Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО

Факультет безопасности информационных технологий

Дисциплина:

«Статистические методы в инженерных исследованиях»

ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №3

«Обработка результатов пассивного эксперимента.
Дисперсионный и корреляционный анализ»

Выполнили:

Бардышев А. А., студент группы N3346
Волощук А. Н., студент группы N3346
Замахов Е. В., студент группы N3346
Пинус И. В., студент группы N3346
Шегай С. Д., студент группы N3346

Проверил:

Федоров А. В.,
Преподаватель, д.т.н., профессор ФБИТ

(отметка о выполнении)

(подпись)

Санкт-Петербург
2025 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Цель работы – Получить умения и отработать навыки методов дисперсионного и корреляционного анализа при исследовании данных о распределении уязвимостей в программном обеспечении на основе данных ФСТЭК РФ (https://bdu.fstec.ru/).

ЗАДАЧИ ЛАБОРАТОРНОЙ РАБОТЫ

1. Провести пассивные эксперименты с данными о уязвимостях
2. Обработать результаты пассивных экспериментов методами дисперсионного и корреляционного анализа
3. Оформить отчет по лабораторной работе

1 МЕТОДЫ РЕШЕНИЯ

1.1 Используемые библиотеки и инструменты

- pandas — для работы с данными
- numpy — для численных вычислений
- statsmodels — для дисперсионного анализа (ANOVA)
- scipy.stats — для расчета корреляций Пирсона
- matplotlib и seaborn — для визуализации результатов
- sklearn — для построения регрессионных моделей
- pingouin — для расчета частных корреляций

1.2 Методы анализа

Дисперсионный анализ (ANOVA)

- Использован однофакторный дисперсионный анализ для оценки влияния отдельных факторов
- Применен двухфакторный дисперсионный анализ для сравнения силы влияния факторов
- Проведен анализ взаимодействия факторов (interaction effects)

Корреляционный анализ

- Расчет коэффициента корреляции Пирсона для парных связей
- Проверка статистической значимости корреляций (p-value)
- Визуализация корреляционных матриц с помощью тепловых карт

Частная и множественная корреляция

- Расчет частных корреляций для устранения влияния третьих переменных
- Построение множественной регрессионной модели для оценки совместного влияния факторов

2 РЕЗУЛЬТАТЫ ВЫПОЛНЕНИЯ ЗАДАЧ

2.1 Задача 1

2.1.1 Анализ влияния каждого фактора отдельно

Влияние фактора "Тип ПО":

- Проведен однофакторный дисперсионный анализ
- F-статистика: $F = 3947.35$
- Коэффициент детерминации $R^2 = 1.00$ (100%)
- P-value: $p = 1.31 \times 10^{-34} < 0.05$
- Результат: Влияние фактора "Тип ПО" СТАТИСТИЧЕСКИ ЗНАЧИМО. 100% вариации в количестве уязвимостей объясняется различиями между типами ПО.

Влияние фактора "Тип ошибки (CWE)":

- Проведен однофакторный дисперсионный анализ
- F-статистика: $F = 127.89$
- Коэффициент детерминации $R^2 = 0.95$ (95%)
- P-value: $p = 2.92 \times 10^{-16} < 0.05$
- Результат: Влияние фактора "Тип ошибки (CWE)" СТАТИСТИЧЕСКИ ЗНАЧИМО. 95% вариации в количестве уязвимостей объясняется различиями между типами ошибок.

2.1.2 Сравнение силы влияния факторов

Проведен двухфакторный дисперсионный анализ

- Фактор "Категория фактора" (ПО vs Ошибка):
  - F-статистика: $F = 33.13$
  - P-value: $p = 4.42 \times 10^{-7} < 0.05$
  - Статистически значимо
- Фактор "Месяц":
  - F-статистика: $F = 0.000817$
  - P-value: $p = 0.999999 > 0.05$
  - Не является статистически значимым

- Результат: Категория фактора (тип ПО vs тип ошибки) оказывает БОЛЕЕ ЗНАЧИМОЕ ВЛИЯНИЕ на распределение уязвимостей, чем календарные месяцы. Влияние месяцев статистически незначимо.

2.1.3 Анализ взаимодействия факторов

- Проведен анализ взаимодействия факторов "Год" и "Месяц" для типов ошибок за период 2023-2025
- Использована формула: `Count ~ C(Year) * C(Month)`
- Влияние фактора "Год":
  - F-статистика: F = 1.42
  - P-value: p = 0.266909 > 0.05
  - Не является статистически значимым
- Влияние фактора "Месяц":
  - F-статистика: F = 0.038
  - P-value: p = 0.999050 > 0.05
  - Не является статистически значимым
- Взаимодействие "Год × Месяц":
  - F-статистика: F = 0.024
  - P-value: p = 1.000000 > 0.05
  - Не является статистически значимым
- Результат: Взаимодействие факторов НЕ является статистически значимым. Влияние месяца примерно одинаково в разные годы. Месячная динамика числа уязвимостей не различается статистически значимо от года к году.

2.2 Задача 2

2.2.1 Расчет корреляционной матрицы

- Построена матрица корреляций Пирсона между различными типами ошибок (CWE)
- Создана визуализация в виде тепловой карты для наглядного представления связей

2.2.2 Проверка значимости корреляций

- Рассчитана матрица p-значений для всех коэффициентов корреляции
- Определены статистически значимые корреляции (p-value < 0,05)

2.2.3 Выявление пар с очень сильной связью

- Найдены пары типов ошибок с очень сильной корреляцией ($|r| \geq 0.9$ по шкале Чеддока)
- Устранены дубликаты пар (A-B эквивалентно B-A)
- Результат: Обнаружена одна пара с очень сильной связью:
  - CWE-119 и CWE-476: коэффициент корреляции r = 0.949178
  - P-value для этой корреляции: p = 0.0038 < 0.05 (статистически значимо)
- Вывод: При возникновении уязвимости типа CWE-119 с очень высокой вероятностью (94.9%) будет обнаружена уязвимость типа CWE-476, и наоборот. Это указывает на тесную взаимосвязь между этими типами ошибок.

2.3 Задача 3

2.3.1 Простая парная корреляция

- Рассчитана матрица простых парных корреляций между:
  - Total_Vulns (общее число уязвимостей)
  - Critical_Vulns (критические уязвимости)
  - Incident_Vulns (уязвимости, связанные с инцидентами)

Результаты простой парной корреляции:
- Total_Vulns ↔ Critical_Vulns: r = 0.593 (умеренная положительная связь)
- Total_Vulns ↔ Incident_Vulns: r = 0.158 (слабая положительная связь)
- Critical_Vulns ↔ Incident_Vulns: r = 0.565 (умеренная положительная связь)

2.3.2 Частная корреляция

- Рассчитана матрица частных корреляций для устранения влияния третьих переменных

- Позволяет оценить "чистую" связь между переменными без учета влияния других факторов

Результаты частной корреляции:

- Total_Vulns ↔ Critical_Vulns (при контроле Incident_Vulns): r = 0.618 (увеличилась с 0.593)

- Total_Vulns ↔ Incident_Vulns (при контроле Critical_Vulns): r = -0.266 (изменилась с 0.158 на отрицательную!)

- Critical_Vulns ↔ Incident_Vulns (при контроле Total_Vulns): r = 0.592 (практически не изменилась с 0.565)

Интерпретация:

- Связь между общим числом уязвимостей и критическими уязвимостями усиливается при устранении влияния инцидентов

- Связь между общим числом уязвимостей и инцидентами становится отрицательной при контроле критических уязвимостей, что указывает на опосредованную связь через критические уязвимости

2.3.3 Множественная корреляция

- Построена множественная регрессионная модель:
  `Total_Vulns ~ Critical_Vulns + Incident_Vulns`
- Коэффициент детерминации: $R^2 = 0.6616$ (66.16%)
- Коэффициент множественной корреляции: R = 0.8134 (81.34%)

Результат: 66% вариации в общем числе уязвимостей можно объяснить линейной зависимостью от числа критических уязвимостей и уязвимостей, связанных с инцидентами. Это указывает на сильную связь между этими метриками.

3 ВЫВОДЫ

3.1 По Задаче 1

1. Влияние отдельных факторов:

- Фактор "Тип ПО" оказывает статистически значимое влияние ($p < 0.05$) и объясняет 100% вариации в количестве уязвимостей ($R^2 = 1.00$)

- Фактор "Тип ошибки (CWE)" оказывает статистически значимое влияние ($p < 0.05$) и объясняет 95% вариации в количестве уязвимостей ($R^2 = 0.95$)

- Оба фактора являются критически важными для распределения уязвимостей

2. Сравнение силы влияния:

- Категория фактора (тип ПО vs тип ошибки) оказывает более значимое влияние, чем календарные месяцы

- Влияние месяцев не является статистически значимым ($p = 0.999999 > 0.05$)

- Вывод: распределение уязвимостей в большей степени зависит от типа ПО/ошибки, чем от времени года

3. Взаимодействие факторов:

- Взаимодействие между факторами "Год" и "Месяц" не является статистически значимым ($p = 1.000000 > 0.05$)

- Влияние месяца примерно одинаково в разные годы

- Вывод: месячная динамика числа уязвимостей не различается статистически значимо от года к году

3.2 По Задаче 2

1. Корреляционные связи:

- Построена матрица корреляций Пирсона между 8 типами ошибок (CWE-79, CWE-119, CWE-416, CWE-476, CWE-125, CWE-120, CWE-20, CWE-787)

- Рассчитана матрица p-значений для проверки статистической значимости корреляций

- Выявлена одна пара с очень сильной связью: CWE-119 ↔ CWE-476 ($r = 0.949$, $p = 0.0038 < 0.05$)

2. Практическое значение:

- При обнаружении уязвимости типа CWE-119 с вероятностью 94.9% будет обнаружена уязвимость типа CWE-476

- Эта пара может рассматриваться как группа взаимосвязанных уязвимостей

- Результаты позволяют прогнозировать появление одних типов уязвимостей при обнаружении других

- Выявленная группа может использоваться для приоритизации проверок безопасности

3.3 По Задаче 3

1. Связи между метриками производителей:

- Простая парная корреляция:
  - Total_Vulns ↔ Critical_Vulns: r = 0.593 (умеренная связь)
  - Total_Vulns ↔ Incident_Vulns: r = 0.158 (слабая связь)
  - Critical_Vulns ↔ Incident_Vulns: r = 0.565 (умеренная связь)
- Частная корреляция (при контроле третьей переменной):
  - Total_Vulns ↔ Critical_Vulns: r = 0.618 (усилилась)
  - Total_Vulns ↔ Incident_Vulns: r = -0.266 (стала отрицательной!)
  - Critical_Vulns ↔ Incident_Vulns: r = 0.592 (практически не изменилась)
- Вывод: Связь между общим числом уязвимостей и инцидентами является опосредованной через критические уязвимости

2. Множественная корреляция:

- Коэффициент множественной корреляции: R = 0.8134 (81.34%)
- Коэффициент детерминации: R^2 = 0.6616 (66.16%)
- 66% вариации общего числа уязвимостей объясняется линейной зависимостью от критических уязвимостей и уязвимостей, связанных с инцидентами
- Результаты могут использоваться для прогнозирования общего числа уязвимостей на основе других метрик

ЗАКЛЮЧЕНИЕ

В ходе выполнения лабораторной работы были успешно применены методы дисперсионного и корреляционного анализа для исследования распределения уязвимостей в программном обеспечении.

Полученные результаты позволяют:

- Оценить влияние различных факторов на распределение уязвимостей
- Выявить корреляционные связи между типами ошибок
- Определить взаимосвязи между различными метриками уязвимостей производителей ПО

Результаты исследования могут быть использованы для:

- Приоритизации проверок безопасности
- Прогнозирования появления уязвимостей
- Разработки стратегий управления информационной безопасностью

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ