

*Горелик Людмила Борисовна*

## ЛАБОРАТОРНЫЕ РАБОТЫ: ИЗУЧЕНИЕ СТАТИСТИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ

*От редакции: эта статья продолжает цикл публикаций Л.Б. Горелик и является второй частью статьи, опубликованной в предыдущем номере журнала.*

Большое количество лабораторных работ в ИУМК «Математика в школе – XXI век» посвящено свойствам функций. О том, как можно выполнить подобную работу, мы уже говорили.

Кроме них, есть интересные сюжеты, где ученик имеет дело с числами и их свойствами. Это, например, работы по изучению признаков делимости, вычислению подходящих дробей, оценке скорости роста факториала, установлению связи между количеством цифр числа, записанного в какой-нибудь системе счисления, и целой частью логарифма по основанию, равному основанию системы счисления. Необычна и крайне интересна лабораторная работа, посвященная производящим функциям. Есть графические работы. Набор лабораторных работ широк и разнообразен.

Все эти работы касаются одного типа закономерностей: абсолютно жестких, однозначных. В физике их называют динамическими. Но современная физика и современная математика исходят из примата *статистических закономерностей*. Именно они являются фундаментальными и по сравнению с динамическими глубже выражают причинно-следственные связи. В школьных программах по математике стохастическая линия появилась совсем недавно, тем интереснее познакомиться с лабораторной работой, посвященной этому необычному

для школьников материалу. Для таких исследований именно лабораторная работа как форма занятия подходит наилучшим образом.

### ЛАБОРАТОРНАЯ РАБОТА «СТАТИСТИЧЕСКИЙ АНАЛИЗ В ЛИНГВИСТИКЕ»

Цель работы – исследование распределения длин предложений в литературных произведениях различных авторов. Это цель для учащихся, выполняющих работу. Цель учителя другая. В ходе выполнения работы нужно привить учащимся навыки обработки статистических данных.

Если придерживаться только этой цели, то учитель должен выполнить следующие задачи:

– познакомить учащихся с такими понятиями, как сплошное и несплошное наблюдение, выборочный метод исследования, генеральная совокупность и выборочная совокупность, случайная выборка, статистическое распределение выборки, группировка, интервалы группировки, частоты интервалов, гистограмма частот, полигон частот;

– обучить некоторым методам получения случайной выборки (в этой работе понадобилось делать не только собственно случайную, но и механическую выборку);

– научить вычислять некоторые характеристики статистического распределения: среднее арифметическое, медиану, моду (если она существует в данном распределении), дисперсию;

– показать способы графического представления статистических данных, научить строить гистограмму и полигон частот.

### 1. ПОСТАНОВКА ПРОБЛЕМЫ

Перечисленные понятия относятся к описательной статистике. Но для того, чтобы работа имела смысл, – вспомним о том, что мы имеем дело с конкретным содержанием, – хорошо бы было также обучить учащихся некоторым методам *аналитической* статистики. Речь идет о формулировке выводов, имеющих прикладное значение для конкретной области человеческой деятельности. В данном случае мы должны получить некоторое заключение о длине предложений в произведениях разных авторов. Очень важно научить делать это правильно, так как, пытаясь сформулировать вывод, учащиеся допускают распространенную ошибку: при разных статистических распределениях выборок они сравнивают непосредственно средние арифметические длин предложений и, естественно, делают неверные выводы. Например, если у одного автора средняя длина предложений равна 22 словам, а у другого 32, то учащиеся делают вывод: у первого автора предложения в среднем короче. На самом деле это не так. Чуть позже мы поговорим об этом.

Вернемся к смыслу работы. Десятиклассники делают анализ художественного текста на уроках литературы. Они говорят о

композиции произведения, о времени и месте действия, перечисляют художественные средства, которые использует автор, анализируют систему образов и делают вывод об авторском стиле. Перечисленные признаки авторского стиля – качественные. Статистика же имеет дело с количественными данными: длиной предложений, частотой использования частей речи, частотой использования тех или иных знаков препинания, количеством инверсий в предложении (обычно подлежащее предшествует сказуемому, в противном случае говорят об инверсии) и так далее. Перечислены далеко не все группы данных, которые могут быть количественными признаками авторского стиля. Кстати, не всегда подобные данные характеризуют авторский стиль. Например, если окажется, что у разных авторов примерно одинаковая средняя длина предложения, то по этой величине невозможно распознать автора произведения или выделить писателя, который по данному параметру отличается от других. Еще один пример. Такой количественный признак, как длина слова, распознает не автора художественного произведения, а язык: всем известен научно доказанный факт, говорящий о том, что средняя длина слова русского языка равна 6 буквам.

В связи с этим возникает проблема: какие произведения следует взять в работу?

Это зависит от того, что мы собираемся исследовать, подсчитывая число слов в предложениях текста того или иного автора. Если мы надеемся на то, что длина предложения является отличительным признаком авторского стиля, то мы должны сделать выборки из произведений различных авторов. По мнению многих, в русскоязычной литературе самые длинные предложения в произведениях Л.Н. Толстого. Кого из писателей взять для сравнения? Всем известен афоризм А.П. Чехова: «Краткость – сестра таланта». Может быть, это относится именно к длинам предложений? Итак, выбираем для сравнения произведения Л.Н. Толстого и А.П. Чехова.

Но может быть, длина предложения связана также и с жанром произведения? Если



*Десятиклассники делают анализ художественного текста...*

мы хотим проверить это предположение, нужно взять произведения разных жанров одного и того же автора. Если же нас интересует и первое, и второе, выполняем двойную работу: сначала берем произведения разных авторов (одного жанра, чтобы один параметр не влиял на другой), а затем анализируем разножанровые произведения одного и того же автора. У А.П. Чехова нет романов, которые могли бы сравниться с романом Л.Н. Толстого «Война и мир» ни по значимости описываемых событий, ни по времени их протекания. Но, тем не менее, в произведениях Чехова есть то, что можно сравнить по размаху с произведениями Толстого. Сошлемся на слова одного из героев романа В.С. Гроссмана «Жизнь и судьба»: «Чехов ввел в наше сознание всю громаду России, все ее классы, сословия, возрасты». Именно демократизм Чехова, давший русской литературе огромное количество образов разных людей, роднит его с Толстым: «Он сказал, как никто до него, даже Толстой не сказал: все мы, прежде всего, люди,... а потом уже архиереи, русские, лавочки, татары, рабочие».

Вполне вероятно, что длина предложения зависит от пространственно-временного контекста художественного произведения. Есть произведения-эпопеи, захватывающие большие события, происходившие на большой территории в течение длительного промежутка времени. Таким произведением является, например, упомянутый раньше роман-эпопея Л.Н. Толстого «Война и мир». Для того чтобы исследовать эту возможную связь, сравним два романа Л.Н. Толстого: «Война и мир» и «Анна Каренина».

Сделаем еще одно предположение. Возможно, длина предложения – признак эпохи и сложившейся в ней культуры, в том числе и речевой. Значит, нужно взять двух разных авторов произведений с примерно одинаковым пространственно-временным контекстом, но таких, которые жили бы в разные исторические эпохи. Для экономии времени и энергии оставляем «Войну и мир» Л.Н. Толстого и подбираем второго автора. Вот что говорит искусствовед В. Лакшин о

романе «Жизнь и судьба» В.С. Гроссмана, русского писателя XX века: «Великая традиция вела Василия Гроссмана. Он сознательно видел перед собою образец двух русских писателей-классиков. Один – уже упомянутый автор «Войны и мира»: его размах, его сплетение дальних линий жизни, судеб в один узел, их схождение из далеких пространств в одно историческое действие, возможность повествования, величавее и крупнее всего явленная Львом Толстым. Но и другое имя стоит для В. Гроссмана рядом с создателем «Войны и мира» – это Чехов». Событийная составляющая романов «Война и мир» и «Жизнь и судьба» одна и та же: это крупнейшие трагедии для русского народа. В первом романе – война 1812 года, во втором – Великая Отечественная война 1941-1945 годов. Но между Л.Н. Толстым и В.С. Гроссманом стоит полуторавековая разница во времени их творчества и описываемых событий. Две разные эпохи. Две разные культуры.

Вполне возможно, что ни эпохальность произведения, ни время его написания не являются причиной удлинения предложений. Тогда нужно посмотреть, как влияет на длину фраз стиль произведения. Идеальным было бы сравнение двух частей романа М.А. Булгакова «Мастер и Маргарита»: античного романа, написанного Мастером, и современного романа, в который и вплетен роман античный. Пафос и высокий стиль первого из них контрастирует с обыденным и ироничным стилем современного романа.

Что еще может характеризовать длина предложения? Может ли она оказаться свой-



*Но может быть,  
длина предложения связана также  
и с жанром произведения?*

ством конкретного живого языка? Тогда нужно анализировать тексты, написанные на разных языках. Интересно было бы взять аутентичный текст (текст подлинный, написанный автором на родном языке) и его литературный перевод на другой язык. Но для этого, возможно, понадобится владение иностранным языком, превышающее школьный уровень.

## 2. ОРГАНИЗАЦИЯ РАБОТЫ

Итак, для проверки гипотез, сформулированных относительно влияния различных параметров на длину предложения в тексте, нужно выполнить большой объем работы. Естественно, она не укладывается в рамки классно-урочной системы, поэтому нужно подобрать подходящую педагогическую технологию. Здесь больше всего подойдет метод проектов. Учащиеся делятся на группы, и каждая выполняет свою часть работы по сбору данных. Технология сбора и представления данных в лабораторной работе описана, но нужно внести необходимые уточнения.

Это касается подсчета слов в предложении. Следует ли считать служебные слова? Как считать сложно-составные предлоги: «в отличие от», «впредь до», «несмотря на», «по пути к»? Составные союзы: «так как», «потому что», «для того чтобы», «после того как»? Как считать наречия и прилагательные, составленные из двух частей речи и соединенные дефисом: «серо-зеленый», «высоко-высоко», «больно-пребольно»? Наречия с приставкой: «по-хорошему», «по-матерински»?

Проблемой подсчета является и положение фразы в тексте. Если речь идет об авторском стиле, то принимать ли во внимание диалоги? Ведь в них могут отражаться особенности речи персонажей, поэтому, например, краткость неразвитой, вульгарной речи, сленга героев художественных произведений никак не может отражать авторский стиль.

Обо всем этом нужно заранее договориться, чтобы во всех группах были одни и те же правила подсчета.

Группы лучше формировать по 2–3–4 человека. После выполнения задания каждая из них демонстрирует свою работу на семинарском занятии, где материал собирается воедино и делаются общие выводы.

В нашем случае можно предложить следующие задания группам (см. табл. 1).

## 3. РАСШИРЕНИЕ ДЕЯТЕЛЬНОСТИ

Как показала практика, даже 60–100 предложений в выборке не гарантируют получения нормального распределения, поэтому сравнение средних не является корректным методом сравнения выборок и ведет к ошибочным выводам. Так, например, если в одной из выборок есть сравнительно большая величина, то она существенно влияет на среднее арифметическое, но в этом случае оно не отражает сути. Часто учащиеся прогнозируют четвертные оценки по предмету, вычисляя именно среднее арифметическое. Предположим, что у ученика стоят в журнале отметки 5, 4, 2, 2, 2. Среднее арифметическое равно 3. Оценка положительная, но знания отсутствуют.

Но даже если распределение частот оказывается нормальным, то простое сравнение средних ведет к неправильным выводам. Одним из направлений в статистическом анализе является правильное сравнение, проверка гипотезы о различии или сходстве двух эмпирических выборок. Для этого нужно пользоваться определенными статистическими критериями. Для нормальных распределений существуют так называемые параметрические критерии: *t-критерий Стьюдента*, *критерий F* и другие. Это критерии, включающие в формулу расчета параметры нормального распределения: дисперсию и среднее арифметическое.

В случае, когда распределение частот не является нормальным, средние не сравниваются. Для этого случая существуют непараметрические критерии, например, критерий знаков, *U-критерий Манна-Уитни*.

В таблице критических значений *U-критерия Манна-Уитни* наибольший объем выборки равен 30, поэтому, убедившись,

Табл. 1

№ гр.	Гипотезы о факторах, влияющих на длину предложений в тексте	Исходный материал для сбора данных	Выводы
1	Эпоха	– Л.Н. Толстой. «Война и мир», – В.С. Гроссман. «Жизнь и судьба»	Приложение 1
2	Эпоха	– Л.Н. Толстой. «Война и мир», – М.А. Булгаков. Роман «Мастер и Маргарита» (античный роман Мастера – роман в романе)	Приложение 2
3	Эпоха	– Л.Н. Толстой. «Война и мир», – М.А. Булгаков. Роман «Мастер и Маргарита» (современный роман)	Приложение 3
4	Жанр произведения	– Л.Н. Толстой. «Война и мир», – Л.Н. Толстой. Рассказы «Метель», «Утро помещика», – Л.Н. Толстой. Очерк «Рубка леса»	Приложения 6–8
5	Жанр произведения	– А.П. Чехов. Рассказы «Женщина без предрассудков», «Случай из судебной практики», «Бенефис соловья», «Дачница» – А.П. Чехов. Повесть «Степь», – А.П. Чехов. Очерк «Остров Сахалин».	Приложения 9–13
6	Жанр произведения	– М.А. Булгаков. Роман «Мастер и Маргарита» (античный роман Мастера – роман в романе), – М.А. Булгаков. Роман «Мастер и Маргарита» (современный роман). – М.А. Булгаков. Рассказы и фельетоны «Торговый ренессанс», «Чаша жизни», «Неделя просвещения», «Сорок сороков», «Московские сцены»	Приложения 14–20
7	Пространственно-временной контекст	– Л.Н. Толстой. «Война и мир», – Л.Н. Толстой. «Анна Каренина»	Приложение 5
8	Стиль повествования	– М.А. Булгаков. Роман «Мастер и Маргарита» (античный роман Мастера – роман в романе), – М.А. Булгаков. Роман «Мастер и Маргарита» (современный роман)	Приложения 14–20
9	Авторский стиль	– Л.Н. Толстой. «Война и мир», – А.П. Чехов. Повесть «Степь», – В.С. Гроссман. «Жизнь и судьба»	Приложения 1 и 4

что распределение частот непараметрическое, мы ограничились выборкой в 30 элементов.

Для сравнения двух выборок объемом  $n = 30$  критическое значение  $U$ -критерия равно 317. При условии, что объем выборок не будет меняться, можно заранее рассчитать то число, с которым мы будем сравнивать

большую из двух ранговых сумм, пользуясь формулой для расчета  $U$ -критерия:

$$U = n_1 n_2 + \frac{n_x (n_x + 1)}{2} - T_x,$$

где  $T_x$  – большая из двух ранговых сумм,  $x$  – номер выборки, которой соответствует большая ранговая сумма.



Для того чтобы подтвердилась гипотеза о существенном различии распределений,  $U$  должно быть меньше или равно 317:

$$U = 30 \cdot 30 + 15 \cdot 31 \leq 317,$$

то есть  $1365 - T_x \leq 317$ , откуда  $T_x \geq 1048$ .

Таким образом, вычислив наибольшую ранговую сумму, мы будем сравнивать ее с числом 1048. Если сумма будет не меньше 1048, то различия распределений значимые.

#### 4. ВЫВОДЫ

В приложениях для каждой выборки вычислены: дисперсия, стандартное отклонение, медиана, мода (если есть), среднее значение и наибольшая сумма рангов. Используя среднюю величину и стандартное отклонение, мы смогли посмотреть характер распределения. Почти в каждой выборке есть значения, которые выходят за пределы  $2U$ . Это значит, что очень длинные, по сравнению с другими, фразы встречаются в выборках редко. Это, скорее, исключение из характерной для автора произведения длины предложений. Так, например, в повести А.П. Чехова «Степь» есть предложения, содержащие соответственно 81, 83 и 85 слов. Разрыв между этими значениями и остальными в выборке значителен. В этом случае они считаются «ошибкой эксперимента» и отбрасываются. Мы не стали их отбрасывать, потому что для сравнения распределений с помощью  $U$ -критерия Манна-Уитни важны не сами значения, а их ранги в едином ранжированном ряду. Но такие числа, которые выходят за пределы двойного стандартного отклонения, влияют на среднее значение выборки, что может послужить поводом для формулирова-

ния неверного вывода. Так, например, средняя длина предложения в романе Толстого «Война и мир» составляет 22 слова, а в повести Чехова «Степь» 28 слов. Но вывод о том, что фразы в повести «Степь» длиннее, неверен. Применение  $U$ -критерия позволило сделать правильный вывод: различия в длине предложений в повести «Степь» и в романе «Война и мир» статистически незначимые.

Подводя итог исследованиям, мы вынуждены признать, что из всех рассмотренных произведений:

– только в произведениях Л.Н. Толстого длина предложений не зависит ни от жанра, ни от пространственно-временного контекста, ни от стиля повествования: в его рассказах и очерках длина предложений статистически сравнима с длиной предложений его разных романов, следовательно, длинные предложения являются особенностью его авторского стиля;

– в рассказах-миниатюрах Чехова и повести «Степь» различия в длинах предложений статистически значимы, то есть у этого автора жанр оказывает влияние на длину предложений;

– то же самое можно сказать о рассказах М.А. Булгакова (у Булгакова длина предложений зависит от стиля повествования, на длину предложений влияет также описываемая в произведениях Булгакова эпоха: в античном романе предложения длиннее, чем в современном).

Исследования нужно продолжить, сделав еще несколько выборок для каждого автора. Возможно, некоторые из сформулированных выводов не подтвердятся.



Наши авторы, 2008.

Our authors, 2008.

*Горелик Людмила Борисовна,  
учитель математики  
МОУ лицей № 102 г. Челябинска.*