НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО

# Отчет по экзаменационному проекту по «Математической статистике»: Анализ факторов, влияющих на победу в футбольных матчах (нa примере чемпионата мира 2022 года)

Выполнили:

Савин Денис – сбор данных, парсинг, предварительная обработка, статистический анализ, построение моделей.

P3267

Зенькова Каролина – визуализация данных, подготовка графиков, написание отчета, формулировка выводов.

Группа: P3266

Преподаватель:

Яворук Т.О.

г. Санкт-Петербург

2025

## 1.Постановка задачи:

## 1.1. Общая постановка:

## Исследовать, какие статистические показатели наиболее значимо влияют на результат футбольного матча (победа/поражение), и разработать модель, позволяющую предсказывать исход игры на основе этих показателей.

## 1.2. Математическая постановка:

Для каждого матча имеется вектор признаков X = (x₁, x₂, ..., xₙ), где:

* x₁ - владение мячом (%)
* x₂ - точность пасов (%)
* xₙ - другие показатели (точность ударов, xG и т.д.)

И целевая переменная Y ∈ {0, 1}, где:

* 1 - победа
* 0 - поражение

Задачи:

1. Найти статистически значимые различия в показателях между победителями и проигравшими
2. Определить корреляции между показателями
3. Построить регрессионную модель Y = f(X)
4. Оценить предсказательную способность модели

## 1.3. Описание данных:

* Источник: fbref.com
* Выборка: 64 матча чемпионата мира 2022 года
* Показатели:
  + 1. Владение мячом (Possession)
    2. Точность пасов (Pass Accuracy)
    3. Точность ударов (Shots Accuracy)
    4. Ожидаемые голы (xG)
    5. Реальные голы (Goals)

## 2. Теория (статистические методы и ключевые понятия):

## 2.1. Описательная статистика: использовалась для первичного анализа данных.

**Основные показатели:**

1. Меры центральной тенденции:

1) Среднее арифметическое: **сумма всех численных значений, делённая на их количество**

2) Медиана: серединное значение в упорядоченном ряду

3) Мода: наиболее часто встречающееся значение

1. Меры изменчивости:

1) Дисперсия

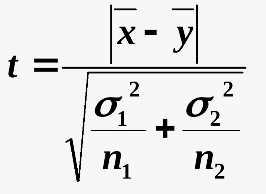
2) Стандартное отклонение

3) Размах: разница между максимальным и минимальным значениями

* 1. Межквартильный размах

1. Графические методы:
   1. Гистограммы распределения
   2. Боксплоты (ящики с усами)
   3. Круговые диаграммы для категориальных данных

**2.2. t-критерий Стьюдента:** использовался для сравнения средних значений между победителями и проигравшими.

 **Формула для независимых выборок:**

**Проверка гипотез:**

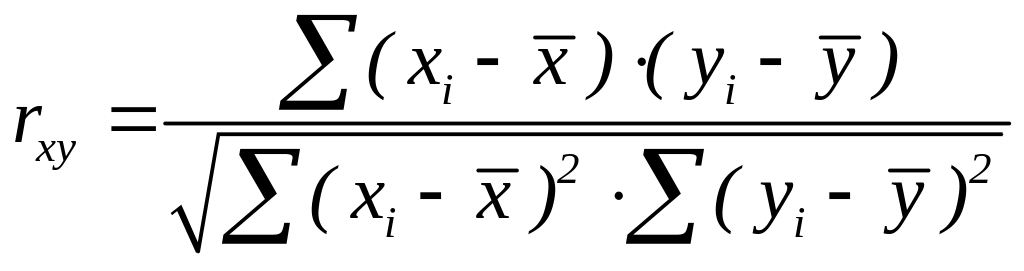
1. Нулевая гипотеза (H₀): μ₁ = μ₂ (средние равны)
2. Альтернативная гипотеза (H₁): μ₁ ≠ μ₂ (средние различаются)

**Интерпретация p-value:**

1. p < 0.05 - различия статистически значимы
2. p ≥ 0.05 - различия не значимы

**2.3. Корреляционный анализ Пирсона:** использовался для оценки линейной зависимости между переменными.

**Формула коэффициента корреляции:**

****

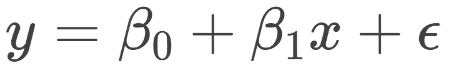
**Интерпретация:**

* 0.9 ≤ |r| ≤ 1.0 - очень сильная корреляция
* 0.7 ≤ |r| < 0.9 - сильная
* 0.5 ≤ |r| < 0.7 - умеренная
* 0.3 ≤ |r| < 0.5 - слабая
* |r| < 0.3 - отсутствие или незначительная корреляция

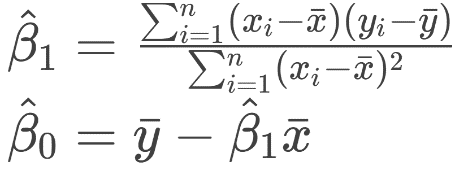
**Проверка значимости:**

* Нулевая гипотеза: ρ = 0 (корреляция отсутствует)
* Альтернативная гипотеза: ρ ≠ 0

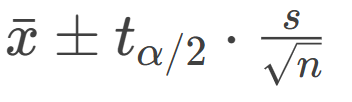
**2.4. Линейная регрессия:** использовалась для моделирования зависимости голов от xG.

**Модель простой линейной регрессии:**

**Оценка параметров (МНК):**



**2.4. Дополнительные методы:**

* + 1. **Доверительные интервалы** для средних и коэффициентов:

## Анализ мощности для оценки достаточности размера выборки

## 2.5. Ключевые понятия:

1. Ожидаемые голы (xG) - вероятностная метрика, оценивающая качество голевых моментов
2. Статистическая значимость - оценка достоверности различий (p-value < 0.05)
3. Коэффициент детерминации (R²) - доля объясненной дисперсии в модели

## Все расчеты проводились с уровнем значимости α = 0.05. Для обработки статистических тестов использовались двусторонние альтернативные гипотезы. В случаях нарушения предположений о нормальности рассматривалось применение непараметрических аналогов (тест Манна-Уитни вместо t-теста, корреляция Спирмена вместо Пирсона).

## 3. Используемые программные средства:

### 3.1. Инструменты разработки:

1. **Язык программирования**: C# (для сбора данных), Python (для анализа)
2. **Библиотеки**:
   * C#: HtmlAgilityPack (парсинг), MathNet.Numerics (статистика)
   * Python: pandas, numpy, matplotlib, seaborn, scipy
3. **Среда разработки**: Visual Studio Code, Jupyter Notebook
4. **Система контроля версий**: GitHub

3.2. Ссылка на данные:

https://fbref.com/en/comps/1/schedule/World-Cup-Scores-and-Fixtures  
https://github.com/sDenisss/MathStatProject2025/blob/main/ParsingResults.txt

3.3. Ссылка на github:

<https://github.com/sDenisss/MathStatProject2025>

## 4. Вывод программы:

## 4.1. Основные статистики:

## === Средние значения по всем матчам ===

## Владение мячом: 50.9% - 49.1%

## Точность пасов: 80.0% - 79.2%

## Всего пасов: 445 - 428

## Точность ударов: 35.7% - 29.7%

## Ударов всего: 4.1 - 3.4

## Голы: 1.58 - 1.11

## Ожидаемые голы (xG): 1.38 - 1.27

## === Победители vs Проигравшие ===

## Победители:

## Владение мячом: 48.8%

## Точность пасов: 79.3%

## Пасы всего: 422

## Точность ударов: 41.2%

## Ударов всего: 4.9

## Голы: 2.41

## Ожидаемые голы (xG): 1.68

## Проигравшие:

## Владение мячом: 51.2%

## Точность пасов: 79.9%

## Пасы всего: 432

## Точность ударов: 26.4%

## Ударов всего: 2.9

## Голы: 0.57

## Ожидаемые голы (xG): 1.03

## === Possession ===

## Среднее (победители): 49.90

## Среднее (проигравшие): 50.10

## t-статистика: -0.056

## p-value: 0.9556

## → Различие НЕ статистически значимо

## === Pass Accuracy ===

## Среднее (победители): 81.17

## Среднее (проигравшие): 80.90

## t-статистика: 0.171

## p-value: 0.8646

## → Различие НЕ статистически значимо

## === Shots Accuracy ===

## Среднее (победители): 42.86

## Среднее (проигравшие): 26.41

## t-статистика: 3.825

## p-value: 0.0003

## → Различие статистически значимо (p < 0.05)

## === Goals ===

## Среднее (победители): 2.59

## Среднее (проигравшие): 0.52

## t-статистика: 6.500

## p-value: 0.0000

## → Различие статистически значимо (p < 0.05)

## === xG ===

## Среднее (победители): 1.71

## Среднее (проигравшие): 0.92

## t-статистика: 3.797

## p-value: 0.0004

## → Различие статистически значимо (p < 0.05)

## 4.2. Графики

## **График 1: Линейная регрессия Голы от xG.**

## 

## График показывает зависимость между ожидаемыми голами (xG) и реально забитыми голами. Линия регрессии имеет уравнение y=0.86x+0.20, а коэффициент детерминации R2=0.35*.*

Анализ:

1. Наклон (0.86): Меньше 1, что означает: команды в среднем забивают чуть меньше, чем предсказывает xG.
2. Разброс точек: Значительный, особенно при высоких значениях xG. Это говорит о том, что xG не всегда точно предсказывает голы.
3. R² (0.35): Только 35% вариации голов объясняется xG.

Вывод: xG — полезный, но не идеальный показатель. На результат влияют и другие факторы (например, мастерство игроков, удача, действия вратаря).

## **График 2: Средние метрики: Победители vs Проигравшие.**

## 📊 1. Разница xG - Goals

🔍 Показывает, насколько команды **реализуют или не реализуют** свои моменты:

Смотреть BoxPlot xg\_minus\_goals.png

* Если xG - Goals < 0 → команда **недореализует** (создаёт больше, чем забивает);
* Если > 0 → **переигрывает ожидания**.

Сравнение средних значений показателей (владение мячом, точность ударов, xG и т.д.) для победителей и проигравших.

Анализ:

1. Победители превосходят проигравших по:
   1. Точности ударов (42.9% vs 26.4%)
   2. Количеству ударов (4.9 vs 2.9)
   3. Голам (2.41 vs 0.57)
   4. xG (1.68 vs 1.03)
2. Владение мячом и точность пасов почти одинаковы у обеих групп.

Вывод: победа зависит не от владения мячом, а от эффективности атак:

1. Больше ударов → больше голов.
2. Выше точность ударов → выше шансы забить.
3. xG у победителей значительно выше → они создают более опасные моменты.

#### **График 3: Ожидаемые голы vs Реальные голы.**

Сравнение xG и реальных голов в виде точечного графика. По оси X — xG, по оси Y — голы.

Анализ:

1. Большинство точек сосредоточено в диапазоне 0–3 для xG и 0–4 для голов.
2. Выбросы: Есть матчи, где голы сильно отклоняются от xG (например, забили намного больше или меньше ожидаемого).
3. Общая тенденция: Чем выше xG, тем чаще команды забивают, но связь не строгая.

Вывод: xG хорошо отражает общую эффективность атак, но в отдельных матчах результат может сильно отличаться из-за случайных факторов или индивидуального мастерства.

## **График 4: Разница между ожидаемыми и реальными голами***.*****

## 

График демонстрирует, насколько реальные голы отличаются от xG.

Анализ:

1. В среднем разница близка к нулю: Это значит, что xG в целом корректно оценивает количество голов.
2. Отклонения: Есть матчи, где разница значительна (например, команда забила на 2 гола больше xG).

Вывод: xG — надежный показатель для анализа в долгосрочной перспективе, но в одном матче возможны аномалии.

### Итоговые выводы по всем графикам:

1. **xG — лучший показатель для прогноза**, но не абсолютно точный.
2. **Победа определяется качеством атак** (удары, xG), а не контролем мяча.
3. **Отдельные матчи могут сильно отклоняться от статистики** из-за случайных факторов.
4. **Для анализа командам стоит фокусироваться на**:
   1. Создании моментов с высоким xG.
   2. Тренировке точности ударов, а не просто на удержании мяча.

## Обсуждения:

### **5.1. Основные выводы:**

1. Наиболее значимые различия между победителями и проигравшими наблюдаются по:
   1. Точности ударов (p = 0.0003)
   2. Реальным голам (p < 0.0001)
   3. Ожидаемым голам (p = 0.0004)
2. Владение мячом и точность пасов не показали статистически значимых различий.
3. xG является лучшим предиктором результата среди анализируемых показателей.

### **5.2. Ограничения исследования:**

1. Размер выборки (64 матча) может быть недостаточным для некоторых выводов.
2. Мы не учитывали дополнительные факторы (состав команд, тактические схемы).
3. Показатели анализировались изолированно, без учета взаимодействий.

**5.3. Перспективы дальнейших исследований:**

1. Увеличение выборки за счет других турниров.
2. Добавление новых показателей (количество фолов, сейвы вратарей).
3. Построение более сложных моделей (логистическая регрессия, случайный лес).
4. Анализ временных характеристик (изменение показателей в течение матча).

**5.4. Практические рекомендации:**

1. Для тренеров: уделять больше внимания качеству моментов (xG), а не владению мячом
2. Для аналитиков: использовать xG как основной показатель эффективности атак.
3. Для скаутов: обращать внимание на игроков с высокой точностью ударов.

**Заключение:**

Проведенное нами исследование показало, что традиционные показатели (владение мячом) менее значимы для предсказания результата, чем современные метрики (xG). Наибольшее влияние на исход матча оказывает эффективность атакующих действий, выражающаяся в точности ударов и качестве создаваемых моментов. Полученные результаты могут быть полезны для профессиональных клубов и аналитиков в области футбольной статистики.