- Александр Волков. Урфин Джюс и его деревянные солдаты urfin.txt
- Александр Волков. Волшебник Изумрудного города volsh.txt
- Андрей Сергеевич Некрасов. Приключения капитана Врунгеля vrungel.txt
- Алексей Волков, Андрей Новиков. Мир спящего колдуна warlock.txt
- Борис Пастернак. Доктор Живаго zhiwago.txt
- Михаил Веллер. Приключения майора Звягина zwqgin.txt

Выбранные тексты исследовались поодиночке и в попарных комбинациях, полученных приписыванием одного текста к другому. Получилось $25 \times 24 = 600$ попарных комбинаций текстов, в том числе $3! + 9 \times 2 = 24$ комбинаций текстов одного автора и 576 комбинаций текстов разных авторов.

Для этих текстов были построены эмпирические мосты Z_n с помощью словаря авторского инварианта. Затем вычислялись максимальные по модулю отклонения эмпирического моста $|M_n| = \sup_{t\in[0;\ 1]} |Z_n(t)|$, и с помощью распределения Колмогорова отыскивался достигаемый уровень значимости

$$\varepsilon^* = 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2|M_n|^2}.$$

Задача состояла в том, чтобы научиться различать:
1) одно произведение одного автора от двух произведений разных авторов;
2) два произведения одного автора от двух произведений разных авторов.
Для этого нужно выбрать уровень значимости ε и соответствующее предельное значение M.

Отметим, что для произведения одного автора значения M не превосходят $3,4758$, что соответствует достигаемому уровню значимости $\varepsilon^* = 6,42 \cdot 10^{-11}$. Низкий достигаемый уровень значимости говорит о неполной адекватности модели выборки для описания появлений слов из выбранного словаря в тексте.

Рассмотрим специфику текстов, давших наибольшие значения M.

Значение $M = 3,4758$ относится к произведению М. Веллера «Все о жизни». Это произведение лежит за рамками чисто литературного жанра и является в значительной мере философским, что и обуславливает существенную неоднородность. Как мы заметим ниже, такие значения отклонений эмпирического моста характерны для собраний сочинений одного автора.

Значения $M = 3,0324$ и $M = 2,5002$ относятся к произведениям «Двенадцать стульев» и «Град обреченный». Отметим, что каждый из этих романов написан двумя авторами, что, по-видимому, и является причиной неоднородности.

Значение $M = 2,5758$ достигается романом «Доктор Живаго»; неоднородность здесь обусловлена наличием стихов в конце романа, то есть, как и в случае «Все о жизни», выходом за границы жанра.

Для всех остальных произведений значения отклонений не превосходят $1,8658$, что соответствует достигаемому уровню значимости $0,0019$. Отметим, что наилучшая однородность достигается для произведений В. Набокова и сравнительно небольших повестей.

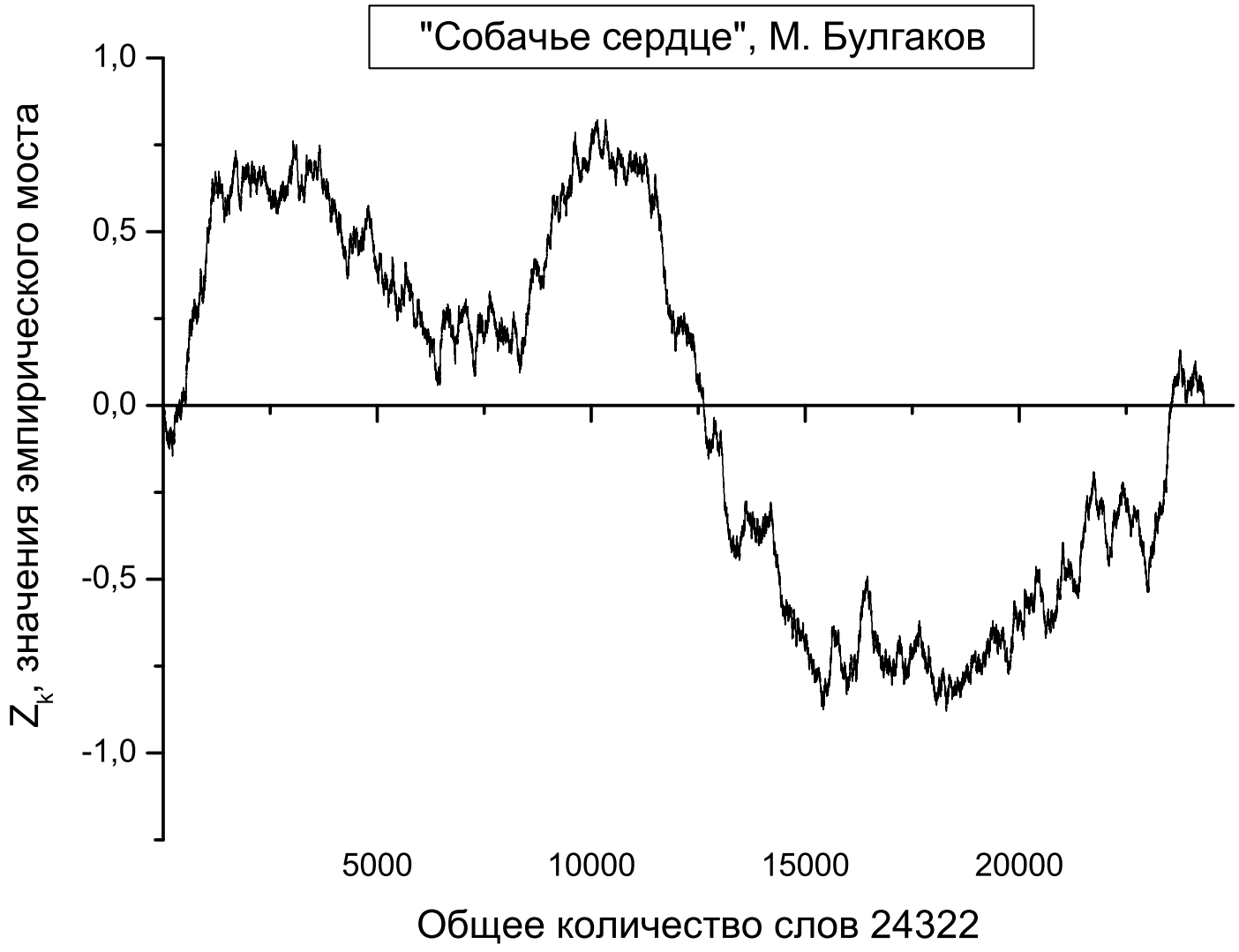

Рис. 2.3. График эмпирического моста для индикаторов служебных слов в повести М. Булгакова «Собачье сердце»

Табл. 2.1. Статистические характеристики разладки в текстах одного автора.

произведение	ν_n	T_n	M_n	n	ε^*
12stul	0,1757	17299	3,0324	81555	2,06013E-08
aelit	0,1616	24585	1,6169	40583	0,01072196
be-god	0,1793	19118	1,4080	48790	0,03793235
chapaev	0,1989	41149	1,3976	90712	0,040225523
dogheart	0,1874	18299	0,8796	24322	0,421550786
elf-klin	0,1948	42600	1,4521	135904	0,029485946
giper	0,1716	20425	1,2353	71760	0,094508689
grad	0,1915	35301	2,5002	107735	7,43849E-06
gramota	0,2163	14929	0,6578	26385	0,779863493
insec	0,2044	30248	1,5020	44423	0,021958702
life	0,2056	50536	3,4758	204368	6,41698E-11
lolita	0,2009	23144	0,7810	105601	0,575354971
luzhin	0,2018	19733	0,7643	51549	0,603220011
master	0,2169	93267	1,5886	112655	0,012853158

произведение	ν_n	T_n	M_n	n	ε^*
nezn-moon	0,2162	60006	1,2138	111824	0,105005271
nezn-sun	0,2173	51916	1,4898	63970	0,02361497
ojkum	0,1744	50047	1,6405	101940	0,009193475
pel-g	0,1935	31973	1,2439	66370	0,090576547
tuman	0,1698	51314	1,8658	85457	0,001893564
urfin	0,1892	31255	0,8909	37286	0,405414339
volsh	0,189	21264	1,6923	31800	0,00651071
vrungel	0,2051	23940	0,8794	34786	0,421761038
warlock	0,2065	16683	0,8966	32251	0,397389406
zhiwago	0,1994	62553	2,5758	152474	3,45377E-06
zwqgin	0,1833	47254	1,3930	112239	0,041259618

Анализ 576 попарных комбинаций текстов разных авторов дал следующие результаты. Только для 99 комбинаций текстов достигнутый уровень значимости превосходит 0,001 (эти комбинации приведены в таблице). Из них для 65 достигнутый уровень значимости больше 0,01, в том числе для 33 больше 0,1.

Табл. 2.2. Статистические характеристики разладки в текстах двух разных авторов.

произведения	ν_n	T_n	M_n	n	ε^*
warlock+vrungel	0,2058	40961	0,7024	67036	0,7073
dogheart+urfin	0,1885	18299	0,7631	61607	0,6052
vrungel+warlock	0,2058	23940	0,7779	67036	0,5804
lolita+chapaev	0,2	94656	0,7888	196312	0,5625
vrungel+lolita	0,2019	43563	0,8326	140386	0,4921
insec+lolita	0,2018	43445	0,8336	150023	0,4905
urfin+pel-g	0,192	37310	0,8416	103655	0,4781
insec+luzhin	0,203	64155	0,8588	95971	0,4521
luzhin+chapaev	0,1999	65359	0,8666	142260	0,4405
lolita+vrungel	0,2019	106445	0,8769	140386	0,4254
urfin+dogheart	0,1885	31255	0,9399	61607	0,3401
vrungel+luzhin	0,2031	54518	0,972	86334	0,3012
master+nezn-moon	0,2166	93267	0,9945	224478	0,2760
volsh+pel-g	0,1922	30647	0,9957	98169	0,2746

произведения	ν_n	T_n	M_n	n	ε^*
luzhin+vrungel	0,2031	42205	0,9985	86334	0,2716
zwqgin+dogheart	0,184	47254	1,0198	136560	0,2494
warlock+lolita	0,2022	27779	1,022	137851	0,2472
warlock+insec	0,2053	62498	1,0355	76673	0,2339
nezn-moon+gramota	0,2164	60006	1,0718	138208	0,2008
gramota+nezn-moon	0,2164	86390	1,0814	138208	0,1927
vrungel+insec	0,2047	65033	1,0854	79208	0,1894
be-god+zwqgin	0,1822	44592	1,0994	161028	0,1782
elf-klin+pel-g	0,1945	42600	1,1	202273	0,1777
warlock+luzhin	0,2036	51983	1,1062	83799	0,1729
nezn-moon+master	0,2166	205090	1,1515	224478	0,1410
zwqgin+volsh	0,1846	90503	1,1532	144038	0,1399
lolita+warlock	0,2022	110803	1,1583	137851	0,1366
giper+tuman	0,1707	123073	1,1628	157216	0,1338
dogheart+pel-g	0,1919	23023	1,165	90691	0,1324

произведения	ν_n	T_n	M_n	n	ε^*
luzhin+warlock	0,2036	42205	1,1806	83799	0,1231
dogheart+volsh	0,1883	45585	1,1909	56121	0,1173
chapaev+lolita	0,2	41149	1,2024	196312	0,1109
gramota+nezn-sun	0,2171	78300	1,2204	90354	0,1017
insec+vrungel	0,2047	30248	1,2286	79208	0,0977
be-god+ojkum	0,176	19308	1,2294	150729	0,0973
nezn-sun+master	0,2171	157236	1,2428	176624	0,0911
elf-klin+volsh	0,1936	123968	1,2624	167703	0,0825
zwqgin+urfin	0,1848	90503	1,269	149524	0,0798
vrungel+chapaev	0,2006	39252	1,3117	125497	0,0641
luzhin+insec	0,203	57470	1,3266	95971	0,0592
lolita+insec	0,2018	111522	1,3411	150023	0,0548
chapaev+elf-klin	0,1963	79231	1,3414	226615	0,0547
elf-klin+dogheart	0,1937	123968	1,3556	160225	0,0507
elf-klin+urfin	0,1935	123968	1,3575	173189	0,0502

произведения	ν_n	T_n	M_n	n	ε^*
master+nezn-sun	0,2171	93267	1,36	176624	0,0495
nezn-sun+gramota	0,2171	51916	1,3768	90354	0,0451
ojkum+giper	0,1733	122364	1,3816	173699	0,0440
pel-g+elf-klin	0,1945	108969	1,385	202273	0,0431
pel-g+urfin	0,192	31973	1,388	103655	0,0424
giper+12stul	0,1738	89058	1,3966	153314	0,0404
insec+warlock	0,2053	30248	1,3997	76673	0,0398
chapaev+luzhin	0,1999	41149	1,408	142260	0,0379
master+gramota	0,2168	93267	1,4094	139039	0,0376
pel-g+volsh	0,1922	31973	1,4149	98169	0,0365
volsh+dogheart	0,1883	21264	1,4448	56121	0,0308
gramota+master	0,2168	119651	1,4618	139039	0,0279
warlock+chapaev	0,2009	36717	1,4853	122962	0,0243
pel-g+dogheart	0,1919	31973	1,5068	90691	0,0213
gramota+warlock	0,211	24392	1,5262	58635	0,0190

произведения	ν_n	T_n	M_n	n	ε^*
zwqgin+be-god	0,1822	47254	1,5287	161028	0,0187
dogheart+zwqgin	0,184	71575	1,5713	136560	0,0143
elf-klin+chapaev	0,1963	59028	1,5714	226615	0,0143
pel-g+zhiwago	0,1978	66752	1,5732	218843	0,0142
be-god+dogheart	0,182	44592	1,5792	73111	0,0136
warlock+gramota	0,211	33692	1,6084	58635	0,0113
dogheart+zhiwago	0,1979	23023	1,6285	176795	0,0099
luzhin+elf-klin	0,1967	50430	1,6454	187452	0,0089
tuman+giper	0,1707	51314	1,648	157216	0,0087
elf-klin+zhiwago	0,1973	135217	1,7037	288377	0,0060
volsh+zhiwago	0,1977	30647	1,7331	184273	0,0049
master+warlock	0,2145	111875	1,7367	144905	0,0048
chapaev+vrungel	0,2006	68786	1,738	125497	0,0048
ojkum+12stul	0,175	119238	1,7541	183494	0,0043
volsh+zwqgin	0,1846	41566	1,7693	144038	0,0038

произведения	ν_n	T_n	M_n	n	ε^*
vrungel+elf-klin	0,1968	34780	1,7841	170689	0,0034
giper+ojkum	0,1733	93039	1,8007	173699	0,0031
warlock+nezn-moon	0,2142	91801	1,811	144074	0,0028
zhiwago+lolita	0,2	62553	1,8127	258074	0,0028
gramota+vrungel	0,21	34416	1,8214	61170	0,0026
elf-klin+luzhin	0,1967	59028	1,8243	187452	0,0026
giper+be-god	0,1748	76634	1,8332	120549	0,0024
nezn-moon+warlock	0,2142	107211	1,8333	144074	0,0024
pel-g+luzhin	0,1972	73393	1,8348	117918	0,0024
zhiwago+insec	0,2007	62553	1,8692	196896	0,0018
urfin+grad	0,191	124449	1,8709	145020	0,0018
chapaev+zhiwago	0,1993	153264	1,8713	243185	0,0018
zhiwago+warlock	0,2007	62553	1,8783	184724	0,0017
chapaev+dogheart	0,1964	79231	1,8856	115033	0,0016
chapaev+volsh	0,1963	79231	1,8879	122511	0,0016

произведения	ν_n	T_n	M_n	n	ε^*
urfin+zhiwago	0,1976	37668	1,8889	189759	0,0016
warlock+elf-klin	0,197	32247	1,8918	168154	0,0016
chapaev+warlock	0,2009	68786	1,8941	122962	0,0015
urfin+zwqgin	0,1848	47052	1,9002	149524	0,0015
dogheart+ojkum	0,177	24115	1,9044	126261	0,0014
dogheart+be-god	0,182	43439	1,9081	73111	0,0014
gramota+insec	0,2088	27472	1,9223	70807	0,0012
lolita+elf-klin	0,1975	95508	1,9381	241504	0,0011
elf-klin+vrungel	0,1968	59028	1,9398	170689	0,0011
volsh+grad	0,1909	67100	1,9453	139534	0,0010

Отметим, что здесь отклонение эмпирического моста от нуля может принимать большие значения, вплоть до 13.
При больших значениях M программа очень точно отыскивает границу между текстами. Так, текст «Аэлита»+«Мастер и Маргарита» делится точкой экстремума эмпирического моста на два фрагмента, первый из которых содержит лишь несколько строк из первой главы «Мастера и Маргариты».

Рассмотрим несколько способов определения критического значения ε, используемого в критерии выбора между гипотезами «текст написан одним автором» и «текст является объединением частей, принадлежащих разным авторам», или, как мы будем обозначать это более кратко, «текст однороден» и «текст неоднороден». Большое значение при выборе критического значения M и ε играют частоты ошибок первого и второго рода. Частота α_i^* ошибки i-го рода — это частота случаев, когда верная i-ая гипотеза отвергается статистическим критерием. Таким образом, для рассматриваемого критерия α_1^* — это отношение числа случаев, когда для текста одного автора значение M_n оказалось не меньше критического, к общему числу текстов 25. Аналогично, α_2^* — это отношение числа комбинаций текстов двух разных авторов, для которых M_n оказалось меньше критического, к общему числу текстов двух разных авторов, то есть к 576.

Если выбрать $M_{max} = 3,4758$ — максимальное для текстов одного автора, то $\alpha_1^* = 0$, но, как показывают подсчеты в этом случае, число текстов двух авторов, для которых значение M_n меньше, равняется 265, и частота ошибки второго рода $\alpha_2^* = 265/576 = 0,460$. Такой выбор M полезен, когда задача состоит в наиболее достоверном выявлении существенной неоднородности текста.

Табл. 2.3. Статистические характеристики разладки в двух текстах одного автора.

произведения	ν_n	T_n	M_n	n	ε^*
aelit+giper	0,1680	35709	2,2519	112342	7,87431E-05
be-god+grad	0,1877	47346	2,7534	156524	5,19998E-07
chapaev+insec	0,2007	96633	1,7162	135134	0,005531114
chapaev+pel-g	0,1966	106004	1,6702	157081	0,007553569
dogheart+master	0,2117	33497	4,1050	136976	4,61903E-15
giper+aelit	0,1680	62746	2,5383	112342	5,06749E-06
grad+be-god	0,1877	35301	2,9651	156524	4,61764E-08
gramota+zhiwago	0,2019	82610	3,6979	178858	2,65059E-12
insec+chapaev	0,2007	48889	1,3334	135134	0,057121241
insec+pel-g	0,1979	59715	2,5414	110792	4,90759E-06
life+zwqgin	0,1978	201188	7,3365	316606	3,55062E-47

произведения	ν_n	T_n	M_n	n	ε^*
lolita+luzhin	0,2011	23144	0,6787	157149	0,74641719
luzhin+lolita	0,2011	19733	0,5288	157149	0,942772469
master+dogheart	0,2117	113799	4,0042	136976	2,36791E-14
nezn-moon+nezn-sun	0,2168	60006	1,0853	175793	0,18951463
nezn-sun+nezn-moon	0,2168	51916	1,1149	175793	0,166378215
ojkum+tuman	0,1723	101292	1,3096	187396	0,064751696
pel-g+chapaev	0,1966	107518	1,7988	157081	0,003094454
pel-g+insec	0,1979	72281	2,6936	110792	9,9729E-07
tuman+ojkum	0,1723	106736	2,1736	187396	0,000157496
urfin+volsh	0,1891	58549	1,1550	69085	0,138740603
volsh+urfin	0,1891	21264	1,1218	69085	0,161325718
zhiwago+gramota	0,2019	152983	2,2754	178858	6,36613E-05
zwqgin+life	0,1978	162651	8,1311	316606	7,48248E-58

Другой подход состоит в выборе такого M, чтобы сумма частот ошибок была минимальна. Это байесовский (с равными априорными вероятностями гипотез) подход к построению статистического критерия. Расчеты показывают, что для этого в рассматриваемой задаче достаточно принять $M_b = 1,8658$, в этом случае $\alpha_1^* = 4/24 = 0,167$, $\alpha_2^* = 83/576 = 0,144$, $\alpha_1^* + \alpha_2^* = 0,311$.
Это же значение $M_{mm} = M_b$ обеспечивает и минимаксный критерий, минимизирущий максимум из α_1^* и α_2^*.

Проанализируем табл. 2.3. Отметим, что из 24 сочетаний текстов одного автора уровень $M_b = 1,8658$ превышен в 13 случаях, то есть частота ошибки первого рода при проверке гипотезы об одном авторе двух текстов против гипотезы о двух разных авторах составляет $13/24 = 0,54$. В двух случаях достигнутые величины M чрезвычайно велики (и, соответственно, достигаемые уровни значимости чрезвычайно малы): при прямом и обратном соединении произведений М. Веллера. Отметим, что причина явления в том, что «Все о жизни» является не художественным произведением, а философским трактатом, и не подчиняется закону сохранения служебных слов (в частности, междометий), характерного для художественных текстов. В остальном достигаемые уровни M не превосходят 4,1. Соответствующие достигаемые уровни значимости не меньше 10^{-15}.

Рассмотрим теперь два романа М. Шолохова.

Табл. 2.4. Статистические характеристики разладки в романах «Поднятая целина» и «Тихий Дон».

произведения	n	ν_n	T_n	M_n	ε^*
Поднятая целина	204955	0,208737	97523	6,728881	9,40152E-40
Тихий Дон	421854	0,182457	210643	9,012559	5,60831E-71

Здесь каждый из романов дает очень большие значения отклонения эмпирического моста, не характерные ни для одного произведения, ни для пары произведений одного автора. Разладка на графиках оказывается чрезвычайно четкой и характерной (рис. 2.4, 2.5). Для того, чтобы более глубоко и последовательно исследовать это явление, мы составили собрания сочинений трех авторов.

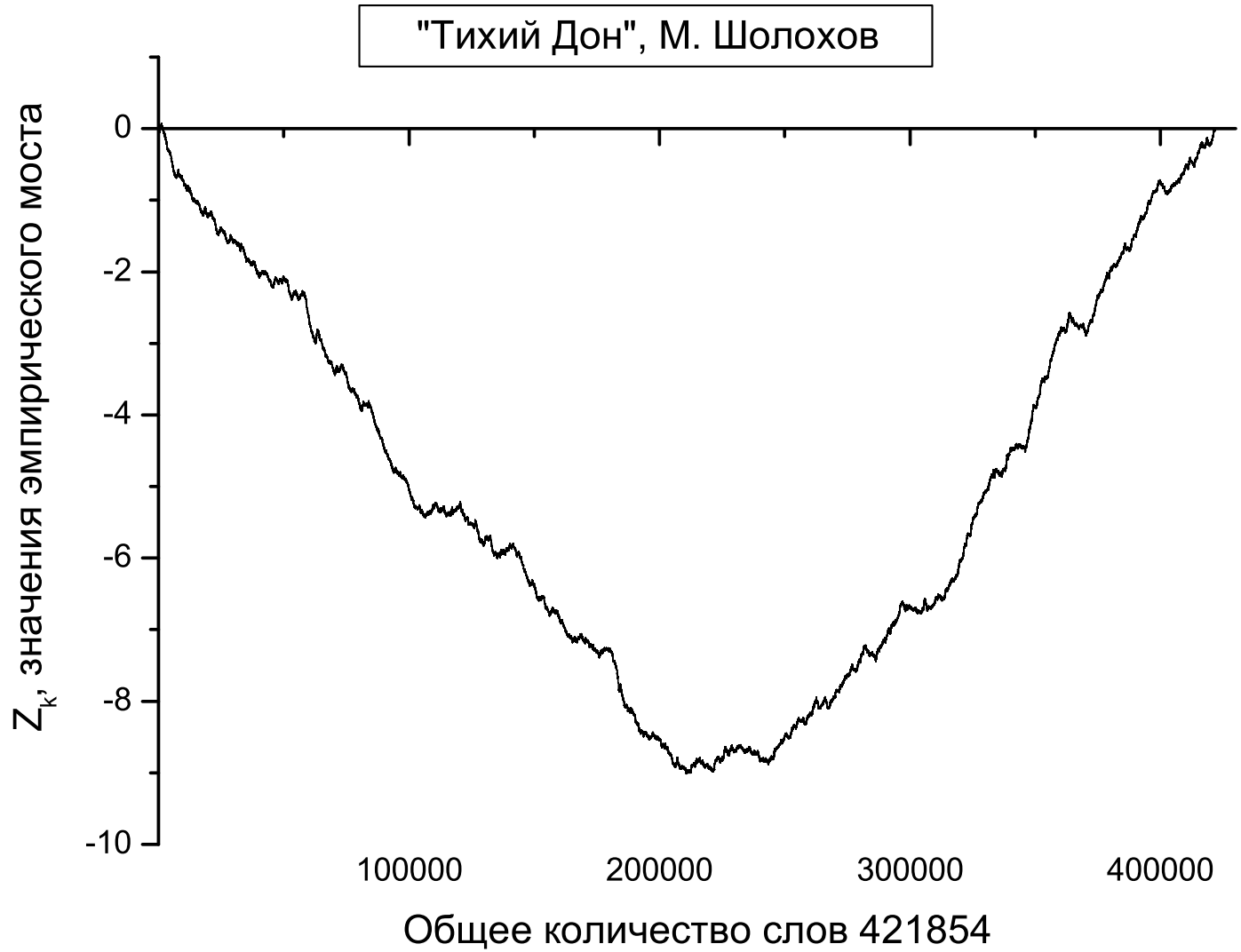

Рис. 2.4. График эмпирического моста для индикаторов служебных слов в романе «Тихий Дон»

Рис. 2.5. График эмпирического моста для индикаторов служебных слов в романе «Поднятая целина»

Нами были составлены следующие три электронных собрания сочинений. Собрания сочинений получены из романов, повестей и рассказов автора, приписанных друг к другу в последовательности, соответствующей биографии автора.
Федор Михайлович Достоевский. «Униженные и оскорбленные», «Преступление и наказание», «Идиот», «Бесы», «Братья Карамазовы», «Подросток».
Лев Николаевич Толстой. «Детство». «Отрочество». «Юность». «Севастопольские рассказы». «Казаки». «Война и мир». «Анна Каренина». «Воскресение».
Михаил Алексеевич Шолохов. «Тихий Дон», книги 1 и 2. «Поднятая целина», часть 1. «Тихий Дон», книги 3,4. «Судьба человека». «Поднятая целина», часть 2.

Исследовались параметры разладки в каждом из собраний сочинений целиком, затем во фрагментах собраний сочинений, полученных разрезанием текста в точках, соответствующих максимальному отклонению эмпирического моста от оси абсцисс. Фрагментам текста присваивались имена приписыванием цифр 1 и 2 к названию исходного текста, цифра 1 соответствовала первому фрагменту, цифра 2 — второму.

Табл. 2.5. Статистические характеристики разладки в собраниях сочинений и их фрагментах.

произведения	n	ν_n	T_n	M_n	ε^*
dostoevski-all	1185242	0,238991	347125	3,480106	6,04546E-11
dostoevski-all1	347126	0,233202	95110	2,354132	3,0716E-05
dostoevski-all11	95111	0,227289	26457	1,90939	0,001362556
dostoevski-all12	252016	0,235564	71341	2,925917	7,32906E-08
dostoevski-all2	838117	0,240361	339021	2,486013	8,56888E-06
dostoevski-all21	339022	0,238375	164295	3,938874	6,6851E-14
dostoevski-all22	499096	0,241673	150422	2,671881	1,25956E-06

произведения	n	ν_n	T_n	M_n	ε^*
sholohov-all	637698	0,191609	407236	13,176486	3,1391E-151
sholohov-all1	407237	0,181841	187777	7,454041	1,09612E-48
sholohov-all11	187778	0,171735	102127	1,340005	0,055127833
sholohov-all111	102128	0,169599	58139	1,493484	0,023101834
sholohov-all112	85651	0,174329	39488	1,81604	0,002731634
sholohov-all12	219460	0,190289	91620	3,135968	5,74209E-09
sholohov-all2	230462	0,209812	109306	6,982154	9,05536E-43
sholohov-all21	109307	0,197459	46789	1,962837	0,000900725
sholohov-all22	121156	0,221169	9642	1,326069	0,059380049
sholohov-all221	9643	0,241027	4905	1,071322	0,201223995
sholohov-all222	111514	0,219466	25802	1,585865	0,01307846

произведения	n	ν_n	T_n	M_n	ε^*
tolstoj-all	1058136	0,223217	127818	3,368991	2,76986E-10
tolstoj-all1	127819	0,23525	41641	3,218909	2,00103E-09
tolstoj-all11	41642	0,223572	25472	1,158508	0,136497001
tolstoj-all111	25473	0,227398	14108	0,776465	0,582863629
tolstoj-all112	16170	0,217506	10984	1,106582	0,172640324
tolstoj-all12	86178	0,240945	59255	1,182742	0,121865597
tolstoj-all121	59256	0,243449	35550	1,831481	0,002440632
tolstoj-all122	26923	0,235466	9035	1,510311	0,020880646
tolstoj-all2	930318	0,221937	252594	3,508171	4,08391E-11
tolstoj-all21	252595	0,216009	59322	1,856277	0,002032718
tolstoj-all211	59323	0,222403	38865	2,723605	7,20802E-07
tolstoj-all2111	38866	0,215343	21260	1,613999	0,010923504
tolstoj-all21111	21261	0,221507	5879	1,502185	0,021928378
tolstoj-all21112	17606	0,207899	9942	1,702325	0,006080468
tolstoj-all2112	20458	0,235845	15930	2,40669	1,8622E-05
tolstoj-all21121	15931	0,226745	7469	2,813355	2,66795E-07
tolstoj-all21122	4528	0,267946	1252	4,255296	3,74125E-16
tolstoj-all212	193273	0,213853	69503	3,055741	1,55072E-08
tolstoj-all22	677724	0,224477	531279	1,54569	0,016820304

Для собрания сочинений Достоевского значение $M = 3,48$ соответствует моменту биографии писателя, когда он начал играть в рулетку в Баден-Бадене. Это произошло во время работы над романом «Идиот». До этого момента характеристики его творчества более однородны ($M = 2,35$), особенно однороден фрагмент 1.1 (первые четыре главы «Униженных и оскорбленных»). После этого момента на графике (рис. 3.9) видны чередования разных периодов. Наименьшую однородность имеет фрагмент 2.1 (конец «Идиота» и «Бесы»), наибольшую — фрагмент 2.2 («Братья Карамазовы» и «Подросток»). Как для всего собрания сочинений, так и для его фрагментов значение M не превосходит 4,26, достигаемый уровень значимости не меньше 10^{-14}.

В творчестве Толстого нет такой резкой смены частоты служебных слов, как в творчестве Достоевского. Для всего собрания сочинений $M = 3,37$, первый фрагмент включает в себя «Детство», «Отрочество», «Юность» и «Севастопольские рассказы». Дальнейшее разбиение позволяет выявить очень однородные фрагменты 1.1 («Детство») и 1.2. Разбиение фрагмента 2 позволяет выявить сравнительно однородные фрагменты 2.1 (до конца третьей части «Войны и мира») и 2.2. Отметим, что по мере дальнейшей фрагментации не происходит снижения отклонений эмпирического моста от оси абсцисс. Как для всего собрания сочинений, так и для его (даже довольно мелких) фрагментов значение M не превосходит 4,26, достигаемый уровень значимости не меньше 10^{-16}.

Рис. 2.6. График эмпирического моста для индикаторов служебных слов в собрании сочинений Ф. Достоевского

Рис. 2.7. График эмпирического моста для индикаторов служебных слов в первом фрагменте собрания сочинений Ф. Достоевского

Рис. 2.8. График эмпирического моста для индикаторов служебных слов во втором фрагменте собрания сочинений Ф. Достоевского

Эмпирический мост для собрания сочинений Шолохова разительно отличается от эмпирических мостов собраний сочинений Достоевского и Толстого. Здесь $M = 13,18$, что соответствует ничтожному уровню значимости порядка 10^{-151}. Визуально можно выделить три периода: первый (фрагмент 1.1) с частотой служебных слов около 0,17, второй (фрагменты 1.2 и 2.1) с частотой 0,19–0,20, третий (фрагмент 2.2) с частотой 0,22. Однородность каждого из фрагментов вполне соответствует тому, что наблюдалось для собраний сочинений Достоевского и Толстого, а неоднородность всего собрания сочинений в целом перешагивает все мыслимые границы. Отметим, что границы фрагментов 1.1 и 2.2 четко привязаны к соответствующим произведениям. Конец фрагмента 1.1 привязан к концу второй книги «Тихого Дона». Из третьей книги во фрагмент вошло около 2 страниц текста, то есть с точностью до двух страниц отыскивается разладка в текстах, содержащих многие сотни страниц.

Аналогично и с фрагментом 2.2. Он включает целиком «Судьбу человека» и вторую часть «Поднятой целины», за исключением нескольких абзацев в начале «Судьбы человека», относящихся к фрагменту 2.1 и составляющих 3,5 страницы текста. Итак, проведенный анализ приводит к предположению, что под именем Шолохова скрываются два писателя: Шолохов-1 — автор двух первых книг «Тихого Дона», и Шолохов-2 — автор «Судьбы человека» и второй части «Поднятой целины». Остальные произведения представляются смесью фрагментов, принадлежащих этим двум авторам. В качестве альтернативы высказывается предположение о том, что Шолохов — уникальный автор, существенно изменивший частоту авторского инварианта в течение жизни.

Результаты анализа согласуются с выводами Г. Ермолаева [26], [27], Г. Кйецаа [37], О. В. Кукушкиной с соавт. [44], а также автора, известного под псевдонимом «Д» [18].

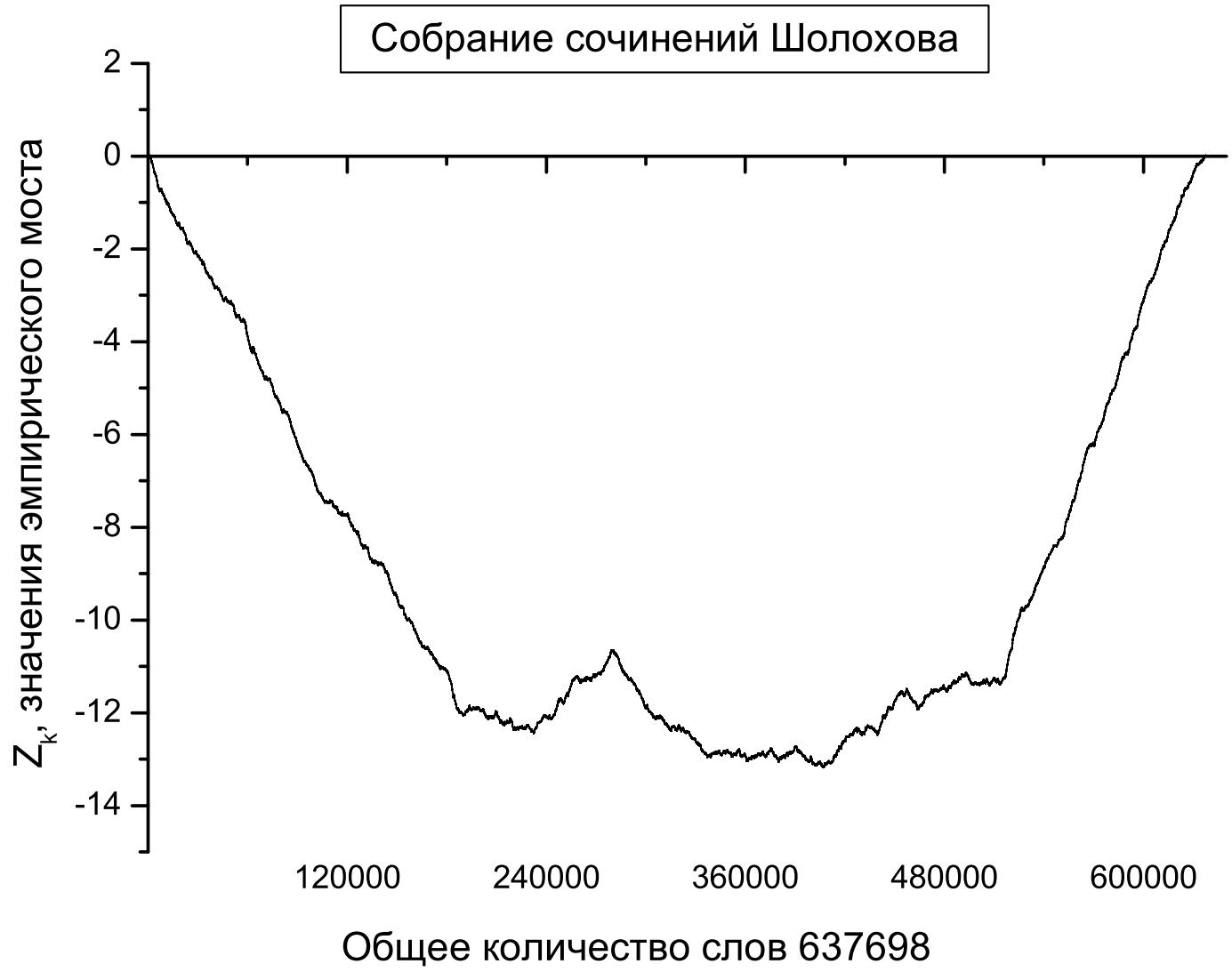

Рис. 2.9. График эмпирического моста для индикаторов служебных слов в собрании сочинений М. Шолохова

Рис. 2.10. График эмпирического моста для индикаторов служебных слов во фрагменте 1.1 собрания сочинений М. Шолохова

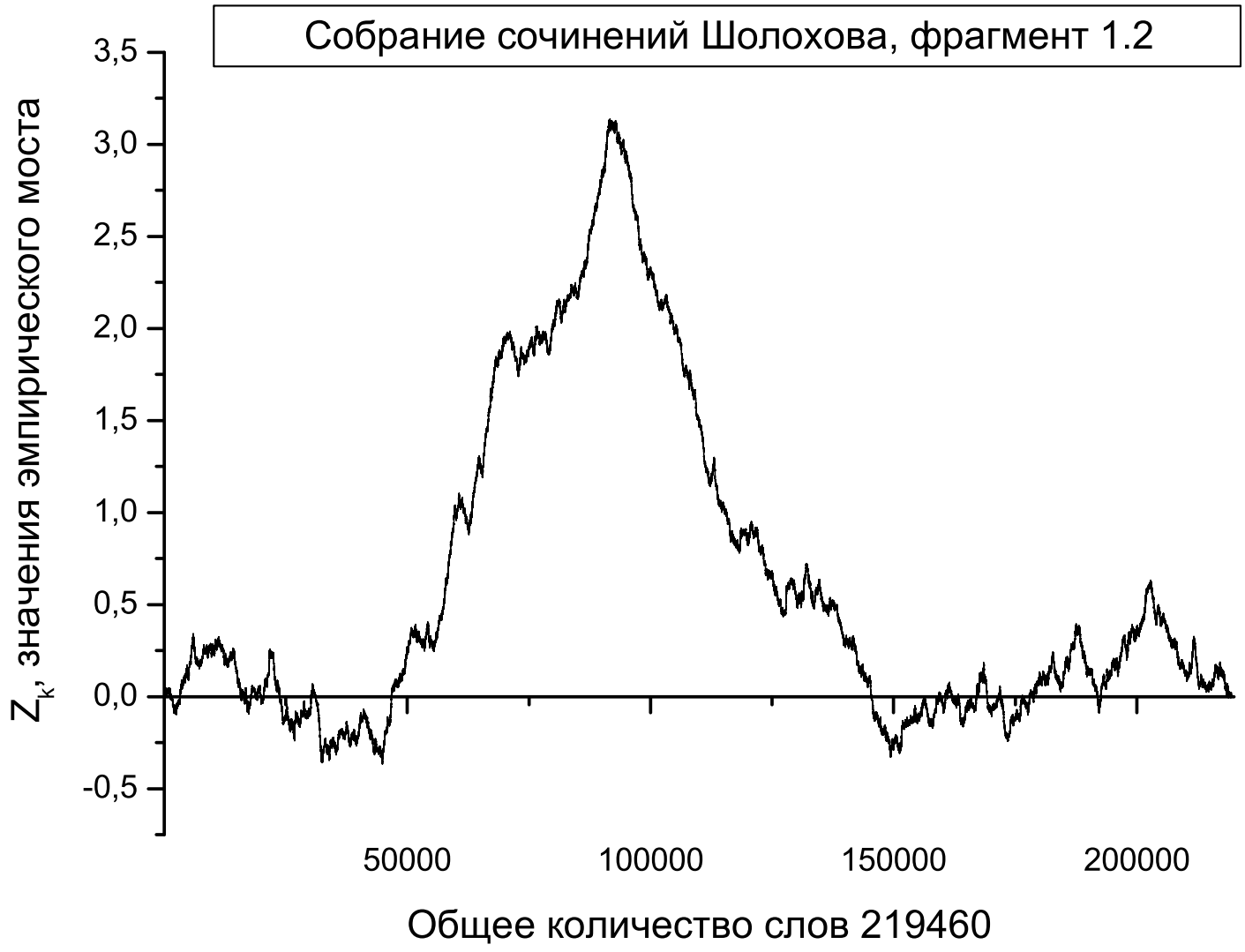

Рис. 2.11. График эмпирического моста для индикаторов служебных слов во фрагменте 1.2 собрания сочинений М. Шолохова

Рис. 2.12. График эмпирического моста для индикаторов служебных слов во фрагменте 2.1 собрания сочинений М. Шолохова

Рис. 2.13. График эмпирического моста для индикаторов служебных слов во фрагменте 2.2 собрания сочинений М. Шолохова

Выбор словаря изучался на следующем материале.
Взяты следующие 10 произведений 5 авторов 20 века:

- А. Толстой. «Аэлита» и «Гиперболоид инженера Гарина»;
- И. Ефремов. «На краю Ойкумены» и «Туманность Андромеды»;
- А. Волков. «Волшебник Изумрудного города» и «Урфин Джюс и его деревянные солдаты»;
- А. и Б. Стругацкие. «Трудно быть богом» и «Град обреченный»;
- В. Пелевин. «Жизнь насекомых» и «Поколение “П” ».

Эти произведения имеют объем от 32 229 до 109 618 слов. Были проанализированы как эти 10 произведений по отдельности, так и все 90 возможных попарных комбинаций, полученных приписыванием одного текста к другому.

Оказалось, что значения $|M_n|$ при использовании словаря предлогов получаются почти одинаковыми для текстов одного и двух авторов — это так называемое явление «слияния» всех авторов, описывающее не инвариант авторского стиля, а общие законы языка.
Напротив, при использовании словаря частиц значения $|M_n|$ меняются в широких пределах для разных классов текстов, что не позволяет разделять тексты разных авторов.

Только весь словарь авторского инварианта и словарь союзов в отдельности позволяют решать поставленные задачи. Именно, для словаря авторского инварианта получилось, что $|M_n|$ у одного произведения одного автора колеблется в пределах от 0,71 («Жизнь насекомых») до 2,23 («Град обреченный»), соответствующие достигаемые уровни значимости 0,698 и $5 \cdot 10^{-5}$. Если исключать «Град обреченный» из рассмотрения как произведение, написанное все же двумя авторами и существенно неоднородное, то следующим за ним будет «Поколение “П”» со значениями $|M_n| = 1,82$ и $\varepsilon^* = 0,0024$. В связи с этим выглядит логичным для анализа текстов полагать $\varepsilon = 0,001$ и соответственно $M = 1,97$. Действительно, тексты не могут обладать такой же однородностью, как последовательности независимых одинаково распределенных случайных величин, и использование стандартных уровней значимости 0,01–0,1 здесь никак не оправдано.

При рассмотрении 80 текстов, составленных из произведений разных авторов, оказалось, что значения $|M_n|$ колеблются в широких пределах от 0,90 до 10,84, но только для 11 текстов оказываются меньше, чем 1,97, то есть частота ошибки второго рода составляет 0,138.

При рассмотрении 10 текстов, составленных из двух произведений одного автора, оказывается, что значения $|M_n|$ меняются в пределах от 0,88 до 3,70, при этом для 5 текстов значения больше 1,97. Отметим, что это тексты действительно существенно разнородные (для Стругацких значения 3,70 и 3,63 в прямом и обратном порядке следования текстов соответственно, для А. Толстого 2,47 и 2,84, для Пелевина 1,27 и 2,19). Итак, при использовании критического уровня 0,001 удается обнаружить разнородные тексты, но при этом к разнородным критерий относит и тексты, написанные одним автором в разное время или в разных стилях.

Рассмотрим теперь, как изменятся результаты при использовании словаря, составленного только из союзов. Для текстов одного автора значения $|M_n|$ меняются от 0,91 («Урфин Джюс и его деревянные солдаты») до 2,37 («Аэлита»). После исключения рекордного значения 2,37 получаем пределы от 0,91 до 1,98 с достигаемыми уровнями значимости от 0,381 до 0,0007 соответственно. Выберем $\varepsilon = 0,0005$ и соответственно $M = 2,04$. В этом случае частота ошибки первого рода равна 0,1. При анализе текстов двух авторов оказывается, что $|M_n|$ меняется в пределах от 0,98 до 12,25, при этом для 11 авторов значения оказываются меньше, чем 2,04, то есть частота ошибки второго рода составляет 0,138, как и при использовании словаря авторского инварианта. При анализе текстов, составленных из двух произведений одного автора, оказывается, что значения супремума модуля эмпирического моста колеблются в пределах от 0,99 до 4,05, при этом только 3 значения превосходят уровень 2,04, то есть частота ошибки составляет 0,3.

Резюмируя результаты анализа, делаем вывод о том, что при использовании всего словаря авторского инварианта можно принять уровень значимости критерия равным 0,001, при использовании словаря союзов — равным 0,0005. При этом в обоих случаях один из 10 однородных текстов принимается за неоднородный, а 11 из 80 неоднородных текстов — за однородные, то есть при решении задачи различения одного текста одного автора и двух текстов двух разных авторов частоты ошибок составляют 0,1 и 0,138 соответственно.

При решении задачи различения двух произведений одного автора от двух произведений разных авторов в 5 случаях из 10 однородные тексты принимаются за неоднородные при использовании словаря авторского инварианта, в 3 случаях из 10 — при использовании словаря союзов. В этом отношении использование словаря союзов представляется предпочтительным.

Lection 3. Number of different words

We analyse correspondence of texts to a simple probabilistic model. The model assumes that the words are selected independently from an infinite dictionary. We count the numbers of different words in the text sequentially and get the process of the numbers of different words.

Our analysis is based on the fact that a text in any natural language can be divided into words. The source material for our analysis is a text with separated words and excluded punctuation. In addition, all capital letters (if any) are replaced by lowercase.
We test the hypothesis H that a text matches a simple probabilistic model. The model satisfies the following two assumptions:
1) the dictionary contains countably many words that are enumerated $i = 1, 2, \ldots$;
2) words are sampled from the dictionary in the i.i.d. fashion according to discrete distribution $F = \{p_i\}_{i \geq 1}$ where $p_i > 0$ is the probability of word i.
First, we study general properties of text statistics under these two assumptions. Second, we develop limit theorems for some classes of distributions F.

Bahadur (1960) and Karlin (1967) studied an infinite urn scheme, that is, n balls distributed to urns independently and randomly; there are infinitely many urns. Each ball goes to urn i with probability $p_i > 0$, $p_1 + p_2 + \ldots = 1$ (without loss of generality $p_1 \geq p_2 \geq \ldots$).
Denote by $R^*_{n,k}$ the total number of urns with at least k balls.
Let $R_n = R^*_{n,1}$ be the amount of different elements.
The total number of urns with k balls exactly is

$$R_{n,k} = R^*_{n,k} - R^*_{n,k+1}.$$

In other words, a text is a sequence of words $X_1, \ldots, X_n$. Variables X_i, $i = 1, \ldots, n$, take values in some infinite set $\mathcal{S}$. We associate the set with the set of all positive integers.
We study statistics that are measurable in sigma-algebra generated by events $\{X_i = X_j\}$, $1 \le i < j \le n$.
We model these statistics by a simple probability scheme:
$X_1, \ldots, X_n$ i.i.d. with distribution $\mathrm{P}(X_1 = i) = p_i$, $i = 1, 2, \ldots$.
We have in these designations

$$R_n = 1 + \sum_{i=2}^{n} \mathrm{I}\{X_i \neq X_j,\ j = 1, \ldots, i-1\};$$

$$R_{n,k} = \sum_{1 \le i_1 < \ldots < i_k \le n} \mathrm{I}\{X_{i_1} = \ldots = X_{i_k} \neq X_j,\ j \in \{1, \ldots, n\} \backslash \{i_1, \ldots, i_k\}\}.$$

Note that $R_n = \sum_{k=1}^{n} R_{n,k}$.

Lemma 3.1

For any $n = 1,\ 2,\ \ldots$:

$$\mathrm{E}R_n = \sum_{i=1}^{\infty}(1 - (1 - p_i)^n),$$

$$\mathrm{E}R_{n,k} = C_n^k \sum_{i=1}^{\infty} {p_i}^k (1 - p_i)^{n-k}, \quad k = 1, \ldots, n.$$

Proof

$$\mathrm{E}R_n = \sum_{i=1}^{\infty} \mathrm{P}(\{X_1 = i\} \cup \ldots \cup \{X_n = i\})$$

$$= \sum_{i=1}^{\infty} (1 - \mathrm{P}(\{X_1 \neq i\} \cap \ldots \cap \{X_n \neq i\})) = \sum_{i=1}^{\infty} (1 - (1 - p_i)^n);$$

$$\mathrm{E}R_{n,k} = \sum_{i=1}^{\infty} C_n^k \mathrm{P}(X_1 = \ldots = X_k = i,\ X_{k+1} \neq i,\ \ldots\ ,\ X_n \neq i)$$

$$= C_n^k \sum_{i=1}^{\infty} {p_i}^k (1 - p_i)^{n-k}.$$

The proof is complete.

Lemma 3.2

$\mathrm{Var} R_n \leq \mathrm{E} R_n$.

Proof

Note that

$$\mathrm{E}\,(R_n)^2 = \mathrm{E}\left(\sum_{i=1}^{\infty}(1 - \mathrm{I}\{X_1 \neq i, \ldots, X_n \neq i\})\right)^2$$

$$= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(1 - \mathrm{P}(X_1 \neq i, \ldots, X_n \neq i) - \mathrm{P}(X_1 \neq j, \ldots, X_n \neq j)$$

$$+\mathrm{P}(X_1 \neq i, \ldots, X_n \neq i, X_1 \neq j, \ldots, X_n \neq j))$$

$$= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\mathrm{I}\{i \neq j\}(1 - (1-p_i)^n - (1-p_j)^n + (1-p_i-p_j)^n) + \mathrm{E}R_n$$

$$\leq \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(1 - (1-p_i)^n - (1-p_j)^n + (1-p_i-p_j+p_ip_j)^n) + \mathrm{E}R_n$$

$$= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(1 - (1-p_i)^n)(1 - (1-p_j)^n) + \mathrm{E}R_n = (\mathrm{E}R_n)^2 + \mathrm{E}R_n.$$

So $\mathrm{Var}R_n \leq \mathrm{E}R_n$.

The proof is complete.

Lemma 3.3

$\mathsf{E}R_n \to \infty$ *as* $n \to \infty$.

Proof
From Lemma 3.1,

$$\mathrm{E}R_n = \sum_{i=1}^{\infty}(1-(1-p_i)^n).$$

Note that

$$1-(1-p_i)^n \geq 1-\varepsilon, \quad 0<\varepsilon<1,$$

if $n > \ln\varepsilon/\ln(1-p_i)$.
So, for any $K<\infty$, $0<\varepsilon<1$, let

$$n \geq \ln\varepsilon/\ln(1-p_{i_0}), \quad i_0 = \lfloor K/(1-\varepsilon)\rfloor + 1.$$

Then

$$\mathrm{E}R_n \geq \sum_{i=1}^{i_0}(1-\varepsilon) \geq K.$$

The proof is complete.

Lemma 3.4

$$\frac{\mathrm{E}R_n}{n} \to 0 \quad as \quad n \to \infty.$$

Proof

For any $\varepsilon > 0$ let i_0 such that

$$\sum_{i=i_0+1}^{\infty} p_i \leq \varepsilon.$$

Then

$$\mathrm{E}R_n = \sum_{i=1}^{i_0}(1-(1-p_i)^n) + \sum_{i=i_0+1}^{\infty}(1-(1-p_i)^n)$$

$$\leq i_0 + \sum_{i=i_0+1}^{\infty} np_i \leq i_0 + n\varepsilon.$$

So

$$\limsup_{n\to\infty} \frac{\mathrm{E}R_n}{n} \leq \varepsilon.$$

The proof is complete.

Theorem 3.1 (SLLN by Karlin, 1967)

$$\frac{R_n}{\mathrm{E}R_n} \to 1 \quad a.s.$$

Proof

From Lemma 3.2 we have

$$\mathrm{E}\left(\frac{R_n}{\mathrm{E}R_n}-1\right)^2=\frac{\mathrm{Var}R_n}{(\mathrm{E}R_n)^2}\le\frac{\mathrm{E}R_n}{(\mathrm{E}R_n)^2}=\frac{1}{\mathrm{E}R_n}. \tag{8}$$

The righthand term tends to 0 by Lemma 3.3.

From Chebyshev inequality we have weak convergence (in probability). Really, for any $\varepsilon>0$

$$\mathrm{P}\left(\left|\frac{R_n}{\mathrm{E}R_n}-1\right|\ge\varepsilon\right)\le\frac{1}{\varepsilon^2\mathrm{E}R_n}\to 0.$$

Bahadur proved it in 1960.

Let $M(t)$ be an extention of $\mathrm{E}R_n$ by linear interpolation for $t\ge 0$ with $\mathrm{E}R_0=0$:

$$M(t)=\mathrm{E}R_{\lfloor t\rfloor}+(t-\lfloor t\rfloor)(\mathrm{E}R_{\lfloor t\rfloor+1}-\mathrm{E}R_{\lfloor t\rfloor}).$$

So $M(t)$ is continuous and strictly acsending for $t\ge 0$ and tends to $+\infty$.

Determine t_k such that

$$M(t_k)=k^2,\quad k\ge 1.$$

Because of (8) and since $M(t)$ is monotonic we have

$$\mathsf{E}\left(\frac{R_{\lfloor t_k\rfloor+1}}{M(\lfloor t_k\rfloor+1)}-1\right)^2 \leq \frac{1}{M(\lfloor t_k\rfloor+1)} \leq \frac{1}{M(t_k)} = \frac{1}{k^2}.$$

It follows that

$$\sum_{k=1}^{\infty} \mathsf{E}\left(\frac{R_{\lfloor t_k\rfloor+1}}{M(\lfloor t_k\rfloor+1)}-1\right)^2 < \infty.$$

from which we may conclude that

$$\frac{R_{\lfloor t_k\rfloor+1}}{M(\lfloor t_k\rfloor+1)} \to 1 \quad \text{a.s.}$$

But $0 \leq R_{n+1} - R_n \leq 1$ for all integer n and consequently

$$\frac{R_{\lfloor t_k\rfloor}}{M(\lfloor t_k\rfloor)} \to 1 \quad \text{a.s.} \tag{9}$$

For each integer $n \geq 1$ we specify k depending on n satisfying

$$\lfloor t_k \rfloor \leq t_k \leq n \leq t_{k+1} \leq \lfloor t_{k+1} \rfloor + 1.$$

Since R_n is monotonic in n it follows that

$$R_{\lfloor t_k \rfloor} \leq R_n \leq R_{\lfloor t_{k+1} \rfloor + 1}$$

and therefore

$$\frac{R_{\lfloor t_k \rfloor}}{M(t_{k+1})} \leq \frac{R_n}{M(n)} = \frac{R_n}{\mathrm{E}R_n} \leq \frac{R_{\lfloor t_{k+1} \rfloor + 1}}{M(t_k)}. \tag{10}$$

But

$$1 \leq \frac{M(t_{k+1}+1)}{M(t_k)} \leq \frac{(k+1)^2 + M(\lfloor t_{k+1} \rfloor + 2) - M(\lfloor t_{k+1} \rfloor)}{k^2}$$

$$\leq \frac{(k+1)^2 + 2}{k^2} \to 1 \quad \text{as} \quad k \to \infty,$$

and this fact with relations (10) clearly imply on account of (9) that

$$\frac{R_n}{\mathrm{E}R_n} \to 1 \quad \text{a.s.}$$

The proof is complete.

Example: Shakespeare's sonnets

$n = 17516$, $R_n = 3258$

Most frequent words:

'and': 489,	'the': 444,
'to': 409,	'of': 371,
'my': 364,	'i': 341,
'in': 322,	'that': 320,
'thy': 266,	'thou': 234,
'with': 181,	'for': 171,
'is': 169,	'not': 167,
'but': 164,	'me': 164,
'a': 163,	'thee': 162,
'love': 160,	'so': 145,

— end of top 20 —

'be': 141,	'as': 121,
'all': 117,	'you': 110,
...	...

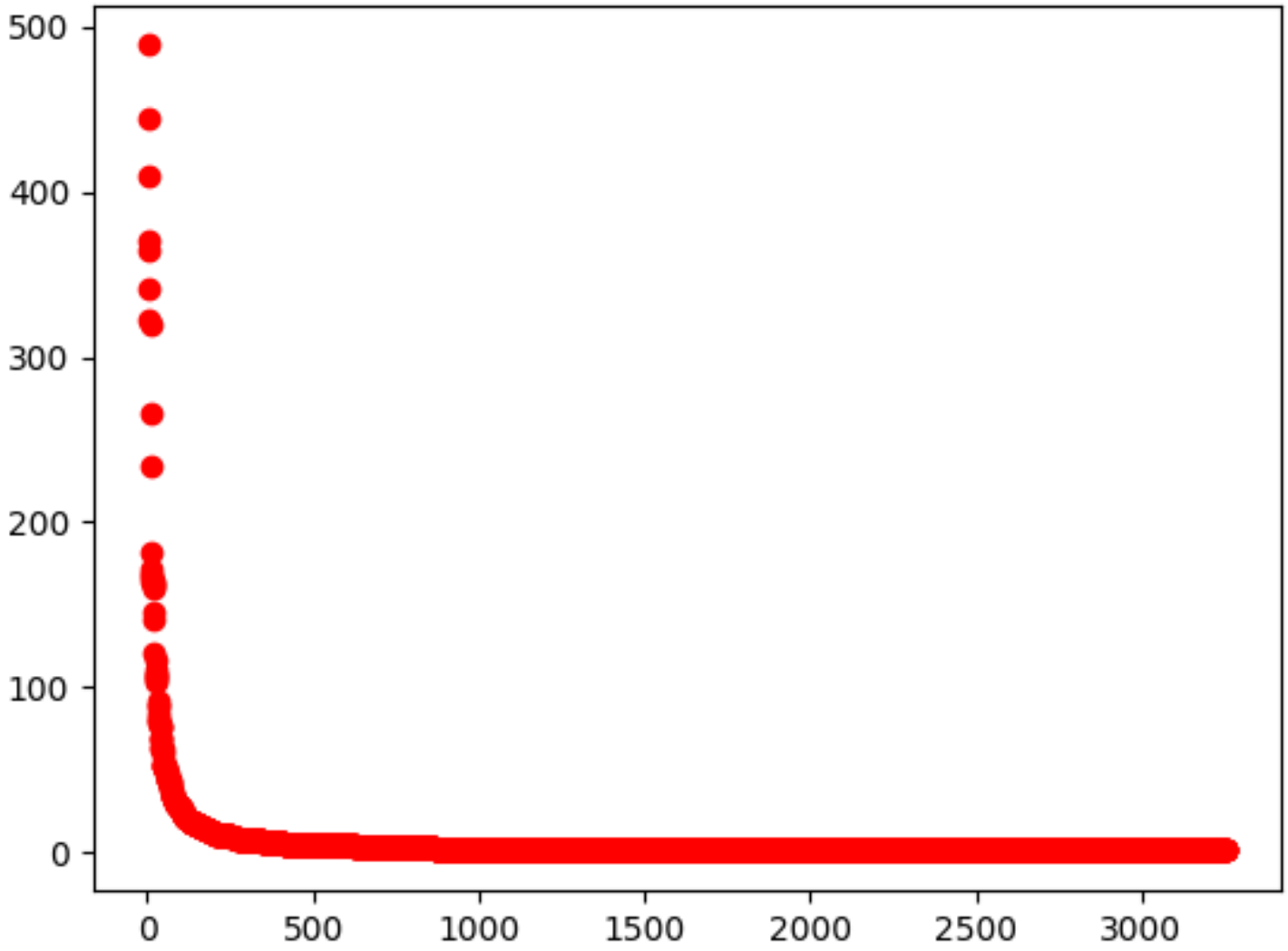

Fig. 3.1. Frequences of words in Shakespeare's sonnets

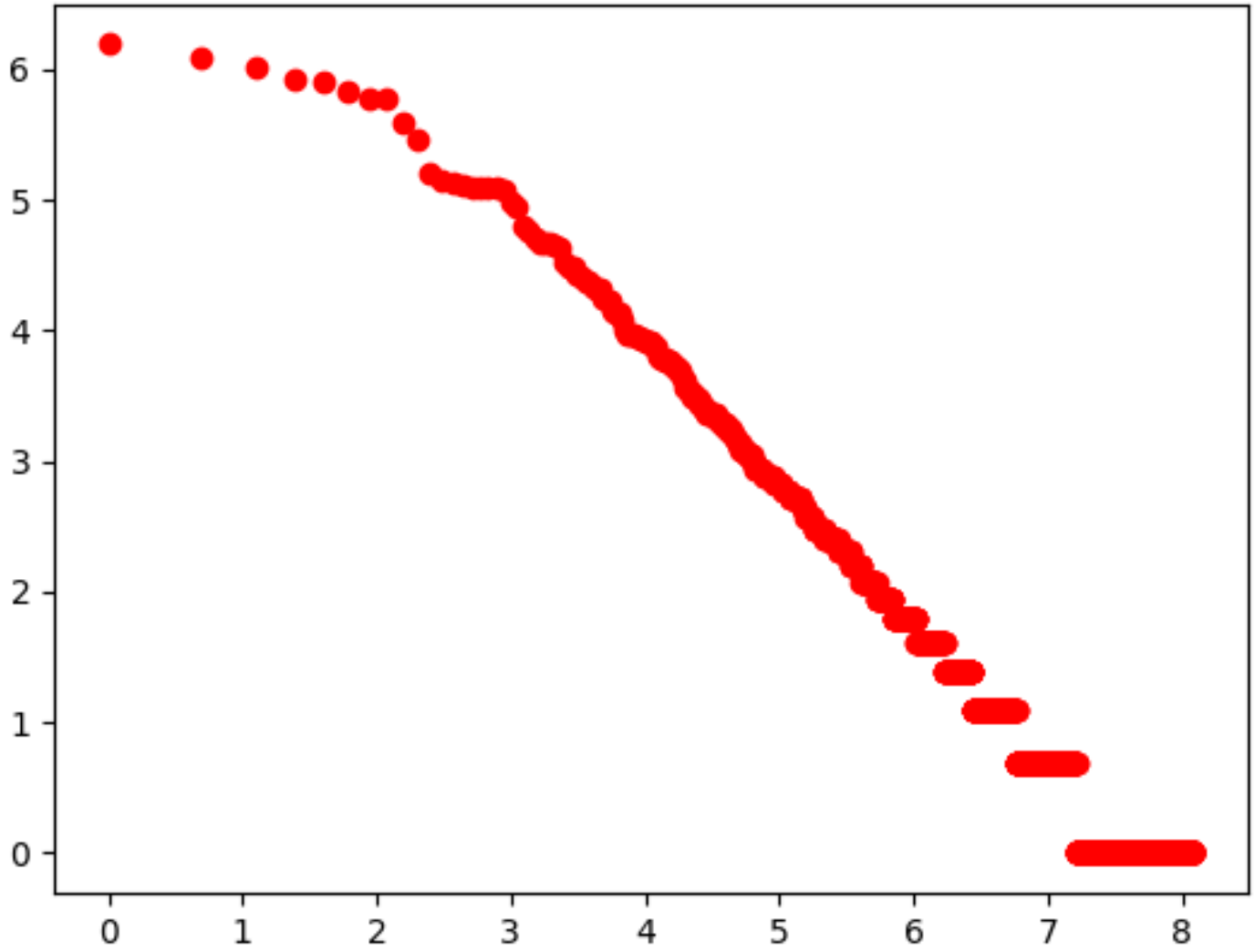

Fig. 3.2. Logs of frequences of words to logs of ranks in Shakespeare's sonnets (Zipfian diagram)

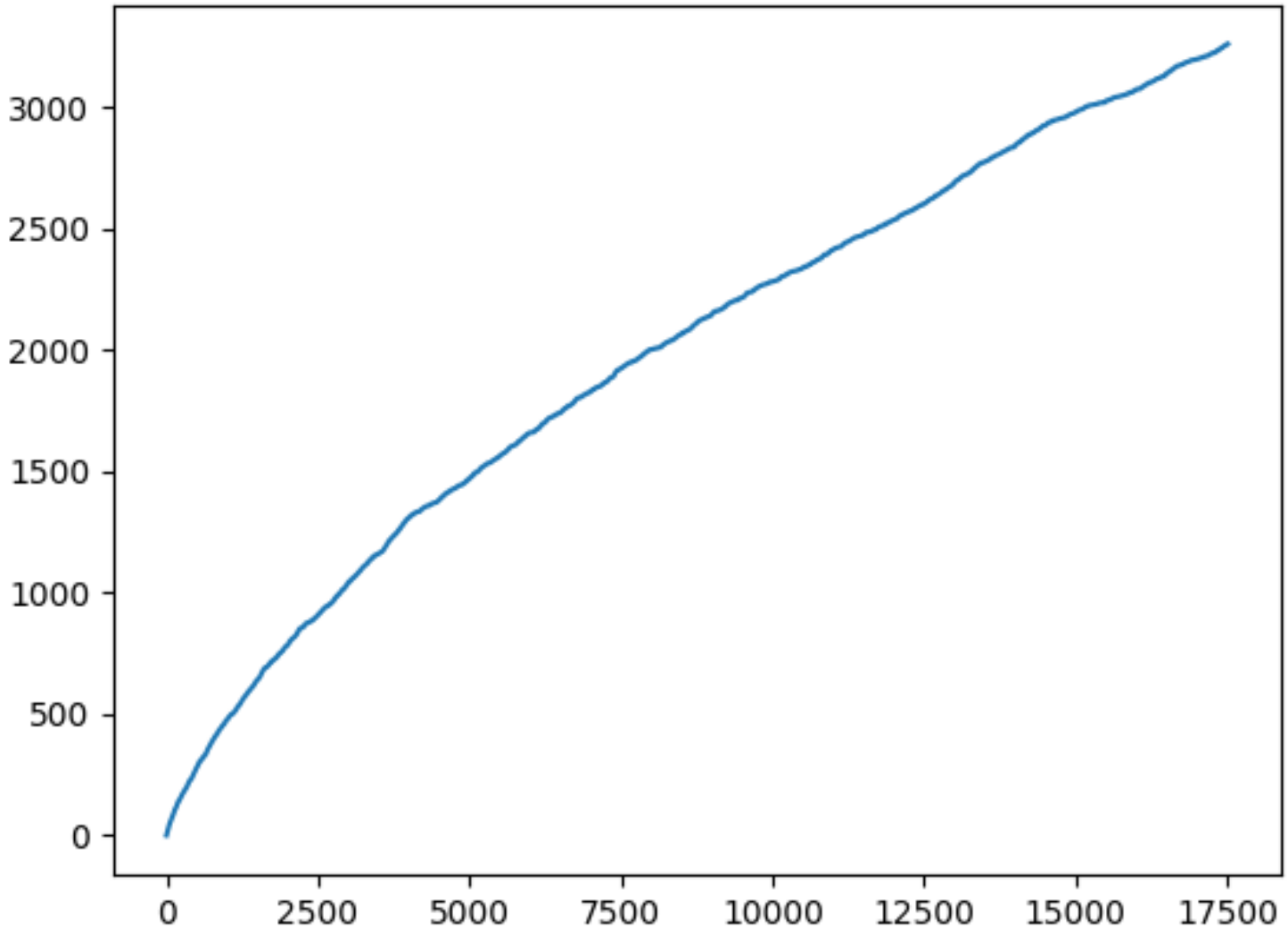

Fig. 3.3. The process of numbers of different words in Shakespeare's sonnets (Heaps' diagram)

From Lemma 3.1,

$$\mathrm{E}R_k = \sum_{i=1}^{\infty}(1-(1-p_i)^k).$$

We estimate the unknown expectation by

$$\widetilde{R}_k = \sum_{i=1}^{R_n}(1-(1-p_i^*)^k)$$

with

$$p_i^* = n_i/n,$$

n_i be the number of occurences of a word with rank i.
The next figure illustrates the badness of this approximation.

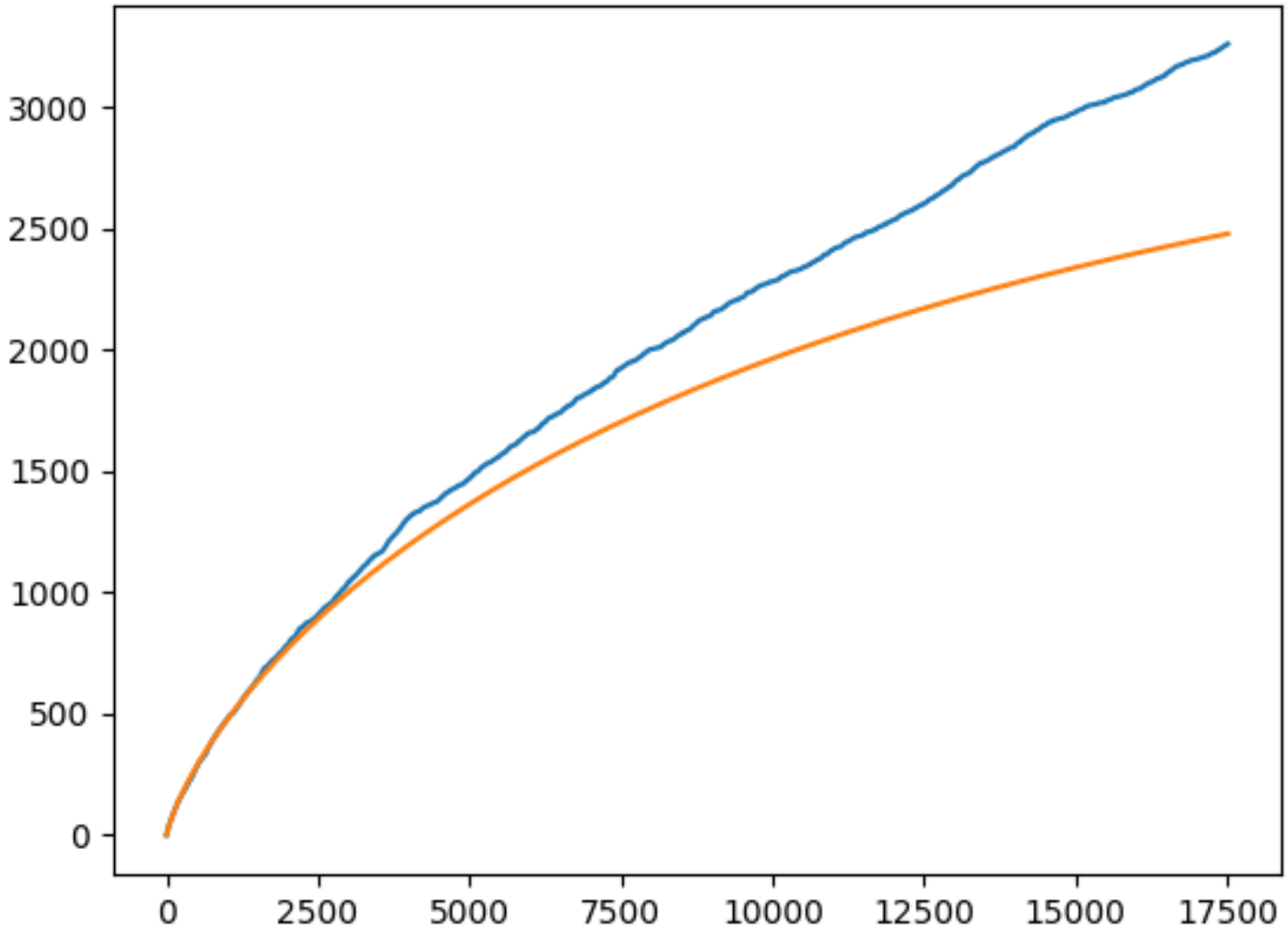

Fig. 3.4. The process of R_k with its empirical approximation $\widetilde{R}_k$

Lection 4. CLT for numbers of different words

Let (see Karlin (1967)) $\Pi = \{\Pi(t),\ t \geq 0\}$ be a Poisson process with parameter 1. We denote by $X_i(n)$ a number of balls in urn i. According to well-known property of splitting of Poisson flows, stochastic processes $\{X_i(\Pi(t)),\ t \geq 0\}$ are Poisson with intensities p_i and are mutually independent for different i's. The definition implies that

$$R^*_{\Pi(t),k} = \sum_{i=1}^{\infty} \mathsf{I}(X_i(\Pi(t)) \geq k), \quad R_{\Pi(t),k} = \sum_{i=1}^{\infty} \mathsf{I}(X_i(\Pi(t)) = k),$$

$$R_{\Pi(t)} = R^*_{\Pi(t),1}.$$

Let $\alpha(x) = \max\{j|\ p_j \geq 1/x\}$ and

$$\alpha(x) = x^{\theta} L(x), \tag{11}$$

$0 \leq \theta \leq 1$, as in Karlin (1967). Here $L(x)$ is a slowly varying function as $x \to \infty$.

Theorem 4.1 (Theorem 1 in Karlin (1967))

Let $\theta \in [0, 1)$. *Then*

$$\mathsf{E}R_{\Pi(t)} \sim \alpha(t)\Gamma(1-\theta)$$

as $t \to \infty$.

Proof

Clearly

$$R_{\Pi(t)} = \sum_{i=1}^{\infty} \mathsf{I}(X_i(\Pi(t)) \geq 1),$$

$$\mathsf{E}R_{\Pi(t)} = \sum_{i=1}^{\infty} \mathsf{P}(X_i(\Pi(t)) \geq 1) = \sum_{i=1}^{\infty} (1 - e^{-p_i t}).$$

In view of the definition of $\alpha(x)$ we may write

$$\mathsf{E}R_{\Pi(t)} = \int_0^{\infty} (1 - e^{-t/x}) d\alpha(x).$$

Integration by parts and a change of variable yelds

$$\mathsf{E}R_{\Pi(t)} = \int_0^{\infty} \frac{t}{x^2} e^{-t/x} \alpha(x)\, dx = \int_0^{\infty} \frac{1}{y^2} e^{-1/y} \alpha(ty)\, dy.$$

Substituting (11) we have

$$\mathsf{E}R_{\Pi(t)} = t^{\theta} \int_0^{\infty} \frac{y^{\theta}}{y^2} e^{-1/y} L(ty)\, dy.$$

A standard Abelian argument produces the result

$$\mathsf{E}R_{\Pi(t)} \sim t^{\theta} L(t) \int_0^{\infty} \frac{y^{\theta}}{y^2} e^{-1/y}\, dy = \alpha(t)\Gamma(1-\theta).$$

See Theorem A6.3.1 in Borovkov (2009) with $V(t) = \alpha(1/t)$ for details.
The proof is complete.

Exercise 4.1

Prove analogs of Lemmas 3.3 and 3.4 for $\mathsf{E}R_{\Pi(t)}$.

Theorem 4.2 (Theorem 1' in Karlin (1967))

Let $\theta \in [0, 1)$. *Then*

$$\mathsf{E}R_n \sim \alpha(n)\Gamma(1-\theta)$$

as $n \to \infty$.

Proof

We start from the elementary inequality

$$0 < e^{-np} - (1-p)^n < np^2 e^{-np}$$

valid for any $0 < p < 1$. So

$$0 < \mathrm{E}R_n - \mathrm{E}R_{\Pi(n)} = \sum_{i=1}^{\infty} (e^{-np_i} - (1-p_i)^n) \leq \sum_{i=1}^{\infty} np_i^2 e^{-np_i}$$

$$= \frac{2}{n} \sum_{i=1}^{\infty} \frac{(np_i)^2}{2!} e^{-np_i} < \frac{2}{n} \sum_{i=1}^{\infty} (1 - e^{-np_i}) = \frac{2}{n} \mathrm{E}R_{\Pi(n)} \to 0$$

as $n \to \infty$.

So $\mathrm{E}R_n - \mathrm{E}R_{\Pi(n)} \to 0$.

The proof is complete.

Theorem 4.3 (Formula (23) in Karlin (1967))

Let $\theta \in (0,1)$. *Then*

$$\mathsf{E}R_{\Pi(t),k} \sim \theta\alpha(t)\Gamma(k-\theta)/k!$$

as $t \to \infty$,

$$\mathsf{E}R_{n,k} \sim \theta\alpha(n)\Gamma(k-\theta)/k!$$

as $n \to \infty$.

Exercise 4.2

Prove Theorem 4.3 by lines of previous theorems using representation

$$\mathsf{E}R_{\Pi(t),k} = \int_0^\infty (t/x)^k e^{-t/x}/k! d\alpha(x).$$

Let for $t \in [0,1],\ k \geq 1$

$$Y^*_{n,k}(t) = \frac{R^*_{[nt],k} - \mathsf{E}R^*_{[nt],k}}{(\alpha(n))^{1/2}}, \qquad Z^*_{n,k}(t) = \frac{R^*_{\Pi(nt),k} - \mathsf{E}R^*_{\Pi(nt),k}}{(\alpha(n))^{1/2}},$$

$$Y_{n,k}(t) = \frac{R_{[nt],k} - \mathsf{E}R_{[nt],k}}{(\alpha(n))^{1/2}}, \qquad Z^{**}_{n,k}(t) = \frac{R^*_{\Pi([nt]),k} - \mathsf{E}R^*_{\Pi([nt]),k}}{(\alpha(n))^{1/2}},$$

$$K_{0,\theta} = -\Gamma(1-\theta),$$

$$K_{k,\theta} = \theta\Gamma(k-\theta) \ \text{ for } \ k > 0.$$

Theorem 4.4 (Theorem 4 in Karlin (1967))

Let $\theta \in (0,1)$. *Then* $(R_n - \mathrm{E}R_n)/B_n^{1/2}$ *converges weakly to standard normal distribution, where*

$$B_n = \Gamma(1-\theta)(2^\theta - 1)n^\theta L(n).$$

Theorem 4.5 (Theorem 5 in Karlin (1967))

Let $\theta \in (0,1)$, $r_1 < \ldots < r_\nu$ be ν positive integers. Then random vector $(Y_{n,r_1}(1), \ldots, Y_{n,r_\nu}(1))$ converges weakly to a multivariate normal distribution with zero expectation and covariances

$$c_{r_i,r_j} = \begin{cases} -\frac{\theta\Gamma(r_i+r_j-\theta)}{r_i!r_j!}2^{\theta-r_i-r_j}, & i \neq j; \\ \frac{\theta}{\Gamma(r_i+1)}\left(\Gamma(r_i-\theta) - 2^{-2r_i+\theta}\frac{\Gamma(2r_i-\theta)}{\Gamma(r_i+1)}\right), & i = j. \end{cases}$$

Dutko (1989) generalized Theorem 4.4 by proving asymptotic normality of R_n if $\mathrm{Var}R_n \to \infty$ as $n \to \infty$. This condition always holds if $\theta \in (0, 1]$ but can hold too for $\theta = 0$. Gnedin, Hansen and Pitman (2007) focus on study of conditions for convergence $\mathrm{Var}R_n \to \infty$. They also collect facts on the issue. Barbour and Gnedin (2009) extend Theorem 4.5 on the case of $\theta = 0$ if variances go to infinity. They found conditions for convergence of covariances to a limit and identified four types of limiting behavior of variances.

Barbour (2009) proved theorems on approximation of the number of cells with k balls by translated Poisson distribution. Key (1992, 1996) have studied the limit behavior of statistics $R_{n,1}$. Hwang and Janson (2008) proved local limit theorems for finite and infinite number of cells. Zakrevskaya and Kovalevskii (2001) proved consistency for one parametric family of an estimator of θ which is implicit function of R_n. Chebunin (2014) constructed R_n-based explicit parameter estimators for a wide range of one-parameter families and proved their consistency.

Note difference between the considered problem statement and the analysis of the behavior of the same statistics in a scheme of series when probabilities of getting in cells depend on the number of cells (in all cells with equal probability as a simplest variant). This is the problem of random allocation studied by Kolchin, Sevast'yanov and Chistyakov (1978). Various modifications of these problems were offered by Mikhailov (1980), Mahmoud (2010), Mahmoud and Smythe (2012), Chebunin (2014b).

However the proof of convergence of the Poisson approximation to the normal distribution can be constructed in a unified manner for growing and infinite numbers of cells. Mirakhmedov (2007) proved convergence of statistics that are more general than the number of different items or the number of elements encountered a determined number of times.

Durieu and Wang (2015) established a functional central limit theorem for a randomization of process R_n: indicators are multiplied randomly by ± 1 before summing. The limiting Gaussian process is a sum of independent self-similar processes in this case.

Theorem 4.6 (Theorem 1 in Chebunin and K. (2016))

Let $\theta \in (0,1)$, $\nu \geq 1$ is integer. Then process $\left(Y^*_{n,1}(t), \ldots, Y^*_{n,\nu}(t),\ 0 \leq t \leq 1\right)$ converges weakly in the uniform metrics in $D(0,1)$ to ν-dimensional Gaussian process with zero expectation and covariance function $(c^*_{ij}(\tau,t))_{i,j=1}^{\nu}$: for $\tau \leq t$, $i,j \in \{1,\ldots,\nu\}$ (taking $0^0 = 1$)

$$c^*_{ij}(\tau,t) = \begin{cases} \sum\limits_{s=0}^{i-1}\sum\limits_{m=0}^{j-s-1} \frac{\tau^s (t-\tau)^m K_{m+s,\theta}}{t^{m+s-\theta} s!m!} - \sum\limits_{s=0}^{i-1}\sum\limits_{m=0}^{j-1} \frac{\tau^s t^m K_{m+s,\theta}}{(t+\tau)^{m+s-\theta} s!m!}, & i < j; \\ t^\theta \sum\limits_{m=0}^{j-1} \frac{K_{m,\theta}}{m!} - \sum\limits_{s=0}^{i-1}\sum\limits_{m=0}^{j-1} \frac{\tau^s t^m K_{m+s,\theta}}{(t+\tau)^{m+s-\theta} s!m!}, & i \geq j; \end{cases}$$

$c^*_{ij}(\tau,t) = c^*_{ji}(t,\tau)$.

Exercise 4.3

Formulate it for R_n. Compare with Theorem 4.4.

The limiting ν-dimensional Gaussian process is self-similar with Hurst parameter $H = \theta/2 < 1/2$. Its first component coinside in distribution with the first component of the limiting process in Theorem 1 in Durieu and Wang (2015).
The above Karlin's theorems are partial cases of the theorem due to $c_{ij} = c^*_{ij}(1,1) - c^*_{i+1,j}(1,1) - c^*_{i,j+1}(1,1) + c^*_{i+1,j+1}(1,1)$.

Lemma 4.1 (Lemma 1 in Chebunin and K. (2016))

(i) *There exist* $n_0 \geq 1$, $C(\theta) < \infty$ *such that*

$$\mathsf{E}R_{\Pi(n\delta)}/\alpha(n) \leq C(\theta)\delta^{\theta/2} \quad \textit{for any} \quad \delta \in [0,1],\ n \geq n_0.$$

(ii) *Let* $\tau \leq t$, *then*

$$\mathsf{E}(R^*_{\Pi(t),k} - R^*_{\Pi(\tau),k}) \leq \mathsf{E}R_{\Pi(t-\tau)}, \quad k \geq 1.$$

(iii) *For any pair* ε, $\delta \in (0,1)$ *there exists integer* n_0 *such that*

$$\mathsf{P}(\forall t \in [0,1]\ \ \exists \tau : |\tau - t| \leq \delta,\ \Pi(n\tau) = [nt]) \stackrel{def}{=} \mathsf{P}(A(n)) \geq 1 - \varepsilon/2$$

for any $n \geq n_0$.

Proof

(i) From Karamata representation (Theorem 2.1, Appendix 6, inequality (A6.2.10) in Borovkov (2013)) exists $C_1(\theta) > 0$ such that for all x and $\delta \in (0, 1]$ under condition $x\delta \geq C_1(\theta)$ there is inequality $L(x\delta)/L(x) \leq 2\delta^{-\theta/2}$. As $\lim\limits_{x\to\infty} \mathsf{E}R_{\Pi(x)}/\alpha(x) = \Gamma(1-\theta)$ (Theorem 1 in Karlin (1967)), there is $C_2(\theta) < \infty$ such that $\mathsf{E}R_{\Pi(x)} \leq C_2(\theta)\alpha(x)$ as $x \geq x_0$.

Let $n\delta > C_1(\theta)$, then $\frac{\mathsf{E}R_{\Pi(n\delta)}}{\alpha(n)} \le C_2(\theta)\frac{(n\delta)^\theta L(n\delta)}{n^\theta L(n)} \le 2C_2(\theta)\delta^{\theta/2}$.
If $n\delta \le C_1(\theta)$ then $\mathsf{E}R_{\Pi(n\delta)}/\alpha(n) \le \mathsf{E}\Pi(n\delta)/\alpha(n) = n\delta/(n^\theta L(n))$.
If n is large enough then

$$n\delta/(n^\theta L(n)) \le n\delta/n^{\theta/2} = (n\delta)^{1-\theta/2}\delta^{\theta/2} \le (C_1(\theta))^{1-\theta/2}\delta^{\theta/2}.$$

(ii)

$$\mathsf{E}(R^*_{\Pi(t),k} - R^*_{\Pi(\tau),k})$$

$$= \sum_{i=1}^{\infty} \sum_{j=0}^{k-1} \mathsf{P}(X_i(\Pi(\tau)) = j)\mathsf{P}(X_i(\Pi(t) - \Pi(\tau)) \geq k - j)$$

$$\leq \sum_{i=1}^{\infty} \mathsf{P}(X_i(\Pi(t-\tau)) \geq 1) = \mathsf{E}R_{\Pi(t-\tau)}.$$

(iii) From SLLN for any $\varepsilon,\ \delta \in (0,1)$ there exists $T_0(\varepsilon,\delta) \geq 2$ such that $\mathsf{P}(\forall T \geq T_0,\ \Pi(T)/T \in [1-\delta/4, 1+\delta/4]) \geq 1-\varepsilon/2$. Let $t \in [0,1]$, then for all $n \geq T_0/\delta$ we have $\Pi(n(t+\delta)) \geq n(t+\delta)(1-\delta/4) \geq n(t+\delta/2) > [nt]$ with probability not lesser then $1-\varepsilon/2$.
If $t \in [\delta, 1]$ then for all $n \geq 2T_0/\delta$ we have $\Pi(n(t-\delta)) \leq \Pi(n(t-\delta/2)) \leq n(t-\delta/2)(1+\delta/4) \leq n(t-\delta/4) \leq nt-1 \leq [nt]$ with probability not lesser then $1-\varepsilon/2$.
The lemma is proved.

Proof of the theorem

Step 1 (covariances) Let $\tau \leq t$,

$$\widetilde{c}_{ij}(\tau, t) = \mathrm{cov}(R^*_{\Pi(\tau),i}, R^*_{\Pi(t),j})$$

$$= \sum_{k=1}^{\infty} \Big(\mathrm{P}(X_k(\Pi(\tau)) < i,\ X_k(\Pi(t)) < j)$$

$$- \mathrm{P}(X_k(\Pi(\tau)) < i)\mathrm{P}(X_k(\Pi(t)) < j) \Big).$$

If $i < j$ then

$$\widetilde{c}_{ij}(\tau, t) = \sum_{k=1}^{\infty} \sum_{s=0}^{i-1} \frac{(\tau p_k)^s}{s!} e^{-\tau p_k}$$

$$\times \left(\sum_{m=0}^{j-s-1} \frac{((t-\tau)p_k)^m}{m!} e^{-(t-\tau)p_k} - \sum_{m=0}^{j-1} \frac{(tp_k)^m}{m!} e^{-tp_k} \right)$$

$$= \int_0^{\infty} \sum_{s=0}^{i-1} \frac{\tau^s x^{-s}}{s!} e^{-\tau/x}$$

$$\times \left(\sum_{m=0}^{j-s-1} \frac{(t-\tau)^m x^{-m}}{m!} e^{-(t-\tau)/x} - \sum_{m=0}^{j-1} \frac{t^m x^{-m}}{m!} e^{-t/x} \right) d\alpha(x).$$

We integrate by parts and divide into two integrals, then we make change $y = x/t$ in the first integral, $y = x/(t+\tau)$ in the second one:

$$\widetilde{c}_{ij}(\tau, t) = \sum_{s=0}^{i-1} \sum_{m=0}^{j-s-1} \frac{\tau^s (t-\tau)^m t^{-m-s}}{s!m!}$$

$$\times \int_0^\infty ((m+s)y^{-m-s-1} - y^{-m-s-2}) e^{-1/y} \alpha(ty) dy$$

$$- \sum_{s=0}^{i-1} \sum_{m=0}^{j-1} \frac{\tau^s t^m (t+\tau)^{-m-s}}{s!m!}$$

$$\times \int_0^\infty ((m+s)y^{-m-s-1} - y^{-m-s-2}) e^{-1/y} \alpha((t+\tau)y) dy.$$

Analogously for $i \geq j$,

$$\widetilde{c}_{ij}(\tau, t) = \sum_{m=0}^{j-1} \left(\frac{1}{m!} \int_0^\infty (m y^{-m-1} - y^{-m-2}) e^{-1/y} \alpha(ty) dy \right.$$

$$- \sum_{s=0}^{i-1} \frac{t^m \tau^s (t+\tau)^{-m-s}}{s! m!}$$

$$\left. \times \int_0^\infty ((m+s) y^{-m-s-1} - y^{-m-s-2}) e^{-1/y} \alpha((t+\tau)y) dy \right).$$

For any integer $r \geq 0$,

$$\int_0^\infty y^{-r-2} e^{-1/y} \alpha(ty) dy \sim \alpha(t) \int_0^\infty y^{\theta-r-2} e^{-1/y} dy = \alpha(t)\Gamma(r+1-\theta)$$

as $t \to \infty$. Note that $\int_0^\infty (ry^{\theta-r-1} - y^{\theta-r-2}) e^{-1/y} dy = K_{r,\theta}$. So (because $\alpha(nt)/\alpha(n) \to t^\theta$ as $n \to \infty$),
$c_{ij}^*(\tau, t) = \lim_{n\to\infty} \widetilde{c}_{ij}(n\tau, nt)/\alpha(n)$.

Step 2 (convergence of finite-dimensional distributions)
Analogously to proof of Theorem 1 in Dutko (1989) we have for any $m \geq 1$, $0 < t_1 < t_2 < \ldots < t_m \leq 1$ triangle array of $m\nu$-dimensional random vectors (independent on k for any n) $\{(\mathsf{I}(X_k(\Pi(nt_j)) \geq i),\ i \leq \nu,\ j \leq m), k \leq n\}_{n \geq 1}$ satisfies Lindeberg condition.

Step 3 (relative compactness) Let for $\tau \le t$,

$$R^*_{\Pi(t),k} - R^*_{\Pi(\tau),k} = \sum_{i=1}^{\infty} \mathsf{I}(X_i(\Pi(t)) \ge k,\ X_i(\Pi(\tau)) < k) \stackrel{def}{=} \sum_{i=1}^{\infty} \mathsf{I}_i,$$

$P_i = \mathsf{P}(\mathsf{I}_i)$. Let $t_1 \le t_2$. Using Lemma 4.1(i,ii) we have

$$\mathsf{E}(Z^*_{n,k}(t_2) - Z^*_{n,k}(t_1))^2 = \sum_{i=1}^{\infty} \mathsf{E}(\mathsf{I}_i - P_i)^2/\alpha(n)$$

$$\le \sum_{i=1}^{\infty} P_i/\alpha(n) \le C(\theta)(t_2 - t_1)^{\theta/2}.$$

Using Step 1 and Theorem 1.4 in Adler (1990), we prove continuity of the limiting Gaussian process. So weak convergence in Skorokhod topology implies the same in the uniform topology. As $R^*_{\Pi(nt),k} - R^*_{\Pi([nt]),k} \leq \Pi([nt]+1) - \Pi([nt])$ a.s., and $\mathsf{E}(\Pi([nt]+1) - \Pi([nt])) = 1$ we have for any $\eta > 0$

$$\mathsf{P}(\sup_{0\leq t\leq 1} |Z^*_{n,k}(t) - Z^{**}_{n,k}(t)| > \eta)$$

$$\leq \mathsf{P}(\sup_{0\leq m\leq n} (\Pi(m+1) - \Pi(m) + 1) > \eta\sqrt{\alpha(n)})$$

$$\leq \sum_{m=0}^{n} \mathsf{E}e^{\Pi(m+1)-\Pi(m)+1} / e^{\eta\sqrt{\alpha(n)}} = (n+1)e^{e-\eta\sqrt{\alpha(n)}} \to 0$$

as $n \to \infty$.

So it is enough to prove relative compactness of $\{Z^{**}_{n,k}\}_{n\geq n_0}$. Let $t_2 - t_1 \geq 1/n$, then $[nt_2] - [nt_1] \leq n(t_2 - t_1) + 1 \leq 2n(t_2 - t_1)$. Let $\gamma = [8/\theta] + 1$. Using independence of terms and Rosenthal inequality we have

$$\mathsf{E}|Z^{**}_{n,k}(t_2) - Z^{**}_{n,k}(t_1)|^\gamma$$

$$\leq \frac{c(\gamma)}{(\alpha(n))^{\gamma/2}} \left(\sum_{i=1}^{\infty} \mathsf{E}|\mathsf{I}_i - P_i|^\gamma + \left(\sum_{i=1}^{\infty} \mathsf{E}(\mathsf{I}_i - P_i)^2 \right)^{\gamma/2} \right)$$

$$\leq \frac{c(\gamma)}{(\alpha(n))^{\gamma/2}} \left(\sum_{i=1}^{\infty} P_i + \left(\sum_{i=1}^{\infty} P_i \right)^{\gamma/2} \right)$$

$$\leq \frac{c(\gamma)}{(\alpha(n))^{\gamma/2}} \left(2n(t_2 - t_1) + (\mathsf{E}R_{\Pi(2n(t_2-t_1))})^{\gamma/2} \right) \leq C(\theta)(t_2 - t_1)^2.$$

Here $C(\theta)$ depends on its argument only. Above we used the fact that variance of an indicator is lesser than its expectation, inequality $\sum_i P_i \leq \mathsf{E}(\Pi([nt_2]) - \Pi([nt_1])) \leq 2n(t_2 - t_1)$, and Lemma 4.1(i,ii). Let $0 \leq t_2 - t_1 < 1/n$, then $[nt_1] = [nt]$ or $[nt_2] = [nt]$ for any $t \in [t_1, t_2]$. So

$$\mathsf{E}(|Z^{**}_{n,k}(t) - Z^{**}_{n,k}(t_1)|^\gamma |Z^{**}_{n,k}(t_2) - Z^{**}_{n,k}(t)|^\gamma) = 0 \leq (t_2 - t_1)^2.$$

So we have (see Billingsley (1999), Theorem 13.5) tightness.

Step 4 (approximation of the original process) From the relative compactness of distributions of processes $\{Z^*_{n,k}\}_{n\geq n_0, k\geq 1}$ we get that for every pair $\varepsilon > 0$, $\eta > 0$ there exist $\delta \in (0,1)$ and integer n_0 such that

$$\mathrm{P}\Big(\sup_{|t-\tau|\leq\delta} \big|Z^*_{n,k}(\tau) - Z^*_{n,k}(t)\big| \geq \eta\Big) \leq \varepsilon/2$$

for all $n \geq n_0$.

Then (as $\mathsf{P}(Y^*_{n,k}(t) = Z^*_{n,k}(\tau) | \Pi(n\tau) = [nt]) = 1$) we have (using Lemma 4.1(iii))

$$\mathsf{P}(\sup_{0 \le t \le 1} \left| Y^*_{n,k}(t) - Z^*_{n,k}(t) \right| \ge \eta)$$

$$\le \mathsf{P}(\sup_{0 \le t \le 1} \left| Y^*_{n,k}(t) - Z^*_{n,k}(t) \right| \ge \eta, A(n)) + \varepsilon/2$$

$$\le \mathsf{P}(\sup_{|t-\tau| \le \delta} \left| Z^*_{n,k}(\tau) - Z^*_{n,k}(t) \right| \ge \eta) + \varepsilon/2 \le \varepsilon.$$

The theorem is proved.

Forward and backward processes of numbers of different words

Hamlet: To be or not to be

hamlet to be or not to be

k	0	1	2	3	4	5	6	7
R_k	0	1	2	3	4	5	5	5

be to not or be to hamlet

k	0	1	2	3	4	5	6	7
R'_k	0	1	2	3	4	4	4	5

An infinite urn scheme

There is a countably infinite dictionary where the words are numbered $1,\ 2,\ \ldots$. Words are chosen one-by-one independently of each other.
Let X_i be the number of the word at ith position, $1 \leq i \leq n$.

$$\mathrm{P}(X_i = j) = p_j > 0, \quad j \geq 1$$

$$p_1 + p_2 + \ldots = 1$$

$$p_1 \geq p_2 \geq \ldots$$

General theorems

Denote by R_n the number of different words in the text of length n

$$\mathrm{E}R_n = \sum_{i=1}^{\infty}(1-(1-p_i)^n)$$

$$\mathrm{Var}R_n \leq \mathrm{E}R_n$$

$$\mathrm{E}R_n \to \infty, \quad \mathrm{E}R_n/n \to 0$$

[Bahadur, 1960]

$$R_n/\mathrm{E}R_n \overset{a.s.}{\to} 1$$

[Karlin, 1967]

A regular case

Regularity condition:

$$\alpha(x) := \max\{k > 0 : \ p_k \geq 1/x\} = x^{\theta} L(x), \quad 0 < \theta < 1,$$

$L(\cdot)$ is the slowly varying function of the real argument:
$L(tx)/L(x) \to 1$ as $x \to +\infty$ for any real $t > 0$.

Equivalent condition:

$$p_i = i^{-1/\theta} l(i),$$

$l(\cdot)$ is the another slowly varying function.
The model is the elementary probability model that corresponds to the Zipf's Law (Zipf, 1936) of power decreasing of word probabilities.

Theorems under the regularity condition

Karlin (1967): $(R_n - \mathrm{E}R_n)/\sqrt{\mathrm{Var}R_n}$ converges weakly to the standard normal distribution,

$$\mathrm{E}R_n \sim \Gamma(1-\theta)\kappa(n),$$

$$\mathrm{Var}R_n/\mathrm{E}R_n \to 2^\theta - 1,$$

$\Gamma(\cdot)$ is the Euler gamma.
So $(R_n - \mathrm{E}R_n)/\sqrt{\mathrm{E}R_n}$ converges weakly to the centered normal distribution with variance $2^\theta - 1$.
Chebunin and Kovalevskii (2016):

$$Z_n = \{Z_n(t),\ 0 \le t \le 1\} = \{(R_{[nt]} - \mathrm{E}R_{[nt]})/\sqrt{\mathrm{E}R_n},\ 0 \le t \le 1\}$$

converges weakly in $D(0,1)$ with uniform metrics to a centered Gaussian process Z_θ with continuous a.s. sample paths and covariance function

$$K(s,t) = (s+t)^\theta - \max(s^\theta, t^\theta).$$

New theorem under the regularity condition

Theorem 4.7 (for joint distribution)

If the regularity condition holds then $(Z_n, Z'_n) = \{(Z_n(t), Z'_n(t)),\ 0 \le t \le 1\}$ *converges weakly in the uniform metrics in* $D(0,1)^2$ *to 2-dimensional Gaussian process* (Z, Z') *with zero expectation and covariance function*

$$\mathrm{E}Z(s)Z(t) = \mathrm{E}Z'(s)Z'(t) = K(s,t), \quad \mathrm{E}Z(s)Z'(t) = K'(s,t),$$

$$K'(s,t) = ((s+t)^{\theta} - 1)1(s+t > 1).$$

Corollary under the regularity condition

Corollary 4.1 (for the difference of processes)

If the regularity condition holds then

$$J_n = \frac{\sum_{k=1}^{n}(R_k - R'_k)}{n\sqrt{R_n}}$$

converges weakly to a centered normal random variable with variance $\frac{\theta}{\theta+2}$.

The Corollary gives the opportunity to test the homogeneity of the sample using any consistent estimate θ^* of parameter θ. Various classes of such estimates have been obtained and analysed by Hill (1975), Nicholls (1978), Zakrevskaya and Kovalevskii (2001, 2019), Guillou and Hall (2002), Ohannessian and Dahleh (2012), Chebunin (2014), Chebunin and Kovalevskii (2019a, 2019b), Chakrabarty et al. (2020).

The p-value is calculated using the tail of the standard normal distribution and the observed value J_{obs} of J_n:

$$\text{p-value} = 2\overline{\Phi}\left(|J_{obs}|\sqrt{1+2/\theta^*}\right).$$

Parameter's estimation

$$\theta_n = \int_0^1 \log^+ R_{[nt]}\, dA(t), \quad \theta'_n = \int_0^1 \log^+ R'_{[nt]}\, dA(t),$$

here $\log^+ x = \max(\log x, 0)$. Function $A(\cdot)$ has bounded variation and

$$A(0) = A(1) = 0, \quad \lim_{x \downarrow 0} \log x \int_0^x |dA(t)| = 0, \quad \int_0^1 \log t\, dA(t) = 1. \tag{12}$$

Let

$$\widehat{\theta} = (\theta_n + \theta'_n)/2.$$

Theorem 4.8 (consistence)

Let $p_i = i^{-1/\theta} l(i, \theta)$, $\theta \in [0, 1]$, and $l(x, \theta)$ is a slowly varying function as $x \to \infty$. Then the estimator $\widehat{\theta}$ is strongly consistent.

Proof

Since $p_i i^{1/\theta}$ is a slowly varying function as $i \to \infty$, we have $\alpha(x) = x^\theta L(x, \theta)$, $L(x, \theta)$ is a slowly varying function as $x \to \infty$ (Karlin, 1967).

Take a sequence $\{\delta_n\}$ such that $n\delta_n \to \infty$, $\delta_n \log n \to 0$ (for example, $\delta_n = \log^{-2} n$). Then for sufficiently large n

$$\left| \int_0^{\delta_n} \log R_{[nt]} \, dA(t) \right| \leq_{a.s.} \max_{0 \leq t \leq \delta_n} |A'(t)| \delta_n \log n \to 0.$$

The rest of the integral is

$$\int_{\delta_n}^1 \log R_{[nt]} \, dA(t) = \int_{\delta_n}^1 \log \frac{R_{[nt]}}{\mathrm{E} R_{[nt]}} \, dA(t) + \int_{\delta_n}^1 \log \mathrm{E} R_{[nt]} \, dA(t).$$

We prove the a.s. convergence of the first integral in RHS to 0 a.s., and then convergence of the second one to θ.

By the SLLN, $\log(R_j / \mathrm{E} R_j) \to 0$ a.s. as $j \to \infty$. Therefore, for any $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathrm{P} \left(\sup_{j \geq n\delta_n} \left| \log \left(\frac{R_j}{\mathrm{E} R_j} \right) \right| \geq \varepsilon \right) = 0.$$

Therefore

$$\mathrm{P}\left(\sup_{k\geq n}\left|\int_{\delta_n}^{1}\log\frac{R_{[kt]}}{\mathrm{E}R_{[kt]}}\,dA(t)\right|\geq\varepsilon\right)$$

$$\leq \mathrm{P}\left(\int_{\delta_n}^{1}\sup_{k\geq n}\left|\log\frac{R_{[kt]}}{\mathrm{E}R_{[kt]}}\right|\,|dA(t)|\geq\varepsilon\right)$$

$$=\mathrm{P}\left(\sup_{k\geq n\delta_n}\left|\log\frac{R_k}{\mathrm{E}R_k}\right|\geq\frac{\varepsilon}{\int_{\delta_n}^{1}|dA(t)|}\right)\to 0\quad\text{as}\quad n\to\infty,$$

that is, $\int_{\delta_n}^{1}\log\frac{R_{[nt]}}{\mathrm{E}R_{[nt]}}\,dA(t)\to 0$ a.s.

We have $\mathrm{E}R_j=j^\theta L(j)$, where $L(j)$ is a slowly varying function (Karlin, 1967). So, uniformly on $t\geq\delta_n>0$,

$$\frac{\mathrm{E}R_{[nt]}}{(nt)^\theta L(nt)}\to 1,\quad \log\frac{\mathrm{E}R_{[nt]}}{(nt)^\theta L(nt)}\to 0.$$

Therefore

$$\int_{\delta_n}^{1} \log \frac{\mathsf{E}R_{[nt]}}{(nt)^{\theta} L(nt)}\, dA(t) \to 0.$$

However, $L(nt) = (nt)^{o(1)}$ for $t \geq \delta_n$, so from (12) we have

$$\int_{\delta_n}^{1} \log((nt)^{\theta} L(nt))\, dA(t) = \int_{\delta_n}^{1} \log((nt)^{\theta + o(1)})\, dA(t)$$

$$= (\theta + o(1)) \int_{\delta_n}^{1} \log(nt)\, dA(t)$$

$$= (\theta + o(1)) \int_{0}^{1} \log(nt)\, dA(t) - (\theta + o(1)) \int_{0}^{\delta_n} \log(nt)\, dA(t) = \theta + o(1).$$

Then

$$\int_{\delta_n}^{1} \log \mathsf{E}R_{[nt]}\, dA(t) \to \theta,$$

and $\theta_n \to \theta$ a.s. Simmetrically, $\theta'_n \to \theta$ a.s. So $\widehat{\theta} \to \theta$ a.s.
The proof is complete.

Corollary 4.2

The estimator

$$\widehat{\theta} = \log_2 \left(R_n / \sqrt{R_{[n/2]} R'_{[n/2]}} \right), \quad n \geq 2.$$

is strongly consistent under the conditions of Theorem 4.8.

Proof

It is enough to put

$$A(t) = \begin{cases} 0, & 0 \leq t \leq 1/2; \\ -(\log 2)^{-1}, & 1/2 < t < 1; \\ 0, & t = 1. \end{cases}$$

The proof is complete.

Shakespeare's sonnets

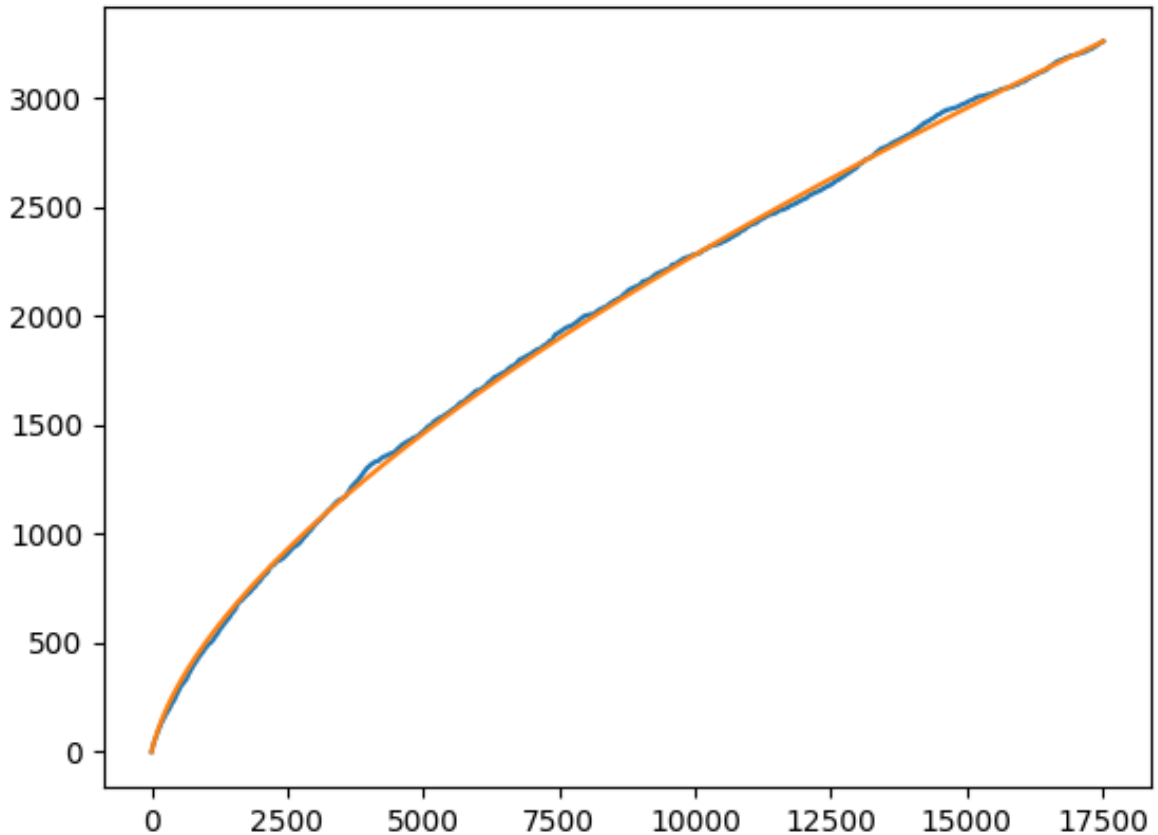

Fig. 4.1. The forward process of numbers of different words for Shakespeare's sonnets and its approximation. $n = 17516$, $R_n = 3258$, $\theta_n = 0.6267$, $q_n = 46.39$.

Zipf-Mandelbrot law

[Zipf, 1936], [Mandelbrot, 1965]

$$p_i = c(i+q)^{-1/\theta}, \quad i \geq 1, \quad 0 < \theta < 1, \quad q > -1.$$

Here

$$c = \left(\zeta(1/\theta, q+1)\right)^{-1},$$

$$\zeta(\alpha, x) = \sum_{i=0}^{\infty} (i+x)^{-\alpha}$$

is the Hurvitz zeta function.

We use approximation

$$r(k) = \sum_{i=1}^{\infty} \left(1 - (1 - \widehat{p}_i)^k\right)$$

$0 \leq k \leq n$.
Here

$$\widehat{p}_i = c(i + q_n)^{-1/\theta_n}, \quad i \geq 1,$$

q_n is such that $r(n) = R_n$.

Sonnets for analysis

Thomas Wyatt (32 sonnets), 1542
William Shakespeare (154 sonnets), 1609
Charlotte Smith, ELEGIAC SONNETS (sonnets I - LIX), 1784

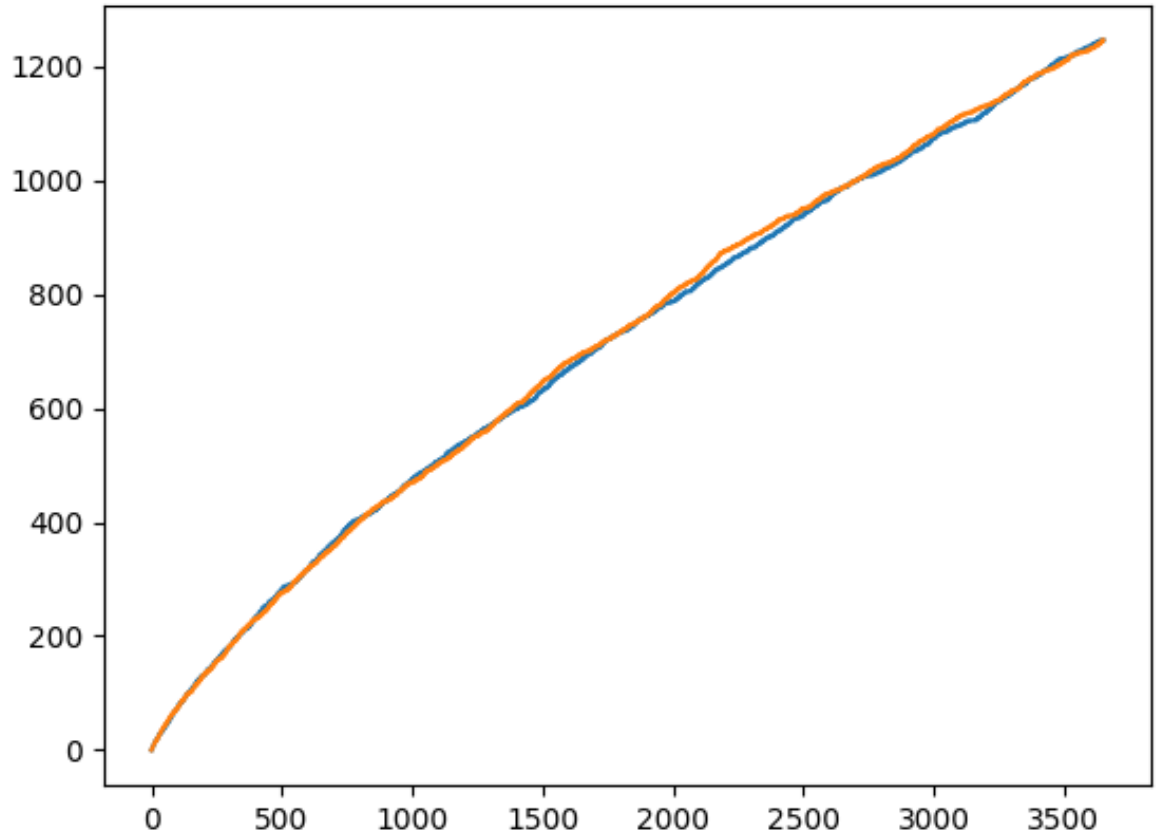

Fig. 4.2. Forward and backward processes of numbers of different words for Wyatt's sonnets

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Wyatt	-0.1139	0.7556	0.7459	0.7507	0.8275	0.0681

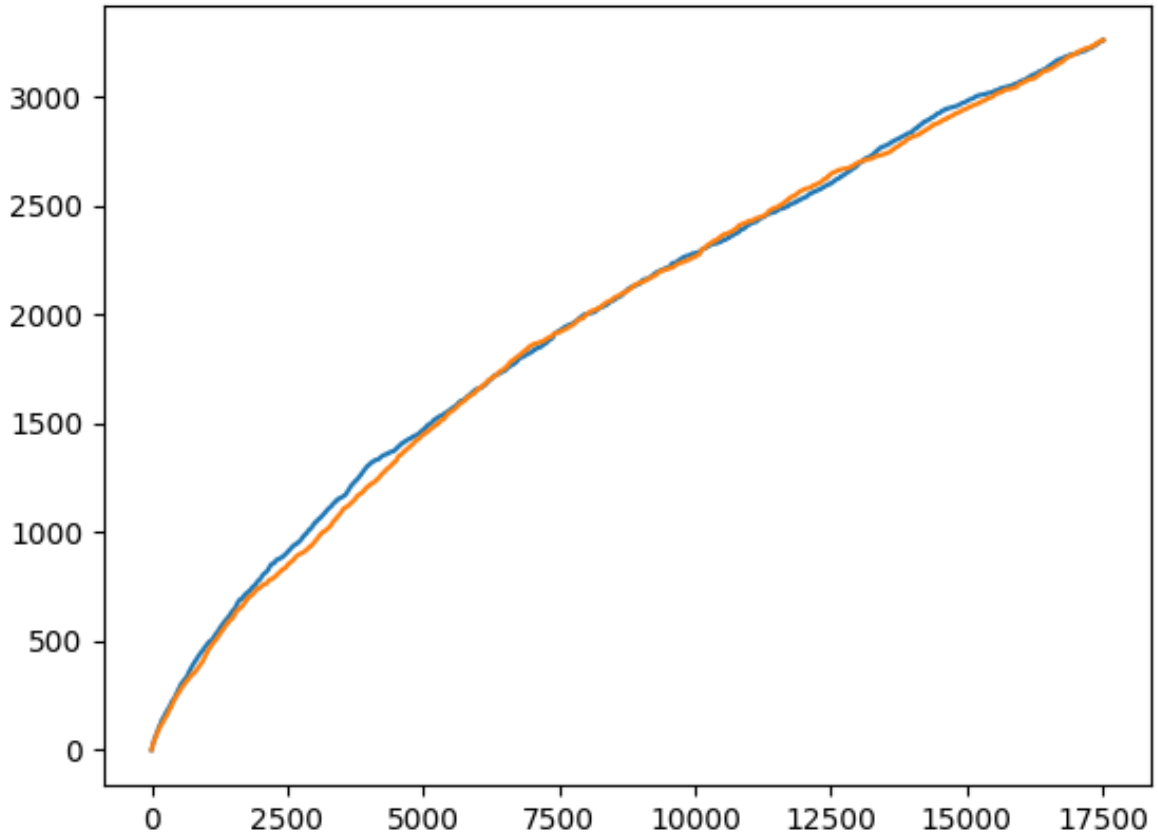

Fig. 4.3. Forward and backward processes of numbers of different words for Shakespeare's sonnets

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Shakespeare	0.2939	0.6267	0.6274	0.6271	0.5475	0.3868

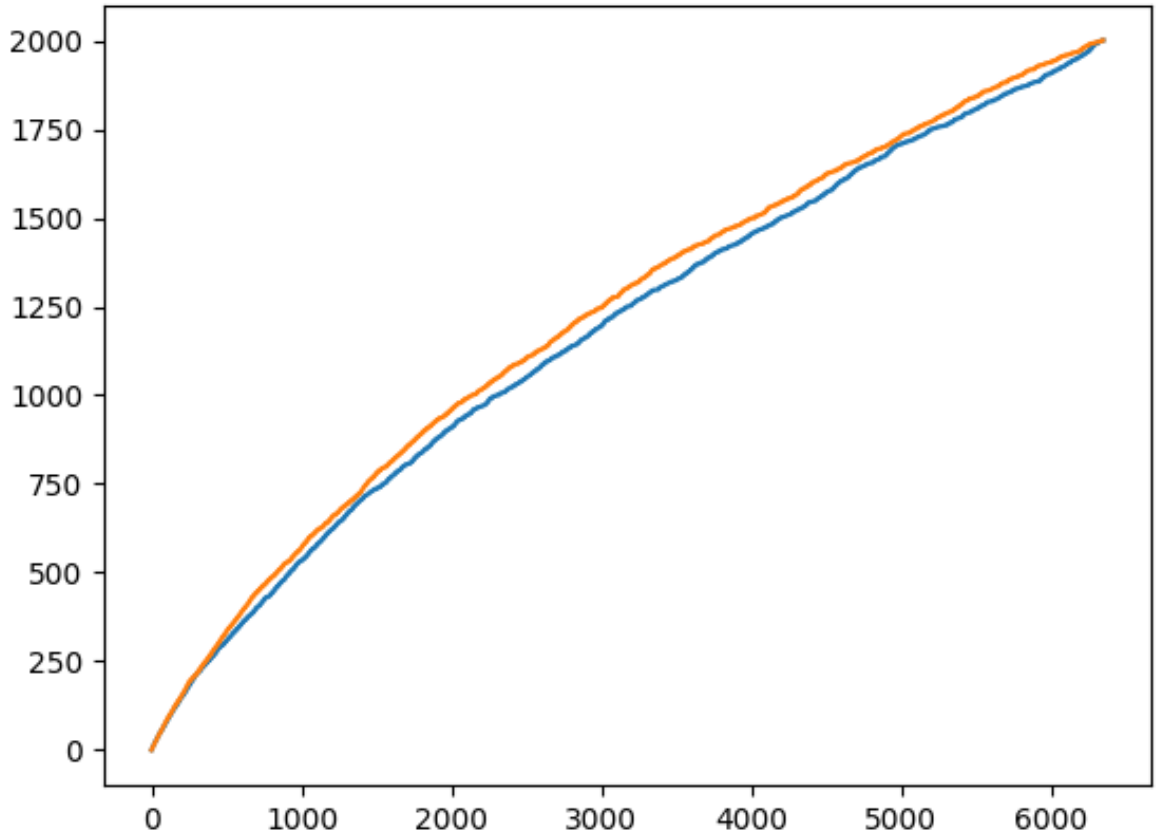

Fig. 4.4. Forward and backward processes of numbers of different words for Smith's sonnets

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Smith	-0.8748	0.6788	0.62	0.6494	0.0772	0.883

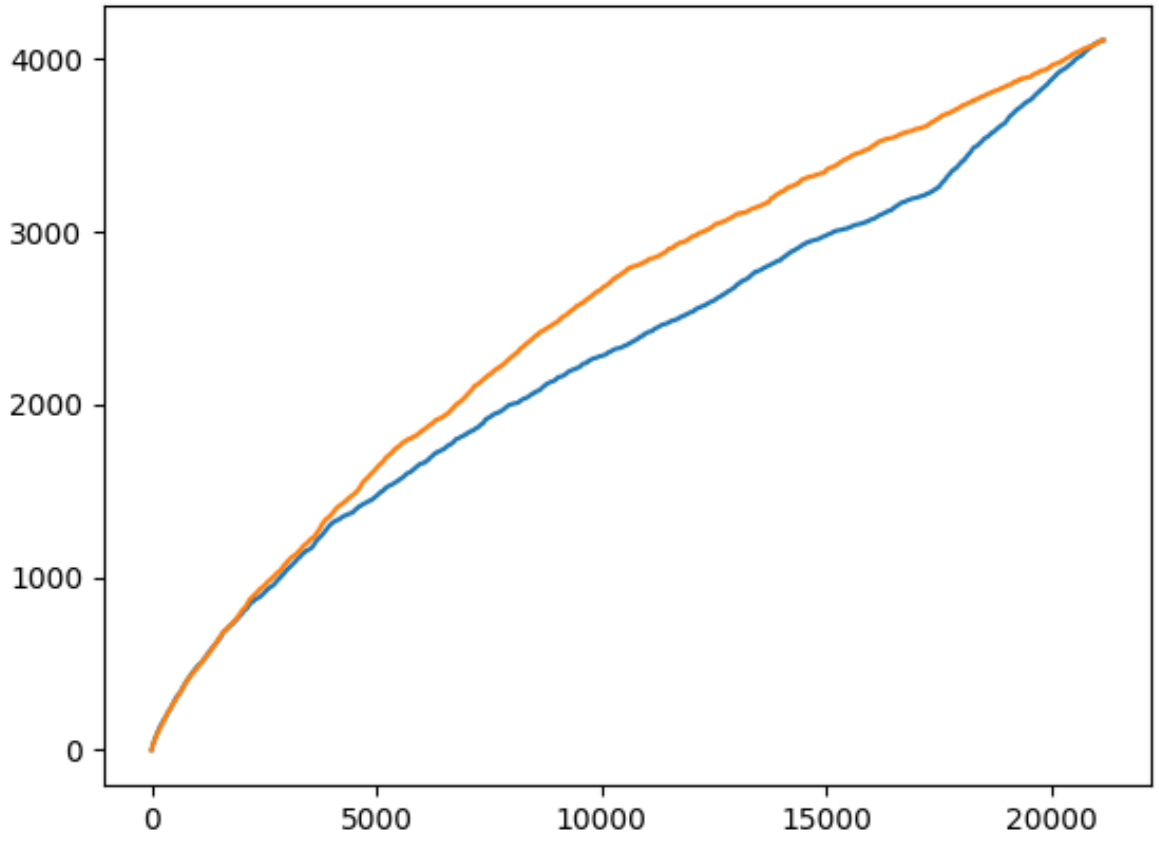

Fig. 4.5. Forward and backward processes of numbers of different words for Shakespeare+Wyatt

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Shakespeare+Wyatt	-3.7886	0.8082	0.5634	0.6858	0.0000	20.3048

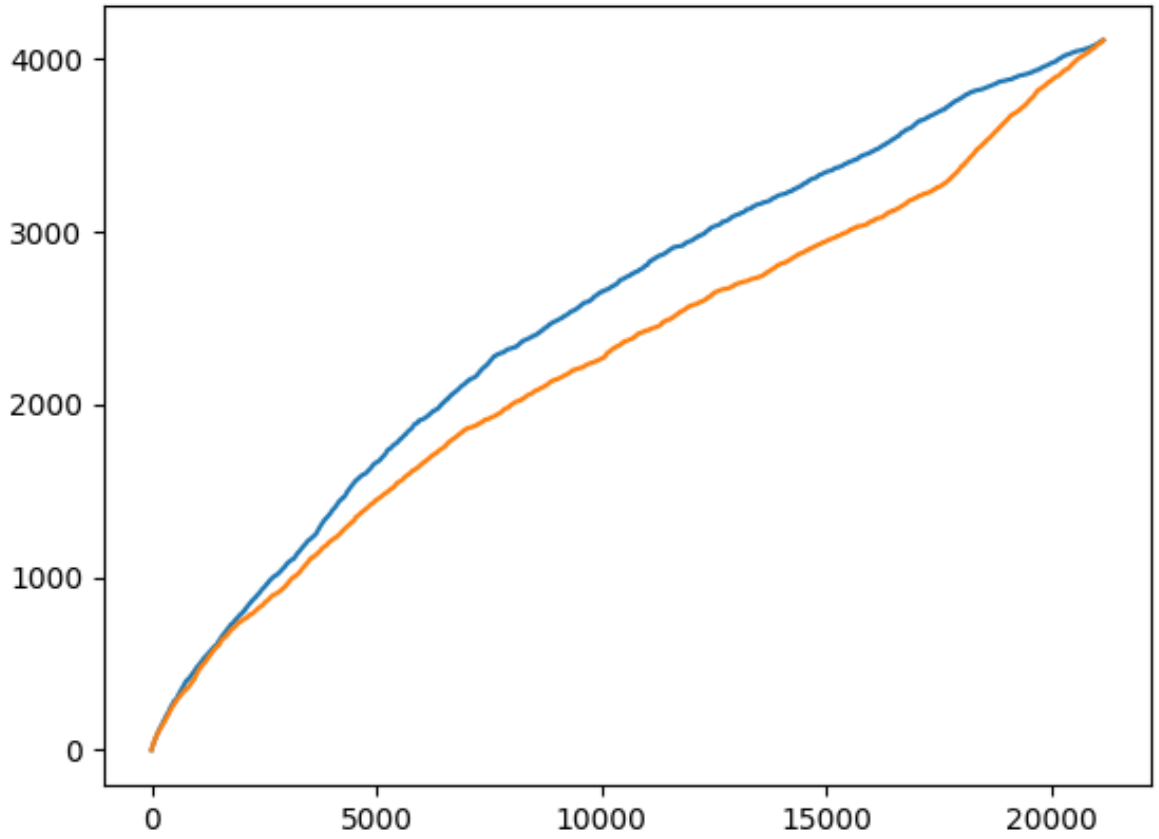

Fig. 4.6. Forward and backward processes of numbers of different words for Wyatt+Shakespeare

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Wyatt+Shakespeare	4.2126	0.5837	0.7948	0.6893	0.0000	22.4295

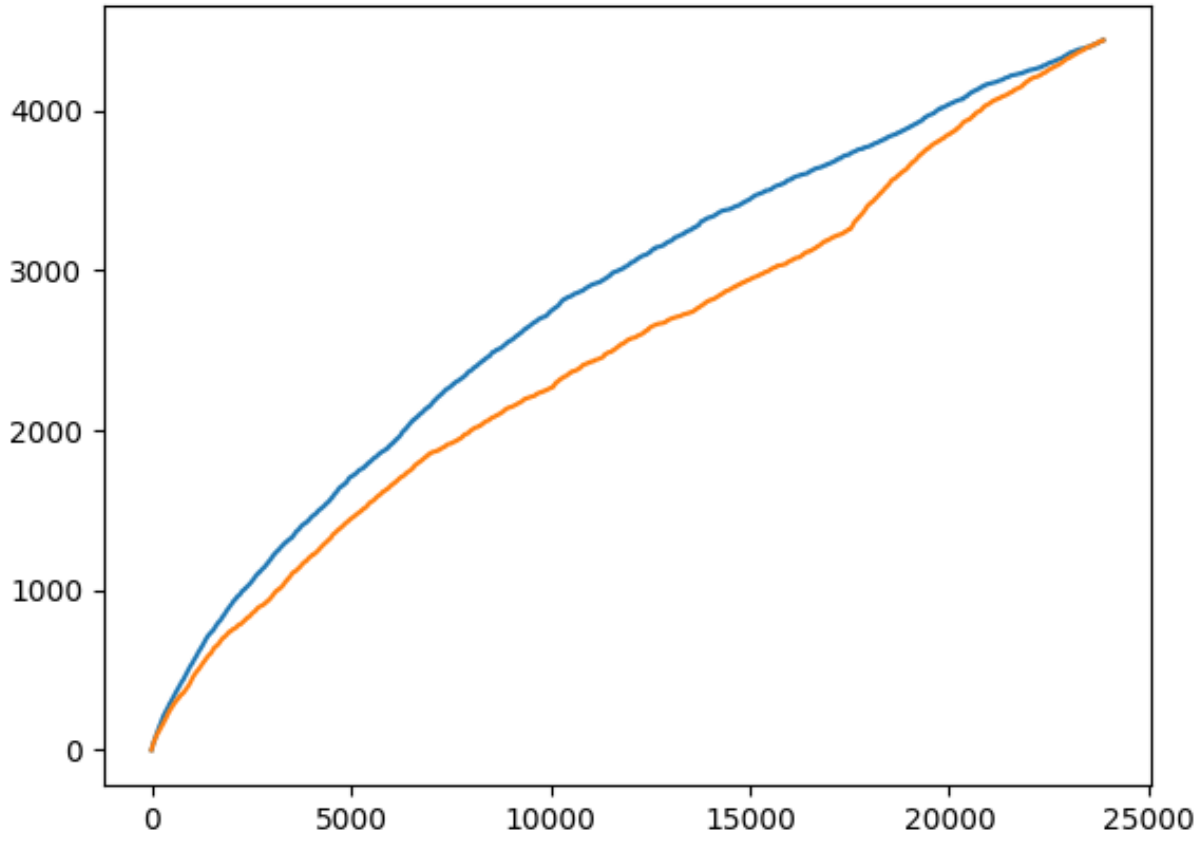

Fig. 4.7. Forward and backward processes of numbers of different words for Smith+Shakespeare

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Smith+Shakespeare	4.6056	0.552	0.7925	0.6723	0.0000	27.3113

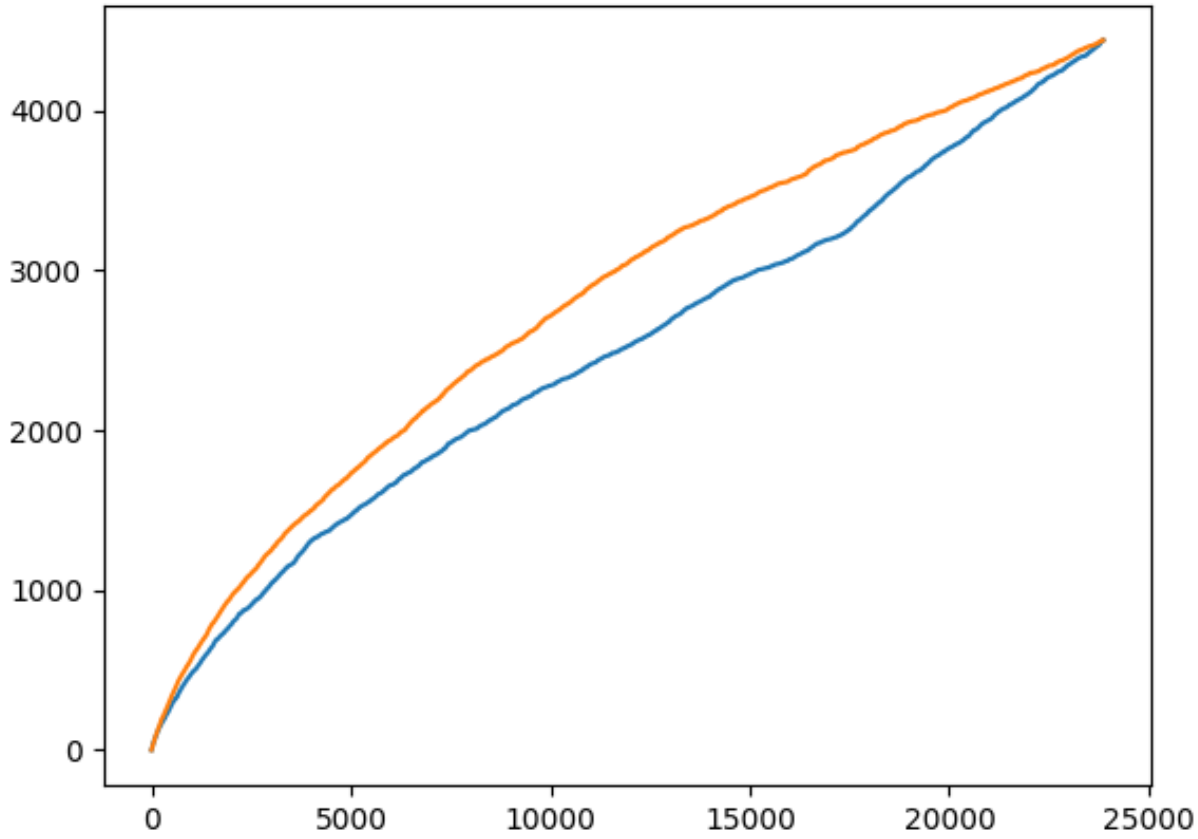

Fig. 4.8. Forward and backward processes of numbers of different words for Shakespeare+Smith

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Shakespeare+Smith	-4.8183	0.8146	0.5444	0.6795	0.0000	28.7613

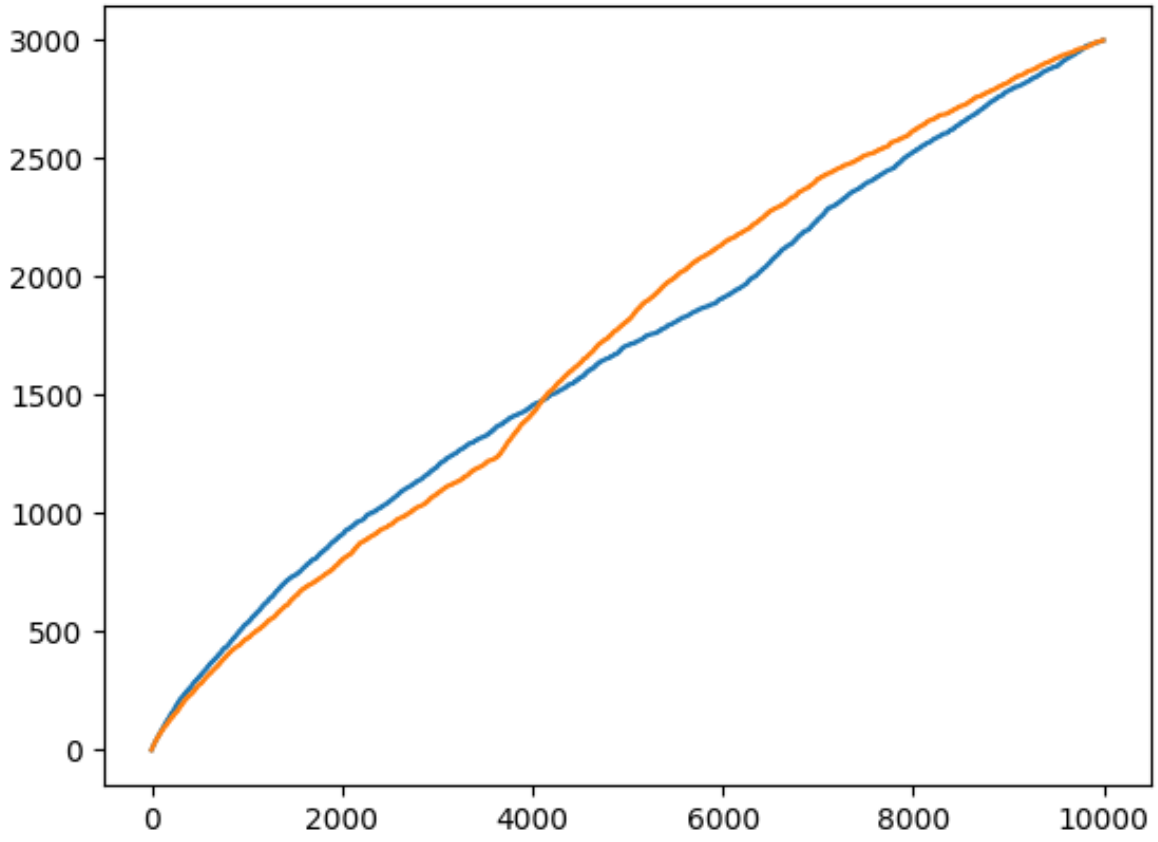

Fig. 4.9. Forward and backward processes of numbers of different words for Smith+Wyatt

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Smith+Wyatt	-0.5909	0.8108	0.7256	0.7682	0.2620	4.5616

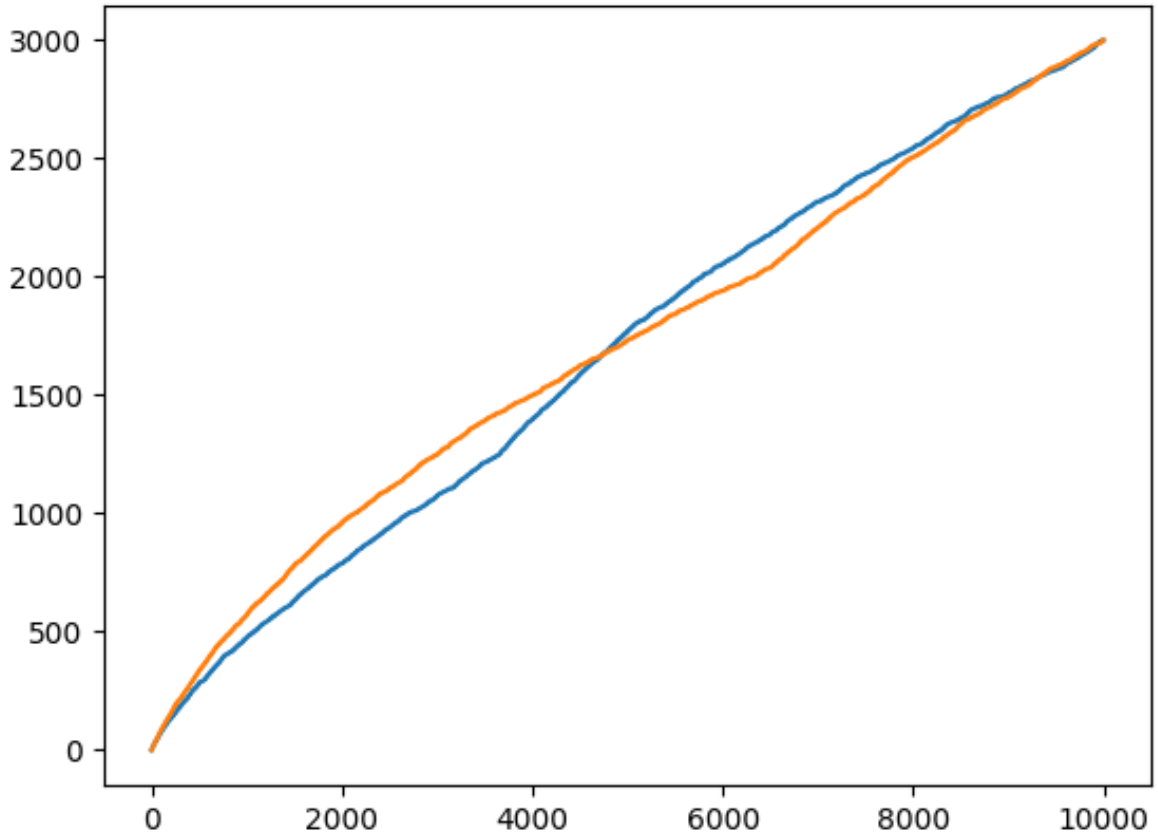

Fig. 4.10. Forward and backward processes of numbers of different words for Wyatt+Smith

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Wyatt+Smith	-0.4583	0.7627	0.7924	0.7775	0.3863	3.7753

Author(s)	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Wyatt	-0.1139	0.7556	0.7459	0.7507	0.8275	0.0681
Shakespeare	0.2939	0.6267	0.6274	0.6271	0.5475	0.3868
Smith	-0.8748	0.6788	0.62	0.6494	0.0772	0.883
Shakespeare+Wyatt	-3.7886	0.8082	0.5634	0.6858	0.0000	20.3048
Wyatt+Shakespeare	4.2126	0.5837	0.7948	0.6893	0.0000	22.4295
Smith+Shakespeare	4.6056	0.552	0.7925	0.6723	0.0000	27.3113
Shakespeare+Smith	-4.8183	0.8146	0.5444	0.6795	0.0000	28.7613
Smith+Wyatt	-0.5909	0.8108	0.7256	0.7682	0.2620	4.5616
Wyatt+Smith	-0.4583	0.7627	0.7924	0.7775	0.3863	3.7753

Лекция 5. Побуквенный анализ цепями Маркова

А. А. Марков (1913) предложил использовать испытания, связанные в цепь, для описания последовательности букв текста, а Д.В. Хмелев (2000a) предложил измерять модификацию расхождения Кульбака—Лейблера для идентификации автора.

Пусть A — множество букв некоторого алфавита,
A^k — множество текстов длины k.
$A^* = \cup_{k=1}^{\infty} A^k$ — множество всех текстов,
$|f|$ — длина текста $f \in A^*$.

Задача классификации:
C_i, $i = 1, \ldots, N$, — классы текстов,
$f_{i,j} \in A^*$, $j = 1, \ldots, m_i$, — тексты соответствующих классов,
$x \in A^*$ надо отнести к одному из классов C_i.

Стохастические предположения:
$f_{i,j}$ являются реализациями цепи Маркова с переходной матрицей Π^i;
x является реализацией цепи Маркова с переходной матрицей Π^θ. где
$\theta \in \{1, \ldots, N\}$ — неизвестный параметр.
Все цепи Маркова независимы в совокупности.

P^i — оценка матрицы Π^i:
$P^i = (P^i_{kl})^{|A|}_{k,l=1}$,
$P^i_{kl} = h_{i,kl}/h_{i,k}$, где
$h_{i,k} = \sum\limits_{l \in A} h_{i,kl}$,
$h_{i,kl} = \sum\limits_{j=1}^{m_i} h_{i,j,kl}$,
$h_{i,j,kl}$ — число переходов букв $k \to l$ в тексте $f_{i,j}$.

ν_{kl} — число переходов букв $k \to l$ в тексте x,
$\nu_k = \sum\limits_{l \in A} \nu_{kl}$,
$Q_{kl} = \nu_{kl}/\nu_k$,
$Z_i(x) = \{(k,l) : \ P^i_{kl} > 0, \ Q_{kl} > 0\}$.

$$L_i(x) = \sum_{(k,l) \in Z_i(x)} \nu_{kl} \ln \frac{Q_{kl}}{P^i_{kl}}.$$

Оценка параметра θ для текста x:

$$\widehat{\theta}(x) = \arg \min_{1 \le i \le N} L_i(x).$$

Удобно использовать нормированные значения

$$L^0_i(x) = L_i(x) / \sum_{k \in A} \nu_k.$$

Отметим, что расхождение Кульбака—Лейблера

$$D_{KL}(Q||P) = \sum_{(k,l) \in Z_i(x)} Q_{kl} \ln \frac{Q_{kl}}{P^i_{kl}}.$$

отличается от каждой из этих статистик.

Theorem 5.1 (неравенство Гиббса)

Пусть $P = \{p_1, p_2, \ldots\}$ *и* $Q = \{q_1, q_2, \ldots\}$ *— вероятностные распределения на одном и том же конечном или счетном множестве, все* $p_i > 0$,

$$D_{KL}(Q||P) = \sum_{i:q_i>0} q_i \ln \frac{q_i}{p_i}.$$

Тогда $D_{KL}(Q||P) \geq 0$, *и*

$$D_{KL}(Q||P) = 0 \;\Leftrightarrow\; Q = P.$$

Доказательство

$$D_{KL}(Q||P) = -\sum_{i:q_i>0} q_i \ln \frac{p_i}{q_i}.$$

Используем неравенство $\ln x \leq x - 1$, верное для $x > 0$:

$$D_{KL}(Q||P) \geq -\sum_{i:q_i>0} q_i \left(1 - \frac{p_i}{q_i}\right) = 1 - \sum_{i:q_i>0} p_i \geq 0.$$

Так как $\ln x = x - 1$ только при $x = 1$, то из равенства $D_{KL}(Q||P) = 0$ следует $p_i = q_i$ для всех i.
Доказательство завершено.

Exercise 5.1

Доказать аналог неравенства Гиббса для статистики Хмелёва.

Подготовка текста (метод *F* Д. В. Хмелева (2000a)):
— превратить все заглавные буквы в строчные;
— удалить все знаки пунктуации;
— удалить излишние знаки пробела, оставляя их по одному между словами;
— вставить пробелы в начале и в конце, если их нет изначально.

Маленький численный эксперимент:
стихотворение «Смерть поэта» М.Ю. Лермонтова (2105 символов в преобразованном тексте, то есть 2104 перехода цепи Маркова) сравнивалось с:
— поэмой М. Ю. Лермонтова «Бородино»;
— первой главой романа «Евгений Онегин» А.С. Пушкина;
— двумя песнями В.С. Высоцкого, описанными в главе 1 (все произведения в нынешней орфографии).
Сформированный общий (содержащийся во всех текстах) алфавит:
[' ', 'а', 'б', 'в', 'г', 'д', 'е', 'ж', 'з', 'и', 'й', 'к', 'л', 'м', 'н', 'о', 'п', 'р', 'с', 'т', 'у', 'х', 'ц', 'ч', 'ш', 'щ', 'ы', 'ь', 'э', 'ю', 'я'].
Исключены буквы *ё, ф, ъ*.

Произведение	$L_i^0(x)$
«Бородино»	0.1992
Глава 1 «Евгения Онегина»	0.1559
Две песни Высоцкого	0.2015


Андерсон Т. Введение в многомерный статистический анализ. — М., 1963.


Андерсон Т. Статистический анализ временных рядов. —- М.: Мир, 1976.


Анго А. Математика для электро- и радиоинженеров. — М., 1964.


Бендерская Е. Н., Жукова С. В. Обработка текстовой информации с использованием хаотической нейронной сети // В сб.: «Квантитативная лингвистика: исследования и модели (КЛИМ - 2005)», материалы Всероссийской научной конференции. Новосибирск, НГПУ. 2005. — С. 271–282.


Биллингсли П. Сходимость вероятностных мер. — М.: «Наука», 1977.


Бойко Ю. П. Селективность авторского стиля на уровне предложения // В сб.: «Квантитативная лингвистика: исследования и модели (КЛИМ - 2005)», материалы

Всероссийской научной конференции. Новосибирск, НГПУ. 2005. — С. 303–315.


Боровков А. А. Теория вероятностей. — М.: «Эдиториал УРСС», 1999.


Боровков А. А. Математическая статистика. — Новосибирск: «Наука», 1997.


Бродский Б. Е., Дарховский Б. С. Асимптотический анализ некоторых оценок в апостериорной задаче о разладке // Теория вероятн. и ее примен., 35:3, 1990. — С. 551–557.


Бродский Б. Е., Дарховский Б. С. Сравнительный анализ некоторых непараметрических методов скорейшего обнаружения момента «разладки» случайной последовательности // Теория вероятн. и ее примен., 35:4, 1990. — С. 655–668.


Бродский Б. Е., Дарховский Б. С. Алгоритм апостериорного обнаружения многократных разладок

случайной последовательности // Автомат. и телемех., N . 1, 1993. — С. 62–67.


Бродский Б. Е., Дарховский Б. С. Проблемы и методы вероятностной диагностики // Автомат. и телемех., N 8, 1999. — С. 3–50.


Вальд А. Последовательный анализ. — М.: «Физматлит», 1960.


Васильченко С. Г. Алгоритм обнаружения моментов разладки случайной последовательности // Фундамент. и прикл. матем., 8:3, 2002. — С. 655-–665.


Вашак П. Длина слова и длина предложения в текстах одного автора // Вопросы статистической стилистики — Киев: Наукова думка, 1974. — С. 314–329.


Виноградов В. В. Проблема авторства и теория стилей. — М., 1961.


Гусарова Г. В., Ковалевский А. П., Макаренко А. Г. Критерии наличия разладки // Сибирский журнал

индустриальной математики. 2005. — Т. VIII, No. 4 (24). — С. 18–33.


«Д». Стремя «Тихого Дона». Загадки романа. — Paris: YMCA Press, 1974.


Дарховский Б. С. Непараметрический метод для апостериорного обнаружения момента «разладки» последовательности независимых случайных величин, Теория вероятн. и ее примен., 21:1,1976. — С. 180–184.


Дарховский Б. С. Задача о неопределенной «разладке» случайной последовательности // Теория вероятн. и ее примен., 56:1, 2011. — С. 30–46.


Дарховский Б. С. Обнаружение разладки случайной последовательности при минимальной априорной информации // Теория вероятн. и ее примен., 58:3, 2013. — С. 585–590.


Дарховский Б. С., Бродский Б. Е. Апостериорное обнаружение момента «разладки» случайной

последовательности // Теория вероятн. и ее примен., 25:3, 1980. — С. 635–639.


Дарховский Б. С., Бродский Б. Е. Непараметрический метод скорейшего обнаружения изменения среднего случайной последовательности // Теория вероятн. и ее примен., 32:4, 1987. — С. 703–711.


Дарховский Б. С., Пирятинская А. Новый подход к проблеме сегментации временных рядов произвольной природы // Стохастическое исчисление, мартингалы и их применения, Сборник статей. К 80-летию со дня рождения академика Альберта Николаевича Ширяева, Тр. МИАН, 287, МАИК, М., 2014. — С. 61–74.


Дрейпер Н.Р., Смит Г. Прикладной регрессионный анализ. — М.: Диалектика, 2007.


Ермолаев Г. Загадки «Тихого Дона» // Slavic and European Journal, V. 18, N 3, 1974.


Ермолаев Г. Кто написал «Тихий Дон» // Slavic and European Journal, V. 20, N 3, 1976.


Ермоленко Г. В. Лингвистическая статистика. Краткий очерк и библиографический указатель. — Алма-Ата, 1970.


Закревская Н. С. Вероятностные модели текста // В сб.: Материалы IXL международной научной студенческой конференции «Студент и научно-технический прогресс». Математика. Новосибирск, НГУ, 2001. — С. 136.


Закревская Н. С. Сравнение марковских и элементарных вероятностных моделей статистик текста // В сб.: Материалы XL международной научной студенческой конференции «Студент и научно-технический прогресс». Математика. Новосибирск, НГУ, 2002. — С. 142–143.


Закревская Н. С. Вероятностные модели текста // В сб.: Информатика и проблемы телекомуникаций. Международная научно-техническая конференция.

Материалы конференции. Новосибирск, СибГУТИ, 2002. — С. 120–121.


Закревская Н.С. Исследование однородности текста с помощью модели скользящего среднего // В сб.: «Квантитативная лингвистика: исследования и модели (КЛИМ - 2005)», материалы Всероссийской научной конференции. Новосибирск, НГПУ, 2005. — С. 26–34.


Закревская Н. С., Ковалевский А. П. Однопараметрические вероятностные модели статистик текста // Сибирский журнал индустриальной математики. 2001. — Т. IV, N 2 (8). — С. 142–153.


Золотарев В.М. Современная теория суммирования независимых случайных величин. — М.: Наука, 1986.


Ибрагимов И. А., Линник Ю. В. Независимые и стационарно связанные величины. М.: Наука, 1965.


Карлин С. Основы теории случайных процессов. — М.: Мир, 1971.


Кйецаа Г. Борьба за «Тихий Дон». — Pergamob Press, USA, 1977.


Клигене Н., Телькснис Л. Методы обнаружения моментов изменения свойств случайных процессов // Автоматика и телемеханика. — 1983. — N 10, с. 5–56.


Ковалевский А. П. Применение принципа инвариантности к анализу однородности текста // В сб.: «Квантитативная лингвистика: исследования и модели (КЛИМ - 2005)», материалы Всероссийской научной конференции. Новосибирск, НГПУ. 2005. — С. 195–204.


Ковалевский А. П. Оценивание параметра закона Ципфа-Мандельброта по последовательности количеств разных элементов выборки // Обозрение прикладной и промышленной математики, 2017. — Т. 24, вып. 4. — С. 348–349.


Колодный Л. Вихри над «Тихим Доном». Фрагменты прошлого: истоки одного навета XX века // Московская правда, 5 и 7 марта 1989.


Крамер Г., Лидбеттер М. Стационарные случайные процессы. — М., 1969.


Кукушкина О.В., Поликарпов А.А., Хмелев Д.В. Определение авторства текста с использованием буквенной и грамматической информации // Проблемы передачи информации, Т. 37, вып. 2, 2001. — С. 96–108.


Кукушкина О. В., Макаров А. Г., Поддубный В. В., Поликарпов А. А., Шевелев О. Г. Анализ количественных характиристик авторского стиля романа «Тихий Дон» и его соотношение с другими текстами М. А. Шолохова на основе иерархической кластеризации // Загадки и тайны "Тихого Дона": двенадцать лет поисков и находок. — М. : «АИРО — XXI», 2010. — С. 127–130.


Леви П. Стохастические процессы и броуновское движение. — М.: Наука, 1972.


Мандельброт Б. Теория информации и психолингвистика: теория частот слов // в сб.: Математические методы в социальных науках. — М., 1973. — С. 316–337.


Марков А. А. Пример статистического исследования над текстом «Евгения Онегина», иллюстрирующий связь испытаний в цепь // Известия Императорской Академии Наук, Сер. 6, Т. 10, вып. 3, 1913. - С. 153.


Марков А. А. Об одном применении статистического метода // Известия Императорской Академии Наук, Сер. 6, Т. 10, вып. 4, 1916. - С. 239.


Мартынов Г. В. Критерии омега-квадрат. М.: Наука, 1978.


Марусенко М. А. Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов — Л.: Филол. ф-т СПбГУ, 1990.


Марусенко М. А., Мельникова Е. Е., Родионова Е. С. Атрибуция анонимных и псевдонимных статей, опубликованных в журналах «Время» и «Эпоха» в 1861–1865 гг // В сб.: «Квантитативная лингвистика: исследования и модели (КЛИМ - 2005)», материалы Всероссийской научной конференции. Новосибирск, НГПУ. 2005. — С. 283–293.


Медведев Р. Кто написал «Тихий Дон»?. — Paris: Christian Bourg. Edit., 1975.


Морозов Н.А. Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного известного автора. Стилеметрический этюд // Известия отд. русского языка и словесности Имп.акад.наук, Т.20, Кн.4, 1915.


Надточий Е. Д. Статистическая диагностика авторских различий в синтаксисе: автореф. ... канд. филол. наук — Л., 1983.


Никитин Я. Ю. Асимптотическая эффективность непараметрических критериев. — М.: «Физматлит», 1995.


Никифоров И. В. Последовательное обнаружение изменения свойств временных рядов. — М.: «Наука», 1983.


Питмен Э. Основы теории статистических выводов. — М.: «Мир», 1986.


Поддубный В.В., Поликарпов А.А. Диссипативная стохастическая динамическая модель развития языковых знаков //Компьютерные исследования и моделирование. 2011. Т. 3, N 2. — С. 103–124.


Поддубный В. В., Шевелев О. Г. Сравнение и кластерный анализ текстов по частотным признакам на основе гипергеометрического критерия // В сб.: «Квантитативная лингвистика: исследования и модели (КЛИМ - 2005)», материалы Всероссийской научной конференции. Новосибирск, НГПУ. 2005. — С. 205–217.


Поддубный В.В., Шевелев О.Г. Сравнительный анализ стилей текстов по частотным признакам на основе гипергеометрического критерия // Информационные технологии и математическое моделирование: Материалы III Всероссийской научно-практической конференции (11-12 декабря 2004 г.) Ч. 2. — Томск: Изд-во Том. ун-та, 2004. — С.48–51.


Рыжиков Ю. Вычислительные методы. — СПб., 2007.


Семянникова Н. В. Верификация некоторых методов атрибуции на переводных текстах // В сб.: «Квантитативная лингвистика: исследования и модели (КЛИМ - 2005)», материалы Всероссийской научной конференции. Новосибирск, НГПУ, 2005. — С. 294–302.


Скороход А. В. Элементы теории вероятностей и случайных процессов. — Киев: Вища школа, 1980.


Торговицкий И. Ш. Методы определения момента изменения вероятностных характеристик случайных

величин // Зарубежная радиоэлектроника. — 1976. — N 1. — С. 3–52.


Феллер В. Введение в теорию вероятностей и ее приложения. Том 1. — Москва, Мир, 1967.


Феллер В. Введение в теорию вероятностей и ее приложения. Т. 2. — М.: Мир, 1967.


Фоменко В.П., Фоменко Т.Г. Авторский инвариант русских литературных текстов // Методы количественного анализа текстов нарративных источников. М.: Ин-т истории СССР, 1983. — С. 86–109.


Хмелев Д.В. Распознавание автора текста с использованием цепей А.А. Маркова // Вестн. МГУ. Сер. 9, Филология. 2000a. N 2. — С.115–126.


Хмелев Д.В. Как определить писателя? // Компьютерра, N 9, 2000b.


Хьетсо Г., Густавссон С., Бекман Б., Гил С. Кто написал «Тихий Дон». (Проблема авторства «Тихого Дона»). — М.: Книга, 1989.


Чебунин М. Г. Оценивание параметров вероятностных моделей по числу различных элементов выборки // Сиб. журн. индустр. матем. – 2014. – Т. 17. – № 3. – С. 135-147.


Чебунин М. Г. Оценивание числа ячеек по числу занятых ячеек при случайном размещении // Вестн. НГУ. Сер. матем., мех., информ. – 2014. – Т. 14. – № 3. – С. 107-113.


Шевелев О.Г. Анализ частоты встречаемости различных длин предложений в литературном тексте как возможной характеристики авторского стиля с помощью самоорганизующихся карт Кохонена // Нейроинформатика и ее приложения: Материалы XII Всероссийского семинара, 1–3 октября 2004 г. — Красноярск: ИВМ СО РАН, 2004. — С.177–178.


Шрейдер Ю. А. О возможности теоретического вывода статистических закономерностей текста (к обоснованию закона Ципфа). Проблемы передачи информации, Т.3, N 1, 1967. — С. 57–63.


Шрейдер Ю. А., Шаров Н. А. Системы и модели. — М., 1982.


Шубик С. А. Размер предложения в немецкой художественной прозе // Сборник статей по методике преподавания иностранных языков и филологии — Вып. 4. Л., 1969. — С. 77–79.


Яглом А. М. Корреляционная теория стационарных случайных функций с примерами из метеорологии. — Л.: Гидрометеоиздат, 1981.


Adler R. J. An introduction to continuity, extrema, and related topics for general Gaussian processes. Inst. Math. Statist. Lecture Notes — Monograph Series, Vol. 12, Inst. Math. Statist. — Hayward, CA, 1990.


Bahadur R. R. On the number of distinct values in a large sample from an infinite discrete distribution // Proceedings of the National Institute of Sciences of India. – 1960. – V. 26A. – № 2. – P. 67-75.


Barbour A. D. Univariate approximations in the infinite occupancy scheme // Alea. – 2009. – V. 6. – P. 415-433.


Barbour A. D., Gnedin A. V. Small counts in the infinite occupancy scheme // Electronic Journal of Probability. – 2009. – V. 14. – № 13. – P. 365-384.


Ben-Hamou, A., Boucheron, S., Ohannessian, M. I., 2017. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications, Bernoulli, V. 23, No. 1, 249–287.


Bhattacharyya G. K., Johnson R. A. Nonparametric tests for shift at an unknown time point // Ann. Math. Statist, V. 39, 1968. — P. 1731–1743.


Brodsky E., Darkhovsky B.S. Nonparametric Methods in Change Point Problems. — Springer, 1993.


Carlstein E. Nonparametric change-point estimation // Ann. of Statist., V. 16, No. 1, 1988. — P. 188–197.


Chakrabarty et al. A statistical test for correspondence of texts to the Zipf - Mandelbrot law / A. Chakrabarty, M. Chebunin, A. Kovalevskii, I. Pupyshev, N. Zakrevskaya, Q. Zhou // Siberian Electronic Mathematical Reports. - 2020. - V. 17. - P. 1959-1974.


Chebunin M., Kovalevskii A. Functional central limit theorems for certain statistics in an infinite urn scheme // Statistics and Probability Letters, V. 119. — 2016. — P. 344-–348.


Chebunin, M. G., Kovalevskii, A. P., 2019a. A statistical test for the Zipf's law by deviations from the Heaps' law, Siberian Electronic Mathematical Reports, V. 16, 1822–1832.


Chebunin, M., Kovalevskii, A., 2019b. Asymptotically Normal Estimators for Zipf's Law, Sankhya A, V. 81, 482–492.


Decrouez, G., Grabchak, M., Paris, Q., 2018. Finite sample properties of the mean occupancy counts and probabilities, Bernoulli, V. 24, No. 3, 1910–1941.


Dümbgen L. The asymptotic behavior of some nonparametric change-point estimators // The Annals of Statist., V. 19, No. 3, 1991. — P. 1471–1495.


Durieu, O., Wang, Y., 2016. From infinite urn schemes to decompositions of self-similar Gaussian processes, Electron. J. Probab., V. 21, paper No. 43, 23 pp.


Durieu, O., Samorodnitsky, G., Wang, Y., 2019. From infinite urn schemes to self-similar stable processes, Stochastic Processes and their Applications (in press).


Dutko M. Central limit theorems for infinite urn models // Ann. Probab. – 1989. – V. 17. – P. 1255-1263.


Eliazar, I., 2011. The Growth Statistics of Zipfian Ensembles: Beyond Heaps' Law, Physica (Amsterdam) 390, 3189.


Gerlach, M., and Altmann, E. G., 2013. Stochastic Model for the Vocabulary Growth in Natural Languages. Physical Review X 3, 021006.


Gnedin A., Hansen B., Pitman J. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws // Probability Surveys. – 2007. – V. 4. – P. 146–171.


Heaps, H. S., 1978. Information Retrieval: Computational and Theoretical Aspects, Academic Press.


Herdan, G., 1960. Type-token mathematics, The Hague: Mouton.


Herdan E. Calculus of legomena. — N.- Y., 1964.


Hurst H. E., Black R. P., Sinaika Y. M. Long term storage in reservoirs. An experimental study. — London: Constable, 1965.

 Hwang H.-K., Janson S. Local Limit Theorems for Finite and Infinite Urn Models // The Annals of Probability. – 2008. – V. 36. – № 3. – P. 992–1022.

 Kallianpur J., Oodaira H. Freidlin—Wentzell type estimates for abstract Wiener spaces // Sankhy$\bar{a}$, Ser. A, Vol. 40, 1978. — P. 116–137.

 *Karlin S.* Central Limit Theorems for Certain Infinite Urn Schemes // Jounal of Mathematics and Mechanics. – 1967. – V. 17. – № 4. – P. 373-401.

 Key E. S. Rare Numbers // Journal of Theoretical Probability, Vol. 5, No. 2, 1992. — P. 375–389.

 Kovalevskii A. Dependence of increments in time series via large deviations // Proceedings of the 7th Korea-Russia International Symposium on Science and Technology. Ulsan, Korea, 2003. — P. 262–267.

 van Leijenhorst, D. C., van der Weide, T. P., 2005. A Formal Derivation of Heaps' Law, Information Sciences (NY) 170, 263.


Mandelbrot, B., 1965. Information Theory and Psycholinguistics. In: B.B. Wolman and E. Nagel. Scientific psychology. Basic Books.


Mikhailov V. G. Asymptotic Normality of the Number of Empty Cells for Group Allocation of Particles // Theory Probab. Appl. – 1980. – V. 25. – № 1. – P. 82–90.


Mirakhmedov S. M. Asymptotic normality associated with generalized occupancy problems // Statistics and Probability Letters. – 2007. – V. 77. – № 15. – P. 1549–1558.


Nicholls, P. T., 1987. Estimation of Zipf parameters. J. Am. Soc. Inf. Sci., V. 38, 443–445.


Petersen, A. M., Tenenbaum, J. N., Havlin, S., Stanley, H. E., Perc, M., 2012. Languages cool as they expand: Allometric scaling and the decreasing need for new words. Scientific Reports 2, Article No. 943.


Sherman L. A. Analitics of literature. A manual for the objective study of English prose and poetry. — Boston, 1893.


Smirnov, N.V., 1937. On the omega-squared distribution, Mat. Sb. 2, 973–993 (in Russian).


Wieand H. S. A condition under which the Pitman and Bahadur approaches to efficiency coincide // The Annals of Statist., V.4, No. 5, 1976. — P. 1003–1011.


Zakrevskaya, N., Kovalevskii, A., 2019. An omega-square statistics for analysis of correspondence of small texts to the Zipf—Mandelbrot law. In: Applied methods of statistical analysis. Statistical computation and simulation. Proceedings of the International Workshop, Novosibirsk, NSTU, 488–494.


Zipf, G. K., 1936. The Psycho-Biology of Language, Routledge, London, 1936.