

Задание на самостоятельную работу по теме «Описательная статистика в MS Excel»

РЕШЕНИЕ ТИПОВОГО ПРИМЕРА

Построим выборку и определим статистические характеристики для количественного показателя X – стоимость принтера. Источник информации – *Яндекс.Маркет*.

1. Формирование выборки

Для построения выборки на сервисе *Яндекс.Маркет* откроем каталог *Компьютеры/Принтеры и МФУ*. Ограничим объем генеральной совокупности максимальной ценой принтера в 10 000 рублей, при этом система сообщает, что объявлений, удовлетворяющих этому требованию на данный момент времени – 326 ,т.е. $N = 326$ (рис. 1.)

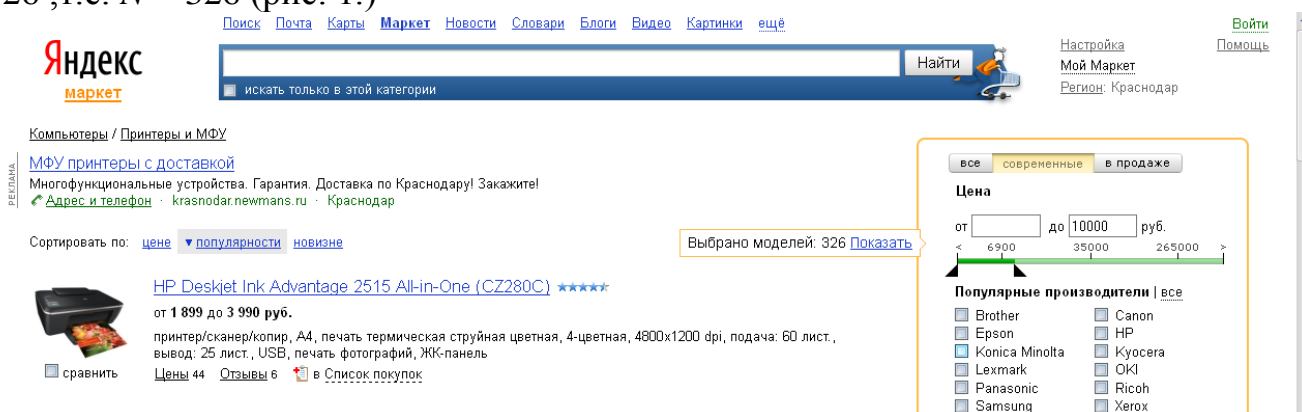


Рис. 1. Установка параметров поиска товара в Яндекс.Маркет

Проведем в MS Excel подготовительную работу. Введем в столбик числа от 1 до 326. Для этого в ячейке A1 введем число 1, а в ячейке A2 формулу: =A1+1. Скопируем протягиванием формулу из ячейки A2 вниз до 326 строки включительно.

С помощью инструмента MS Excel «Выборка» отберем из генеральной совокупности принтеры в выборку. Указанный инструмент входит в надстройку «Анализ данных» (рис. 2).

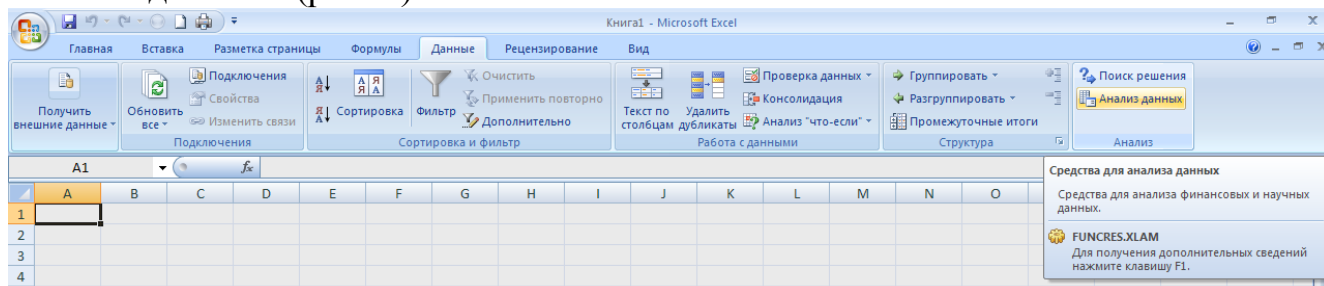


Рис. 2. Надстройка «Анализ данных» на вкладке «Данные»

После выбора указанной надстройки открывается окно диалога, представленное на рис. 3. В нем надо выбрать инструмент «Выборка» и нажать «Ок».

Если пользователь ранее не использовал надстройку «Анализ данных», то ее сначала надо добавить на панель инструментов MS Excel. Для этого необходимо

открыть основное меню MS Excel (рис. 4 – кнопка в левом верхнем углу) и нажать в нем кнопку «Параметры Excel» (рис. 4 – левая кнопка внизу меню).

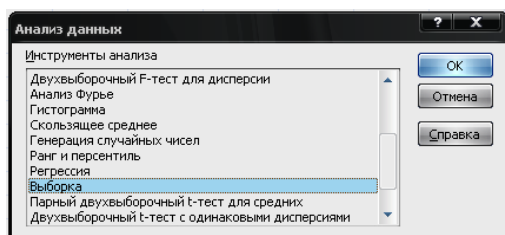


Рис. 3. Окно диалога «Анализ данных»

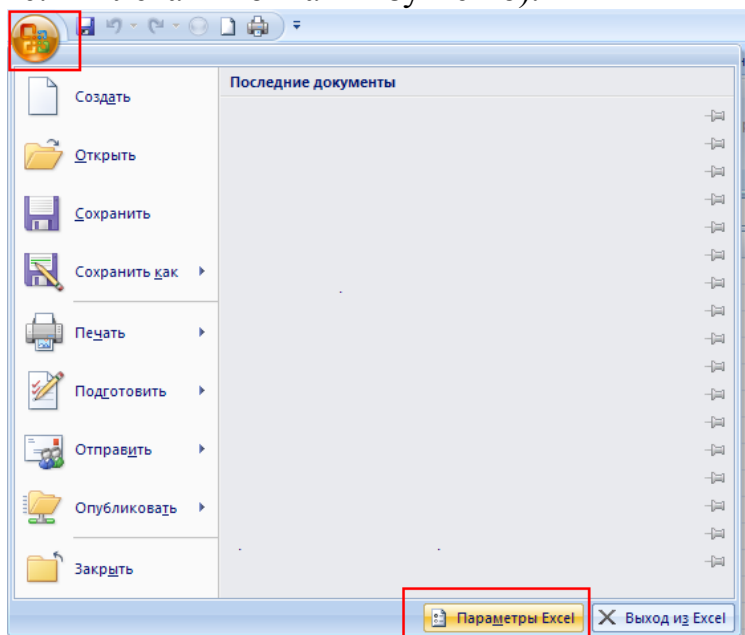


Рис. 4. Основное меню MS Excel

В появившемся окне диалога (рис. 5) слева надо выбрать пункт «Настройка» и нажать кнопку «Перейти» внизу окна, рядом со списком «Настройки Excel». Откроется окно диалога «Настройки» (рис. 6), в котором надо поставить флажок на пункте «Пакет анализа» и нажать «Ок». После этого кнопка «Анализ данных» появится на вкладке «Данные» (рис.2).

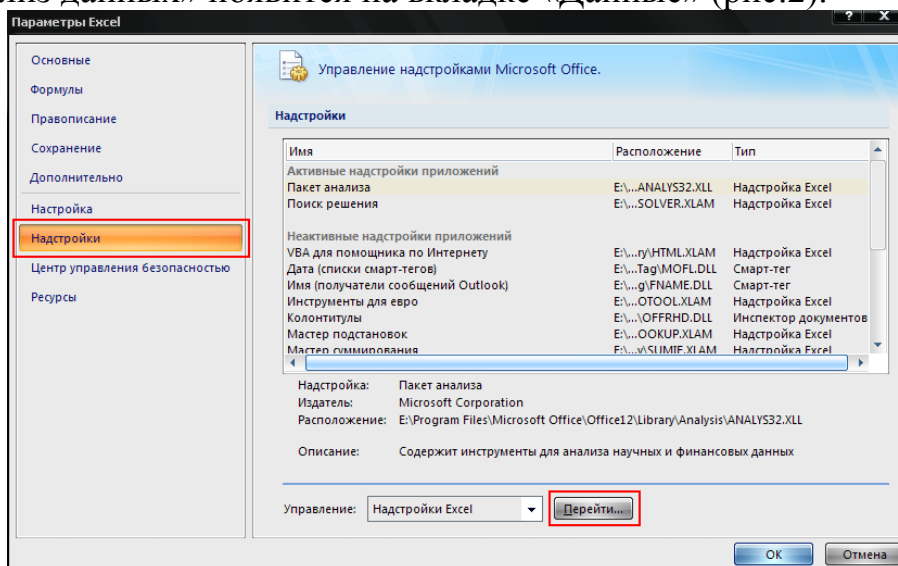


Рис. 5. Окно диалога «Параметры Excel».

Итак, выбираем инструмент «Выборка». Окно мастера работы с этим инструментом представлено на рисунке 7. В поле «Входной интервал» необходимо указать диапазон номеров, из которых формируется выборка. В нашем случае это диапазон A1:A326. В поле «Число выборок» необходимо указать объем выборки – в нашем случае 40.

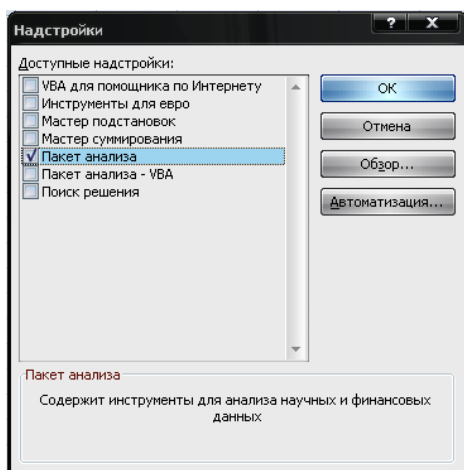


Рис. 6. Окно диалога «Надстройки»

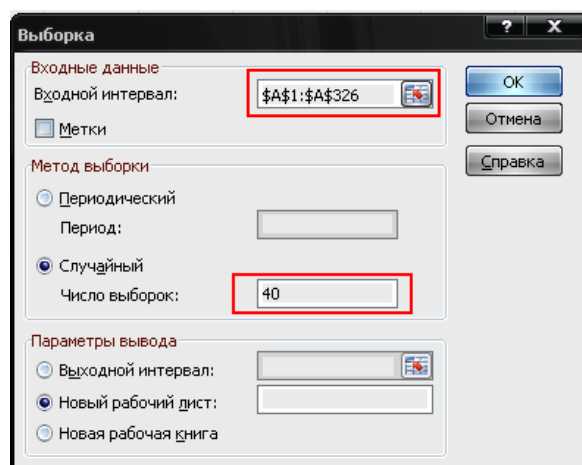


Рис. 7. Мастер работы с инструментом «Выборка»

Инструмент «Выборка» формирует собственно случайную повторную выборку. При этом повторяющиеся номера мы будем игнорировать. Чем больше объем генеральной совокупности, тем меньше в выборке, построенной MS Excel, будет повторяющихся номеров. Желательно чтобы в итоге оказалась выборка объемом не менее 30 элементов.

Остальные параметры в окне диалога «Выборка» задавать или менять не нужно. После нажатия кнопки «Ок» на новом рабочем листе в столбце *A* появятся случайные номера из диапазона от 1 до 326. Упорядочим их по возрастанию (инструмент «Сортировка» на вкладке «Данные»). Повторяющиеся номера будем игнорировать. В нашем случае повторились случайные номера 39, 78 и 277.

Таким образом, итоговая выборка будет содержать 37 элементов, т.е. окончательный объем нашей выборки $n = 37$.

Случайные номера, полученные с помощью инструмента «Выборка» в MS Excel: 9, 16, 39, 40, 41, 54, 64, 78, 94, 95, 108, 122, 138, 143, 154, 160, 165, 174, 186, 189, 191, 192, 197, 203, 217, 222, 230, 238, 248, 249, 259, 277, 278, 279, 292, 306, 320.

Следующий шаг: выписать цены на принтер в объявлениях на Яндекс.Маркет с перечисленными номерами. Как правило, на одной странице выводится по 10 объявлений. Для нахождения объявления с данным номером, например, с номером 138, можно:

- разделить 138 на 10, получим 13,8;
- на 14-й странице (увеличили на 1 целую часть числа 13,8) берем 8-е сверху объявление (дробная часть числа 13,8).

Если у номера последняя цифра 0, например, номер 40 – то на 4-й странице (40 разделили на 10) берем 10 объявление.

Рассмотрим объявление №9 (рис.8), там указано две цены: минимальная и максимальная. Мы будем работать с минимальной ценой, т.е. выписываем 5015 руб. Целесообразно, данные вводить в MS Excel в столбик *B* рядом с выборкой номеров. Эту часть работы надо выполнить максимально внимательно и аккуратно.



Canon i-SENSYS MF3010 ★★★★★

от 5 015 до 7 900 руб.

принтер/сканер/копир, А4, печать лазерная черно-белая, 18 стр/мин ч/б, 1200x600 dpi, подача: 150 лист., вывод: 100 лист., память: 64 Мб, USB

сравнить

[Цены 52](#) [Отзывы 16](#) [в Список покупок](#)

Рис. 8. Объявление №9 в генеральной совокупности о продажах принтера

Результат построения выборки:

	A	B							
1	9	5015	11	108	6600	21	191	7246	
2	16	1290	12	122	9250	22	192	6050	
3	39	9989	13	138	2510	23	197	8239	
4	40	6978	14	143	7806	24	203	5291	31
5	41	6599	15	154	4150	25	217	3234	259
6	54	5829	16	160	6070	26	222	1865	277
7	64	9730	17	165	3595	27	230	5970	278
8	78	7079	18	174	6842	28	238	9466	279
9	94	1195	19	186	7805	29	248	2782	292
10	95	3539	20	189	9500	30	249	4935	306
									320
									3782
									7600
									4420
									5200
									4660
									7990
									3348

Рис. 9. Результаты получения выборки: первый столбик – случайный номер, второй столбик – минимальная цена из объявления с этим номером.

Дальнейший анализ будем проводить для цен (скопируем их на новый лист и упорядочим по возрастанию): 1195, 1290, 1865, 2510, 2782, 3234, 3348, 3539, 3595, 3782, 4150, 4420, 4660, 4935, 5015, 5200, 5291, 5829, 5970, 6050, 6070, 6599, 6600, 6842, 6978, 7079, 7246, 7600, 7805, 7806, 7990, 8239, 9250, 9466, 9500, 9730, 9989

2. Построение интервального вариационного ряда

Для построения интервального вариационного ряда воспользуемся инструментом «Гистограмма» надстройки «Пакет анализа», который вызывается аналогично инструменту «Выборка».

На последнем шаге на новом листе книги MS Excel в столбце A в строках, начиная с 1-й строки до 37-й строки включительно, мы разместили выборку цен, упорядоченную по возрастанию.

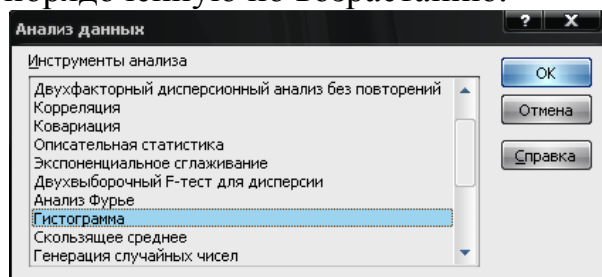


Рис. 10. Выбор инструмента «Гистограмма»

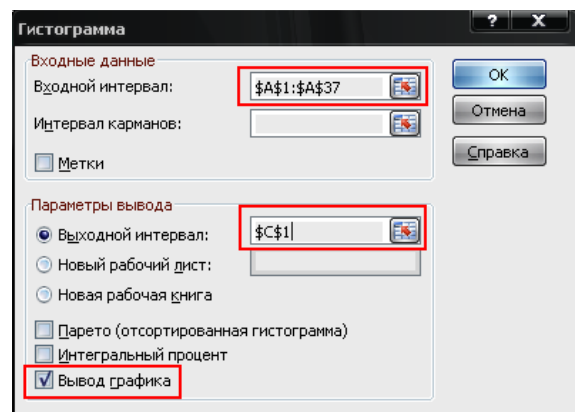


Рис. 11. Мастер работы с инструментом «Гистограмма»

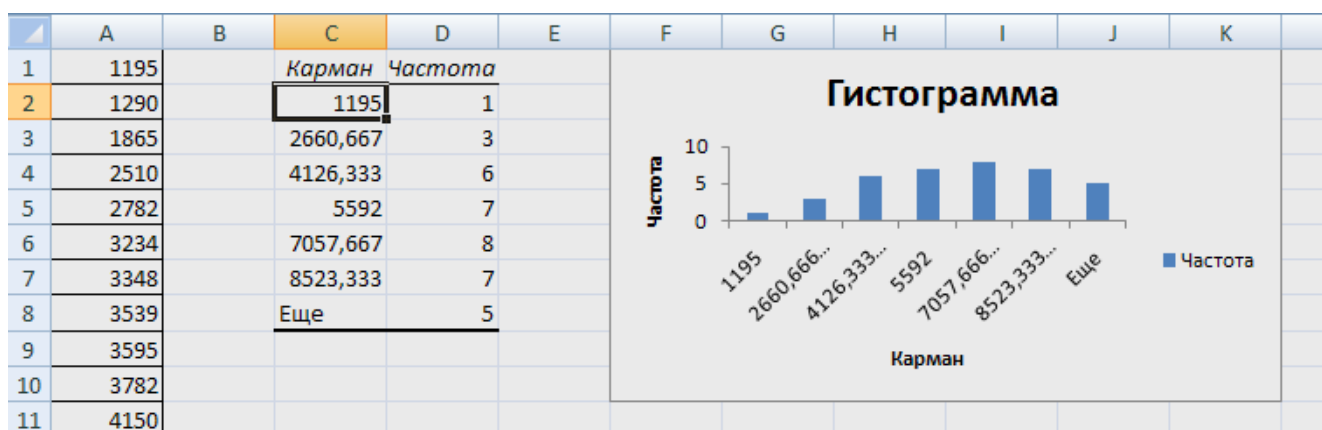


Рис. 12. Результат применения инструмента «Гистограмма»

В окне мастера работы с инструментом «Гистограмма» необходимо указать (рис. 11): *Входной интервал* – диапазон ячеек, в котором находится выборочная совокупность (в нашем случае, A1:A37). Остальные параметры работы мастера можно оставить по умолчанию. В нашем случае мы поставили флажок «Вывод графика» - программа построила гистограмму (рис. 12), а также результаты вывели не новый рабочий лист (по умолчанию), а на этом же листе, начиная с ячейки C1 (рис. 11).

Проведем интерпретацию данных в диапазоне C1:D8 (рис. 12), здесь выведен интервальный вариационный ряд. Число интервалов программа определяет по формуле Стерджеса. Карман – это промежуток вида $a < x \leq b$. На основе результатов на рис. 12, получаем интервальный вариационный ряд:

Интервал	$(-\infty; 1195]$	$(1195; 2660,7]$	$(2660,7; 4126,3]$	$(4126,3; 5592]$	$(5592; 7057,7]$	$(7057,7; 8523,3]$	$(8523,3; +\infty]$
Частота	1	3	6	7	8	7	5

Отформатировав гистограмму, получим:



Рис. 13. Гистограмма цен на принтеры

3. Определим числовые характеристики вариационного ряда

Для вычисления числовых характеристик по выборке воспользуемся инструментом «Описательная статистика» надстройки «Пакет анализа», который вызывается аналогично инструментам «Выборка» и «Гистограмма».

В качестве исходных данных снова используем разместили выборку цен, упорядоченную по возрастанию и размещенную на отдельном листе книги MS Excel в столбце A в строках, начиная с 1-й строки до 37-й строки включительно.

Вызываем надстройку «Анализ данных» и выбираем инструмент «Описательная статистика» (рис. 14). В открывшемся окне мастера работы с инструментом «Описательная статистика» (рис. 15) устанавливаем входной интервал – A1:A37 и ставим флажки «Итоговая статистика» и «Уровень надежности» (значение уровня надежности оставляем по умолчанию равным 95%). Остальные параметры можно не менять.

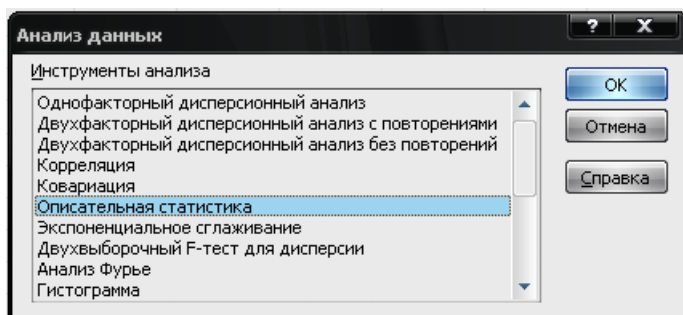


Рис. 14. Выбор инструмента «Описательная статистика»

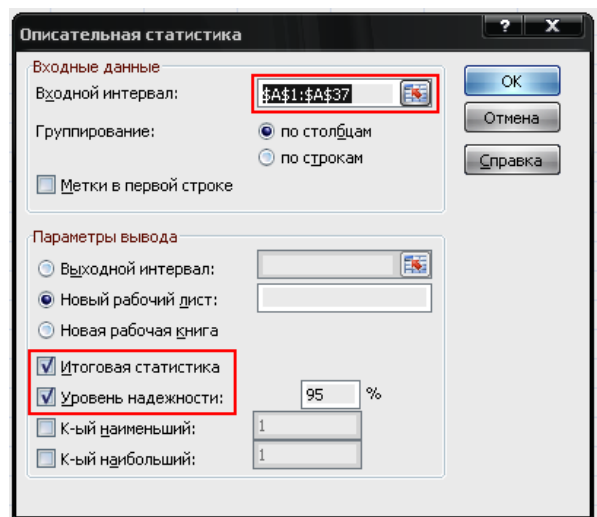


Рис. 15. Мастер работы с инструментом «Описательная статистика»

	A	B
1	<i>Столбец1</i>	
2		
3	Среднее	5768,891892
4	Стандартная ошибка	400,0538961
5	Медиана	5970
6	Мода	#Н/Д
7	Стандартное отклонение	2433,432849
8	Дисперсия выборки	5921595,432
9	Экссесс	-0,809510843
10	Асимметричность	-0,070560995
11	Интервал	8794
12	Минимум	1195
13	Максимум	9989
14	Сумма	213449
15	Счет	37
16	Уровень надежности(95,0%)	811,3469011

Рис. 16. Результаты работы инструмента «Описательная статистика»

На основании проведенного выборочного исследования и рассчитанных по выборке показателей описательной статистике можно сделать следующие выводы:

- Средняя цена принтера: $\bar{x} = 5768,9$ руб.
- Медиана цены: $Me = 5970$ руб.
- Дисперсия: $D = 5921595,4$.

Среднее квадратическое отклонение (стандартное отклонение): $\sigma = 2433,4$.

Экцесс: $E = -0,8$.

Асимметрия: $A = -0,07$.

Предельная ошибка в выборке (уровень надежности(95,0%)): $\delta = 811,3$.

Небольшое относительное отклонение выборочного среднего (\bar{x}) и медианы (Me):

$$\frac{\bar{x} - Me}{\bar{x}} = \frac{-201,1}{5768,9} \cdot 100\% = -3,5\%$$

говорит о надежности среднего в выборке.

Коэффициент вариации:

$$V = \frac{\sigma}{\bar{x}} \cdot 100\% = \frac{2433,4}{5768,9} \cdot 100\% = 42,2\%.$$

То, что коэффициент вариации практически равен 40%, свидетельствует о заметной колеблемости цены принтеров в выборке¹.

Доверительный интервал для средней цены на принтеры с надежностью 95%:

$$\begin{aligned} \bar{x}_B - \delta \leq \bar{x} \leq \bar{x}_B + \delta, \\ 5768,9 - 811,3 \leq \bar{x} \leq 5768,9 + 811,3, \\ 4957,6 \leq \bar{x} \leq 6580,2 \end{aligned}$$

Т.е. среднюю цену на принтеры с надежностью 95% покрывает интервал (4957,6; 6580,2).

ОБРАЗЕЦ ОТЧЕТА ПО САМОСТОЯТЕЛЬНОЙ РАБОТЕ:

Отчет по самостоятельной работе «Описательная статистика в MS Excel»

Выполнил: *Иванов Иван Иванович* (ВТ-ЭП-2)

1. Построение выборки

Построим выборку и определим статистические характеристики для количественного показателя X – **стоимость принтера**. Источник информации – *Яндекс.Маркет*.

Объем генеральной совокупности ограничили максимальной ценой принтера в 10 000 рублей, $N = 326^2$.

С помощью инструмента «Выборка» получена повторная собственно-случайная выборка номеров объявлений объемом 40³. После удаления повторяющихся номеров, остались следующие номера ($n = 37$): 9, 16, 39, 40, 41, 54, 64, 78, 94, 95, 108, 122, 138, 143, 154, 160, 165, 174, 186, 189, 191, 192, 197, 203, 217, 222, 230, 238, 248, 249, 259, 277, 278, 279, 292, 306, 320.

По перечисленным номерам получены значения интересующего нас показателя (цены на принтер, указаны в порядке возрастания): 1195, 1290, 1865, 2510, 2782, 3234, 3348, 3539, 3595, 3782, 4150, 4420, 4660, 4935, 5015, 5200, 5291,

¹ Будем считать, что если коэффициент вариации существенно меньше 40%, то колеблемость признака в выборке небольшая, а если он существенно больше 40%, то колеблемость признака в выборке большая

² Желательно получить объем генеральной совокупности не менее 300.

³ В данной работе всем необходимо указывать параметр «Число выборок» равный 40.

5829, 5970, 6050, 6070, 6599, 6600, 6842, 6978, 7079, 7246, 7600, 7805, 7806, 7990, 8239, 9250, 9466, 9500, 9730, 9989.

2. Построение интервального вариационного ряда

С помощью инструмента «Гистограмма» получен интервальный вариационный ряд цен на принтеры:

Интервал	$(-\infty; 1195]$	$(1195; 2660,7]$	$(2660,7; 4126,3]$	$(4126,3; 5592]$	$(5592; 7057,7]$	$(7057,7; 8523,3]$	$(8523,3; +\infty]$
Частота	1	3	6	7	8	7	5



Гистограмма цен на принтеры

3. Числовые характеристики вариационного ряда

Для вычисления числовых характеристик по выборке воспользуемся инструментом «Описательная статистика».

На основании проведенного выборочного исследования и рассчитанных по выборке показателей описательной статистике получены следующие результаты:

Средняя цена принтера: $\bar{x} = 5768,9$ руб.

Медиана цены: $Me = 5970$ руб.

Дисперсия: $D = 5921595,4$.

Среднее квадратическое отклонение (стандартное отклонение): $\sigma = 2433,4$.

Экцесс: $E = -0,8$.

Асимметрия: $A = -0,07$.

Предельная ошибка в выборке (уровень надежности(95,0%)): $\delta = 811,3$.

Небольшое относительное отклонение выборочного среднего (\bar{x}) и медианы (Me):

$$\frac{\bar{x} - Me}{\bar{x}} = \frac{-201,1}{5768,9} \cdot 100\% = -3,5\%$$

говорит о надежности среднего в выборке.

Коэффициент вариации:

$$V = \frac{\sigma}{\bar{x}} \cdot 100\% = \frac{2433,4}{5768,9} \cdot 100\% = 42,2\%.$$

То, что коэффициент вариации практически равен 40%, свидетельствует о заметной колеблемости цены принтеров в выборке⁴.

Доверительный интервал для средней цены на принтеры с надежностью 95%:

$$\begin{aligned}\bar{x}_B - \delta &\leq \bar{x} \leq \bar{x}_B + \delta, \\ 5768,9 - 811,3 &\leq \bar{x} \leq 5768,9 + 811,3, \\ 4957,6 &\leq \bar{x} \leq 6580,2\end{aligned}$$

Т.е. среднюю цену на принтеры с надежностью 95% покрывает интервал (4957,6; 6580,2).

ЗАДАНИЕ НА САМОСТОЯТЕЛЬНУЮ РАБОТУ

По объявлениям на сервисе Яндекс.Маркет провести выборочное исследование количественного показателя – стоимость товара (товар выбирает сам студент):

1. Получить выборку объемом не менее 30 наблюдений.
2. Построить интервальный вариационный ряд и гистограмму
3. Найти числовые характеристики вариационного ряда

Расчеты оформить в MS Excel в соответствии с приведенным образцом, оформить отчет в MS Word. Файл с расчетом в MS Excel и отчет в электронном виде выслать на электронную почту преподавателя. Также необходимо сдать отчет в распечатанном виде.

⁴ Будем считать, что если коэффициент вариации существенно меньше 40%, то колеблемость признака в выборке небольшая, а если он существенно больше 40%, то колеблемость признака в выборке большая