**Дэвид Ланган. Прогнозирование результатов футбольных матчей**

Это сокращенный перевод книги, объясняющей, как повысить качество прогнозов результатов футбольных матчей путем использования различных вероятностных моделей. Автор подробно объясняет, как построить Пуассоновское, отрицательное биномиальное и ряд других распределений на основе данных о забитых голах. Результаты моделирования также предоставят вам много интересной статистической информации о клубах английской Премьер-лиги.

S. David Langan. Predict Football Matches. Using Spreadsheet Models to Become a Winning Sports Bettor

Изображение выглядит как текст, спорт

Автоматически созданное описание

### Глава 2. Модели распределения

#### Нормальное распределение

В 1738 году во втором издании книги «Учение о шансах» (The Doctrine of Chances) математик [Абрахам де Муавр](https://ru.wikipedia.org/wiki/%D0%9C%D1%83%D0%B0%D0%B2%D1%80,_%D0%90%D0%B1%D1%80%D0%B0%D1%85%D0%B0%D0%BC_%D0%B4%D0%B5) впервые представил нормальное распределение. С тех пор оно получило широкое распространение для описания явлений реального мира. Однако, нормальное распределение не является универсальным. Так Эрнест О'Бойл и Герман Агинис показали,[[1]](#footnote-1) что 71% игроков НБА ниже среднего показателя по набранным очкам. Авторы отказались от модели распределения, основанной на всех данных. Вместо этого они использовали модели, учитывающие уникальные характеристики отдельных наборов данных.

Выводы О'Бойла и Агиниса особенно важны в контексте футбольных матчей. В течение многих лет для прогнозирования забитых мячей аналитики полагались на распределение Пуассона. Т.е. считали, что шансы забить у всех клубов равны. Но это не так. В этой книге мы используем различные модели распределения, и выберем те, что отражают уникальные показатели каждого футбольного клуба. Благодаря этому мы надеемся улучшить качество прогнозов.

Если за выбором той или иной модели распределения не стоит теория, т.е., объяснение, почему ожидаемое число забитых голов отдельных команд подчиняется различным распределениям, то такой подход приведет к переподгонке. Всё дело в том, что исторические данные включают, как закономерности, так и случайности. Выбор модели распределения, наилучшим образом объясняющей исторические данные, предполагает, что все эти данные основаны на закономерностях, и оставляет меньше места волатильности, свойственной данным. Переподгонка приводит к тому, что модель лучше объясняет исторические данные, но обладает меньшей предсказательной силой. – *Здесь и далее текст, набранный с отступом – комментарии Сергея Багузина.*

#### Какая модель распределения лучше всего предсказывает результаты футбольных матчей?

Модели распределения вероятностей для прогнозирования исходов футбольных матчей применяются давно. Некоторые исследователи использовали распределение Пуассона. Другие – применяли модель Пуассона с избыточным количеством нулевых значений или отрицательное биномиальное распределение. Все эти исследователи находили, что их модели неплохо справляются с прогнозированием. К сожалению, модель основанная лишь на одном виде распределения вероятностей (неважно каком), будет ненадежной.

Каждый футбольный клуб уникален. Состоит из игроков разного уровня, клубом управляют тренеры с разной философией. Клуб, исповедующий атакующий футбол, состоящий из элитных игроков, забьет голов больше, чем клуб, предпочитающий защитную философию с игроками среднего уровня. Следовательно, элитный клуб будут забивать чаще, и характерная для него модель распределения может быть иной.

Автор иллюстрирует все примеры результатами матчей Премьер-лиги сезона 2012–2013, сыгранных до 20 января 2013 года. Я использую [итоги](https://www.flashscore.ru/football/england/premier-league-2020-2021/results/) Премьер-лиги сезонов 2019/2020 и 2020/2021.

Изображение выглядит как стол

Автоматически созданное описание

Рис. 1. Как часто команды забивали то или иное число голов в сезоне 2020–2021

Сумма 760 говорит, что было сыграно 38 туров по 10 матчей для 2 команд: 38\*10\*2 = 760. 224 раза команды не смогли забить. Столбец E содержит усеченные (или объединенные) данные. В ячейке Е9 указано число голов 5 и более. Усечение будет использоваться во всей книге чтобы свести к минимуму влияние выбросов.

Далее будут подробно описаны три модели: обычное распределение Пуассона, распределение Пуассона с избыточным количеством нулевых значений (сокращенно ZIP) и отрицательное биноминальное распределение. Модель распределения Пуассона наиболее точно соответствует фактическим данным. Т.е., дельта между фактическими значениями частоты голов и прогнозными значениями на основе распределения Пуассона меньше, чем при использовании иных распределений. Означает ли это, что мы можем основывать наши прогнозы забитых голов в сезоне 2021–2022 на распределении Пуассона? Нет. С началом нового сезона может оказаться, что лучшее соответствие обеспечивает иное, например, равномерное распределение. Более того, тот факт что распределение Пуассона было наилучшим для совокупности команд Премьер-лиги в сезоне 2020–2021, не означает, что для каждой отдельной команды оно также будет наилучшим. Чтобы лучше предсказывать результаты матчей, важно анализировать исторические значения по каждому клубу раздельно.

Результаты такого раздельного анализа обобщены в таблице. Видно, что модель Пуассона с избыточным количеством нулевых значений лучше всего подходит для семи клубов, распределение Пуассона – для восьми, а отрицательное биномиальное распределение – для пяти.

Изображение выглядит как стол

Автоматически созданное описание

Рис. 2. Виды распределений для 20 команд Премьер-лиги в сезоне 2020–2021

Далее в главах 3–5 будет показано, как моделировать значения частоты голов для отдельных команд, используя пуассоновское, модифицированное пуассоновское и отрицательное биномиальное распределения. В главе 9, вы узнаете, как использовать эти вероятности для прогнозирования результатов матчей. В главе 10 будет представлена модель прогнозирования основанная на разнице забитых голов двух команд.

## Глава 3. Распределение Пуассона

#### Математическая теория

Автор концепции – [Симеон Дени Пуассон](https://ru.wikipedia.org/wiki/%D0%9F%D1%83%D0%B0%D1%81%D1%81%D0%BE%D0%BD,_%D0%A1%D0%B8%D0%BC%D0%B5%D0%BE%D0%BD_%D0%94%D0%B5%D0%BD%D0%B8) – в 1837 году проанализировал возникновение дискретных независимых событий в течение фиксированного интервала времени. Исследования Пуассона оставались непроверенными в течение почти шестидесяти лет, пока экономист и математик [Владислав Борткевич](https://ru.wikipedia.org/wiki/%D0%91%D0%BE%D1%80%D1%82%D0%BA%D0%B5%D0%B2%D0%B8%D1%87,_%D0%92%D0%BB%D0%B0%D0%B4%D0%B8%D1%81%D0%BB%D0%B0%D0%B2_%D0%98%D0%BE%D1%81%D0%B8%D1%84%D0%BE%D0%B2%D0%B8%D1%87) не понял, что распределение Пуассона можно использовать для анализа неудачных событий, постигших прусских солдат. Борткевич определил, что число прусских солдат, убитых брыкающимися лошадьми за двадцатилетний период, соответствовало форме распределения Пуассона. Придя к этому выводу, Борткевич показал, как теоретические модели распределения могут быть использованы для объяснения статистической обусловленности редких событий. Сейчас распределение Пуассона используется для прогнозирования числа сбоев в день в сетях связи, количества банкротств, поданных в месяц, количества звонков клиентов на телефонную линию поддержки в час, и, что важно для наших целей, количества голов, забитых клубом за 90 минут.

Распределение Пуассона является подходящим распределением вероятностей для моделирования значений частоты голов, поскольку каждый гол является дискретным и независимым событием. Дискретное событие – это событие, которое может принимать только целые значения. Футбольный клуб может забить 3 гола в матче, но не 3,45.

Концепция ожидаемых голов низвергает этот принцип)). Подробнее см. [Джеймс Типпет. xG – философия ожидаемых голов](https://baguzin.ru/wp/?p=23909).

Независимое событие – это событие, которое не изменяет вероятность возникновения последующих событий. Гол считается независимым событием, поскольку он не влияет на вероятность того, сколько голов будет забито в дальнейшем. Предположение о независимости действует в течение всего матча. Хотя в концовке оно и не такое точное.

Каждая модель распределения описывается функцией плотности вероятности. Эта функция определяет связь между ее параметрами и результирующим значением вероятности. Плотность вероятности для распределения Пуассона:

где *P(x)* – вероятность того, что клуб забьет *х* голов за матч, *е* – основание натурального логарифма (число Эйлера), *е* = 2,72, но мы будем использовать Excel-функцию EXP(), чтобы избежать ошибок округления; *λ* (лямбда) – математическое ожидание случайной величины (среднее количество событий за фиксированный промежуток времени); с увеличением лямбды распределение Пуассона стремится к нормальному распределению.

#### Фулхэм

На примере выступления Фулхэма в Премьер-лиге сезона 2020–2021, покажем, как строить распределение Пуассона:

Изображение выглядит как стол

Автоматически созданное описание

Рис. 3. Частота голов Фулхэма в сезоне 2020–2021 описывается распределением Пуассона

Шаг 1. С использованием сводной таблицы постройте частотное распределение голов (столбец С). Для этого выберите Фулхэм в области фильтров сводной таблицы в ячейке С2.

Шаг 2. Постройте область D4:E10, чтобы отразить весь набор голов от 0 до 5, и соответствующую им частоту голов.

Шаг 3. В ячейке Е11 найдите матожидание (среднее число) голов в одном матче. Используйте формулу =СУММПРОИЗВ(D5:D10;E5:E10)/E11. Получившееся значение 0,71 отражает тот факт, что в среднем за матч Фулхэм забивал менее одного гола.

Шаг 4. В ячейке F5 найдите значение функции плотности вероятности для распределения Пуассона для нуля голов. Используйте формулу =ПУАССОН.РАСП(D5;$E$12;ЛОЖЬ). Протащите формулу до ячейки F10. В ячейке G5 рассчитайте ожидаемое число матчей, в которых Фулхэм не забьет: =F5\*$E$11.

Шаг 5. Как понять, насколько хорошим является соответствие между фактическим числом матчей с тем или иным количеством голов (столбец Е) и моделью Пуассона (столбец G)? Воспользуемся статистическим методом, называемым «критерий хи-квадрат»:

Здесь *Ei* фактическое число раз, когда было забито *i* голов; значение берется из столбца Е, а число голов из столбца D; например, значение в ячейке Е6, равное 14, означает, что один гол был забит в 14 матчах. *Gi* – ожидаемое значение голов в соответствии с распределение Пуассона; берется из столбца G. Для вычисления хи-квадрат (отклонения) в ячейке H5 используйте формулу =(E5-G5)^2/G5. Протащите ее до ячейки H10. В ячейке H11 найдите сумму значений ячеек H5:H10.

Статистика хи-квадрат в нашем примере равна 0,32. Что означает этот результат? Хи-квадрат является мерой соответствия распределения Пуассона фактическим значениям. Чтобы определить, является ли соответствие между наблюдаемыми и прогнозируемыми значениями хорошим или плохим, мы должны сравнить получившееся значение 0,32 с теоретическим. Для последнего в Excel есть функция ХИ2.ОБР(вероятность;степени\_свободы).

Здесь *Вероятность* – степень точности, которую мы хотим получить от результатов наших прогнозов. Стандартная точности, используемая многими математиками, равна 95%. Т.е., мы ожидаем, что с вероятностью 95% прогнозируемые данные (столбец G на рис. 3) и фактические данные (столбец Е) описываются одной и той же закономерностью. Если это так, то любая разница между наблюдаемыми и прогнозируемыми значениями объявляется результатом случайности.

Второй параметр функции ХИ2.ОБР *Степени свободы* – число элементов в выборке минус 1. В нашей выборке присутствуют 6 элементов (голы 0, 1, 2, 3, 4 и 5). Так что число степеней свободы равно 5. Ниже представлены теоретические значения статистики хи-квадрат для степеней свободы от 1 до 10 и точности 0,95:

Изображение выглядит как стол

Автоматически созданное описание

Рис. 4. Теоретическая статистика хи-квадрат

Статистика хи-квадрат в нашем примере – 0,32 – существенно меньше теоретического предела, для пяти степеней свободы, равного 11,07. Это означает, что распределение забитых мячей Фулхэма почти идеально моделируется распределением Пуассона. Соответственно, мы можем использовать распределение Пуассона, для прогнозирования исхода последующих матчей с участием Фулхэма.

### Глава 4. Распределение Пуассона с избыточным количеством нулевых значений

#### Математическая теория

Модель *распределения Пуассона* *с избыточным количеством нулевых значений*(zero-inflated Poisson distribution, ZIP) является модификацией распределения Пуассона. ZIP-распределения используются для моделирования наборов данных, которые имеют больше нулевых значений, чем при обычном распределении Пуассона. Наличие дополнительных нулевых значений в данных изменяет форму распределения Пуассона, так что лямбда больше не будет равняться среднему значению распределения. Также мы не сможем применить стандартную функцию плотности вероятности ПУАССОН.РАСП().

Вместо этого мы используем две функции: первую – для ожидаемой вероятности нуля голов, вторую – для одного и более голов:

Параметр *φ* объясняет чрезмерную дисперсию, вызванную наличием дополнительных нулевых значений. Когда *φ* = 1, ZIP переходит в распределение Пуассона.

#### Ливерпуль

Рассмотрим построение ZIP модели на примере Ливерпуля образца сезона 2020–2021.

Шаг 1. С использованием сводной таблицы постройте частотное распределение голов (столбец С). Для этого выберите Ливерпуль в ячейке С2.

Шаг 2. Постройте область D4:E10, чтобы отразить весь набор голов от 0 до 5, и соответствующую им частоту голов. Добавьте столбец F с относительной частотой голов. Введите в ячейке F5 формулу =E5/$E$11, и протащите ее до F10.

Шаг 3. Введите в столбец G значения ZIP в соответствии с формулами (3). В ячейке G5 введите формулу для х=0: =(1-E13)+$E$13\*EXP(-$E$12)\*$E$12^D5/ФАКТР(D5). В ячейке G6 введите формулу для х=1: =$E$13\*EXP(-$E$12)\*$E$12^D6/ФАКТР(D6). Протащите формулу до ячейки G10. Установите начальные (тестовые) значения *φ* и *λ* в ячейки Е12 и Е13, чтобы убедиться, что функции Excel работают должным образом. Значение *φ* должно быть больше нуля, но меньше единицы. Значение *λ* должно быть больше нуля. Установим 0,5 и 2.

Изображение выглядит как стол

Автоматически созданное описание

Рис. 5. Функция плотности вероятности для ZIP (столбец G) при тестовых значениях *φ* = 0,5 и *λ* = 2

Подготовим оптимизационную модель. В ней мы будем искать такие значения *φ* и *λ*, которые будут минимизировать отличие ожидаемой частоты (ZIP) от фактической частоты (столбец F).

Шаг 4. Рассчитайте вес каждого элемента выборки. Т.е., вклад нуля голов, одного гола и т.д.:

где *rf(x)* – относительная частота гола; значение из столбца F. В ячейку Н5 введите формулу =ЕСЛИОШИБКА(1/F5/(1-F5);0). Эта формула возвращает вес в соответствии с формулой (4), если частота гола больше нуля, и ноль, если частота гола равна нулю.

Изображение выглядит как стол

Автоматически созданное описание

Рис. 6. Взвешенные значения квадратов разности

Шаг 5. Рассчитайте квадрат ошибки между ожидаемым значением (ZIP) и фактической частотой. В ячейке I5 введите формулу =(F5-G5)^2. Протащите ее до I10.

Шаг 6. Используя реальное число голов, вес частоты голов и значения квадратов ошибок, вычислите взвешенные квадраты ошибок. Введите в J5 формулу =E5\*H5\*I5. Протащите ее до J10. В ячейке J11 найдите сумму взвешенных квадратов ошибок. Получилось значение 8,180.

Шаг 7. Теперь с помощью инструмента *Поиск решения* мы найдем такие *φ* и *λ*, при которых значение суммы взвешенных квадратов ошибок будет минимальным. Найдите на вкладке *Данные* в правой части ленты кнопку *Поиск решения*. (Если этой кнопки нет, пройдите по меню *Файл* –> *Параметры*, перейдите на вкладку *Надстройки*, и в нижней части экрана в поле *Управление* выберите *Надстройки Excel*. Нажмите кнопку *Перейти*. В открывшемся окне *Надстройки* поставьте галочку напротив *Поиск решения*, нажмите Ok.)

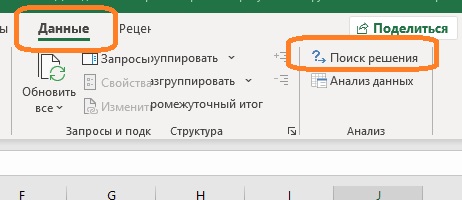


Рис. 7. Надстройка Excel *Поиск решения*

Откроется окно *Поиск решения*, сделайте установки, как на рисунке:

Изображение выглядит как текст

Автоматически созданное описание

Рис. 8. Настройка *Поиска решения*

Укажите, значение в какой ячейке мы будем оптимизировать (1). Выберите цель оптимизации. В нашем случае – поиск минимума (2). Укажите, какие ячейки Excel может изменять для поиска оптимума (3). Задайте ограничение. Для этого нажмите кнопку *Добавить*. В нашем случае ограничение одно – *φ* ≤ 1. Поставьте галочку *Сделать переменные без ограничений неотрицательными* (5). Это задаст *φ* ≥ 0 и *λ* ≥ 0. Выберите метод решения *Поиск решения нелинейных задач методом ОПГ* (6). Нажмите *Найти решение*. Если вы все сделали верно, поработав несколько секунд, Excel выведет на экран окно *Результат поиска решения*. Прочтите все сообщения, и нажмите Ok. Excel автоматически заменит начальные значения *φ* и *λ* на оптимальные.

Изображение выглядит как стол

Автоматически созданное описание

Рис. 9. *φ* и *λ*, при которых сумма взвешенных квадратов ошибок минимальна

Сравните рис. 6 и 9. Сумма взвешенных квадратов ошибок уменьшилась с 8,180 до 0,102. Это произошло благодаря изменению значений *φ* и *λ*.

Иногда работа надстройки *Поиск решения* закончится сообщением, что приемлемое решение найти не удалось. Это означает, что модифицированное распределение Пуассона (ZIP) не подходит для описания ваших данных. Попробуйте иные распределения, описанные ниже.

Шаг 8. Используя найденные значения *φ* и *λ* в столбце К рассчитайте ожидаемое количество матчей, в которых будет забито *х* голов, а в столбце L – значение хи-квадрат, как делали это для распределения Пуассона в главе 3. Сумму значений хи-квадрат в ячейке L11 сравните с теоретическим значением распределения хи-квадрат для вероятности 95% и 5 степеней свободы (см. рис. 4). В нашем случае значение хи-квадрат равно 0,35, что существенно меньше теоретического значения 11,07.

Изображение выглядит как стол

Автоматически созданное описание

Рис. 10. ZIP-распределение статистически значимо для описания частоты голов Ливерпуля в сезоне 2020–2021

### Глава 5. Отрицательное биномиальное распределение

#### Математическая теория

Отрицательное биномиальное распределение описывается двумя функциями плотности вероятности. Первая – дает вероятность того, что клуб не забьет голов, а вторая вычисляет вероятностей забить один гол или более:

По определению *k* ≥ 0. При *k* ≥ 1000 отрицательное биномиальное распределение становится неотличимым от распределения Пуассона. Однако столь большие значения *k* нас интересовать не будут.

Мы можем записать плотность вероятности для каждого *х*:

Изображение выглядит как текст

Автоматически созданное описание

Рис. 11. Функции плотности вероятности для *х* = 0, 1, 2, 3, 4 и 5

#### Метод моментов

При применении метода моментов предполагается, что среднее значение (*х̅*) и дисперсия (*s2*) фактических голов равны среднему и дисперсии отрицательного биномиального распределения. Исходя из этого, параметр *k* можно рассчитать с помощью формулы:

Здесь *s2* – дисперсия голов, и рассчитывается по формуле:

Здесь *f(x)* – число голов после усечения; *х* = 0, 1, 2, 3, 4 и 5; *х̅* – среднее число голов за матч; *N* – общее число матчей. Метод моментов позволяет понять, соответствует ли отрицательное биномиальное распределение фактическим данным или нет. Поскольку параметр *k* должен быть больше нуля, вам не следует использовать отрицательное биномиальное распределение, если метод моментов вернул *k* < 0. Обратитесь к другим моделям.

Классическое описание отрицательного биноминального распределения несколько иное. Пусть проводятся испытания, в каждом из которых может произойти только 2 события: событие «успех» с вероятностью *p* или событие «неудача» с вероятностью *q = 1 – p* (схема Бернулли). Пусть проведено *x + r* испытаний, из которых *х* закончились неудачей, а *r* – успехом. В этом случае случайная величина *x* будет иметь отрицательное биномиальное распределение с плотностью вероятности:

В Excel есть функция ОТРБИНОМ.РАСП(число\_неудач;число\_успехов;вероятность\_успеха;интегральная)

Мне не удалось сопоставить классический подход и предложенный автором книги.

#### Бёрнли

Рассмотрим концепцию на примере Бёрнли:

Изображение выглядит как стол

Автоматически созданное описание

Рис. 12. Метод моментов на примере Бёрнли сезона 2020–2021

Шаг 1. С использованием сводной таблицы постройте распределение голов (столбец С). Для этого выберите Бёрнли в ячейке С2.

Шаг 2. Постройте область D4:E10, чтобы отразить весь набор голов от 0 до 5, и соответствующее им число матчей. В ячейке Е11 подсчитайте общее количество матчей, в ячейке Е12 – среднее за матч количество голов =СУММПРОИЗВ(D5:D10;E5:E10)/E11, а в ячейке Е13 дисперсию голов =СУММПРОИЗВ((D5:D10-E12)^2;E5:E10/E11).

Шаг 3. В ячейке Е14 рассчитайте параметр дисперсии методом моментов в соответствии с формулой (9) и формулой в Excel =E12^2/(E13-E12).

Шаг 4. В ячейке Е15 рассчитайте вероятность *p* по формуле (8).

Шаг 5. В столбце F, используя формулы с рис. 11, рассчитайте частоту реализации *х* голов. В столбце G рассчитайте ожидаемое число матчей, в которых будет забито *х* голов для сезона из 38 матчей.

Шаг 6. В столбце Н рассчитайте значение хи-квадрат, а в ячейке Н11 – сумму значений хи-квадрат. Сравните её с теоретическим значением распределения хи-квадрат для вероятности 95% и 5 степеней свободы (см. рис. 4). Поскольку наше значение (0,53) меньше теоретического (11,07), можно сделать вывод о пригодности отрицательного биноминального распределения для прогнозирования числа ожидаемых голов для Бёрнли образца 2020–2021.

\*\*\*

В главе 6 рассматривается оценка параметра дисперсии для отрицательного биномиального распределения методом минимального взвешенного квадратичного отклонения. А в главах 7 и 8 моделирование числа забитых голов с помощью геометрического и равномерного распределения.

### Глава 9. Пришло время сделать некоторые прогнозы!

#### Условные вероятности и правило умножения вероятностей

Условная вероятность используется для прогнозирования результатов нескольких событий. Если мы знаем вероятности двух событий *A* и *B*, мы можем рассчитать вероятность наступления события *A*, при условии наступления события *B*:

*P(A│B)* – условная вероятность наступления события *А* при условии, что наступило событие *В*. Числитель уравнения – вероятность наступления обоих событий A и В. Знаменатель – вероятность наступления события В.

Для независимых событий (а голы являются независимыми событиями) вероятность одновременного возникновения событий *A* и *В* может быть рассчитана с использованием правила умножения вероятностей. Это правило гласит, что совместная вероятность независимых событий *A* и *В* равна вероятности события *A*, умноженной на вероятность события *B*:

*(13) P(A∩B) = Р(А) ∙Р(В)*

Например, вероятность победы Фулхэма со счетом 2:1 в матче против Бёрнли может быть рассчитана с использованием правила умножения. Если мы знаем, что вероятность того, что Фулхэм забьет два гола в матче, равна 12,4%, а вероятность того, что Бёрнли забьет один гол, равна 33,0%, вероятность победы Фулхэма со счетом 2:1 равна 12,4% \* 33% = 4,1%. Вы можете спросить: «Откуда мы взяли значения 12,4% и 33%?». Они появились в результате моделирования тем или иным видом распределения (см. рис. 3 и 12).

Рассмотрим построение таблицы вероятностей всех исходов на примере.

#### Фулхэм – Бёрнли

Шаг 1. Постройте шаблон таблицы. Клетки серого цвета соответствуют победе Бёрнли, а клетки синего цвета – ничьей. В столбце В введите значения вероятностей забить *x* голов, рассчитанные ранее для Фулхэма с распределением Пуассона. В столбце С введите значения вероятностей для Бёрнли, используя отрицательное биноминальное распределение.

Шаг 2. Вычислите условные вероятности результатов матчей. Для этого в ячейке G3 введите формулу =$B3\*ВПР(G$2;$A$2:$C$8;3;ЛОЖЬ). Протяните ее на диапазон G3:L8. Можно вычислить вероятности исходов: победы, поражения и ничьей, суммировав соответствующие ячейки.

Я для этого применил пользовательскую функцию [СуммЦвет](http://baguzin.ru/wp/?p=317):

Function СумЦвет(диапазон As Range, критерий As Range) As Double

' Определяет сумму значений в ячейках "диапазона",

' цвет которых совпадает с цветом в ячейке "критерий"

Application.Volatile True

Dim i As Range

For Each i In диапазон

If i.Interior.Color = критерий.Interior.Color Then

СумЦвет = СумЦвет + i

End If

Next

End Function

Изображение выглядит как стол

Автоматически созданное описание

Рис. 13. Расчет вероятностей результатов матчей

#### Преобразование процентных коэффициентов в другие форматы

Большинство сайтов онлайн-ставок предоставят вам возможность изменить формат коэффициентов по умолчанию на формат, основанный на процентах. Тем не менее, полезно знать, как конвертировать процентные вероятности. Используя вероятность победы Бёрнли, равную 0,35441 (см. рис. 13) можно рассчитать десятичный коэффициент:

Например, если букмекерская контора разместила десятичные коэффициенты 2,82, ваша ставка в 100 фунтов стерлингов в случае успеха принесет 282 фунта стерлингов (включая первоначальную ставку в 100 фунтов стерлингов).

Дробные коэффициенты практически идентичны десятичным коэффициентам, за исключением того, что сумма, поставленная игроком, не включена в значение коэффициентов.

*(15) Дробный коэф. (Бёрнли) = Десятичный коэф. (Бёрнли) – 1 = 2,82 – 1 = 1,82*

Если игрок сделал ставку в 100 фунтов стерлингов на основе дробного коэффициента 1,82, чистая прибыль, которую он должен получить в случае успеха, составит 182 фунта стерлингов (не включая сумму первоначальной ставки).

Ставки на победу (Moneylines) обычно используются для присвоения коэффициентов спортивным событиям, для которых спреды по очкам не имеют значения. Например, автомобильные гонки или боксерские поединки.

Расчет будет иным, если вероятность более 50%. Например, когда вероятность победа Бёрнли оценивается в 60%...

Ставки на победу часто выражаются в ставках на 100 фунтов стерлингов. Следовательно, коэффициент +182,16 означает, что игрок должен рискнуть 100 фунтами стерлингов, чтобы заработать 182,16 фунтов стерлингов. И наоборот, коэффициент –150 означает, что игрок, делающий ставки на спорт, должен рискнуть 150 фунтами стерлингов, чтобы заработать 100 фунтов стерлингов.

#### Итак…

Теперь в вашем распоряжении статистически обоснованная модель для прогнозирования результатов матчей Премьер-лиги. Но мы еще не закончили. Ранее я рекомендовал никогда не принимать решения о ставках на основе одной модели распределения. У вас всегда должна быть под рукой вторая модель, чтобы подтвердить результаты вашей основной модели. Заключительная глава будет посвящена описанию независимой модели прогнозирования вероятности победы. Эта вторичная модель основана на разнице забитых голов двумя клубами во всех предыдущих матчах сезона для прогнозирования результатов будущих матчей.

### Глава 10. Модель прогнозирования на основе разницы забитых голов (margin-of-victory)

Термин margin-of-victory (MoV) относится к разнице между количеством голов, забитых победившими и проигравшими в матче. Модель учитывает, где проходит игра: дома или в гостях. Существует множество различных вариантов модели. Мы используем модель, описанную Уэйном Уинстоном в его замечательной книге *Mathletics: как игроки, менеджеры и любители спорта используют математику в бейсболе, баскетболе и футболе*.

Автор использовал результаты Премьер-лиги сезона 2011/2012, чтобы предсказать результаты матчей в сезоне 2012/2013. Я использовал итоги сезона 2019/2020, чтобы предсказывать результаты матчей сезоны 2020/2021.

Шаг 1. Для начала создайте сводную таблицу на основе листа *Исходный*. Выведите в область строк названия команд, в область фильтра – сезон 2019–2020, а в область значений – голы:

Изображение выглядит как стол

Автоматически созданное описание

Рис. 14. Итоги Премьер-лиги сезона 2019–2020

Перенесите данные, как значения в новый диапазон, исключите команды покинувшие Премьер-лигу в конце сезона, и добавьте команды, поднявшиеся из Чемпионшипа. Назначьте новым командам число голов, как у наихудшего клуба, оставшегося в Премьер-лиге после сезона, а именно – Кристал Пэлас = 31.

Изображение выглядит как стол

Автоматически созданное описание

Рис. 15. Таблица с голами сезона 2019–2020 для команд-участниц сезона 2020–2021; командам, поднявшимся из Чемпионшипа указано по 31 голу

Шаг 2. В столбце C укажите ранг команды. В ячейке C4 поместите формулу =РАНГ.СР(B4;$B$4:$B$23;0). Протащите ее до C23. Если команды забили одинаковое число голов будет найден их средний ранг.

Шаг 3. Метод ранжирования не отражает фактическую разницу между голами, забитыми командами. Поэтому вместо ранга мы будем использовать коэффициент, основанный на отклонении числа голов команды от среднего числа голов. В ячейке D4 введите формулу =B4/$B$24. Протащите ее до D23.

Шаг 4. Значения рейтинга в столбце D усреднены вокруг 1. Предпочтительно, чтобы оценки были усреднены около нуля. Такие значения будут более интуитивно понятными: клуб с положительным рейтингом забьет больше голов за сезон, чем в среднем по лиге. И наоборот, клуб с отрицательным рейтингом забьет меньше голов за сезон, чем в среднем по лиге. В столбце E просто вычтите 1 из значений в столбце D.

Шаг 5. Чтобы включить в модель фактор преимущества домашнего матча в ячейке F4 введите предварительную оценку такого преимущества, равную 0,300. Это значение предполагает, что, играя дома, клуб за матч будет забивать на 0,300 гола больше. Это всего лишь предположение. От сезона к сезону это значение колеблется и обычно находится в диапазоне от 0,200 до 0,400. Чуть позже с помощью надстройки *Поиск решения* мы рассчитаем фактические значения, используя результаты матчей сезона Премьер-лиги 2019/2020.

Шаг 6. Столбец G дублирует столбец E. Во-первых, это нужно для работы надстройки *Поиск решения*. Надстройка может изменять значения в ячейках только, если они являются константами. Во-вторых, по мере того, как вы будете обновлять свою модель каждую неделю, вы сможете отслеживать изменения рейтинга каждого клуба, сравнивая значения в столбцах Е и G. Обратите внимание, что столбец G содержит значения, а не формулы.

Шаг 7. Далее мы рассчитаем прогнозируемую разницу забитых и пропущенных голов (MoV) для матчей сезона 2020/2021, используя усредненные по нулю оценки в столбце G. Для удобства присвоим имя диапазону G4:G23. Для этого выделите ячейки G4:G23, и введите в поле *Имя*, расположенное слева от строки формул, *Рейтинг* (см. левый верхний угол на рис. 14).

Шаг 8. Постройте модель MoV. На том же листе Excel добавьте новую таблицу (рис. 16). Обе таблицы (рис. 15 и рис. 16) должны быть размещены на одном листе Excel, иначе надстройка *Поиск решения* не сможет одновременно обработать значения Рейтинга и MoV. В таблице MoV укажите параметры матчей Премьер-лиги сезона 2020–2021: дату проведения матчей, названия команд и число забитых голов (столбцы I:M).

Изображение выглядит как стол

Автоматически созданное описание

Рис. 16. Исходные данные в таблице MoV

Шаг 9. В столбце N рассчитайте фактическую разницу забитых и пропущенных голов. В ячейке N4 введите формулу =L4-M4. В столбце О рассчитайте прогнозируемую MoV. Для каждого матча прогноз MoV состоит из трех компонентов: надбавки за игру дома, рейтинга домашнего клуба и рейтинг клуба гостей. В ячейке O4 введите формулу:

=$F$4+ИНДЕКС($G$4:$G$23;ПОИСКПОЗ(J4;$A$4:$A$23;0))-ИНДЕКС($G$4:$G$23;ПОИСКПОЗ(K4;$A$4:$A$23;0))

Изучим её. Часть $F$4 – это домашнее преимущество. Во втором слагаемом ИНДЕКС($G$4:$G$23;ПОИСКПОЗ(J4;$A$4:$A$23;0)) сначала функция ПОИСКПОЗ() ищет клуб из ячейки J4 (это Вест Хэм) в диапазоне $A$4:$A$23. Причем ищет точное совпадение (третий аргумент ПОИСКПОЗ равен 0). Функция ПОИСКПОЗ возвращает значение 6, т.к. Вест Хэм является шестым в списке. Затем функция ИНДЕКС() ищет шестое значение в диапазоне $G$4:$G$23 и возвращает значение рейтинга, равное -0,044. Третье слагаемое выполняет аналогичную работу для гостевой команды Ньюкасл Юнайтед, и возвращает её рейтинг -0,259. Значение в ячейке О4 =0,300-0,044-(-0,259)=0,515.

Протащите формулу из ячейки О4 на весь столбец с данными.

Шаг 10. Определите ошибку между фактическими и прогнозируемыми значениями MoV. В ячейке P4 введите формулу =N4-O4. Протащите формулу до конца области данных. В ячейке Q4 введите формулу =P4^2, которая найдет квадрат ошибки. Протащите формулу до конца области данных. В ячейке R4 найдите сумму квадратов ошибок, введя формулу =СУММ(Q:Q). Её протаскивать не надо.

Шаг 11. Рассчитайте актуальные рейтинги клубов. Для этого с помощью надстройки *Поиск решения* найдите такие значения в столбце G и ячейке F4, при которых сумма квадратов ошибок будет минимальной. Запустите *Поиск решения*, и настройте его, как указано ниже:

Изображение выглядит как текст

Автоматически созданное описание

Рис. 17. Настройка параметров *Поиска решения*

Значение в ячейке R4 уменьшилось с предварительных 1198 (см. рис. 16) до 1042.

Изображение выглядит как стол

Автоматически созданное описание

Рис. 18. Таблица MoV после оптимизации

Значение домашнего преимущества (F4) и рейтинги всех команд (G4:G23), соответствующие минимальной сумме квадратов ошибок, автоматически обновлены. Фактическое значение домашнего преимущества оказалось ничтожно малым = 0,011. Это означает, что в сезоне 2020–2021 команды на домашнем поле практически не имели преимущества. Возможно, это связано с тем, что почти весь сезон игры шли при пустых трибунах.

Изображение выглядит как текст, квитанция

Автоматически созданное описание

Рис. 19. Фактическое домашнее преимущество в сезоне 2020–2021 и актуализированные рейтинги команд

Я перенес рейтинги на отдельный лист Excel и сравнил рейтинги по итогам двух сезонов:

Изображение выглядит как стол

Автоматически созданное описание

Рис. 20. Сравнение рейтингов двух сезонов; клубы отсортированы по дельте

Наибольший рост продемонстрировали Лидс, МЮ, Астон Вилла, Вест Хэм и Тоттенхэм. Эти клубы были довольно высоко в турнирной таблице.

#### Сравнение прогнозов через модель распределения и MoV

В главе 9 мы рассчитали вероятности голов в гипотетическом матче Фулхэм – Бёрнли (см. рис. 13). Ожидалось, что Фулхэм забьет 0,710 гола, в Бёрнли 0,860. Т.е., преимущество Бёрнли оценивалось в 0,150 гола. Проверим эти прогнозы моделью MoV. Коэффициент на победу Фулхэма = преимущество хозяев плюс рейтинг Фулхэма минус рейтинг Бёрнли = 0,011+(–0,650)–(–0,550)=–0,089. Модель MoV ожидает, что Бёрнли забьет на 0,089 гола больше, чем Фулхэм. Что ж обе модели дают близкие результаты.

#### Заключительные мысли

Теперь в вашем распоряжении две модели прогнозирования результатов матчей. Если вы будете обновлять значения по итогам каждой игровой недели, вы сможете последовательно прогнозировать результаты будущих матчей Премьер-лиги. Если вы опытный пользователь Excel, вам следует подумать об автоматизации рабочих листов, чтобы упростить еженедельное обновление значений. Если вы не являетесь опытным пользователем Excel, посетите веб-сайт [Practical Sports Prediction Series](http://www.sportspredictionseries.com/), чтобы узнать больше об автоматизации Excel (необходима регистрация). Веб-сайт также служит форумом, на котором вы можете оставить комментарии или задать автору вопросы, касающиеся методов прогнозирования, описанных в этой книге.

1. Ernest O’Boyle Jr. and Herman Aguinis, "Hie Best and die Rest: Revisiting die Norm of Normality of Individual Performance" Personnel Psychology 65 (2012): 79-119. [↑](#footnote-ref-1)