СОЗДАНИЕ ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКОЙ МОДЕЛИ КАЗАХСКОГО ТЕКСТА

Тоқмырзаев Дархан
Программист
Казахский национальный унивкерситет имени аль-Фараби

Жұбанова Еңлік
Программист
Международный университет информационных технологий

Пірманова Айнұра
Математик-статист, PhD докторант
Казахский национальный унивкерситет имени аль-Фараби

Аннотация: В настоящее время лингвистическая статистика переходит от первоначального этапа описания языковых явлений к разработке теорий, способных количественно предсказывать языковые закономерности. Основная задача исследования заключается в анализе математического распределения языковых единиц в текстах, что является ключевым для моделирования языковых феноменов. В рамках данной работы было исследовано статистическое распределение основных частей речи в текстах на казахском языке, включая тысячи словоупотреблений из романа М. Ауэзова «Путь Абая» и газетных текстов. Используя методы математической статистики, такие как нормальное распределение, распределение Пуассона и распределение Шарлье, были проверены и подтверждены теоретические закономерности распределения данных частей речи. Эксперименты показали значительные особенности в распределении языковых элементов, что способствует дальнейшему пониманию структуры языка и его функционирования в различных контекстах.

Ключевые слова: лингвистическая статистика, математическое распределение, тексты на казахском языке, части речи, распределение Пуассона, Нормальное распределение, моделирование языка.

Annotation: Currently, linguistic statistics is moving from its initial stage of describing linguistic phenomena to developing theories capable of quantitatively predicting language patterns. The main task of the study is to analyze the mathematical distribution of linguistic units in texts, which is key to modeling language phenomena. This research investigated the statistical distribution of major parts of speech in texts in the Kazakh language, including thousands of word usages from M. Auezov's novel "The Path of Abai" and newspaper texts. Using methods of mathematical statistics such as normal distribution, Poisson distribution, and Charlier distribution, theoretical patterns of distribution of these parts of speech were examined and confirmed. The experiments revealed significant features in the distribution of linguistic elements, contributing to a deeper understanding of the structure of language and its functioning in various contexts.

Keywords: linguistic statistics, mathematical distribution, Kazakh language texts, parts of speech, Poisson distribution, Normal distribution, language modeling.

Лингвистическая статистика сейчас завершает свой первоначальный этап, а именно описание лингвистических явлений, и переходит к следующему, более сложному этапу. Это — проблема создания теории, которая могла бы предсказывать языковые закономерности в количественном измерении.

Таким образом, определение математического распределения языковых единиц в тексте становится одной из наиболее важных задач при создании модели языкового явления [Фрумкина Р., 1962: 124-133].

По данной проблеме были проведены исследования, касающиеся проверки соответствия статистического распределения частей речи в текстах на казахском языке некоторым теоретическим закономерностям. В качестве теоретических закономерностей распределения рассматривались «нормальное распределение», «распределение Пуассона» и «распределение Шарлье» в математической статистике. В качестве объектов исследования были выбраны тексты: 10 тысяч словоупотреблений из романа М. Ауэзова «Путь Абая» и 50 тысяч словоупотреблений из газетного текста. Было проверено статистическое распределение наиболее основных частей речи, таких как существительные, глаголы, прилагательные и наречия, в этих текстах. Важным условием исследования становится определение минимального объема текста, при котором не нарушается закономерность распределения конкретного лингвистического элемента (например, определенной части речи) при его использовании в языке.

Эксперимент, проведенный с такой целью, представлен следующим образом:

Разделение слов в тексте по частям речи и их соответствие условным обозначениям;

Разделение общего объема текста на микрочасти, состоящие из 25, 50, 100, 200 и 500 словоупотреблений, то есть на серийные выборочные части;

Соблюдение переменных мер: общий объем текста по каждому стилю – N; объем микрочасти (серии) – k; количество микрочастей на каждую серию – n;

Принятие обобщенной записи каждого случая – B =k (знак равенства условный). Здесь N=n. Например, если рассмотреть текст романа "Путь Абая", состоящий из 10000 словоупотреблений, разделенный на три части: N1=2500, N2=5000, N3=10000, и если каждый Ni разделить на микрочасти объемом k=25, 50, 100, 200, 500 словоупотреблений (n), их количество будет меняться. Например, если k=25, то: N=2500, то B=100 частей; N=5000, то B=200 частей; N=10000, то B=400 частей. Аналогично, если k=50, то B=50; B=100; B=200 или если k=100, то B=25; B=50; B=100. Текст газеты, выбранный для эксперимента, состоящий из 50000 словоупотреблений, также был разделен и исследован по серийным частям, как показано выше;

Для каждого варианта была подсчитана частота необходимой части речи в микрочастях, и было определено количество серий, в которых часть речи встречается с одинаковой частотой. На основе этих данных был создан следующий дискретный вариационный ряд распределения, состоящий из частоты и количества микрочастей:

x_i	x_1	x_2	x_3	…	x_s
n_i	n_1	n_2	n_3	…	n_s

Здесь xi (i=1,2,3,...,s) - частота определенной части речи, ni (i=1,2,3,...,s) - количество микрочастей, соответствующих каждой частоте xi, n = сумма всех ni. $\sum_{i=1}^{s} n_i$ =n.

По результатам проведенного эксперимента изучалась степень соответствия вариационных рядов распределения основных частей речи в текстах художественной литературы и газетных материалов с некоторыми теоретическими закономерностями. В качестве теоретических закономерностей были рассмотрены законы распределения Пуассона, Шарлье (типы А и В), а также нормальное и логарифмически-нормальное распределения.

В качестве примера создания дискретного и непрерывного вариационного ряда рассмотрим выборочную часть текста романа "Путь Абая", состоящую из N=10000 словоупотреблений. Предположим, что длина внутренних частей (микровыборок) или серийных выборок составляет k=100 словоупотреблений, что делает количество серий B =k=B =100. Далее, определим частоту использования существительных в каждой серийной части текста, и предположим, что минимальное количество употреблений – 19, а максимальное – 43. На основе этого интервала (19; 43) можно создать следующий дискретный вариационный ряд, увеличивая каждую частоту (число использований) на "1" [Джубанов А., 1987: 131]:

x_i	19	20	21	22	23	24	25	…	40	41	42	43	Сумма ni:
n_i	1	0	2	3	3	3	5	…	1	0	0	1	100

Для перехода от дискретного вариационного ряда к непрерывному вариационному ряду выбирается шаг равный Δx, и формируется интервальный ряд (xi; xi+Δx). Арифметическое среднее граничных значений интервала обозначается как xj′, а соответствующее количество nj′ для каждого интервала можно найти по формулам: xj′ = (xi + xi+Δx) / 2, (1) nj′ = количество элементов в интервале (xi; xi+Δx). (2) Здесь i=1,2,3,...,n — порядковый номер в дискретном ряду, j=1,2,3,...,m — порядковый номер в непрерывном ряду.

Если Δx=3, тогда непрерывный вариационный ряд может быть представлен следующим образом (пример в Таблице 1):

Таблица 1. Пример непрерывного вариационного ряда

Интервал (xi; xi+3)	Арифметическое среднее xj′	Количество nj′
(19; 22)	20.5	количество_1
(22; 25)	23.5	количество_2
(25; 28)	26.5	количество_3
...	...	...
(40; 43)	41.5	количество_n

Здесь количество_1, количество_2, ... количество_n — это значения nj′ для каждого интервала, вычисленные в соответствии с указанными формулами. Эти данные помогут проиллюстрировать как частоты определенной части речи распределяются в тексте, что существенно для анализа с помощью теоретических закономерностей, таких как законы распределения Пуассона и Шарлье, а также нормального и логарифмически-нормального распределений.

Таким образом, сформированные дискретные и непрерывные вариационные ряды помогают определить эмпирические закономерности распределения частот частей речи в тексте. Эти данные необходимы для использования в специализированных компьютерных программах, которые позволяют вычислять теоретические частоты и статистические параметры.

Давайте теперь рассмотрим формулы для двух типов теоретических законов распределения, которые мы обсуждали ранее.

Закон распределения Пуассона

Основным теоретическим законом в лингвистических исследованиях является закон распределения Пуассона. Этот закон часто используется в экспериментах, когда вероятность каждого отдельного события (p) очень мала, а количество экспериментов очень велико.

Можно предположить, что эти условия также выполняются для частот распределения частей речи в текстах на казахском языке.

Для использования закона Пуассона, как уже упоминалось выше, нам необходимо создать дискретный вариационный ряд на основе статистических данных о частях речи. В эксперименте, определяя частоту возникновения желаемого события, например, встречаемость существительных (x), "математическое ожидание" (a) или его арифметическое среднее и количество экспериментов (n), формула закона Пуассона записывается следующим образом [Смирнов Н., 1969: 280]:

$$n^T = \frac{a^x \cdot e^{-a}}{x!} \cdot n. \qquad (3)$$

В этом (3) уравнении формула арифметического среднего, которая заменяет "математическое ожидание", записывается следующим образом:

$$a \approx \bar{x} = \frac{1}{n} \cdot \sum n_i \cdot x_i' \ . \qquad (4)$$

Закон нормального распределения

Можно сказать, что закон нормального распределения широко используется в статистических исследованиях. Основная причина этого — простота графической формы этого закона и его близость по форме к многим «одногорбым» распределениям. Еще одно преимущество закона нормального распределения заключается в обширности опыта математическо-статистической обработки данных во время экспериментальных наблюдений. Одним из важнейших аспектов закона является возможность его

использования для определения объема части текста и статистической оценки (тестирования).

Переменная и непрерывная величина x принимает значения в интервале от минус бесконечности до плюс бесконечности. Распределение переменной "x" по нормальному закону выражается следующим образом:

$$n^T = \varphi(x) \cdot n = \frac{n \cdot \Delta x}{\sigma} \cdot \frac{1}{\sqrt{2\Pi}} \cdot \ell^{-\frac{(x-a)^2}{2\sigma^2}}. \qquad (5)$$

Здесь x — длина интервала, необходимая для перехода от дискретного вариационного ряда к непрерывному ряду, а — "математическое ожидание" (определяется уравнением 4, упомянутым выше). 2σ2 — это дисперсия, также называемая "центральным моментом второго порядка", и её математическая формула выглядит следующим образом:

$$\sigma^2 = \mu_2 = \frac{1}{n-1} \cdot \sum_i x_i \cdot (x_i - a)^2. \qquad (6)$$

Математическое ожидание (a) и дисперсия (σ^2) являются ключевыми статистическими параметрами, описывающими законы распределения. Например, с их помощью можно определить рассеянность значений переменной x и форму кривой распределения на графике. В практических исследованиях для определения степени соответствия полученных данных теоретическому "нормальному распределению" важно знать параметры "асимметрии" и "эксцесса":

Асимметрия (Skewness) - показывает степень асимметрии распределения данных относительно среднего значения. При нормальном распределении ассиметрия равна нулю. Положительное значение указывает, что распределение имеет длинный правый хвост, а отрицательное - что хвост длиннее слева.

Эксцесс (Kurtosis) - мера остроты пика распределения данных. Для нормального распределения эксцесс равен 3. Значения эксцесса, превышающие 3, указывают на более острый пик по сравнению с нормальным распределением, а значения меньше 3 - на более плоский.

Эти параметры помогают более точно оценить, насколько эмпирическое распределение данных соответствует нормальному распределению, что важно для многих статистических тестов и анализа данных.

$$A = \frac{\mu_3}{\sigma^3}, \qquad (7)$$

$$E = \frac{\mu_4}{\sigma^4} - 3. \qquad (8)$$

Здесь σ обозначает стандартное отклонение, 3μ3 и 4μ4 — это центральные моменты третьего и четвёртого порядков соответственно:

Стандартное отклонение (σ) — это мера разброса значений вокруг среднего (математического ожидания). Оно показывает, насколько широко значения варьируются от среднего значения распределения. Рассчитывается как квадратный корень из дисперсии.

Третий центральный момент (3μ3) — используется для оценки асимметрии распределения. Положительное значение 3μ3 указывает на асимметрию распределения с длинным правым хвостом (правосторонняя асимметрия), отрицательное значение — с длинным левым хвостом (левосторонняя асимметрия).

Четвёртый центральный момент (4μ4) — используется для оценки эксцесса распределения. Это мера того, насколько плоским или острым является пик распределения по сравнению с нормальным распределением. Если 4μ4 больше, чем у нормального распределения, пик будет более выраженным, а если меньше — более плоским.

Эти параметры важны для полного характеристического описания распределения, особенно в статистическом анализе, где точное понимание формы распределения может влиять на выводы исследования.

$$\mu_3 = \frac{1}{n} \cdot \sum_i n_i \cdot (x_i - a)^3 , \qquad (9)$$

$$\mu_4 = \frac{1}{n} \cdot \sum_i n_i \cdot (x_i - a)^4 . \qquad (10)$$

Для нормального распределения значения третьего центрального момента (асимметрия, А) и четвёртого центрального момента (эксцесс, Е) равны нулю. Поэтому, если значения асимметрии (А) и эксцесса (Е) для анализируемого распределения также близки к нулю, можно предположить, что эмпирическое распределение приближается к нормальному. Наоборот, чем больше отклонения значений А и Е от нуля, тем меньше соответствие между экспериментальным распределением и нормальным законом распределения.

Для более глубокого изучения теоретических законов распределения, таких как Шарлье типов А и В, а также логарифмически нормального распределения, можно обратиться к следующей литературе:

Статистические методы анализа и обработки данных - этот источник может предоставить подробные объяснения и примеры использования различных типов распределений, включая их применение в реальных исследованиях.

Теория вероятностей и математическая статистика - книга, где обсуждаются основы теории вероятностей, включая различные типы распределений и их свойства.

Прикладная статистика и основы метрологии - источник, который может быть полезен для понимания применения статистических методов в научных и инженерных задачах, включая анализ распределений [Романовский В., 1963: 94], [Хальд Л., 1956: 61].

Эти ресурсы помогут получить более глубокое понимание различных распределений и методов их анализа в контексте прикладной статистики и данных.

Список использованных источников:

1. Джубанов А.Х. Квантитативная структура казахского текста (опыт лингвистического анализа на ЭВМ). Алма-Ата, 1987. С. 147.
2. Романовский В.И. Математическая статистика. Ташкент, 1961. Кн. 1. 637 с.; 1963. Кн. 2. 794 с.
3. Смирнов Н.В., Дунин-Барковский И.В. Курс теории вероятностной и математической статистики. М., 1969. 511с.
4. Фрумкина Р.М. О законах распределения слов и классов слов // Структурно-типологические исследования. М., 1962. С. 124-133.
5. Хальд Л. Математическая статистика с техническим приложением. М., 1956. 644 с.