МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ

«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

Кафедра «Програмна інженерія та інформаційні технології управління»

Звіт з лабораторної роботи №1

з дисципліни «Математична статистика»

Виконав:

Студент групи КН-36а

Антонеко В. В.

Перевірив:

Голоскоков О. Є.

Харків – 2018

**Лабораторная работа 1.**

ОЗНАКОМЛЕНИЕ С ПРОГРАММНЫМ ПАКЕТОМ ДЛЯ СТАТИСТИЧЕСКОГО АНАЛИЗА

**Общие сведения компании**

Пакет STATISTICA разработан фирмой StatSoft (США). Первоначально он входил в качестве модуля в состав самых популярных в то время электронных таблиц Lotus 1-2-3. Как самостоятельный продукт пакет впервые заявил о себе в 1991 г. Последняя версия продукта совместима с Windows Vista, в ней поддерживаются графический интерфейс пользователя и динамический обмен данными. Благодаря этому пакет может работать в сочетании с другими Windows-приложениями. В последние версии включен также язык программирования Statistica-BASIC, позволяющий расширять возможности пакета в соответствии с потребностями пользователя.

**Общие сведения о пакете STATISTICA**

STATISTICA - Статистическая графическая программа для системы WINDOWS, включающая широкий диапазон статистических функций с большими графическими возможностями. Доступ к графическим процедурам осуществляется в процессе статистической обработки данных. STATISTICA включает различные процедуры для ввода и модификации данных, сохранения текстовых и графических файлов.

Процедуры STATISTICA построены таким образом, что Вы можете начать с использования тех возможностей системы, которые Вам наиболее понятны и далее расширить их применение в соответствии с потребностями.

**Возможности пакета STATISTICA**

Возможности. STATISTICA позволяет проводить исчерпывающий, всесторонний анализ данных, представлять результаты анализа в виде таблиц и графиков, автоматически создавать отчеты о проделанной работе. С помощью удобной системы подсказок можно обучаться не только работе с самим пакетом, но и современным методам статистического анализа.

Файлы данных. Данные в системе STATISTICA организованы в виде электронных таблиц, как в привычной для пользователей программе Excel. Файл содержит наблюдения и переменные. Наблюдения можно рассматривать как эквивалент записей в базах данных (или строк электронной таблицы), а переменные - как эквивалент полей (столбцов электронной таблицы). Каждое наблюдение состоит из набора значений переменной.

В пакете STATISTICA все операции, включая копирование, перетаскивание и автоматическое заполнение ячеек, производятся так же, как в популярных электронных таблицах. При нажатии правой кнопки мыши появляется всплывающее меню, где точно так же предлагается перечень операций, которые можно выполнить над выделенным объектом.

Общее число переменных в стандартном файле STATISTICA может быть до 4092, количество наблюдений ограничено лишь объемом жесткого диска. В системе имеется также менеджер мегафайлов (доступный из модуля Управление данными), который позволяет

работать с очень большими файлами, содержащими до 32000 переменных.

Система STATISTICA предоставляет всесторонние возможности по импорту и экспорту данных, в том числе и из таблиц Excel.

Архитектура и интерфейс системы. Система STATISTICA состоит из отдельных модулей, каждый из которых является полноценным Windows-приложением. Система постоянно обновляется, в нее вводятся новые модули и вычислительные процедуры.

Интерфейс системы может быть настроен на конкретный пользовательский проект: можно задать отображение стольких диалоговых окон, таблиц результатов, графиков, сколько в данном случае необходимо.

Методы анализа. Система включает следующие модули: Основные статистики и таблицы. Исчерпывающий набор описательных статистик, таблицы сопряженности, таблицы флагов и заголовков, кросстабуляция многомерных откликов и многомерных дихотомий, вычисление корреляционных матриц, обработка пропущенных

данных, t-критерии для зависимых и независимых выборок, критерии однородности дисперсии, однофакторный дисперсионный анализ.

Непараметрическая статистика. Непараметрические критерии, ранговые корреляции, подгонка распределений.

Множественная регрессия. Пошаговая регрессия с включением и исключением предикторов, нелинейная регрессия, риджрегрессия, построение прогнозов, всесторонний анализ остатков, вычисление прогнозов и доверительных интервалов для прогнозируемых значений (можно анализировать очень большие модели, до 500 переменных).

Нелинейное оценивание. Подгонка любой задаваемой пользователем функции, задаваемая пользователем функция потерь, разрывная регрессия.

Временные ряды и прогнозирование. Широкий выбор моделей анализа временных рядов, включая модели АРПСС - авторегрессии и проинтегрированного скользящего среднего, модели с интервенцией, анализ распределенных лагов, спектральный анализ чрезвычайно длинных временных рядов, преобразования рядов, включая быстрое преобразование Фурье и многие другие процедуры углубленного анализа.

Кластерный анализ. Широкий набор процедур кластерного анализа, включая иерархическое объединение, двухвходовое объединение, метод к-средних; алгоритмы оптимизированы для анализа очень больших проектов, например, методом к-средних можно анализировать 400000 наблюдений с 10 переменными.

Факторный анализ. Процедуры факторного анализа и анализа главных компонент, ортогональные и косоугольные факторы, иерархический анализ косоугольных факторов и др.

Канонический анализ. Вычисление канонических переменных и канонических корней.

Многомерное шкалирование. Анализ расстояний, матриц сходств и различия, диаграмма Шепарда и др.

Деревья классификации. Современные методы построения деревьев классификации с категориальными и порядковыми предикторами и различными функциями потерь.

Дискриминантный анализ. Процедуры всестороннего дискриминантного анализа, разнообразные статистики и графическое представление результатов.

Логлинейный анализ. Всесторонний анализ многовходовых таблиц сопряженности, автоматическое построение лучшей модели.

Анализ выживаемости. Анализ таблиц жизни, оценки Каплана-Мейера, регрессионные модели: Кокса, логнормальная, экспоненциальная, зависящие от времени ковариаты, разнообразные статистики и критерии.

Дисперсионный анализ. Полный набор методов одномерного и многомерного дисперсионного анализа, фиксированные и переменные ковариаты, апостериорные критерии, контрасты, проверка предположений дисперсионного анализа, планы с повторными измерениями, иерархически вложенные планы, планы с пропущенными ячейками и многое другое.

Компоненты дисперсии. Смешанные модели дисперсионного анализа, оценка компонент дисперсии.

Графические возможности. STATISTICA обладает огромными возможностями для построения графиков непосредственно из таблиц исходных данных и таблиц результатов, причем графика и анализ

данных тесно интегрированы. Например, если после вычисления корреляционной матрицы у пользователя возникает потребность в графическом представлении корреляционной зависимости, то достаточно поместить курсор на соответствующий коэффициент

корреляции, нажать правую кнопку мыши и в появившемся меню выбрать пункт Быстрые статистические графики, а затем одну из диаграмм рассеяния. На экране появится требуемый график. В разных модулях системы имеются свои специальные графики, учитывающие особенности получаемых в них результатов.

**Преимущества и недостатки пакета STATISTICA**

*Преимущества пакета STATISTICA:*

– широкий выбор разнообразных инструментов для статистического анализа данных;  
– позволяет обработать массивные объемы данных;  
– возможность проведения кластерного, дискриминантного, канонического, дисперсионного и факторного анализа;  
– наличие функции группировки данных;  
– возможность построения рядов, линейных и нелинейных моделей;  
– большой набор инструментов для исследования корреляции между различными переменными;  
– поддержка нейронных сетей;  
– возможность импортирования данных из Excel-документов;  
– наличие блока инструментов data mining;  
– возможность построения диаграмм и 2D/3D графиков (более 10000 типов);  
– возможность категоризации по переменным (иерархическая группировка);  
– качественная техническая поддержка;  
– интеграция с Visual Basic (можно использовать DLL-библиотеки);  
– включает обучающие материалы для новичков;

*Недостатки пакета STATISTICA:*

– применение пакета требует высокой теоретической подготовки в

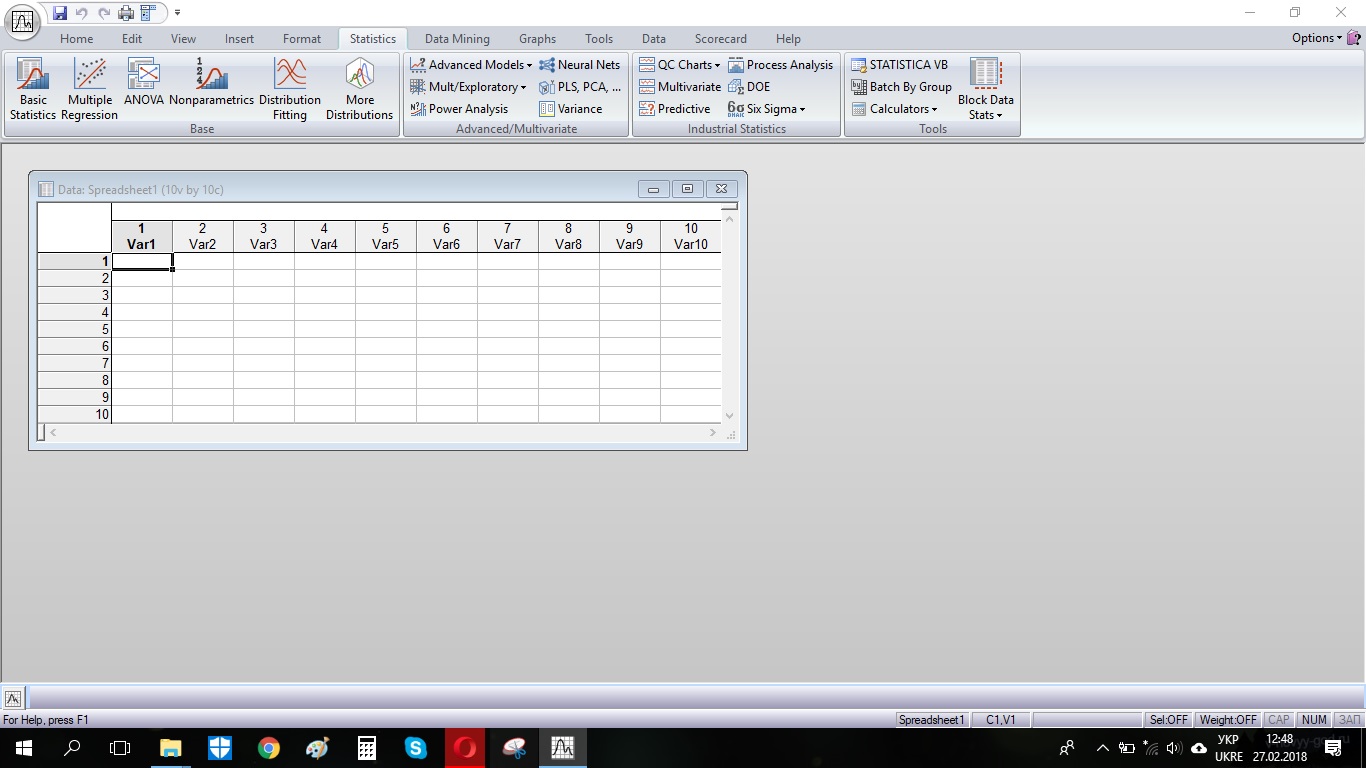
ТВиМС. статистический обработка данные пакет;

– отсутствие реализации некоторых важных тестов временных рядов (в частности - тестов на стационарность);

**Построение графиков в системе STATISTICA**

Некоторая компания SST по разработке мобильных ввела учет количества скачанных раз их продукции и суммарную прибыль за 2017 год по месяцам. Мы оперируем следующими данными: месяц, кол-во скачанных раз и прибыль.

Запустим программу STATISTICA (из меню «Пуск», или нажав соответствующий ярлычок на Рабочем столе). Появится основной интерфейс программы пакета STATISTICA(рисунок 1.0)

Рисунок 1.0 – Основной интерфейс

Для создания нового проекта этого воспользоваться одним из трех способов:

* в пункте основного меню File (Файл) выбрать New (Новый);
* нажать кнопку  (белая страничка – «создать файл по умолчанию») на панели инструментов;
* применить сочетание клавиш «Ctrl + N».

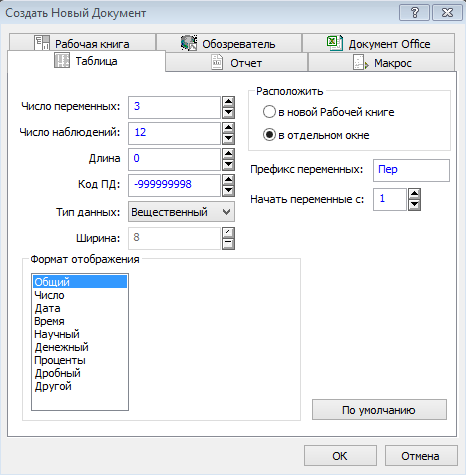


Рисунок 1.1 – Меню настроек нового проекта

Создадим новую таблицу с данными. В которой создадим 3 переменных, и 12 строк для данных. Вид созданной таблице приведены на рисунке 1.2

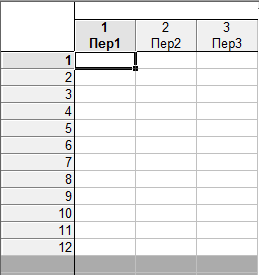


Рисунок 1.2 – Созданная таблица

Изменяем названия переменных. Для этого дважды кликаем по Пер 1 и вводим желаемое название в поле Имя (рисунок 1.3)

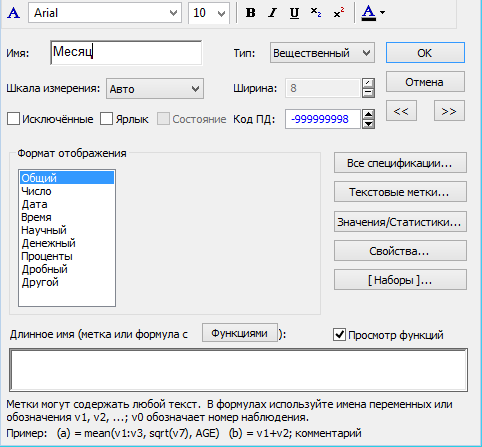


Рисунок 1.3 – Меню настроек переменной

Вид таблицы с изменениями приведен на рисунке 1.4.



Рисунок 1.4 – Итоговый вид таблицы

Так как сильной стороной приложения Statistica есть графика, построим гистограммы распределения по году, количеству, и прибыли. Также приведем пример круговой диаграммы.

Для построения графиков переходим во вкладку «Графика». Затем кликаем по кнопке «2М Графики» и выбираем график. Параметры «Тип графика» и «Тип подгонки» оставить без изменений. Для определения переменных, по которым будет построен график нажать кнопку «Переменные»(рисунок 2.1). В появившемся окне (рисунок 2.2) выбрать переменную «Месяц» и «Прибыль» нажать кнопку «OK».

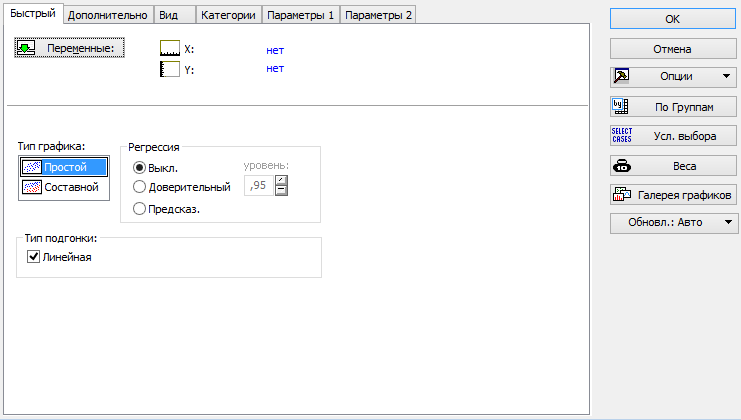


Рисунок 2.1 - Окно задания параметров построения графика

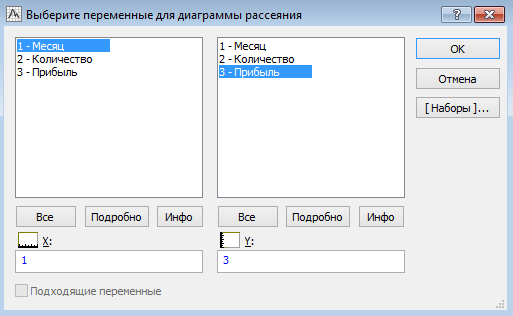


Рисунок 2.2 - Выбор переменной для построения графика

По окончании задания параметров построения графика нажать кнопку «OK». В результате система отобразит графическое распределение исходных данных (рисунок 2.3).

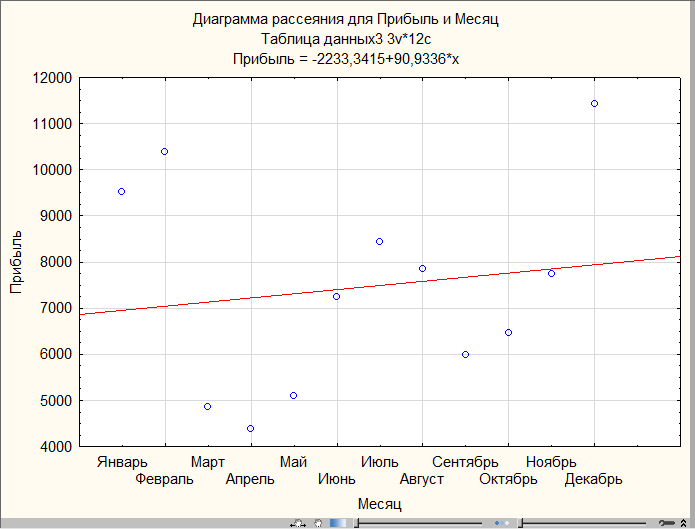


Рисунок 2.3 – Диаграмма рассеяния

На рисунке 2.3 мы видим график распределения прибыли за каждый месяц 2017 года. В декабре прибыль составила 11500. Это была максимальная прибыль компании SST за месяц.

Сделаем еще несколько других графиков.

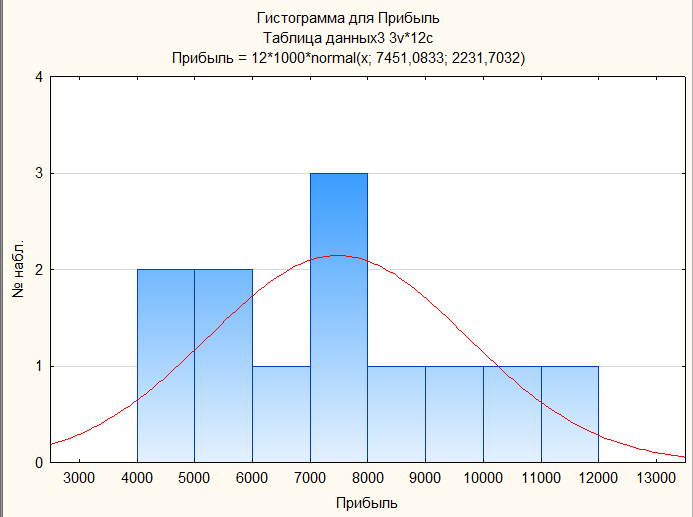


Рисунок 2.4 – Гистограмма

На рисунке 2.4 мы видим гистограмму количества раз наблюдаемой заработанной прибыли за месяц. Наглядно видно, что между 7000 и 8000 прибыли наблюдалось самое большое количество раз – 3. Самое малое количество раз наблюдаемой прибыли было между 6000 и 7000 прибыли, а также в промежутке с 8000 до 12000.

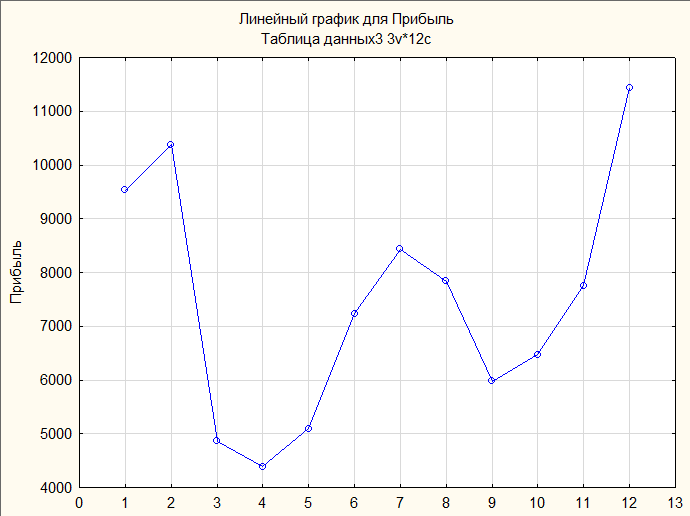


Рисунок 2.5 – Линейный график

На рисунке 2.5 мы видим линейный график для прибыль. Проанализировав данный график сразу можно заметить, что самая большая прибыль была в январе, феврале и декабре. Отсюда сделаем вывод, что пользователи приложения качают больше качают приложение зимой.

Создадим новую таблицу самых скачиваемых приложений за месяц разработанной компанией SST.





Рисунок 2.6 – Круговая диаграмма

На рисунке 2.6 мы видим, что самая скачиваемое приложение «Jump Gump».

**Висновок:** В результаті виконаної лабораторної роботи ми ознайомилися з программним пакетом STATISTICA, його перевагами і недоліками. Також навчилися створювати таблиці, заповнювати їх і будувати різноманітні графіки по значенням з цих таблиць.