Received: 16 March 2022 | Revised: 6 July 2022 | Accepted: 9 July 2022

DOI: 10.1111/test.12315

ORIGINAL ARTICLE

WILEY

UEFA EURO 2020: An exciting match between football and probability

Giulia Fedrizzi[1] | **Luisa Canal**[2] | **Rocco Micciolo**[2,3]

[1]EPSRC Centre for Doctoral Training in Fluid Dynamics, University of Leeds, Leeds, UK

[2]Department of Psychology and Cognitive Sciences, University of Trento, Rovereto, Italy

[3]Centre for Medical Sciences, University of Trento, Trento, Italy

Correspondence
Rocco Micciolo, Department of Psychology and Cognitive Sciences, University of Trento, Corso Bettini, 31 - 38068 Rovereto, Italy.
Email: rocco.micciolo@unitn.it

Abstract

Football, as one of the most popular sports, can provide exciting examples to motivate students learning statistics. In this paper, we analyzed the number of goals scored in the UEFA EURO 2020 final phase as well as the waiting times between goals, considering censored times. Such a dataset allows us to consider some aspects of count data taught at an introductory level (such as the Poisson distribution), as well as more advanced topics (such as survival analysis taking into account the presence of censored times). Employing data from the final phase of UEFA EURO 2020, depending on the course level, the student will acquire knowledge and understanding of a range of key topics and analytical techniques in statistics, develop knowledge of the theoretical assumption underlying them and learn the skills needed to model count data.

KEYWORDS

exponential model, football match results, goal waiting times, Poisson distribution, teaching statistics

1 | INTRODUCTION

Association football, simply known as football or soccer, is one of the world's most popular sports, attracting millions of spectators and involving thousands of players of all genders.

Statistics in the world of football has been widely used to calculate a number of descriptive indicators. For example, one can retrieve information on passing accuracy, possession, free-kicks or corners taken, offsides, and so on. On the other hand, probability is more widely employed in the field of betting, as football is a game that is well suited to different types of gambling.

In this paper, we want to show how UEFA EURO 2020 can be an exciting example to introduce students to count data and count processes. In a basic-level course, a teacher can present the simplest count model, that is, the Poisson process evaluating the number of goals scored in a given time period. According to this model, there is the assumption that events (goals in our case) occur at random times with a constant average rate, which can be estimated from data. In an introductory course, the comparison between the observed and expected frequencies of goals in a match could be limited at a descriptive level. On the other hand, in a more advanced course, formal goodness of fit tests can be performed; in the case at hand, both the widely-employed chi-square test and a more specific test can be presented.

Furthermore, UEFA EURO 2020 provides more data than just the number of goals per match: it provides the times between goals, which in the case of a Poisson process are independent and exponentially distributed.

For teaching, this means the opportunity to introduce the exponential distribution and, depending on the student cohort, the non-parametric estimate of the survival function.

However, this paper is not intended to be a "how-to-use" statistics for football models, nor for football models implying attacking/defensive abilities or for statistical predictions. In considering the Poisson, we consider a match as the statistical time period for the Poisson, where the outcome is the *total* number of goals scored in a match. In considering times between goals in matches, we allow for time censoring.

At a more advanced level, this dataset can be used as a nice example to illustrate statistical techniques such as weighted regression or generalized linear models for count data. When employing these data, depending on the course level, the student will acquire knowledge and understanding of a range of key topics and analytical techniques in statistics, develop knowledge of the theoretical assumption underlying them and learn the skills required to model count data.

2 | DATA SOURCE

The final phase of the 60th edition of the European Football Championship, called UEFA EURO 2020, took place between June 11, 2021 and July 11, 2021. A total of 51 matches were played: 36 in the *group stage* and 15 in the *knockout stage* (round of 16, quarter-finals, semi-finals and final). All matches in the group stage lasted 90 minutes, while, for matches in the knockout phase, if a match was level at the end of normal playing time, extra time was played for a total of 120 minutes (90 minutes of normal playing time plus 30 minutes of extra time).

The times at which the goals were scored were retrieved from the UEFA website (https://www.uefa.com/uefaeuro/history/seasons/2020/). Goals scored in the extra time of the first half of a match were considered to be scored at time $t = 45$ minutes, while goals scored in the extra time of the second half of a match were considered to be scored at time $t = 90$ minutes.

Data were analyzed employing R, version 4.1.1 [15].

3 | ANALYSIS OF THE NUMBER OF GOALS SCORED IN THE UEFA EURO 2020 FINAL PHASE

3.1 | Normal playing time

A key assumption of the Poisson model is that the number of events needs to be independent observations in the same process over a selected time period. In the group stage, the 24 admitted teams were divided into six groups of four teams; within each group, six games were played. Therefore, under the assumption of the same Poisson process, we would expect the same total number of goals in each group. When considering the number of goals per group stage match, the following totals were observed: 16, 22, 8, 15, 14 and 19.

A discussion could be stimulated as to whether these data can be considered as observations of the same process. Depending on the students' background and study level, possible answers can be given by remaining at a descriptive level or by resorting to inferential procedures. For advanced courses, where generalized linear models are taught, an ANOVA for Poisson observations can be performed. In addition, if students have experience with R and simulations, one can propose to the classroom a simulation to evaluate how frequent the observed 14-goal difference (or more) can be under the assumption of the same Poisson process. In our case, neither of the approaches gave sufficient evidence to reject this assumption.

As a result, we can now consider the frequency distribution of the number of goals per match observed in the 51 matches played both in the group and knockout phases, considering only normal playing time (Table 1). A total of 135 goals were scored, so the average number of goals per match was 2.65. The teacher may propose that students also calculate the variance of the number of goals, finding a value (2.43) quite similar to the mean; such a result does not support any evidence against the possibility that the data can follow a Poisson distribution. Therefore, it is possible to calculate the probabilities of observing x goals employing the Poisson distribution with a mean μ equal to the observed one (2.65) for x ranging between 0 and 6. The sum of these probabilities is approximately 0.981 so the teacher can discuss the necessity to consider also more extreme cases than those observed (ie, six goals in a match). These probabilities are shown in the third column of Table 1, where the last probability refers to the case of scoring 6 or more goals.

TABLE 1 Observed and expected frequencies of scores for UEFA EURO 2020 final phase (normal playing time only)

No. of goals	Obs.	*P*	Exp.
0	3	.071	3.61
1	9	.188	9.57
2	15	.248	12.66
3	10	. 219	11.17
4	7	.145	7.39
5	4	.077	3.91
6+	3	.053	2.68

The expected frequencies can be calculated by multiplying the probabilities by 51 (the total number of matches played) and are shown in the last column of Table 1.

From a descriptive point of view, observed and expected frequencies are in good agreement. Depending on the students' background, a formal goodness of fit test can be calculated; for example, the chi-square statistic for the data shown in Table 1 is 0.753 (with 5 degrees of freedom), which gives no evidence against the hypothesis of a Poisson distribution.

At this point, the teacher must strongly emphasize that such a result does not mean that data support Poisson; depending on the students' background, there is also an opportunity for discussing the chi-square test as an omnibus test for evaluating the goodness of fit. There is another test, more specific for the Poisson distribution, which relies on the equality between the mean and the variance. In particular, Fisher [8] discussed "the special test for discrepancy of the variance," that is, the *dispersion index*. This index is used to test for homogeneity of the observations and is also referred to as the variance test for homogeneity [1].

Let $y_1, y_2, \ldots, y_n$ denote n observations from a Poisson distribution; the dispersion index is defined as $D = \sum_{i=1}^{n} (y_i - \bar{y})^2 / \bar{y}$, where $\bar{y}$ is the arithmetic mean, and it is approximately distributed as a χ^2 with $(n-1)$ degrees of freedom. In our case, where $n = 51$ and $\bar{y} = 2.65$, the dispersion index is $D = 45.96$ (with 50 degrees of freedom) with an associated *P-value* of 0.636, which gives no evidence in these data to reject the Poisson model.

3.2 | Normal and extra playing time

So far, we have only considered the goals scored during normal playing time. However, in the final phase of UEFA EURO 2020, there were eight matches in which extra time was necessary. In this section, we consider the number of goals scored in each match irrespective of its duration (90 or 120 minutes).

Table 2 considers all of the 51 matches, both in the group stage and in the knockout phase. The frequency distribution of the observed number of goals per match is shown separately for the 43 matches played without the extra time and for the 8 matches played with extra time. The corresponding expected frequencies were calculated assuming a Poisson distribution with the same parameter μ, where μ is the number of goals scored per minute. The estimate of this parameter ($\hat{\mu} = 0.0294$) is the ratio between the total number of goals scored (142) and the total number of minutes played (4830). The expected frequencies corresponding to x goals for the 43 matches played without extra time were calculated employing a Poisson distribution with a mean of 2.65 (ie, 0.0294×90) and are reported in the fourth column of Table 2; those for the 8 matches that ended after extra time was calculated employing a mean of 3.53 (ie, 0.0294×120) and are reported in the seventh column of Table 2. The last two columns of Table 2 show observed and expected frequencies of the number of goals in the 51 matches. Frequencies are in good agreement again. Depending on the students' background, a chi-square test for the goodness of fit can be calculated; for the data that are shown in Table 2 this test yields (after grouping the last three frequencies) a value of 1.032 (with 5 degrees of freedom), which gives no evidence against the hypothesis of a Poisson distribution.

In more advanced courses, where the theory of generalized linear model is taught, the goodness of fit of a Poisson distribution can be evaluated by fitting a model, specifying the family as Poisson and the link as "log". The number of goals is the dependent variable, while the natural logarithm of the time length of the match is specified as an offset. In our case, the deviance of the fitted model was 47 with 50 degrees of

TABLE 2 Observed and expected frequencies of goals for UEFA EURO 2020 final phase

	Matches lasting 90 minutes			Matches lasting 120 minutes			All matches	
Goals	Obs.	*P*	Exp.	Obs.	*P*	Exp.	Obs.	Exp.
0	2	.071	3.05	0	.029	0.23	2	3.29
1	9	.188	8.07	0	.104	0.83	9	8.90
2	10	.248	10.68	3	.183	1.46	13	12.14
3	10	.219	9.42	3	.215	1.72	13	11.14
4	7	.145	6.23	0	.190	1.52	7	7.75
5	4	.077	3.30	0	.134	1.07	4	4.37
6	1	.034	1.45	1	.079	0.63	2	2.08
7	0	.013	0.55	0	.040	0.32	0	0.87
8+	0	.006	0.25	1	.028	0.22	1	0.48

freedom, therefore showing no evidence against the Poisson model.

The Poisson is the simplest distribution for modeling count data. Advanced courses can consider cases where overdispersion is present in the data, and the negative binomial or the Waring distribution can be employed [2]. However, in UEFA EURO 2020 the variance of the number of goals was lower than the mean. In doctoral courses, a model based on the generalized Poisson distribution [5,11] can be discussed. Such a model can accommodate both overdispersed and underdispersed count data through a parameter δ, which can assume positive (overdispersion), null (equidispersion) or negative (underdispersion) values. In the case of UEFA EURO 2020, the estimate of this parameter was -0.055, with an associated 95% confidence interval between -0.25 and 0.15 and a *P-value* for testing the hypothesis $\delta = 0$ of 0.299, giving no evidence against the more parsimonious Poisson model.

4 | ANALYSIS OF THE TIMES BETWEEN GOALS

In addition to the number of goals in a match, we can also consider times between goals. In a Poisson process, these are independent and exponentially distributed with probability density function $f(t) = \lambda e^{-\lambda t}$, where t is the "time to event" and λ is the reciprocal of the mean time to event. In this case, the survival probability $P(T > t)$ is given by $S(t) = e^{-\lambda t}$. This relationship can be employed to evaluate the goodness of fit of an exponential distribution. Applying logarithms to $S(t) = e^{-\lambda t}$, we get $-\log(S(t)) = \lambda t$, and a plot of the estimated values of $-\log(S(t))$ against time t will exhibit a linear trend (with a slope equal to λ).

In our case, the "time to event" of a goal (in minutes) is the waiting time between two successive goals, or between the start of the match and the first goal. A distinctive characteristic of survival data is that the event of interest may not be observed in every statistical unit. This feature is known as censoring. Censoring can arise because of time limits and other restrictions depending on the nature of the studies. In our case, if a game ends without goals scored, it is not possible to measure the time elapsed between the start of the match and the first goal. Therefore, we define the measured time, as 90 or 120 minutes, as a censored time. Similarly, if the last goal was scored after for example, 75 minutes and there was no extra time, we consider a censored time of 15 minutes. In the case of UEFA EURO 2020, we observed a total number of 142 goals during a total of 4830 minutes considering both normal playing and extra time. Therefore, the average rate is 0.0294 goal/minute which is the estimate of the parameter λ of the exponential distribution. The reciprocal of this value (ie, 34.0) measures how many minutes elapse, on average, between one goal and the next one.

When teaching survival analysis, after introducing both waiting and censored times, it is possible to illustrate the nonparametric product limit estimator [12] and estimate the survival curve. In our case, this curve is shown in Figure 1.

When interpreting the results shown in Figure 1, we have to remember that the statistical unit of this analysis is the time between two successive goals (or the censored time associated with each match, if any). In our analysis, we have a total of 188 observations (ie, times): 142 are waiting times, while the remaining 46 are censored times (there were five matches with a goal scored just at the end of the match). The survival curve shown in Figure 1 gives the estimated cumulative probability of observing no goal after having waited t minutes from the previous goal. This probability is 1 when $t = 0$ and decreases monotonically as time goes on.

The lowest waiting time was 1 minute, which was observed during the quarter-final match between Belgium and Italy: Belgium scored a goal at the end of the first play time (45 minutes), one minute after a goal scored by Italy. The highest observed waiting time was 95 minutes, which occurred during the round-of-16 match between Italy and Austria: the first goal was scored by Italy in the extra time, 95 minutes after kick-off. The estimated median time was 24 minutes.

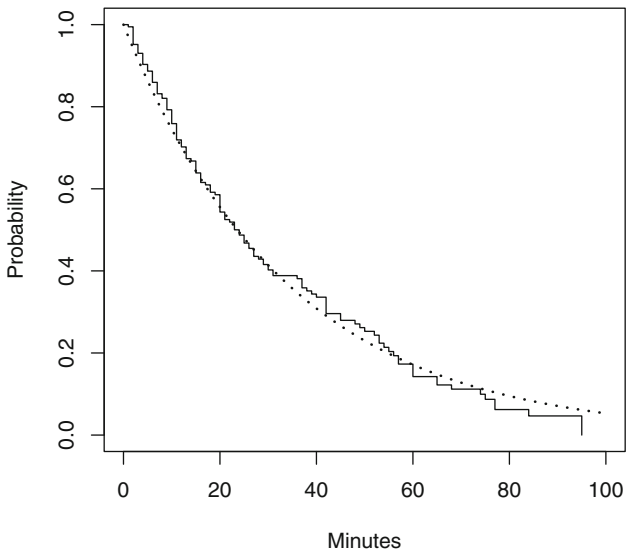

FIGURE 1 Kaplan–Meier survival probability associated with the waiting times of the goals scored during the final phase of UEFA EURO 2020 (solid line) and survival probability of an exponential random variable with $\lambda = 0.0294$ (dotted line)

Therefore, the estimated probability that one had to wait at least 24 minutes to observe the next goal is 0.5. Considering an average rate of goal scoring of 0.0294 goal/min as a value for λ, we can plot the expected exponential survival curve against the empirical one (Figure 1). There is a good agreement between expected and observed results. The median survival time is $-\log(0.5)/\lambda$, that is, 23.6 minutes in our case, which is in good agreement with the observed value (24 minutes). Furthermore, the regression line is shown in Figure 2, where the estimated values of $-\log(S(t))$ are plotted against time t, has an estimated slope of 0.0331 goal/min, very similar to the observed average rate of goal scoring (0.0294). When proposing this approach, there is a good opportunity for discussing the assumptions underlying the ordinary least squares, wondering, in particular, if the homoscedasticity can be considered valid. The number of observations decreases with time, causing the precision of the estimates of $S(t)$ to decrease with time. Therefore, a better estimate of the slope can be obtained by employing the weighted least squares, with weights inversely proportional to the variance associated with each point. In our case, the estimated slope is 0.0300, which is closer to the expected one (0.0294).

For a Poisson process, the times between events are exponentially distributed and independent. Above we have investigated the exponentially assumption. As a step to explore independence, we can plot times to the next event against the previous time between events, but this is not feasible for each match. Instead, we can consider the 51 matches as a sequence [4]. In this case, the analysis takes into account the cumulated times of goal scoring (ie, from the first goal, scored after 53 minutes, to the last goal, scored after 4777 minutes after the first kick-off). In the case of a Poisson process, plotting each time vs the previous time results in a straight line of a unitary slope with an intercept given by the average time between one goal and the next. One may wonder how to handle the order of the two matches played simultaneously, within each of the six groups, on the last day of the corresponding group stage. In a Poisson process, this result does not depend on the chosen order. One can check this result with UEFA EURO 2020 data by generating the 64 possible different time sequences. The above could also be used to analyze times between goals without considering censored times (and a match as a statistical time period).

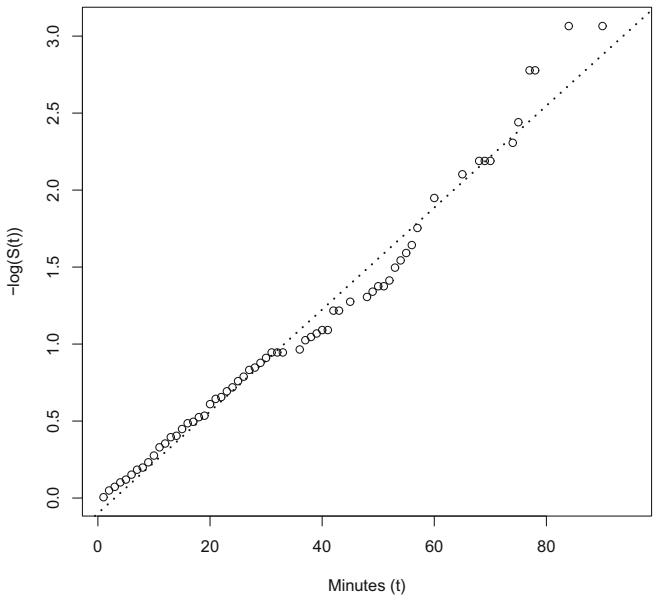

FIGURE 2 Plot of the estimated values of $-\log(S(t))$ against time t for the waiting times of the goals scored during the final phase of UEFA EURO 2020 (circles) and of the corresponding regression line

5 | DISCUSSION AND CONCLUSIONS

UEFA football data can be employed as an engaging context to consider some aspects of count data at the introductory level as well as at a more advanced level. If we look at the distribution of the number of goals per match, a simple model like the Poisson, which is the first count process introduced to students, appears in good agreement with the data.

However, from a teaching point of view, it must be emphasized that even if there is no evidence in these data to reject the Poisson model, this does not mean the data support Poisson. Therefore, one cannot claim that the 51 matches of UEFA EURO 2020 are 51 independent homogeneous Poisson processes with a constant probability per unit of time of scoring a goal (which does not vary from match to match) since football matches are a system of highly co-operative intercorrelated entities.

The good fit of the exponential distribution when modeling the waiting times between goals is expected from a theoretical point of view but could be considered somewhat counter-intuitive and unrealistic given the memoryless property of the exponential model. In fact, Figure 2 reports some small, but consistent, deviations from the exponential model. However, it should be considered that the measurement of time may be inaccurate since actual playing times are not recorded and censored times are also present.

We remark that UEFA EURO 2020 data are presented as a *fil rouge* in a "journey" through some key topics and analytical techniques in statistics. Our analyses are not

intended to be good modeling of football data. The outcome of a match can be thought of as depending on a number of factors such as the ability of the teams (particularly in terms of attacking/defensive abilities), the availability of more or less strong and trained players, and the tactics defined by the coaches, etc. Being able to model the outcome of a match with a probabilistic model might seem a very difficult task, given the complexity of the game and the variety of factors involved. Many authors have focused on models for predicting the outcome of a match (home win, draw, away win) in national leagues taking into account an entire season over a rather long time span (several months or years) [17,20]. For example, goal-based team performance covariates were used by Goddard [10] to forecast win-draw-lose match results. Koopman and Lit [13] developed a statistical model for the analysis and forecasting of football match results employing time series analysis with intensity coefficients that change stochastically over time. Modeling FIFA World Cup football data, van der Wurp et al [19], employed copula regression to include dependency into models, obtaining match outcome probabilities as well. Rue and Salvesen [16] suggested a Bayesian dynamic generalized linear model to estimate the time-dependent skills of all teams in a league. In addition, several authors have focused their attention on the football betting market [3,6,7].

A fundamental assumption of the exponential model and of the resulting Poisson model is the concept of independence between events. Independence can be thought of as a useful way to model outcomes (like the result of tossing a coin or of a game of chance) that are not inherently random but can be predicted exactly (at least in principle) once the relevant parameters are specified. However, in various situations in the real world, people often discard independence, suffering from cognitive illusions [18]; one famous case can be found in basketball and is known as the "hot hand" [9] and, more recently, also in football ("hot shoe") [14].

The good fit of the Poisson model can be at least partly explained by considering some features of the case study: the teams participating in the final phase came from a selection phase, so their skills were relatively homogeneous; the time span in which the final phase of the competition took place was short, which prevents teams from greatly improving their skills; no distinction between home or away games can be made; the abilities of the two teams were averaged by considering only the total number of goals scored in each match rather than counting them separately.

In this paper, we hope to have shown how data from the final phase of UEFA EURO 2020 can be considered an engaging context to be used in teaching statistics both at an introductory level as well as at more advanced levels.

ACKNOWLEDGMENT

Open Access Funding provided by Università degli Studi di Trento within the CRUI-CARE Agreement.

ORCID

Giulia Fedrizzi https://orcid.org/0000-0003-1610-0068
Luisa Canal https://orcid.org/0000-0002-1493-108X
Rocco Micciolo https://orcid.org/0000-0002-8299-9879

REFERENCES

1. A. Agresti, *Foundations of Linear and Generalized Linear Models*, John Wiley & Sons Inc., Hoboken, New Jersey, 2015.
2. L. Canal and R. Micciolo, *Probabilistic models to analyze psychiatric contacts*, Epidemiol. Psychiatr. Sci. **8** (1999), 47–55.
3. M. Chalikias, E. Kossieri, and P. Lalou, *Football matches: Decision making in betting*, Teach. Stat. **42** (2020), 4–9.
4. S. Chu-Chun-Lin, *Rendezvous of the Poisson and exponential distributions at the World Cup of Soccer*, Teach. Stat. **21** (1999), 60–62.
5. P. C. Consul and G. C. Jain, *A generalization of the Poisson distribution*, Technometrics **15** (1973), 791–799.
6. M. J. Dixon and S. G. Coles, *Modelling Association football scores and inefficiencies in the football betting market*, Appl. Stat. **46** (1997), 265–280.
7. M. J. Dixon and M. E. Robinson, *A birth process model for association football matches*, Statistician **47** (1998), 523–538.
8. R. A. Fisher, *The significance of deviations from expectations in a Poisson series*, Biometrics **6** (1950), 17–24.
9. T. Gilovich, R. Vallone, and A. Tversky, *The hot hand in basketball: On the misperception of random sequences*, Cogn. Psychol. **17** (1985), 295–314.
10. J. Goddard, *Regression models for forecasting goals and match results in association football*, Int. J. Forecast. **21** (2005), 331–340.
11. T. Harris, Z. Yang, and J. W. Hardin, *Modeling underdispersed count data with generalized Poisson regression*, Stata J. **12** (2012), 736–747.
12. E. L. Kaplan and P. Meier, *Nonparametric estimation from incomplete observations*, J. Am. Stat. Assoc. **53** (1958), 457–481.
13. S. J. Koopman and R. Lit, *A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League*, J. R. Stat. Soc. A **178** (2015), 167–186.
14. M. Ötting and A. Groll, *A regularized hidden Markov model for analyzing the 'hot shoe' in football*, Stat. Model. (2021): 1471082X211008014. [First Published May 19, 2021]. https://doi.org/10.1177/1471082X211008014.
15. R Core Team. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. 2021. https://www.R-project.org/
16. H. Rue and O. Salvesen, *Prediction and retrospective analysis of soccer matches in a league*, Statistician **49** (2000), 339–418.
17. E. F. Saraiva, A. K. Suzuki, C. A. O. Filho, and F. Louzada, *Predicting football scores via Poisson regression model: applications to the National Football League*, Commun. Stat. Appl. Methods **23** (2016), 297–319.
18. A. Tversky and D. Kahneman, *Judgment under uncertainty: heuristics and biases*, Science **185** (1974), 1124–1131.

19. H. van der Wurp, A. Groll, T. Kneib, G. Marra, and R. Radice, *Generalised joint regression for count data: a penalty extension for competitive settings*, Stat. Comput. **20** (2020), 1419–1432.
20. E. Wheatcroft, *Forecasting football matches by predicting match statistics*, J. Sports Anal. **7** (2021), 77–97.

How to cite this article: G. Fedrizzi, L. Canal, and R. Micciolo, *UEFA EURO 2020: An exciting match between football and probability*, Teach. Stat. **44** (2022), 119–125. https://doi.org/10.1111/test.12315