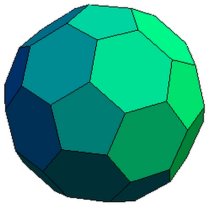
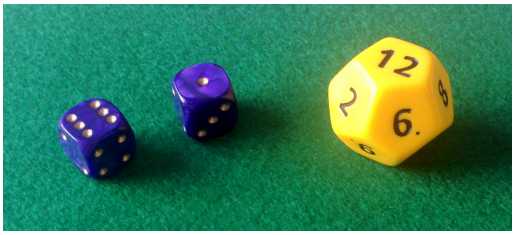
Probability, statistics and football

Franka Miriam Brückler
Paris, 2015.

Please read this before starting! Although each activity can be performed by one person only, it is suggested that you work in groups of three: one performing one part of each experiment (e.g., rolling one die), the other performing the second part of the experiment (e.g., rolling the second die) and the third keeping track of the results. It is also suggested that you discuss and agree on answers to questions before continuing with the tasks.

Note that at each workplace you have pencils and calculators: throughout the tasks, you will be required to fill in some answers and to do some calculations.

1 Level 1

Discovering probability

Introduction. Probabilities of events are described by numbers between 0 and 1 (i.e. percentages between 0 % and 100 %: you obtain the percentage by multiplying the decimal value with 100, e.g., 0,25 is 25 %). The larger the probability, the oftener the event is expected to happen in a series of trials with equal conditions.

⚀ At your workplace you can find a 12-sided (dodecahedral) die and two regular 6-sided dice. *Guess* the answers to the following questions:

1. The probability of rolling a ⚀ with a regular die is _____ and with a 12-sided die it is _____.
2. The probability of not rolling a ⚀ with a regular die is _____.
3. The probability of rolling ⚀⚀ with two dice is _____.
4. The probability of rolling a total of 5 with two regular dice is _____.
5. The probability of rolling a total smaller than 5 with two regular dice is _____.
6. If you roll the pair of "normal" dice (and check the totals) a number of times and then the same number of times roll the 12-sided die, do you expect that the corresponding numbers will appear approximately the same number of times? _____

⚀ Roll all the tree dice and make notes of the totals obtained on the two regular dice and the numbers on the 12-sided die. Repeat at least 30 times. Here is the table to help you keep track of the results:

Roll	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
12-sided die															
2 dice															
Roll	16.	17.	18.	19.	20.	21.	22.	23.	24.	25.	26.	27.	28.	29.	30.
12-sided die															
2 dice															
Roll	31.	32.	33.	34.	35.	36.	37.	38.	39.	40.	41.	42.	43.	44.	45.
12-sided die															
2 dice															

⚁ Now, determine how many times appeared which result on the two regular dice, and how many times appeared which result on the 12-sided die.

Result . . . appeared on	1	2	3	4	5	6	7	8	9	10	11	12
12-sided die:												
2 regular dice:												

⚂ Try to represent your results graphically in such a way that someone not seeing the tables above could understand which results you obtained in your experiment!

⚃ You have rolled the dice ⚽ = _____ times; this is your total number of trials. The fraction of the occurrences of a result and ⚽ is called the results' (relative) frequency. Calculate the frequencies for your results; you can express them as fractions, decimally or as percentages, it is up to you!

(Relative) frequency of result	1	2	3	4	5	6	7	8	9	10	11	12
On 12-sided die:												
On 2 regular dice:												

Use the relative frequencies to answer the following sentences and compare your answers to your guesses from the beginning:

1. The ⚀ with a regular die had frequency _____ and with a 12-sided die frequency _____.
2. A result different from ⚀ with a regular die appeared with frequency _____.
3. The result ⚀⚀ with two dice appeared with frequency _____.
4. The total of 5 with two regular dice appeared with frequency _____.
5. A total smaller than 5 with two regular dice appeared with frequency _____.
6. The frequencies of same results on the two dice and the 12-sided die are (not) similar. (Cross the "not" if you think the results are similar).

🎲 The previous activities show that relative frequencies, particularly when calculated from a large number of trials, can be used as estimates of theoretical probability. In season 2014/15, the Ligue 1 teams had the following relative frequencies of goals per shot:[1]

Equipe	Buts	Tirs	Freq. Rel.	Meilleur cube
Paris SG (M, L)	83	491	16,90%	
Olympique Lyon	72	524	13,74%	
Olympique Marseille	76	579	13,13%	
FC Évian Thonon Gaillard	41	334	12,28%	
SM Caen (N)	54	462	11,69%	
HSC Montpellier	46	396	11,62%	
SC Bastia	37	320	11,56%	
AS Saint-Étienne	51	443	11,51%	
EA Guingamp (P)	41	359	11,42%	
AS Monaco	51	452	11,28%	
Stade Reims	47	441	10,66%	
OGC Nizza	44	418	10,53%	
FC Toulouse	43	431	9,98%	
Girondins Bordeaux	47	488	9,63%	
Stade Rennes	35	365	9,59%	
OSC Lille	43	463	9,29%	
FC Lorient	44	486	9,05%	
RC Lens (N)	32	411	7,79%	
FC Metz (N)	31	450	6,89%	
FC Nantes	29	424	6,84%	

One may imagine that rolls of a die represent a team's shots and that rolling a specific result, e.g. a 1, represents the team scoring a goal. If you know that there exist dice with 4, 6, 8, 10, 12, 14, 16, 18 and 20 sides, complete the table above by entering the number of sides of the die you would use to model shots!

⚽⚽⚽ CONCLUSIONS ⚽⚽⚽

The more trials you make, the nearer the calculated (relative) frequency of an event is to its (theoretical) probability.

If the probability of an event happening is p, the probability of it not happening is $1 - p = 100\% - p$.

If some events cannot happen simultaneously, the probability of at least one of them happening is the sum of their probabilities.

[1] Data from `http://www.squawka.com/football-team-rankings`.

If two events are independent (one does not affect the outcome of the other and vice versa), the probability of both of them happening simultaneously is the product of their probabilities.

⚽⚽⚽

The birthday paradox

Introduction. Sometimes one can intuitively make a good guess about probability. But, watch out: probability is full of not so intuitive results. One of them is known as the birthday paradox.

⊡ What do you think, how many people should at least be in a group so that it is more probable than not (i.e., that the probability is just above 50 %) that someone shares your birthday? _____
If there are 23 persons, do you think it is probable that someone shares your birthday? _____

⚂ How many do you think should be assembled in order that at least two of them share a (not specified) birthday? _____ If there are 23 persons, do you think it is probable that there is a pair with a common birthday? _____

⚂ Now, as the first activity suggested, correct guesses about probability should not differ significantly from relative frequencies in a reasonably large number of trials. For the above questions, trials would be taking (randomly chosen) groups of people of a given size, checking if there is a person sharing your birthday, or if there are two sharing a birthday. If you were right in your guesses, then if you check a sufficiently large number of groups of 23 people, in more (or less, depending on what your choice was) than half of the groups there would be a person sharing your birthday or two persons sharing a birthday.

On the table you can find the tables with data on birthdays of players from the World Cup 2014. In each team there were 23 players and you decide to check _____ teams (the more, the better). In how many teams can you find a person sharing *your* birthday? _____ So, the estimate for the probability of someone sharing your birthday in a group of 23 people is _____

Now check in how many teams there are two players sharing a birthday: _____ Thus the estimated probability of two people sharing a birthday in a group of 23 people is _____ Do the frequencies meet your expectations? _____

⚂ Something to think about at home: Here is the table of probabilities corresponding to the two questions for various sizes of groups. By considering the probabilities for small groups (of 2, 3, 4, 5 people), and using conclusions from previous sections, try to figure out how the probabilities were calculated!

Number of people in the group	2	3	4	5	10	20	23	50
Prob. of someone sharing your birthday	0,5 %	0,8 %	1,1 %	1,4 %	2,7 %	5,3 %	6,1 %	12,8 %
Prob. of 2 of them sharing a birthday	0,3 %	0,8 %	1,6 %	2,7 %	11,7 %	41,1 %	50,7 %	97 %

⚽⚽⚽ Conclusions ⚽⚽⚽

One has to be very careful about details when calculating or estimating probabilities. It is not the same to ask if someone in a group shares *your* birthday, and if some two people in the group share *any* birthday. You need at least 253 people to make it more likely than not that someone shares your birthday, but only 23 to make it more likely than not that at least two of them have a common birthday.

⚽⚽⚽

2 Level2

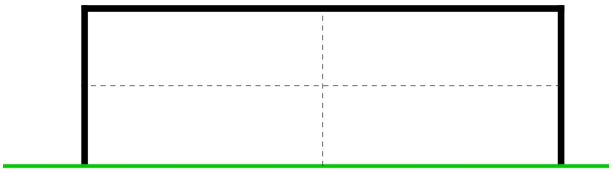
Football penalties: Introduction to probabilistic modeling

Introduction. It has been noted that in most leagues and championships about 70 % to 80 % of penalties result in a goal. A common notion on penalties is that they are a kind of lottery, both for the shooter and the keeper. By simple probabilistic modeling one can support the reasonableness of the mentioned percentages if this lottery assumption is true. By the way, the numbers can also be supported by geometric and physical arguments, but this is not our concern here.

⚀ In the table below you can find the data on penalties for the last three years of the French Ligue 1 and the last three World Cup Finals. Calculate the corresponding frequencies.

	Ligue 1 14/15	Ligue 1 13/14	Ligue 1 12/13	WC2014	WC2010	WC2006
Penalties awarded:	126	79	71	36	32	26
Penalties scored:	103	66	64	26	23	38
Frequency (in %)						

⚁ Imagine the following scenario: The keeper holds the shot if he guesses rightly into which part of the goal the shooter sends the ball, otherwise the penalty results in a goal. For example, one could consider the goal divided into 4 areas, so the keeper has probability $p =$_____ to guess correctly. In other words the probability of scoring a goal in this simple scenario is _____. Now, are the assumptions of the scenario realistic? What is missing? Think about that before reading further.



⚂ Is it sensible to consider the goal to be divided, from the keeper's point of view, into just 2 or 3 parts? _____ And into 100 parts? _____ Write the range of numbers of parts you think reasonable to dived the goal into: _____. Choose a number ⚽ = _____ from that range. At your workplace you can find plastic pearls in two colors: 60 in color 1 and 6 in color 1. Take ⚽−1 pearls in color 1 and one in color 2 and put them in the first box. E.g., if you decide to divide the goal into 4 parts, you use 3 pearls in color 1 and one in color 2).

The pearl in color 2 represents the keeper saving the goal. If you would close your eyes and take a pearl from the box, what is the probability of drawing a pearl in color 1 (i.e., the goal being scored)? ⚽ = _____.

⚃ Besides the keeper holding the ball, what other reason can prevent the scoring of a goal? Think of Zlatan Ibrahimović or David Trezeguet ☺. _____

Obviously, in professional football it is not very probable that a player misses the goal. Choose a percentage from the first row of the table below and take pearls accordingly. Put them in the second box. (this box represents possible misses by the The pearls in color 2 represent the misses.

Chosen percentage of miss	2 %	4 %	5 %	6 %	8 %	10 %
Take . . . pearls in color 1	49	24	19	47	23	9
Take . . . pearls in color 2	1	1	1	3	2	1

⚄ Now, close your eyes and draw one pearl from each of the boxes. Take a look at them: if neither of them is in color 2 (i.e. neither represents an unsuccessful shot), make a mark on the paper. Return each pearl in its original box. Repeat the experiment at least 30 times; do not forget to count how many times you have made the draw! So, in _____ draws ("penalty shots"), there were _____ marks ("goals"). Thus in your model, the frequency of successfully realized goals is _____. Compare this to the percentages from the beginning!

⚅ If the probability of the keeper making the wrong choice is p, and the probability of the shooter not missing the goal is q (in your case, this was the difference from 100 % to your chosen percentage of miss, i.e. _____), what is the right guess for the probability of scoring a goal: $p+q$ or pq or _____? Use the frequency obtained in ⚄ above to help you answer this question!

⚽⚽⚽ Conclusions ⚽⚽⚽

If two events are independent (one does not affect the outcome of the other and vice versa), the probability of both of them happening simultaneously is the product of their probabilities. Although in real life true independence is rare, it can often be assumed without making a big error (for example, in the above model we assume that the keeper's choice is independent on the shooter's miss or non-miss).

⚽⚽⚽

Odds and betting coefficients

Introduction. Betting houses offer various bets on football. If one places a bet, the numbers called coefficients tell how much this person will win in case he made the right guess. Since the outcome is not known in advance, the coefficients are calculated from probabilities. In this activity you will learn the basic connection between probability, odds and coefficients.

⚀ If you would play a game of tossing a coin with somebody and you bet on "heads" and your opponent on "tails", what is the probability that you make the right guess? _____ And the probability of your opponent guessing rightly? _____ How many possible outcomes are there? _____ Complete the following equality:

the number of possible outcomes in your favor : the number of outcomes in your opponent's favor
= _____ : _____.

This ratio we call your **odds** in this game. If, before tossing the coin, you place a bet of value ☺, what amount should your opponent place if both of you want a fair game: ☺, less than ☺ or more than ☺? _____ If both of you bet the amount 1, how much do you win in the case you guessed correctly, if you count the return of your bet as win? _____. The last number is your betting **coefficient**. How is this number related to the probability you winning? _____

⚁ There are several ways to write the coefficients. For example, for the match Paris SG - Bordeaux played on Sep 11th 2015, one site offered the coefficient 9/2 and another 5,10. The last is the coefficient in the sense above: if it were fair, it would mean that it is reciprocal to the probability of the match ending in a draw. The fractional coefficient represents just the profit, i.e. it has value 1 less than our decimal coefficient.

You could have also found a coefficient like +533. This is called moneyline odds and represents the profit per 100-unit bets (+) or how much you would have to wager to win 100 units (if the number is stated with a − in front). If the number is +, then you just divide it by 100 to get the fractional odds. Use the table below to determine if all the three betting houses made the same estimate for the probability of a draw in the match Paris SG : Bordeaux!

Type of coefficient	Value	Decimal value	Probability
Decimal	5,10	5,10	
Fractional	9/2		
Moneyline	+533		

⚂ In the last season of League 1, 88 of the 380 matches played ended in a draw — this makes _____ %. A year before it was 107 of 380 matches or _____ %. These numbers is not unusual: in professional football, about 20 % to 30 % of all matches (i.e., 20 to 25 of 100 of matches) end in a draw. Imagine that you are offered to bet on a match to end in a draw, but you do not know which teams play that match. Choose 20 %, 25 % or 30 % as the probability of you winning that bet: _____. This corresponds to odds[2] _____:_____. What would be the fair coefficient? ⚽=_____

⚄ Betting houses usually offer coefficients lower than fair in order to make sure they earn money. Now you will use the 20-sided die found on your workplace to test how much you would win or loose if you take ⚽ as the betting coefficient and how much if the coefficient had value ⚽−0,4. Your chosen probability of bet was _____ $= \frac{\_\_\_\_}{20}$. This means that of 20 numbers on the die, the first ☺_____ numbers can model the cases when matches really end in a draw. For example, if your probability was 5/20, the numbers 1 to ☺= 5 would represent a draw.

Now roll your die 30 times and count the number of times and note your results:

Roll	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
Result															
Roll	16.	17.	18.	19.	20.	21.	22.	23.	24.	25.	26.	27.	28.	29.	30.
Results															

So you obtained numbers from 1 to ☺ _____ times in 30 rolls (i.e., _____ of the 30 matches ended in a draw). If you had betted a value of 1 on each roll (match), how much you would have won or lost with the betting coefficient ⚽? _____ And with ⚽−0,4? _____

⚽⚽⚽ Conclusions ⚽⚽⚽

Decimal betting coefficients represent the reciprocal values of the corresponding probabilities. By decreasing the true value of coefficients betting houses ensure profit for themselves.

⚽⚽⚽

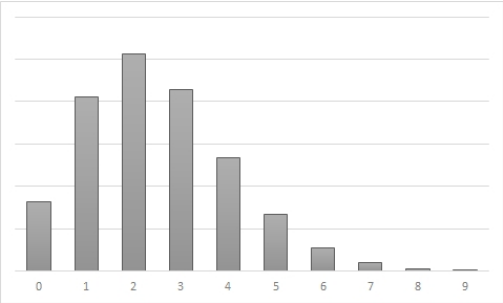
[2]If you are not sure if you solved part ⚀ above correctly: the odds are "the number of results in the favor of your bet : the number of the remaining possible results".

3 Level 3

Poisson (not fish) in football

Introduction. When making random experiments or observing results of non predictable events (e.g., results in football matches), one will often discover typical patterns how specific results appear in a given number of events. Such patterns are called distributions: a distribution consists of pairs of type "result - number of appearances of the result" and is most conveniently represented by a bar chart.

For example, it has been noted that in most championships and cups the numbers of matches with 0, 1, 2, ... goals approximately follow a specific pattern, known as the Poisson distribution. The specific form of this pattern is determined by the average number of results considered, in our case by the average number of goals per match in a championship. A typical example would be described by the diagram below.

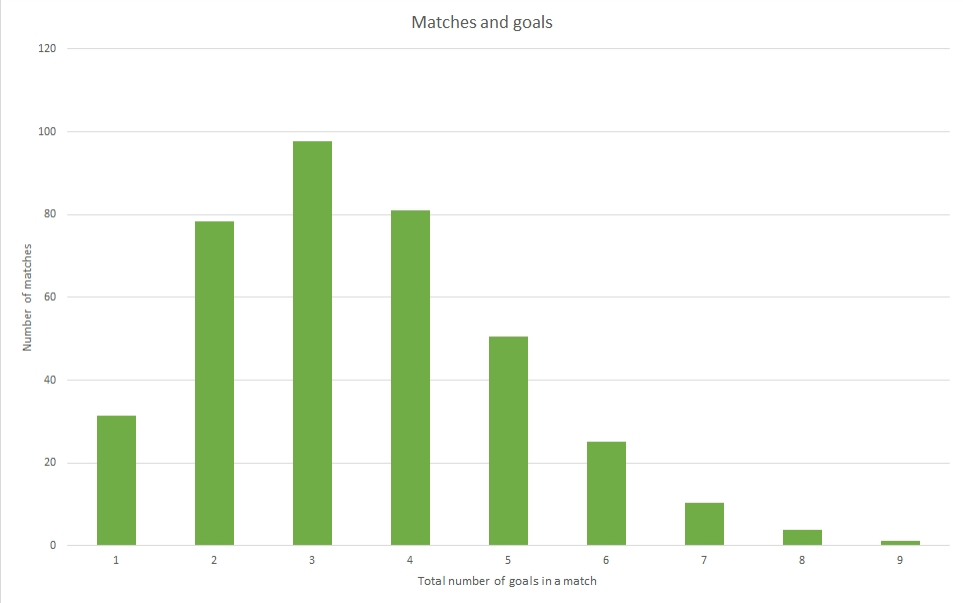


⊡ The French Ligue 1 has 20 teams. Each pair plays twice, so in a whole season there are _____ matches. In season 2014/15, each of the results appeared as many times as indicated by the table below. Fill in the correct numbers in the last two columns (legend: R = result; F = frequency (number of matches with given result); N = number of goals in a match; T = total number of matches with N goals).

R	F	R	F	R	F	R	F	R	F	R	F	R	F	R	F	R	F	N	T
0:0	34																		
1:0	50	0:1	31																
2:0	34	1:1	43	0:2	18														
3:0	15	2:1	37	1:2	22	0:3	5												
4:0	4	3:1	16	2:2	7	1:3	11	0:4	4										
5:0	2	4:1	5	3:2	10	2:3	11	1:4	2	0:5	2								
6:0	1	5:1	2	4:2	3	3:3	4	2:4	1	1:5	1	0:6	0						
7:0	0	6:1	1	5:2	0	4:3	0	3:4	0	2:5	0	1:6	1	0:7	0				
8:0	0	7:1	0	6:2	1	5:3	0	4:4	0	3:5	1	2:6	0	1:7	0	0:8	0		
9:0	0	8:1	0	7:2	0	6:3	0	5:4	0	4:5	0	3:6	0	2:7	1	1:8	0	0:9	0

⚀ In the whole Ligue 1 2014/15 season there were scored _____ in total. Thus, the average number of goals per match was ⚽=_____.

⚁ Below you can see the bar-chart which would be calculated for Ligue 1 2014/15 from ⚽ using the Poisson distribution (the calculated number of matches with 0 goals would be 31, those with 1 goal 78, etc.). Next to the printed bars draw the bars corresponding to true data from the table above and compare with the calculated results!



⚽⚽⚽ Conclusions ⚽⚽⚽

For repeated independent experiments, e.g. football matches, that can have 0, 1, 2, ... successes, e.g. goals, the distribution of numbers of experiments (matches) with a given number of successes (goals) is often more or less Poisson distribution. A Poisson distribution is determined by just one parameter, and that is the average number of successes per experiment (goals per match). ⚽⚽⚽

Introduction to prediction of results

Introduction. If the average number of events of a specific sorts (e.g., goals per match) is m and if you assume that the Poisson distribution is appropriate for your situation, then the probability of a trial with a specific number of events (e.g., a match with a specific number of goals) is given by

Nr. of events	0	1	2	3
Probability	$\frac{1}{e^m}$	$\frac{m}{e^m}$	$\frac{m^2}{2e^m}$	$\frac{m^3}{6e^m}$
Approx. prob.	$0{,}3679^m$	$m \cdot 0{,}3679^m$	$m^2 \cdot 0{,}1839^m$	$m^3 \cdot 0{,}0613^m$
Nr. of events	4	5	6	...
Probability	$\frac{m^4}{24e^m}$	$\frac{m^5}{120e^m}$	$\frac{m^6}{720e^m}$	...
Approx. prob.	$m^4 \cdot 0{,}0153^m$	$m^5 \cdot 0{,}003^m$	$m^6 \cdot 0{,}005^m$	...

In the table above, e denotes a mathematical constant with approximate value 2,718. To simplify your calculations we have entered a third line with approximate formulas for the probabilities.

⊡ Can you detect the general rule for the Poisson probability of k events in a trial?

⚀ For predicting football results, i.e. determining the betting coefficients, betting houses use variations of Poisson distribution with estimates of m based on previous data and other data available. Here you will see the essential idea of the approach. Below you can find all the results from Ligue 1 in season 2014/15.

Home : Away	BAS	BOR	CAE	EVI	GUI	RCL	LIL	LOR	OL	OM	MET	ASM	MHS	NAN	NIC	PSG	REI	REN	STE	TFC
Bastia		0–0	1–1	1–2	0–0	1–1	2–1	0–2	0–0	3–3	2–0	1–3	2–0	0–0	2–1	4–2	1–2	2–0	1–0	1–0
Bordeaux	1–1		1–1	2–1	1–1	2–1	1–0	3–2	0–5	1–0	1–1	4–1	2–1	2–1	1–2	3–2	1–1	2–1	1–0	2–1
Caen	1–1	1–2		3–2	0–2	4–1	0–1	2–1	3–0	1–2	0–0	0–3	1–1	1–2	2–3	0–2	4–1	0–1	1–0	2–0
Evian	1–2	0–1	0–3		2–0	2–1	0–1	1–0	2–3	1–3	3–0	1–3	1–0	0–2	1–0	0–0	2–3	1–1	1–2	1–0
Guingamp	1–0	2–1	5–1	1–1		2–0	0–1	3–2	1–3	0–1	0–1	1–0	0–2	0–1	2–7	1–0	2–0	0–1	0–2	2–1
Lens	1–1	1–2	0–0	0–2	0–1		1–1	0–0	0–2	0–4	2–0	0–3	0–1	1–0	2–0	1–3	4–2	0–1	0–1	1–0
Lille	1–0	2–0	1–0	1–0	1–2	3–1		2–0	2–1	0–4	0–0	0–1	0–0	2–0	0–0	1–1	3–1	3–0	1–1	3–0
Lorient	2–0	0–0	2–1	0–2	4–0	1–0	1–0		1–1	1–1	3–1	0–1	0–0	1–2	0–0	1–2	0–1	0–3	0–1	0–1
Lyon	2–0	1–1	3–0	2–0	3–1	0–1	3–0	4–0		1–0	2–0	2–1	5–1	1–0	1–2	1–1	2–1	2–0	2–2	3–0
Marseille	3–0	3–1	2–3	1–0	2–1	2–1	2–1	3–5	0–0		3–1	2–1	0–2	2–0	4–0	2–3	2–2	3–0	2–1	2–0
Metz	3–1	0–0	3–2	1–2	0–2	3–1	1–4	0–4	2–1	0–2		0–1	2–3	1–1	0–0	2–3	3–0	0–0	2–3	3–2
Monaco	3–0	0–0	2–2	2–0	1–0	2–0	1–1	1–2	0–0	1–0	2–0		0–0	1–0	0–1	0–0	1–1	1–1	1–1	4–1
Montpellier	3–1	0–1	1–0	2–0	2–1	3–3	1–2	1–0	1–5	2–1	2–0	0–1		4–0	2–1	1–2	3–1	0–0	0–2	2–0
Nantes	0–2	2–1	1–2	2–1	1–0	1–0	1–1	1–1	1–1	1–0	0–0	0–1	1–0		2–1	0–2	1–1	1–1	0–0	1–2
Nice	0–1	1–3	1–1	2–2	1–2	2–1	1–0	3–1	1–3	2–1	1–0	0–1	1–1	0–0		1–3	0–0	1–2	0–0	3–2
Paris Saint-Germain	2–0	3–0	2–2	4–2	6–0	4–1	6–1	3–1	1–1	2–0	3–1	1–1	0–0	2–1	1–0		3–2	1–0	5–0	3–1
Reims	2–1	1–0	0–2	3–2	2–3	0–0	2–0	1–3	2–4	0–5	0–0	1–3	1–0	3–1	0–1	2–2		1–0	1–2	2–0
Rennes	0–1	1–1	1–4	6–2	1–0	2–0	2–0	1–0	0–1	1–1	1–0	2–0	0–4	0–0	2–1	1–1	1–3		0–0	0–3
Saint-Étienne	1–0	1–1	1–0	3–0	2–1	3–3	2–0	2–0	3–0	2–2	1–0	1–1	1–0	1–0	5–0	0–1	3–1	0–0		0–1
Toulouse	1–1	2–1	3–3	1–0	1–1	0–2	3–2	2–3	2–1	1–6	3–0	0–2	1–0	1–1	2–3	1–1	1–0	2–1	1–1	

Each team played 38 matches. Choose a team: _____. Fill in the data on how many matches your team has scored 0, 1, 2, ... goals:

Goals scored	0	1	2	3	4	5	6	7
In ... matches								
Relative frequency								

In total, your team has scored _____ goals in 38 matches, which gives an average of $m =$_____ per match. Remembering that decimal betting coefficients are reciprocals of probabilities, use the formulas for the Poisson distribution to determine the probabilities and the decimal coefficients for your team scoring a specific number of goals:

Goals scored	0	1	2	3	4	5	6	7
Probability								
Decimal coefficients								

Compare your results to the true ones by drawing simultaneous bar charts for the obtained true relative frequencies and calculated probabilities!