Проектно - исследовательская работа «Статистическая обработка данных»

Выполнили ученицы 8 «Б» класса МОУ Калачеевская гимназия№1

Кучеровы Ж. Л. Козюберда Е. Шмигирилова Д.

ЦЕЛИ :

- Собрать информацию о том, что такое статистика, что изучает статистика, где применяются результаты статистических исследований.
- Определить место статистики в изучении окружающего мира, различных общественных и социально-экономических явлений.
- На конкретных примерах осуществить статистические исследования и наглядно представить статистическую информацию.
- Выяснить значимость компьютерных технологий в решении статистических задач.

ПРОЕКТНОЕ ЗАДАНИЕ:

Выбрав отрывок из произведения русского писателя и отрывок из стихотворения русского поэта, провести статистический анализ текстов.

- **Составить таблицу распределения по частотам и относительным частотам всех букв русского алфавита. Сравнить полученный результат с частотной таблицей букв русского языка из книги А.М. Яглома. Разбить алфавит на 3 участка:№1 - от «а» до «й», №2 - от «к» до «ф»,№3 - от «х» до «я». Составить таблицу распределения частот участков. Указать участок наибольшей частоты. Построить гистограмму с выбранным распределением на участки.**

- **Определить в тексте количество глаголов и существительных. Считая рассмотренную выборку репрезентативной, определить примерное количество глаголов и существительных в отрывке этого же произведения объёмом 2000 слов.**

- **Составить таблицы распределения по частотам и по процентным частотам количества букв в слове. Найти среднее арифметическое, размах, моду, медиану выборки. Построить гистограмму и полигон распределения.**

Введение

Наша группа взяла для исследования отрывок из повести «Дубровский» и стихотворение «Я помню чудное мгновенье» А.С.Пушкина. Этот выбор был неслучаен. На занятиях кружка мы узнали, что, хотя Пушкин и очень не любил математику , но в его библиотеке имелись две книги по теории вероятностей и статистике. Одна из книг представляла собой знаменитый труд великого французского математика и механика Лапласа «Опыт философии теории вероятностей», вышедший в Париже в 1825 году. А книгу французского математика , инженера-кораблестроителя и статистика Шарля Дюпена «Производительные и торговые силы Франции», изданную в 1827 году, Пушкин называл «философическими таблицами».В ней приводятся сравнительные статистические таблицы по экономике некоторых европейских стран, в том числе и России.

Сбор информации (проза)

Похороны[8] свершились[11] на[2] третий[6] день[4]. Тело[4] бедного[7] старика[7] лежало[6] на[2] столе[5], покрытое[8] саваном[7] и[1] окружённое[10] свечами[7]. Столовая[8] полна[5] была[4] дворовых[8]. Готовились[10] к[1] выносу[6]. Владимир[8] и[1] трое[4] слуг[4] подняли[7] гроб[4]. Священник[9] пошёл[5] вперёд[6], дьячок[6] сопровождал[11] его[3], воспевая[8] погребальные[12] молитвы[7]. Хозяин[6] Кистенёвки[10] последний[9] раз[3] перешёл[7] за[2] порог[5] своего[6] дома[4]. Гроб[4] понесли[7] рощею[5]. Церковь[7] находилась[10] за[2] нею[3]. День[4] был[3] ясный[5] и[1] холодный[8] Осенние[7] листья[6] падали[6] с[1] дерев[5] При[3] выходе[6] из[2] рощи[4] увидели[7] кистенёвскую[12] деревянную[10] церковь[7] и[1] кладбище[8], осенённое[9] старыми[7] липами[6]. Там[3] покоилось[9] тело[4] Владимировой[12] матери[6]; там[3], подле[5] могилы[6] её[2], накануне[8] вырыта[6] была[4] свежая[6] яма[3]. Церковь[7] полна[5] была[4] кистенёвскими[13] крестьянами[11], пришедшими[10] отдать[6] последнее[9] поклонение[10] господину[9] своему[6]. Молодой[7] Дубровский[10] стал[4] у[1] клироса[7]; он[2] не[2] плакал[6] и[1] не[2] молился[7], но[2] лицо[4] его[3] было[4] страшно[7]. Печальный[9] обряд[5] кончился[8]. Владимир[8] первый[6] пошёл[5] прощаться[9] с[1] телом[5], за[2] ним[3] и[1] все[3] дворовые[8]. Принесли[8] крышку[6] и[1] заколотили[10] гроб[4]. Бабы[4] громко[6] выли[4]; мужики[6] изредка[7] утирали[7] слёзы[5] кулаком[7]. Владимир[8] и[1] тех[3] же[2]…

1) Посчитав частоту появления каждой буквы в данном тексте, составим выборку , вариантами которой являются буквы русского алфавита.

2) Посчитав количество букв в каждом слове отрывка, составим выборку , вариантами которой являются количество букв в слове.

3)Упорядочив данные значений вариант, составим вариационные ряды и заполним соответствующие таблицы распределения частот

Таблица частот появления букв русского алфавита (проза)

Буквы	а	б	в	г	д	е	ё	ж	з	и	й
Кратность	54	15	40	15	32	68	10	6	9	72	9
Относительная частота	54/877	15/877	40/877	15/877	32/877	68/877	10/877	6/877	9/877	72/877	9/877
Частота %	6.2	1.7	4.6	1.7	3.6	7.8	1.1	0.7	1	8.2	1
Частота % (част.табл. Яглома)	6,2	1,4	3,8	1,3	2,5	7,2	1	0,7	1,6	6,2	1

Буквы	к	л	м	н	о	п	р	с	т	у	ф
Кратность	36	64	30	54	100	34	51	45	30	15	7
Относительная частота	36/877	64/877	30/877	54/877	100/877	34/877	51/877	45/877	30/877	15/877	7/877
Частота %	4.1	7.3	3.4	6.2	11.4	3.9	5.8	5.1	3.4	1.7	0,8
Частота % (част.табл. Яглома)	2,8	3,5	2,6	5,3	9	2,3	4	4,5	5,3	2,1	0,2

Буквы	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
Кратность	7	4	4	8	5	0	25	16	0	4	15
Относительная частота	7/877	4/877	4/877	8/877	5/877	-	25/877	16/877	-	4/877	15/877
Частота %	0.5	0.5	0.9	0.6	2.9	-	0.5	1.7	-	0.55	1.7
Частота % (част.табл. Яглома)	0,9	0,4	1,2	0,6	0,3		1,6	1,4	0,2	0,6	1,8

Сравнив полученный результат , видим , что таблица частот для данной выборки отличается от частотной таблицы из книги А.М. Яглома , но общая закономерность прослеживается. Для букв « а, е, ё, ж, й, ц, ш, ь, ю, я » частоты данных таблиц одинаковы.

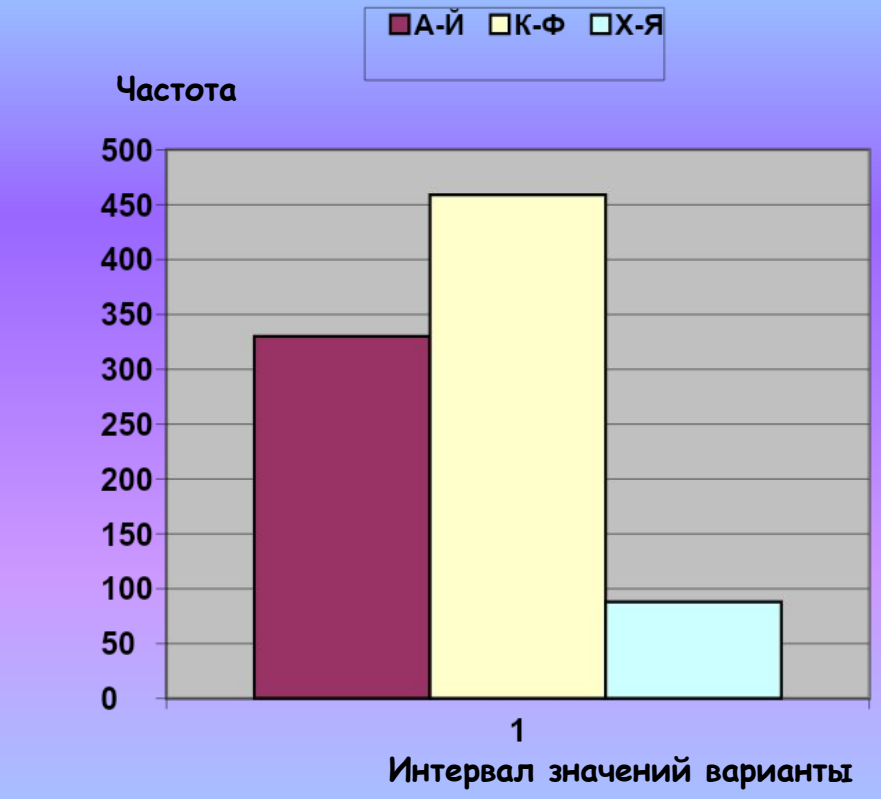
Числовые характеристики

- Мода: буква «о» -(чаще всего встречающаяся в произведении буква)
- Количество глаголов-28
- Количество существительных-53

Таблица частот интервального ряда

Интервал значений варианты	А - Й	К - Ф	Х – Я
Частоты	330	459	88
Относительные частоты	330/877	459/877	88/877
Частоты %	37,6	52,4	10

Гистограмма и полигон частот появления букв русского языка (проза)



В отрывке из 156 слов мы получили 28 глаголов и 53 существительных. Количество глаголов и существительных в отрывке этого же произведения объёмом 2000 слов находили, используя определение репрезентативной выборки, по формуле

$$Si/S=Mi/N,$$

где N-объём репрезентативной выборки,

S- объём генеральной выборки,

Mi- частота (кратность) варианты репрезентативной выборки,

Si- частота (кратность) варианты генеральной выборки.

В нашем случае N=156,M1=28,M2=53,S=2000, поэтому

- для глаголов: S1=374;
- для существительных: S2=707.

Таблица распределения частот количества букв в слове (проза)

Варианты	1	2	3	4	5	6	7	8	9	10	11	12	13
Кратность	13	12	13	21	13	24	22	15	8	9	3	2	1
Частота %	8,3	7,6	8,3	13,5	8,3	15,4	14,2	9,6	5,1	5,8	1,9	1,4	0,6

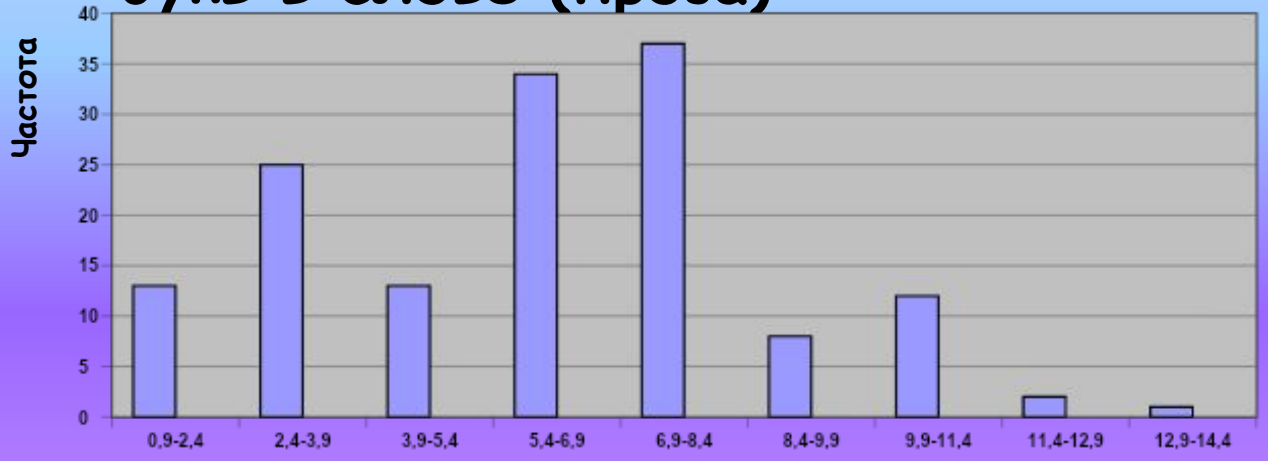
Числовые характеристики

- **Мода M(0)=6** (чаще всего в этом произведении встречаются слова, состоящие из 6 букв)
- **Медиана M(e)= 6** (половина слов состоит не менее чем из 6 букв)
- **Среднее арифметическое** $\overline{X}$ = **7** (среднее количество букв в слове)

Таблица частот интервального ряда

Интервальное значение варианты	0,9-2,4	2,4-3,9	3,9-5,4	5,4-6,9	6,9-8,4	8,4-9,9	9,9-11,4	11,4-12,9	12,9-14,4
Частота	25	13	34	24	37	8	12	2	1
Частота %	16	8,3	21,8	15,4	23,2	5,1	7,7	1,9	0,6

Гистограмма и полигон частот количества букв в слове (проза)



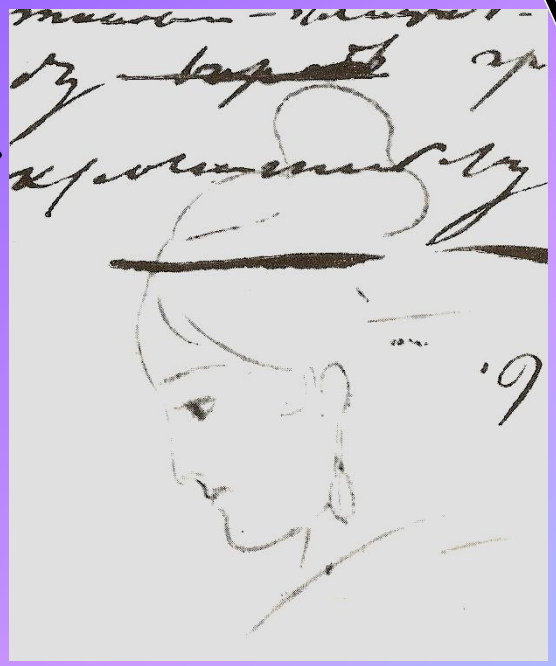
Сбор информации (стихотворение)

1 5 6 9
- Я помню чудное мгновенье.

6 4 7 2
- Передо мной явилась ты.

3 10 7
- Как мимолетное виденье,

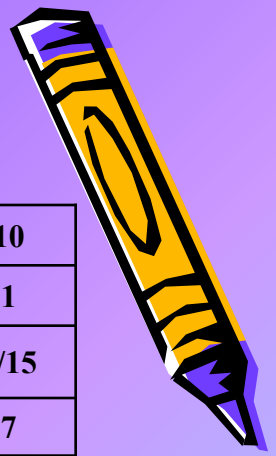
3 5 6 7
- Как гений чистой красоты.



1) Посчитав количество букв в каждом слове отрывка, составим выборку , вариантами которой являются количество букв в слове.

2) Посчитав частоту появления каждой буквы в данном стихотворении, составим выборку , вариантами которой являются буквы русского алфавита.

Таблица распределения частот количества букв в слове (стихотворение)



Варианта	1	2	3	4	5	6	7	9	10
Кратность	1	1	2	1	2	3	3	1	1
Отн.частота	1/15	1/15	2/15	1/15	2/15	3/15	3/15	1/15	1/15
Частота,%	7	7	14	7	14	20	20	7	7

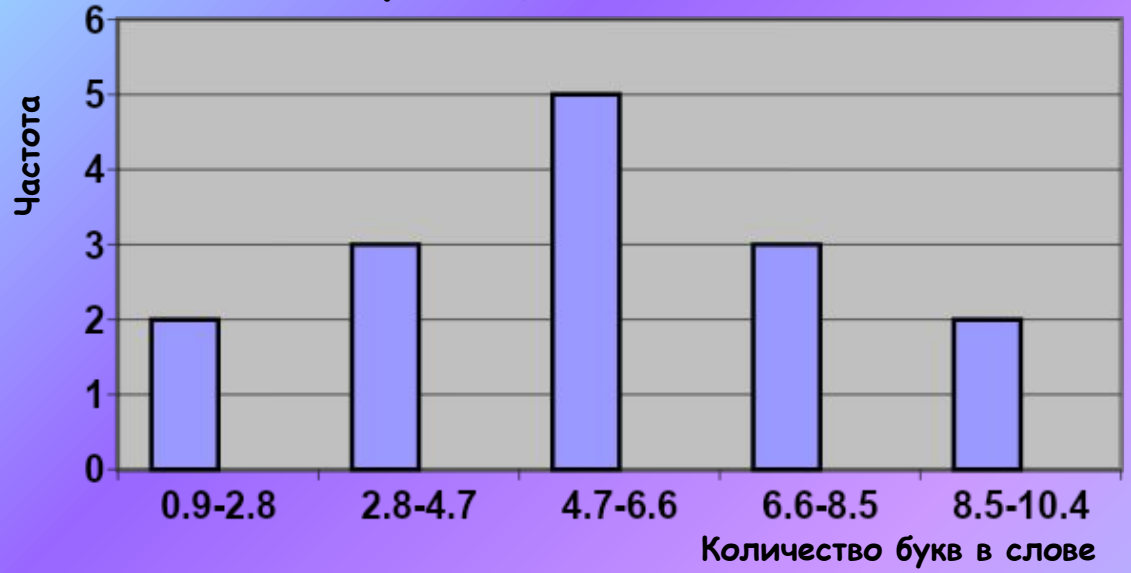
Числовые характеристики

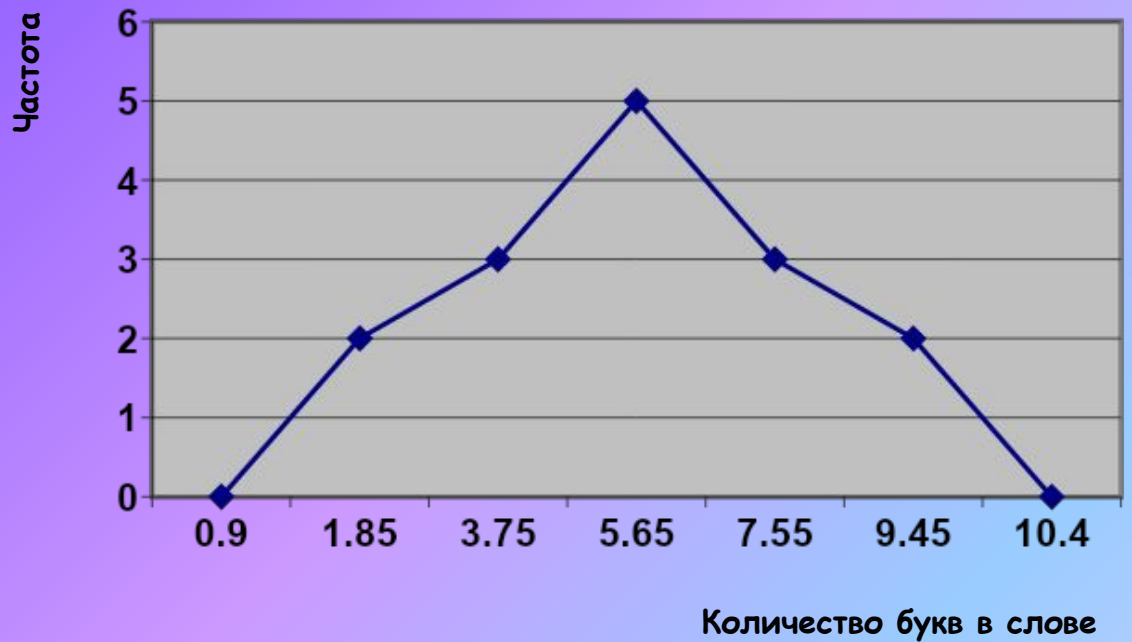
- **Среднее арифметическое**: $\overline{X}$ = 5,4 (среднее количество букв в слове)
- **Размах: R = x (max) – x (min);** R = 9 (разность между наибольшим и наименьшим количеством букв в словах)
- **Мода:** M1 (0) = 6; M2 (0) = 7 (чаще всего встречаются слова, состоящие из 6 и 7 букв)
- **Медиана** = 6 (половина слов состоит не менее чем из 6 букв)

Таблица частот интервального ряда

Интервальное значение варианты	0,9 – 2,8	2,8 – 4,7	4,7 – 6,6	6,6 – 8,5	8,5 – 10,4
Частота	2	3	5	3	2
Относительная частота	2/15	3/15	5/15	3/15	2/15
Частота %	14	20	34	20	14

Гистограмма и полигон частот количества букв в слове (стихотворение)





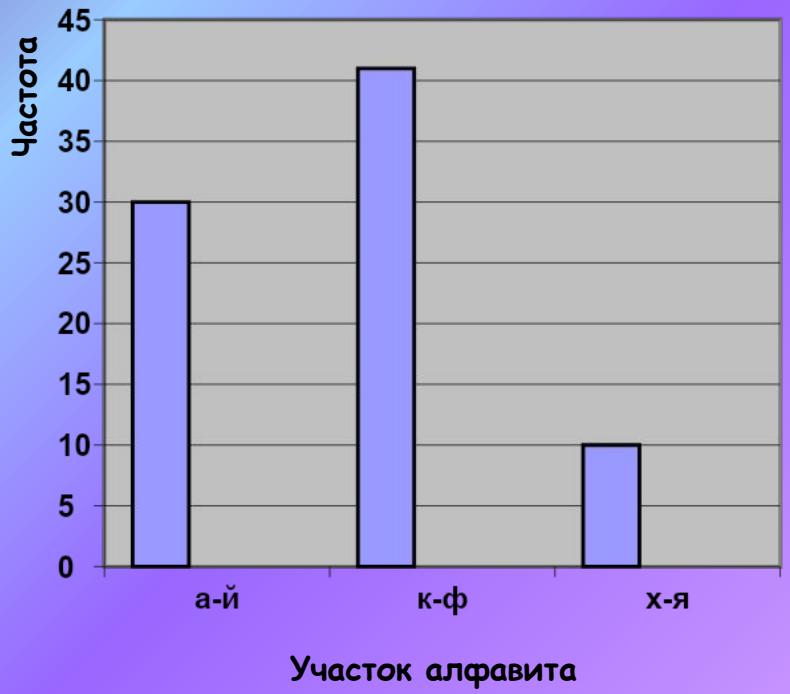
Числовые характеристики

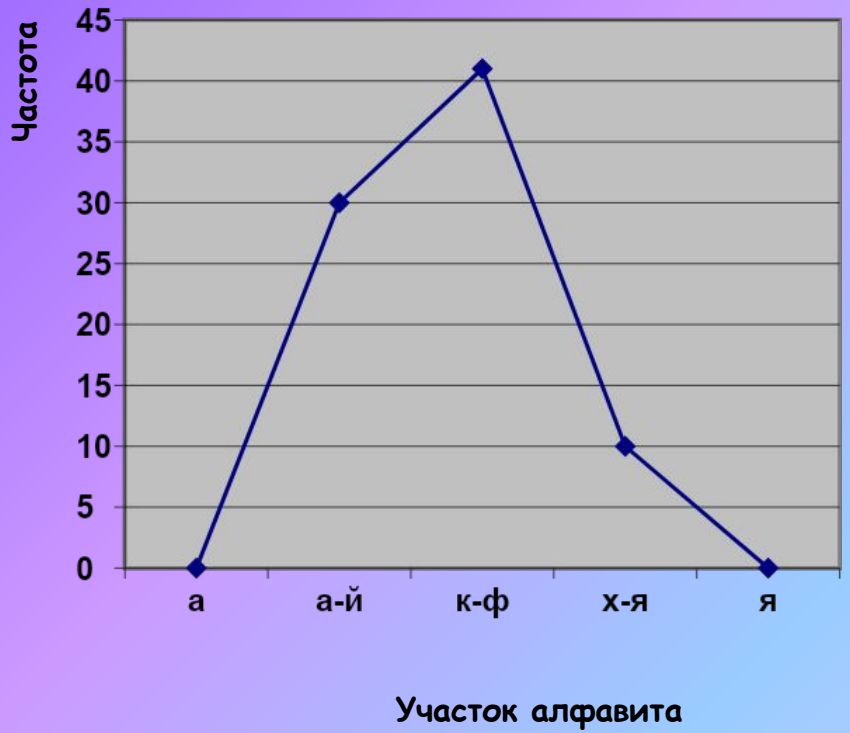
- Мода: буква «о» -(чаще всего встречающаяся в стихотворении буква)

Таблица частот интервального ряда появления букв русского алфавита (стихотворение)

Интервал знач. варианты	«а»-«й»	«к»-«ф»	«х»-«я»
Частота	30	41	10
Частота, %	37	51	12

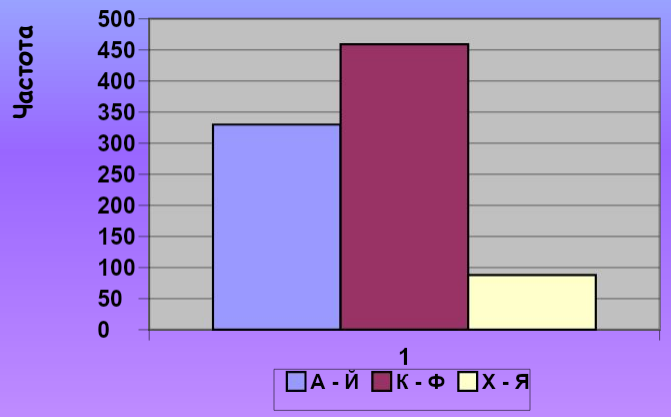
Гистограмма и полигон распределения частот появления букв русского алфавита (стихотворение)

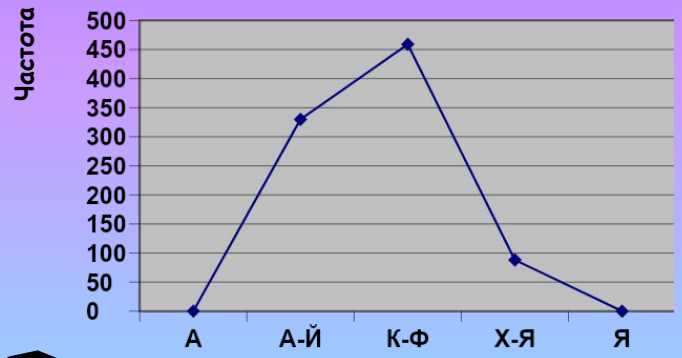




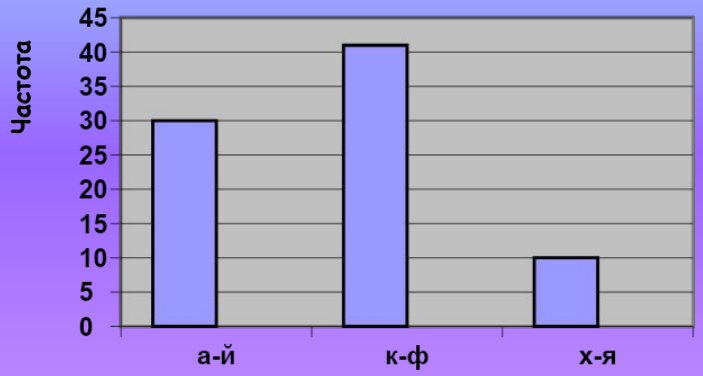
Сравнительная характеристика распределения частот появления букв русского алфавита

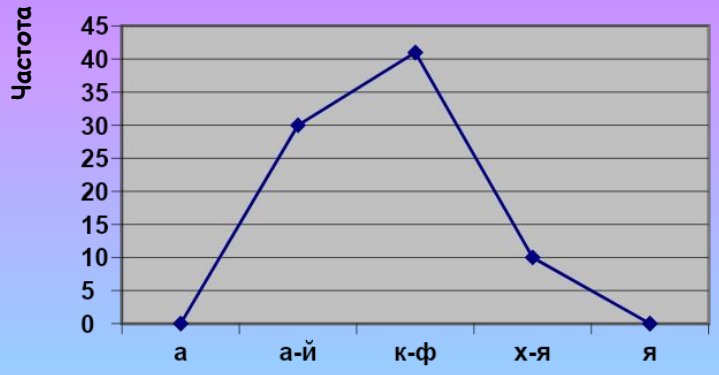
Проза





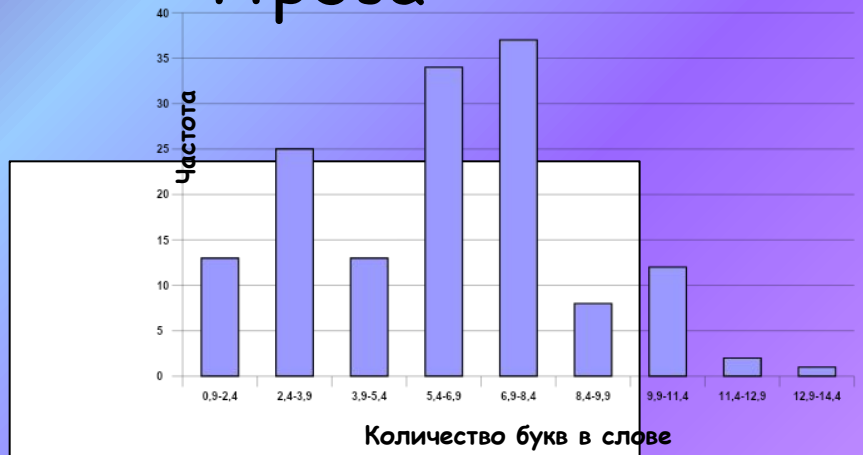
Стихотворение

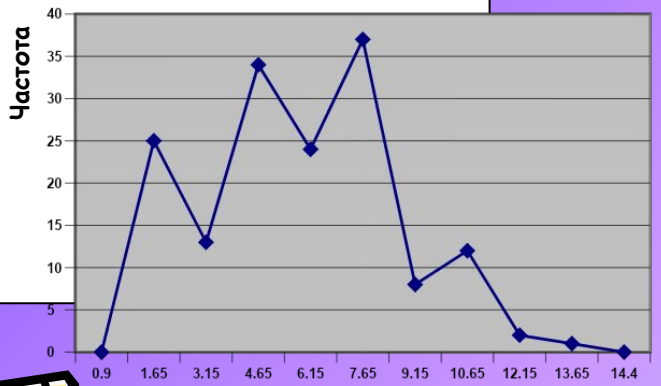




Сравнительная характеристика распределения частот количества букв в слове

Проза





Стихотворение

ВЫВОД:

Сравнив гистограммы и полигоны распределения частот появления букв русского алфавита для отрывка из повести и стихотворения, мы видим, что:

- процентные частоты двух данных распределений практически одинаковы;
- участком наибольшей частоты как для первого, так и для второго распределения является участок алфавита от «к» до «ф» ;
- модой как для первого, так и для второго распределения является буква «о»;
- обе выборки являются унимодальными (одна мода).

Это подтверждает, что у каждого автора есть своя частотная таблица использования букв, слов, специфических оборотов, неповторимая для других писателей, как отпечатки пальцев.

Сравнив гистограммы и полигоны распределения частот количества букв в слове для отрывка из повести и стихотворения, мы видим, что:

- распределение частот количества букв в слове для стихотворения имеет одну моду М(о)=6букв (унимодальная выборка) ;
- распределение частот количества букв в слове для отрывка из повести имеет две моды М(о)=6букв и М (о)=7букв (бимодальная выборка) .