Статистический вывод на основе критерия Колмогорова

Эта заметка родилась на стыке трех моих увлечений: футбол, Excel и статистика)) Известно, что число голов, забитых каждой командой в одном матче подчиняется распределению Пуассона. Я решил проверить это на результатах матчей английской премьер-лиги сезона 2021/2022. Всего было 38 туров по 10 матчей в туре, по две команды в одном матче. Итого 760 исходных значений.

	A	B	C	D	E	F	G	H
1	Тур	Хозяева	Голы хозяев	Гости	Голы гостей		Голов	Случаев
2	1	BRE	2	ARS	0		0	212
3	1	MUN	5	LEE	1		1	244
4	1	BUR	1	BHA	2		2	164
5	1	CHE	3	CRY	0		3	88
6	1	EVE	3	SOU	1		4	32
7	1	LEI	1	WOL	0		5	15
8	1	WAT	3	AVL	2		6	3
9	1	NOR	0	LIV	3		7	2
10	1	NEW	2	WHU	4		Всего голов	1071
11	1	TOT	1	MCI	0		Всего случаев	760
12	2	LIV	2	BUR	0		Голов в среднем	1,409
13	2	AVL	2	NEW	0			
14	2	CRY	0	BRE	0			

Рис. 1. Распределение числа забитых голов

Распределение Пуассона

Распределение Пуассона имеет один параметр λ – среднее количество успешных испытаний в заданной области возможных исходов. Количество успешных испытаний *X* пуассоновской случайной величины изменяется от 0 до бесконечности. Распределение Пуассона описывается формулой:

$$(1)\quad P(X) = \frac{e^{-\lambda}\lambda^{X}}{X!}$$

где *P(X)* – вероятность *X* успешных испытаний, λ – среднее ожидаемое количество успехов, *e* – основание натурального логарифма, равное 2,71828, *X* – количество успехов.

В Excel распределение Пуассона можно задать формулой =ПУАССОН.РАСП(X; λ; ЛОЖЬ). Чтобы сравнивать одинаковые сущности, я разделил число случаев (столбец H на рис. 1) на общее число случаев. И получил вероятности (столбец B на рис. 2). Также я подсчитал значения распределения Пуассона для X = 0, 1, …, 7 при λ = 1,409 (столбец C на рис. 2). Здесь 1,409 – среднее число голов, забитых одной командой в матче в сезоне 2021/2022. Например, вероятность не забить ни одного гола X = 0 задается формулой =ПУАССОН.РАСП(0;1,409;ЛОЖЬ) = 0,244 или 24,4%.

	A	B	C	D
1	Распределение числа забитых голов			
2	Голов	Факт	Пуассон	
3	0	0,279	0,244	
4	1	0,321	0,344	
5	2	0,216	0,243	
6	3	0,116	0,114	
7	4	0,042	0,040	
8	5	0,020	0,011	
9	6	0,004	0,003	
10	7	0,003	0,001	
11				
12				
13				
14				
15				

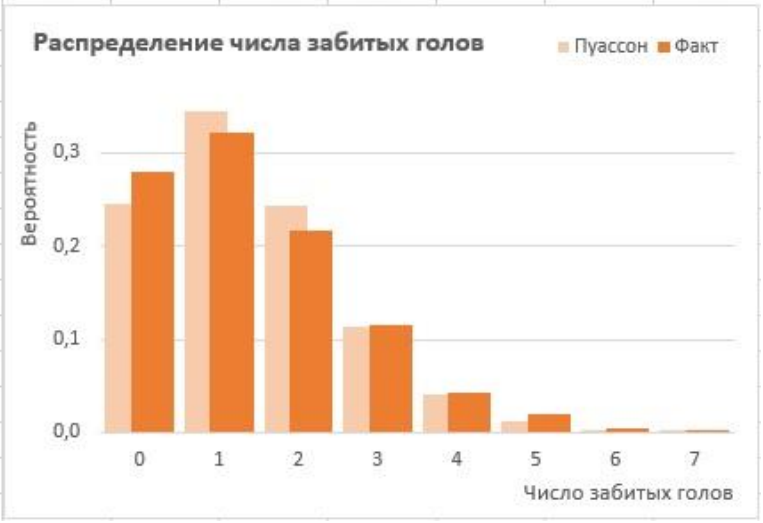

Рис. 2. Вероятности забить *n* голов в сезоне 2021/22 и распределение Пуассона для λ = 1,409

Статистический вывод

Глядя на рис. 2 можно заметить, что фактические вероятности забить n голов и вероятности, соответствующие распределению Пуассона для среднего λ = 1,409 неплохо совпадают. Статистический вывод позволяет количественно оценить, насколько «неплохо».

Итак, в качестве нулевой гипотезы H_0 примем, что наша выборка по результатам сезона 2021/22 происходит из генеральной совокупности, подчиняющейся распределению Пуассона. В качестве альтернативной гипотезы H_1 будем считать, что выборка происходит из генеральной совокупности, описываемой иным распределением.

Осталось выбрать статистику, которая позволит сравнить с одной стороны расхождения между фактическим и распределением Пуассона, а с другой – с критическим значением статистики, соответствующим α = 5% или, что еще строже, α = 1%. t-статистика не подходит, и я решил впервые в своей практике воспользоваться статистикой Колмогорова.

Статистика критерия Колмогорова

Критерий согласия Колмогорова служит для проверки гипотезы о принадлежности значений выборки определённому теоретическому закону распределения. В нашем случае мы хотим проверить принадлежит ли фактическое распределение частоты голов в сезоне 2021/22 распределению Пуассона.

Статистика критерия задается формулой

$$(2)\ D_n = \sup_{|x|<\infty} |F_n(x) - F(x,\Theta)|$$

где $F_n(x)$ – эмпирическая интегральная функция распределения на участке от 0 до x; $F(x, \Theta)$ – теоретическая интегральная функция распределения с параметром Θ на участке от 0 до x; x – значения, для которых получено эмпирическое распределение, в нашем случае – число голов от 0 до 7; n – объем выборки, в нашем случае – 760; sup – супремум, почти синоним максимума.

В нашем примере теоретическая функция с параметром Θ – это распределение Пуассона с неизвестным параметром λ. Мы заменяем неизвестный параметр λ, значением 1,409, полученным из экспериментальных данных (см. дополнение от 10.12.2022 ниже).

Изобразим наши данные в терминах уравнения (2):

	A	B	C	D	E	F
1		Плотность вероятности		Интегральная функция		
2	x	$F_n(x)$	F(x,Θ)	$F_n(x)$	F(x,Θ)	D_n
3	0	0,279	0,244	0,279	0,244	0,035
4	1	0,321	0,344	0,600	0,589	0,011
5	2	0,216	0,243	0,816	0,831	-0,015
6	3	0,116	0,114	0,932	0,945	-0,014
7	4	0,042	0,040	0,974	0,985	-0,012
8	5	0,020	0,011	0,993	0,997	-0,003
9	6	0,004	0,003	0,997	0,999	-0,002
10	7	0,003	0,001	1,000	1,000	0,000

Рис. 3. Разница интегральных функций распределения: фактической и Пуассона

Здесь в столбцах D и E я отразил интегральные (накопленные) частоты, как сумму частот для отдельных значений из столбцов B и C. В столбце F подсчитана разность значений соответствующих строк столбцов D и E. Видно, что максимальная разница между интегральными функциями фактического и распределения Пуассона достигается в первой точке при x = 0.

Распределение Колмогорова

Распределение Колмогорова имеет вид (k – целое):

$$(3)\ K(x) = \begin{cases} \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2}, & x > 0 \\ 0, & x \leqslant 0 \end{cases}$$

Современный Excel позволяет построить распределение Колмогорова на основании формулы (3) без обращения к таблицам из справочников (см. Excel-файл лист «Рис. 4»)

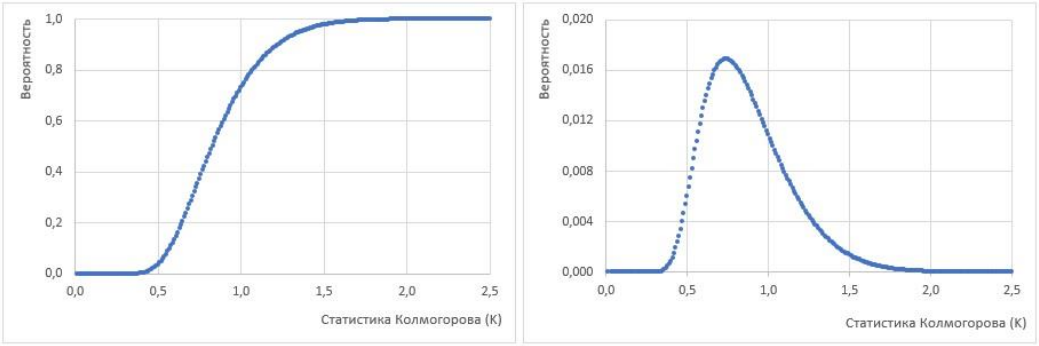

Рис. 4. Функция распределения Колмогорова: а) интегральная; б) плотность вероятности[1]

Кривая плотности вероятности распределения Колмогорова подобна нормальной, но с ярко выраженным правым хвостом.

Статистика Колмогорова является правосторонней, и в соответствии с теоремой Колмогорова…

$$(4)\ \lim_{n\to\infty} P\left(\sqrt{n}\cdot D_n \le t\right) = K(t)$$

… позволяет находить доверительные интервалы теоретической функции распределения F(x,Θ).

Критерий Колмогорова

Следуя традиции, можно использовать два доверительных интервала для отклонения нулевой гипотезы H_0: 95%-ный и 99%-ный:

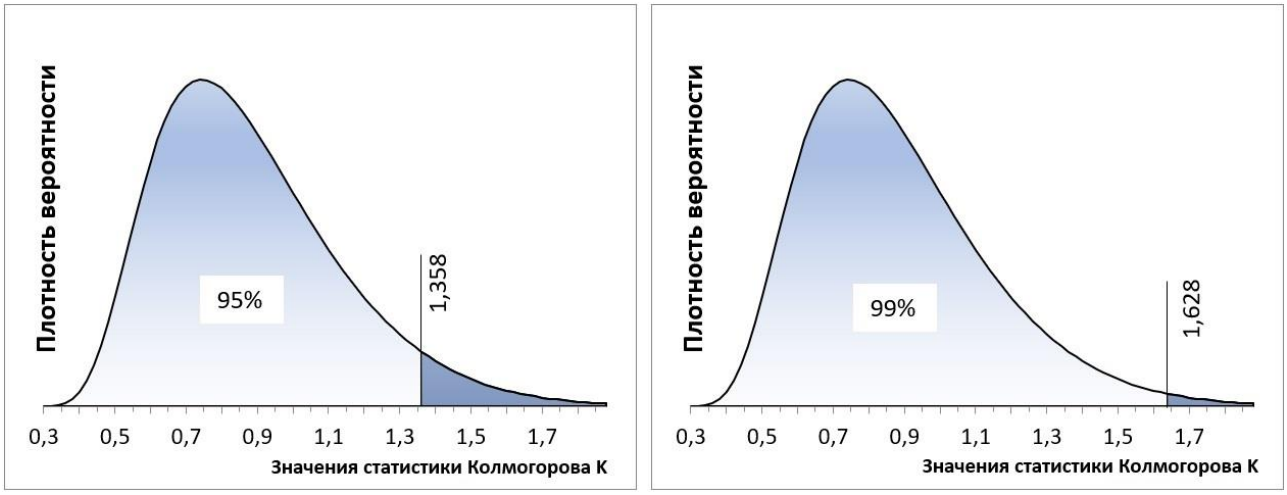

Рис. 5. Области отклонения нулевой гипотезы (более темные), и соответствующие им значения статистики Колмогорова (K)

Таким образом, для отклонения нулевой гипотезы с достоверностью 95% необходимо, чтобы эмпирическое значение статистики Колмогорова (K) превысило $K_{95\%}$ = 1,358. Для отклонения нулевой гипотезы с достоверностью 99% необходимо, чтобы эмпирическое значение статистики Колмогорова (K) превысило $K_{99\%}$ = 1,628.

Проверка соответствия частоты голов распределению Пуассона

Теперь осталось сравнить D_n = 0,035 (см. рис. 3) со значениями критерия Колмогорова для n = 760 при уровне значимости α = 0,05 и α = 0,01. Для этого надо перейти от D_n к K, используя формулу (4).

$$(5)\ K = \sqrt{n}\cdot D_n = \sqrt{760}\cdot 0{,}035 = 0{,}954$$

Результат сравнения удобно изобразить на числовой прямой:

[1] Так совпало, что два совершенно разных понятия в заметке обозначены одним символом – λ. В первом случае это была статистика распределения Пуассона, во втором – статистика Колмогорова.

Рис. 6. Зоны отклонения нулевой гипотезы

В соответствии с критерием Колмогорова нулевую гипотезу о соответствии распределения частот голов в сезоне 2021/22 распределению Пуассона с λ = 1,409 отклонить нельзя.

Дополнение от 10.12.2022. Дмитрий в комментариях обратил мое внимание, что использование распределения Колмогорова для статистического вывода о $K_{эксп.}$ = 0,954 не корректно. Т.е., сравнение $K_{эксп.}$ = 0,954 нужно вести не с $K_{95\%}$ = 1,358 и $K_{99\%}$ = 1,628, а с другими значениями, полученными не на основании распределения Колмогорова (3), а на основании иного распределения.

В нашем примере теоретическая функция с параметром Θ – это распределение Пуассона с неизвестным параметром λ. Если бы мы знали параметр λ, то могли бы сравнить статистику критерия Колмогорова (2) с распределением Колмогорова (3). Что я и сделал в заметке. Мы же знаем только оценку параметра λ, подсчитанную по выборке за сезон 2021/22 и равную 1,409. Когда по анализируемой выборке оцениваются параметры теоретического закона, согласие с которым проверяется, статистика критерия Колмогорова может существенно отличаться от распределения Колмогорова. Как считает Александр Иванович Орлов если пренебрегать этим отличием, согласие с проверяемым законом будет подтверждаться чаще, чем следует. Математический аппарат, который используется в этом случае, выходит за рамки уровня моего блога))

При беглом знакомстве с работами А.И. Орлова я нашел лишь критику использования распределения Колмогорова, когда по анализируемой выборке оцениваются параметры теоретического закона, согласие с которым проверяется. Позитивную программу, что делать в этом случае, я нашел в работах Бориса Юрьевича Лемешко, к которым отсылаю заинтересованных читателей.