Original Article

Learning the principles of simulation using the birthday problem

Rui Manuel da Costa Martins
Escola Superior de Saúde Egas Moniz, Centro de Investigação Interdisciplinar Egas Moniz (CiiEM), Almada, Portugal
e-mail: ruimartins@egasmoniz.edu.pt

Summary Using the famous Birthday problem, we present here a practical activity that allows students to perceive the basic reasoning behind simulation and explore its potential. Through a playful approach with probabilities, students are led along a path that illustrates difficulties with intuition and introduces them to theoretical results for sample proportions.

Keywords: Teaching statistics; Simulation; Birthday paradox; Football; Sample proportion; Law of large numbers.

INTRODUCTION

Students love games and hands-on discovery and simulation facilitates engagement in these while illustrating results that may be non-intuitive as well as general theory such as the Law of Large Numbers.

Simulation has an enormous preponderance in modern statistics, and the advantages of simulation in teaching statistics have long been known. In one of its first numbers, this journal published articles alluding precisely to that. Thomas and Moore (1980) stated that 'The introduction of the computer into the school classroom has brought a new technique to teaching, the technique of simulation'. Zieffler and Garfield (2007) and Tintle et al. (2015) discuss the role and the importance of simulation-based learning on the undergraduate statistics curriculum. However, others (e.g. Hodgson and Burke 2000) discuss some problems that may arise while teaching a subject by simulation, namely, the development of some misconceptions in students' mind.

The activity discussed here is the well-known and much-publicized birthday problem (see, e.g. Falk 2014). Here, we follow the example of Matthews and Stones (1998) in considering two football (soccer) teams and hence birthday coincidences in 22 players. A major positive outcome of this activity is the discussion that will naturally arise among students, with the teacher acting as a mediator.

THE ACTIVITY

To ensure that the activity achieves its purposes, one must first present the problem to the class and promote a discussion without giving the solution.

The problem: *In a football (soccer) game, what is the probability that at least two of the 22 players have the same birthday?*

The context of the problem is purposely chosen because football is very popular in Portugal and because the resulting probabilities are not intuitive. It is absolutely critical before discussion to state that we will assume that all 365 days are equally likely for any one birthday and that we assume the players' birthdays occur independently of each other.

At the beginning of the discussion, most of the hunches are well below 0.476, the correct value. The answers can range from 0.01% to 10%. The usual explanation students give relates the number of players to the number of days in a year, 22/365 = 0.06. However, their intuition is very inconsistent. Occasionally, when a colleague says he knows someone, a relative or a friend, whose birthday is on the same day, the hunches, like by magic, start to increase and can go to values close to 90%, but never 100%.

If we ask students to predict how many players we should have on the pitch to be more than 50% likely that any two players have the same birthday, even though it is a slightly different

problem, they state something like half of 365. This was already noted by Falk (2014).

Moreover, they rarely think of the number of pairs that can be made with 22 people (231 in this case). When, in some classes, I suggest this number, the students immediately start giving totally different answers. Initially, some respond 231/365, but others say 0.5 or close to that.

After the initial period of discussion, students should simulate and explore data. If then desired, students can try to analytically solve the question (see, e.g. Falk 2014).

This is a practical activity that relies on computers. As advice, I would say that it is mandatory to provide a 'roadmap' to the students with instructions on how to use the software, otherwise the activity might be counterproductive. Meanwhile, we should introduce the idea and reasoning behind statistical simulation. It is advantageous for me to make an analogy with the idea of taking balls out of an urn with replacement, because it is the mental sketch that most of my students had since high school. The difference is that we use a random numbers generator as a surrogate of the urn. Any computational tool with simple functions for generating (pseudo) random numbers can be used, for instance, R. For example, in R, if we want to obtain a random number between 1 and 10, we simply do

```
sample(1:10,1,replace=TRUE).
```

For obtaining specific birthdays for the game, all we need is to sample, with replacement, from the set A={1,…,365}, ignoring leap years and twins, and store the values in a variable, e.g. 'birthdays':

```
birthdays=sample(1:365,22,replace=TRUE).
```

Each student simulates several, e.g. 30, birthday sets of size 22 (11 players each team) and count how many have at least two repeated elements, which are the common birthdays. At this stage of my course, they are not familiarized with complex procedures, such as loops or non-trivial functions. Usually, their counting strategy is somewhat naïve, because they are beginning to learn R. Generally, they take note on paper about the number of sets that are being simulated and have repetitions. To help a little with counting, we can consider tabulating the birthdays for each simulated dataset,

```
table(birthdays),
```

to obtain the frequencies and identify coincidences. It is part of discovery for every student to follow his/her own strategy for counting. To make the experience a little bit more realistic, we can convert the simulated birthdays to the exact day of the current year, by doing

```
curtyearday=as.Date(birthdays,ori-
gin="2017-01-01")
table(curtyearday).
```

Students then share and compare their obtained percentages (the relative frequencies) for the number of sets with birthday repetitions. Generally, their initial beliefs about the problem tend to be updated towards a value closer to the correct answer (0.476). It is instructive to plot the students' observed percentages, as a dotplot, histogram or stem-and-leaf, for the class to see for themselves both the variation of the percentages and what value tends to be in the 'centre' of the graph. The students should compute the class average, e.g. for a class of 25, by considering a vector of length 25. For example,

```
pctes=c(0.37,0.23,0.63,0.57,0.57,0.67,
0.43,0.33,0.43,0.57,0.50,0.57,0.47,0.53,
0.50,0.43,0.43,0.50,0.63,0.60,0.47,0.50,
0.40, 0.43,0.60)
p.hat=sum(pctes)/25.
```

We expect the class average (in this case 0.49) to be a closer estimate of the theoretical answer than most of the individual estimates.

Generally, at this point, the ideas underpinning estimation and the LLN start emerging. Some students realize that if they had repeated the experiment a few more times, probably they would end up with an estimate even closer to the theoretical value (and we must encourage them to do so). Sometimes, a student says, 'what we are doing here is calculating the average number of successes. If we increase the sample we can get a better result …'. Generally, almost all are aware of the frequentist definition of probability and, therefore, most of them comprehend that, in the limit, they will end up with a value for the relative frequency of successes (obtaining a set with repeated birthdays is considered the success) very close to the theoretical answer. The name – Law of large numbers – can be introduced (or reinforced if students have seen it) here. If desired, students can simulate larger samples and explore the graph of class observed percentages, and

these could also be used, if appropriate to a course, to explore the central limit theorem for sample proportions.

After reaching a consensus about an answer to the problem, some students perceive that 'their intuition was cheating them'. We are then in a position to move from the empirical (simulated) to the analytical solution.

ON THE SIMULATION

The simulation procedure here described is fairly basic, requiring very little technical background, but is entirely justified by my students' study program (undergraduate in health sciences). In the case of our institution, there is only a semester on a quantitative subject – Biostatistics.

Comparisons of the simulated results obtained by each student causes them to update their intuition and generally in the right direction. They begin noticing a certain pattern, a certain value around which those frequencies seem to vary. Even when their reasoning is based only on the 30 individual simulations, some start to realize that the probability must be higher than they thought, because they are getting too many sets with birthday repetitions! However, some say that they must be doing something wrong or the simulation is not working well because they consider that they are getting a higher percentage of simulated sets containing repetitions than expected according to their initial beliefs, which predicted a low probability.

STUDENTS' REACTION TO SIMULATION

For most of our students, this is the first time they hear about statistical simulation. Sometimes, it happens that some of them do not become fully convinced of the possibility of solving this problem of probabilities by using computer simulation. Once, I had a student who said 'how do I know that the computer is not cheating?' In that situation, as commented earlier, we should explain that, in simple terms, statistical simulation via a computer is the same as having a human being taking balls out of an urn (with or without replacement as appropriate), who does not get tired!

Some of the most interesting student comments over the years reveal the students' surprise with the potentialities of the simulation and at same time their trust in the process: 'we can use statistical simulation for almost anything that we do not know but we intend to come to know!'; 'Simulation encourages discovery'; 'The use of statistical software is much more motivating than calculating by hand!'; 'Will classes always be like this?'

These statements provide further illustration of the potential that a simulation activity can have in the attitude of health science students and the way to motivate them to appreciate statistics. Yet, in the same class, some are attached to a pen-paper-based behaviour. They claim they like calculations because they are used to and feel comfortable with them. One of the reasons frequently verbalized has to do with the assessment system. If exams do not pay any attention to simulation, why should they have any interest in self-discovery activities! This is an interesting point, because we cannot introduce new ways of learning while keeping the assessment system unchanged.

CONCLUSION

Simulation is an integral and very important feature of modern statistics. There are several scenarios where its role is paramount. It is a very effective learning tool to help students acquire a conceptual, and not purely mechanical, understanding of a subject.

In addition, we can introduce concepts like distribution of the sample proportion and CLT.

Our practical classes last for 2 h and, usually, the task can be accomplished in about 45 min (simulation plus the analytical part), so it is not too time consuming and is a good investment of time. After the exercise, everyone has gained an idea of what statistical simulation is. In case you try this, I hope your students enjoy it as much as mine do!

Acknowledgements

The author thanks the Editor and the reviewers whose insightful comments led to a substantially improved presentation of this work.

References

Falk, R. (2014). A closer look at the notorious birthday coincidences. *TEST*, **36**(2), 41–46. https://doi.org/10.1111/test.12014.

Hodgson, T. and Burke, M. (2000). On simulation and the teaching of statistics. *Teaching Statistics*, **22**(3), 91–96. https://doi.org/10.1111/1467-9639.00033.

Matthews, R. and Stones, F. (1998). Coincidences: The truth is out there. *Teaching Statistics*, **20**(1), 17–19. https://doi.org/10.1111/j.1467-9639.1998.tb00752.x.

Thomas, F.H. and Moore, J.L. (1980). CUSUM: Computer simulation for statistics teaching. *Teaching Statistics*, **2**(1), 23–28. https://doi.org/10.1111/j.1467-9639.1980.tb00374.x.

Tintle, N., Chance, B., Cobb, G., Roy, S., Swanson, T. and VanderStoep, J. (2015). Combating anti-statistical thinking using simulation-based methods throughout the undergraduate curriculum. *The American Statistician*, **69**(4), 362–370. https://doi.org/10.1080/00031305.2015.1081619.

Zieffler, A. and Garfield, J. (2007). Studying the role of simulation in developing students' statistical reasoning. *Bulletin of the International Statistical Institute 56th Session*, Lisbon.