Федеральное государственное бюджетное образовательное учреждение

высшего образования

«Саратовский государственный технический университет имени Гагарина Ю.А.»

Кафедра «Информационные системы и технологии»

# Отчет по лабораторной работе №5

по дисциплине

### «Основы интеллектуального анализа данных»

на тему

«Подготовка набора данных и анализ корреляционной зависимости переменных.»

Выполнил: студент группы б1-ИФСТ-21

Яновский Евгений Валерьевич

Преподаватель:

Печенкин Виталий Владимирович

Саратов-2022

**Задание.**

**I. ПЕРВЫЙ ЭТАП.**

**Формирование набора случайных данных в программе MS Excel.**

Для выполнения задания запустите программу MS Excel, установленную на компьютере.

В процессе работы сохраняйте файлы в папку, которая обсуждалась в разделе «Совет» выше под именем Report\_5\_FIO с соответствующим расширением. При написании фамилии используйте транслитерацию латинскими буквами. Например, если ваша фамилия «Иванов», имя файла будет Report\_5 \_Ivanov.

Как и на прошлых занятиях, мы используем ВАШ ИНДИВИДУАЛЬНЫЙ порядковый номер в журнале, обозначим его ***N***.

Создайте пустую электронную таблицу и сохраните файл на свой сетевой диск в папку с отчётами. В первой строке таблицы поместите следующие заголовки столбцов.

|  |  |  |  |
| --- | --- | --- | --- |
| Х=X1\*A | X2\*4-2 | Y=X\*A+N | Z=Y+X2 |

Далее используем две константы *A* и *N*.

*A* равно 4, если Ваш порядковый номер *N* в журнале является чётным числом, в противном случае *A = 7*.

*N* – ваш порядковый номер в журнале

**ВАЖНО.** Для констант A и N выделите специальные ячейки правее от трех столбцов, заполняемых данными, используйте в формулах не значения, а имена соответствующих ячеек.

***В первом столбце* (Х=X1\*A)** будут находиться случайно сгенерированные числа переменной X в диапазоне от 0 до A (используем формулу «=СЛЧИС()\*A»).

Мы используем для получения случайных чисел функцию MS Excel СЛЧИС(), которая генерирует случайное число из диапазона [0, 1]. Вставьте предложенную формулу в первую свободную ячейку первого столбца X1\*A и, выделив её, растяните на последующие 100 ячеек таблицы в столбце.

**ВАЖНО.** Для предотвращения пересчёта случайных чисел после генерации столбца X скопируйте весь диапазон сгенерированных случайных чисел и вставьте их обратно специальной вставкой, используя опцию «ЗНАЧЕНИЯ».

Выделите диапазон сгенерированных случайных чисел, используйте контекстное меню «Копировать», используйте контекстное меню «Специальная вставка…» – выбрать «Значения».

Поступаем так во всех случаях, когда в определении значения ячейки фигурирует случайное число

***Во втором столбце* (X2)** с помощью формулы и распространения значения на соответствующий диапазон из 100 последующих ячеек введите значение, которое определяется по формуле

*X2= СЛЧИС()\*4-2*,

Полученное значение будет изменяться случайным образом в диапазоне [-2, 2].

Это значение необходимо нам **для внесения случайных изменений** в одну из переменных, вычисленных так, как это показано далее.

***В третьем столбце* (Y)** с помощью формулы «=X\*A+N», где X – адрес первой ячейки со значением в **первом** столбце. Распространяем значение формулы на соответствующий диапазон введите значение, которое определяется по формуле

*ЗНАЧЕНИЕ = X \* A + N*  (формула в первой ячейке должна иметь вид «*=A2+N*»)

где *N –*Ваш порядковый номер в журнале.

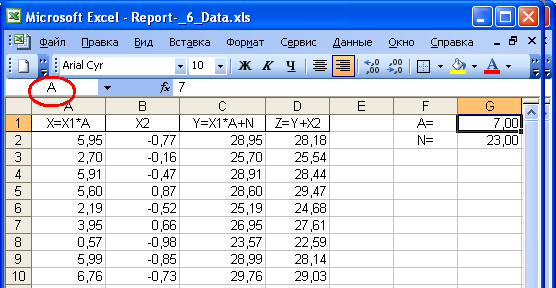
Вставьте соответствующую формулу в первую свободную ячейку третьего столбца и, выделив её, растяните на последующие 100 ячеек таблицы в столбце.

***В четвёртом столбце* (Z)** вычислим значение переменной *Y* с добавлением случайных колебаний X2. Используйте следующую формулу для вычисления значения переменной **Z**

Z = X2+Y (формула Excel для первой ячейки будет иметь такой вид «=C2+B2»)

Измените формат представления чисел, так, чтобы у всех чисел было ДВА десятичных разряда.

Полученный набор данных должен выглядеть примерно так, как это показано на следующем рисунке

****

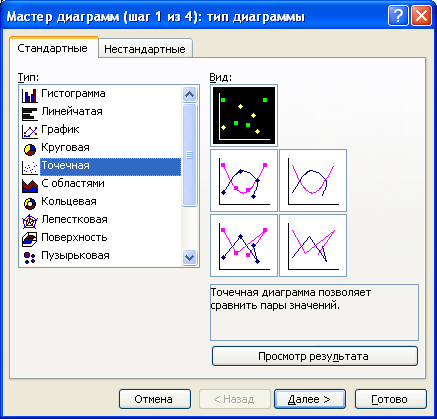
Обратите внимание на имя, заданное для ячейки, содержащей значение A (в левом верхнем углу – обведено красным овалом). Аналогично нужно поступить и с ячейкой, в которой записан ваш индивидуальный номер.

**ВАЖНО.** Сделайте скриншот результатов работы и поместите его в файл отчёта. Обратите внимание на то, что на скриншоте **должны быть видны значения констант *A* и *N***, которые определяются в соответствии с вашим номером в журнале группы по правилам, сформулированным выше.

**II. ВТОРОЙ ЭТАП. Построение диаграммы рассеяния в программе MS Excel**

Для того, чтобы убедиться, что между переменными имеется линейная связь (условия использования коэффициента корреляции Пирсона), построим средствами MS Excel диаграмму рассеяния, которая показывает значения для двух переменных как точку на плоскости в евклидовой системе координат.

Вставьте в лист MS Excel диаграмму типа «Точечная», которая доступна в диалоге, представленном ниже



В качестве переменных для осей X и Y, соответственно, используйте значения столбцов A (вычисленное значение X1\*A) и D (значение Z) электронной таблицы.

Установите минимальные и максимальные значения для отображаемых диапазонов шкал так, чтобы отсечь области, не содержащие значений.

В заголовке диаграммы укажите, как это показано на следующем рисунке, свой индивидуальный номер по журналу *N*.

Диаграмма будет иметь примерно такой вид (*зависит от значений параметров A и N*)



**ВАЖНО.** Скопируйте диаграмму в файл отчёта. На изображении должен быть виден ваш **индивидуальный номер** **N** в заголовке.

**Покажите результаты работы преподавателю.** Он отметит факт выполнения первой части задания

**III. ТРЕТИЙ ЭТАП.**

**Экспорт полученных данных в программу PSPP.**

Сохраните данные в файл в своей сетевой папке и выполните процедуру экспорта полученных данных из программы MS Excel в программу PSPP. Процедура экспорта подробно описана в задании 4, которое мы выполняли раньше.

В программе PSPP задайте имена переменных для четырё столбцов как X, X2 Y, Z. При определении атрибутов переменных в PSPP сделайте так, чтобы при отображении значения выводилось 2 знака после запятой.

**Покажите определения новых переменных и их значения в программе PSPP преподавателю.**

**IV. ЧЕТВЁРТЫЙ ЭТАП.**

**Определение коэффициента корреляции между переменными *X, Y, Z***

Обратите внимание на тот факт, что мы **ЗАРАНЕЕ знаем**, что переменные X и Y связаны между собой как линейным уравнением. Поэтому они должны иметь максимальный коэффициент корреляции, показывающий силу связи между ними.

В значение переменной Z мы ***внесли изменения***, которые являются случайными по своему характеру и находятся в определённом нами диапазоне. Соответственно, значение коэффициента корреляции между переменными X и Z должно уменьшиться. Покажем это с помощью процедуры корреляционного анализа, доступного в программе PSPP.

Начните процедуру с помощью команды меню, показанной на следующем рисунке

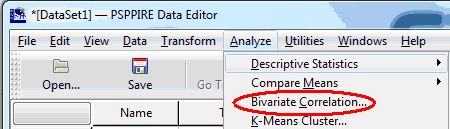


Рис. Вызов команды корреляционного анализа

Добавьте в поле переменных имена X, Y, Z переменных из набора данных импортированного из MS Excel.

Выполните построение корреляционной матрицы.

Сделайте скриншот корреляционной таблицы и поместите его в файл отчёта.

**Обратите внимание на**

* значения коэффициентов корреляции Пирсона,
* симметричность корреляционной матрицы относительно главной диагонали,
* значение уровня значимости для каждой ячейки в матрице.

**Скриншоты экрана.**

