http://doi.org/10.14393/ER-v30a2023-47

Explaining the frequentist interpretation of probability with simulations in R software[1]

Felipe Rafael Ribeiro Melo[2]

ABSTRACT

The frequentist interpretation of probability may not be so clear to students when they are only presented to texts and formulas. In this context, the professor can use ludic strategies in order to the students to absorb such content in a more enlightening way. Through scripts created by the author in the R language, an activity was performed out with students of the discipline of probability from the Production Engineering course at a university, in which they could visualize this concept through simulations of draws and graphical and tabular representations. Each person present in the classroom simulated 100 replications of the same random experiment, recording the relative frequency of occurrence of the event of interest after 100 replications. The average of these relative frequencies resulted in a value very close to the probability of the chosen event (obtained via classical interpretation), making the frequentist interpretation more enlightening for students.

KEYWORDS: Probability. Frequentist interpretation. R language. Playful activity.

[1] English version by Felipe Rafael Ribeiro Melo. *E-mail:* felipe.ribeiro@uniriotec.br.

[2] PhD in Statistics. Federal University of the State of Rio de Janeiro, Rio de Janeiro, RJ, Brazil. Orcid: https://orcid.org/0000-0002-1482-8533. *E-mail*: felipe.ribeiro@uniriotec.br.

http://doi.org/10.14393/ER-v30a2023-47

Explicando a interpretação frequentista da probabilidade com simulações no software R

RESUMO

A interpretação frequentista da probabilidade pode não ser tão clara aos discentes quando são apresentados apenas a textos e fórmulas. Neste contexto, o professor pode usar estratégias lúdicas a fim de os discentes absorverem tal conteúdo de maneira mais esclarecedora. Por meio de *scripts* criados pelo autor na linguagem R, foi realizada uma atividade com alunos da disciplina de probabilidade do curso de Engenharia de Produção em uma universidade, na qual puderam visualizar este conceito por meio de simulações de sorteios e representações gráficas e tabulares. Cada pessoa presente em sala de aula simulou 100 replicações de um mesmo experimento aleatório, com registro da frequência relativa de ocorrência do evento de interesse após 100 replicações. A média destas frequências relativas resultou em valor bem próximo à probabilidade do evento escolhido (obtida via interpretação clássica), tornado a interpretação frequentista mais clara para os discentes.

PALAVRAS-CHAVE: Probabilidade. Interpretação frequentista. Linguagem R. Atividade lúdica.

Explicando la interpretación frecuentista de probabilidad con simulaciones en software R

RESUMEN

La interpretación frecuentista de la probabilidad puede no ser tan clara para los estudiantes cuando solo se les presentan textos y fórmulas. En este contexto, el docente puede utilizar estrategias lúdicas para que los estudiantes absorban dichos contenidos de una manera más esclarecedora. A través de guiones creados por el autor en lenguaje R, se realizó una actividad con estudiantes de la disciplina de probabilidad del curso de Ingeniería de Producción de una universidad, en la que pudieron visualizar este concepto a través de simulaciones de sorteos y representaciones gráficas y tabulares. Cada persona presente en el aula simuló 100 repeticiones del mismo experimento aleatorio, registrando la frecuencia relativa de ocurrencia del evento de interés después de 100 repeticiones. El promedio de estas frecuencias relativas dio como resultado un valor muy cercano a la probabilidad del evento elegido (obtenido mediante interpretación clásica), lo que hace que la interpretación frecuentista sea más esclarecedora para los estudiantes.

PALABRAS CLAVE: Probabilidad. Interpretación frecuentista. Lenguaje R. Actividad lúdica.

* * *

Man is nothing in itself. He is only his own possibility.
But he is, nonetheless, the responsible infinite.
Albert Camus

Introduction

Since the earliest civilizations, humanity has grappled with the concepts of chance and uncertainty. The use of probability to measure uncertainty and variability dates back hundreds of years. According to DEGROOT & SCHERVISH (2012, p.1), probability theory has been constantly evolving since the 17th century, believed to have been initiated by French mathematicians Blaise Pascal (1623-1662) and Pierre Fermat (1601-1665) when they were able to derive exact probabilities for certain gambling problems involving dice. However, the practice of gambling dates back many centuries.

> "By about the year 3500 b.c., games of chance played with bone objects that could be considered precursors of dice were apparently highly developed in Egypt and elsewhere. Cubical dice with markings virtually identical to those on modern dice have been found in Egyptian tombs dating from 2000 b.c. We know that gambling with dice has been popular ever since that time and played an important part in the early development of probability theory." (DEGROOT & SCHERVISH, 2012, p.1).

Today, the theory of probability is an important tool in most areas of engineering, science, and administration. Because of this, several university courses include a probability course (or a probability and statistics course combined), in which concepts such as probability calculus, random variables, and probability distributions are explored. This paper, in particular, focuses on the frequentist interpretation of probability, providing a classroom experience report of an activity based on drawing simulations in the R software (R CORE TEAM, 2023) to illustrate this interpretation. It shows numerical and graphical outputs in a class of the Bachelor of Production Engineering program at a university in the state of Rio de Janeiro. A playful and, at times, enjoyable way to explore the frequentist interpretation of probability through drawing simulations using the R software. It's worth noting that the frequentist interpretation of probability establishes that, for a "large number" of replications of a random experiment, the probability of an event is reasonably well approximated by the relative frequency of the event occurring in these many replications.

There are several studies on the use of playfulness in childhood learning, but there is little discussion about its application in the university teaching and learning process (REIS; SILVA; DEMO, 2020, p.717). Learning solely through texts and formulas in higher education can become a lengthy process, especially for students with greater difficulties in mathematics, which often originates from high school or even elementary school. A simulation activity presented by the professor, using free and open-source software, and with active student participation, as a supplementary activity, tends to make the understanding of the concept faster and more enjoyable.

In addition to this Introduction, this paper is divided into six more sections. The first one introduces some important concepts that should be covered in a probability course before addressing the frequentist interpretation of probability, such as random experiments, sample space, and events. Next, a section discusses the classical and frequentist interpretations of probability, with a greater focus on the latter, along with a brief discussion of convergence in probability. The following two sections feature the R programming language: first, a section explains what the R software is, the integrated development environment RStudio, and the concept of scripts to be compiled in the R language. This is followed by a section dedicated to the two scripts created by the author for the classroom activity. The results section provides a comprehensive account of the application of the activity in the classroom, including graphical output and a table with relative frequencies of interest obtained through simulation. Finally, the conclusion section raises the discussion about the executed activity, its positive aspects, and relevant reflections regarding the frequentist interpretation of probability and the generation of simulations using a programming language.

Preliminary concepts in a probability course

From a didactic perspective, it is expected that a probability course at the undergraduate level introduces the concepts of random experiment, sample space, and events before formally defining probability. This approach is taken because probability is a function applied to events, which are subsets of a sample space, and the sample space is associated with a random experiment.

An experiment whose outcome cannot be predicted with certainty is called a random experiment. Tossing a coin, rolling a die, drawing an element from a group of elements, the score of your favorite team's next game, and the maximum temperature that will be recorded in your city tomorrow are some examples of random experiments. Despite the impossibility of predicting the outcome of a random experiment with certainty, it is possible to define the set of all its possible outcomes. This set is called the sample space of this random experiment and is often denoted by Ω, with its elements commonly referred to as sample points. Finally, any subset of a sample space is an event from that sample space. Events are typically denoted by uppercase letters from the Latin alphabet, often starting at the beginning of the alphabet. In particular, if Ω has n elements (that is, n sample points), then there are 2^n different events in this sample space, including the empty set (representing an impossible event) and Ω itself (called a certain event).

To illustrate the concepts discussed in the paragraph above, consider the (random) experiment of rolling a die and determining which face will land facing up. Its sample space is Ω = {1, 2, 3, 4, 5, 6} and here are some examples of events from this sample space: A = {2, 4, 6}, B = {5, 6} and C = {3}. While event A represents "the face facing up is an even number", event B consists of "the face facing up is greater than 4", and event C represents "the face facing up is 3". It is said that an event occurs if and only if the outcome of the random experiment is contained within the event. In the example above, when it is said that event A has occurred, then the outcome of the random experiment was 2, 4, or 6. The reciprocal is also true: if the outcome was 2, 4, or 6, it is said that event A has occurred.

Based on the definition of events, it becomes reasonable to begin presenting the probability measure in its various approaches, in addition to its axiomatic definition. In fact, from a probability perspective, the primary motivation for defining events is to obtain their respective probabilities of occurrence. For any event A within a sample space Ω, the probability of event A occurring will be denoted as P(A). In other words, probability is a real function P applied to events whose range is the interval [0;1].

The frequentist interpretation of probability

Before studying into the frequentist interpretation of probability, it is common to discuss the classical interpretation of probability. Quite popular due to its ease of application, the classical interpretation is only valid when it is assumed that all sample points are equally likely. Under this condition, for any event A in a sample space Ω, the probability of event A occurring is given by the ratio of the number of elements in A to the number of elements in Ω, which is:

$$P(A) = \#A \,/\, \#\Omega.$$

The numerator of this fraction is commonly referred to as the "number of favorable cases for the event", while the denominator is referred to as the "number of possible cases".

While the classical interpretation of probability is practical, it cannot be extended to any sample space. In this more general scenario, the so-called frequentist interpretation of probability is more interesting since it applies to any sample space. In this broader context, let's assume that n replications (under similar conditions) of a random experiment are carried out - with sample space Ω - and the interest lies in obtaining the probability of the occurrence of an event A in Ω. Denoting by n_A the number of times that event A occurred in these n replications (i.e., the absolute frequency of event A's occurrence), and by $f_A = n_A/n$ the relative frequency of event A's occurrence in these n replications, the frequentist interpretation of probability establishes that, if n is sufficiently large, f_A is a good approximation for P(A), i.e.,

$$P(A) \approx f_A.$$

In other words, f_A converges in some probabilistic sense to P(A) when $n \to \infty$.

There are some criticisms regarding this type of interpretation. Firstly, it provides only an approximation. Furthermore, it suggests that the number of replications should be "sufficiently large", but there is no definite indication of a real number that would be considered large enough. Moreover, the similar conditions under which the experiments should be replicated are not clear. Considering a random experiment of coin tossing (which may not be fair),

> “...it is stated that the coin should be tossed each time "under similar conditions", but these conditions are not described precisely. The conditions under which the coin is tossed must not be completely identical for each toss because the outcomes would then be the same, and there would be either all heads or all tails. In fact, a skilled person can toss a coin into the air repeatedly and catch it in such a way that a head is obtained on almost every toss. Hence, the tosses must not be completely controlled but must have some “random” features. Furthermore, it is stated that the relative frequency of heads should be “approximately 1/2,” but no limit is specified for the permissible variation from 1/2. If a coin were tossed 1,000,000 times, we would not expect to obtain exactly 500,000 heads.” (DEGROOT & SCHERVISH, 2012, p.3).

It is important to clarify the asymptotic behavior that establishes the convergence of f_A to P(A) when $n \to \infty$. This convergence occurs in a probabilistic sense and is not a convergence process as commonly seen in courses on Differential and Integral Calculus for deterministic functions. In this latter type of convergence, from a certain point k, there exists an interval [a; b], which depends on k and contains the limit of the function, such that for every $n > k$, the function always lies within this interval [a; b]. However, it is not possible to guarantee, with 100% probability, that f_A always belongs to the same interval [a; b] for $n > k$. What we can guarantee is that, for any $\varepsilon > 0$, the larger n becomes, the more likely the difference between f_A and P(A) is smaller than ε. In other words, f_A converges in probability to P(A) as $n \to \infty$. For further details, see MEYER (1975, p.285).

Still within the scope of convergence in probability, consider the realization of n replications of the same random experiment and the subsequent calculation of f_A not just once, but m times, independently of each other. Let $n_{A,j}$ and $f_{A,j}$ be the absolute and relative frequencies of event A at the end of the j-th sequence of n replications (disregarding previous sequences). As the Law of Large Numbers establishes, the expression

$$(1/m)(f_{A,1} + f_{A,2} + \ldots + f_{A,m})$$

converges in probability to P(A) when $m \to \infty$, since the terms of the sequence $\{n_{A,j}\}$ are independent and identically distributed random variables following a binomial distribution with parameters n and $p = P(A)$, and therefore the expected value of each $f_{A,j}$ is the probability of the occurrence of event A:

$$E[f_{A,j}] = E[n_{A,j}/n] = (1/n) \times n \times P(A) = P(A).$$

To illustrate, suppose a class of 50 students, each with a fair die. Each of these students will roll their die 100 times and record how many times they get the 6 face, so it’s possible to calculate the relative frequency of the "6 face" result in the 100 rolls for each of the 50 students. With that said, the arithmetic mean of the 50 observed relative frequencies tends to be close to $1/6 \approx 0.1667$.

The R software and the RStudio IDE

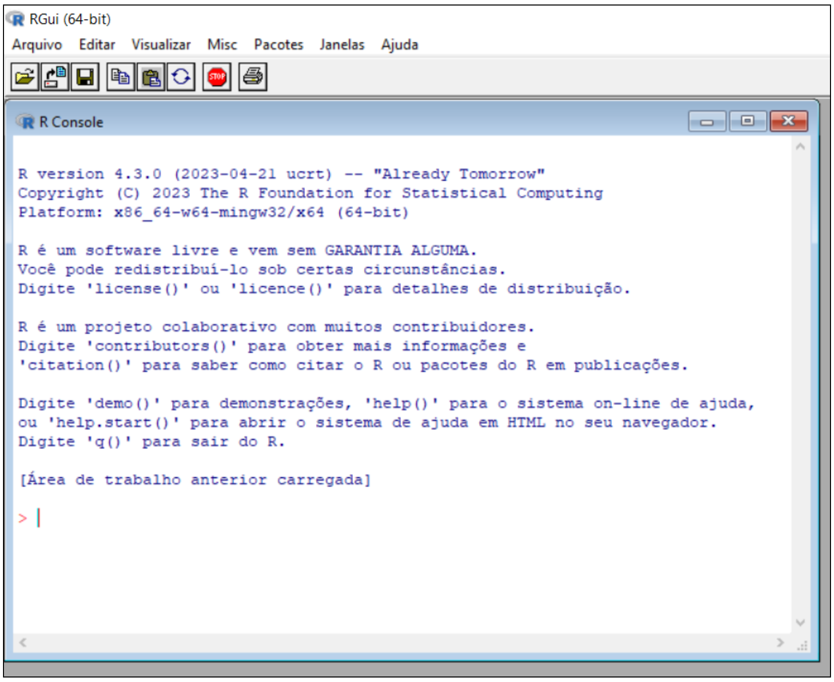
The following paragraphs provide some clarifications on the R software/programming language, the RStudio integrated development environment, and what scripts are from the perspective of the R language. In particular, this section is primarily aimed at readers interested in simulation activities to explain the frequentist interpretation of probability but who have little to no knowledge about R, RStudio, and script creation/compilation in a programming language.

The R software is an open-source program commonly used for data processing and statistical analysis, but not limited to these purposes. It can also be thought of as a programming language, particularly an object-oriented language. The R language is script-based, and its default interface is designed to accept command lines, despite the existence of some packages that provide a "point and click" interface with certain limitations (SONDEREGGER, 2018, p.9). Increasingly prevalent in academic circles, R software has a history of more than 20 years and, despite its unattractive interface, offers numerous advantages. It is a free, powerful, and stable program available for Windows, Linux, and Mac. It is supported by a large team of developers worldwide and has a wealth of available packages that provide specific functionalities. Furthermore, cutting-edge methodologies are initially developed in R (and made available in the form of packages), and there are numerous free materials, tutorials, and discussion forums available on the Internet.

The R software can be downloaded for free at https://cloud.r-project.org/. For the Windows operating system, a faster method is to access https://cloud.r-project.org/bin/windows/base/. After downloading and completing the installation process, simply open the newly installed program to view its command line interface, as shown in Figure 1: a window called R Console, within a window named R Gui.

In Figure 1, please note the red ">" symbol. It is called the command prompt. It indicates that R is ready to receive a command line (which can be as simple as 2+3 or a more extensive and complex command line). In this R Console window, simply press Enter after typing the command line to compile it, and the result will be displayed. Give it a try by typing 2+3, press Enter, and disregard the number 1 enclosed in square brackets on the left side of the result.

FIGURE 1: The default interface of the R software.



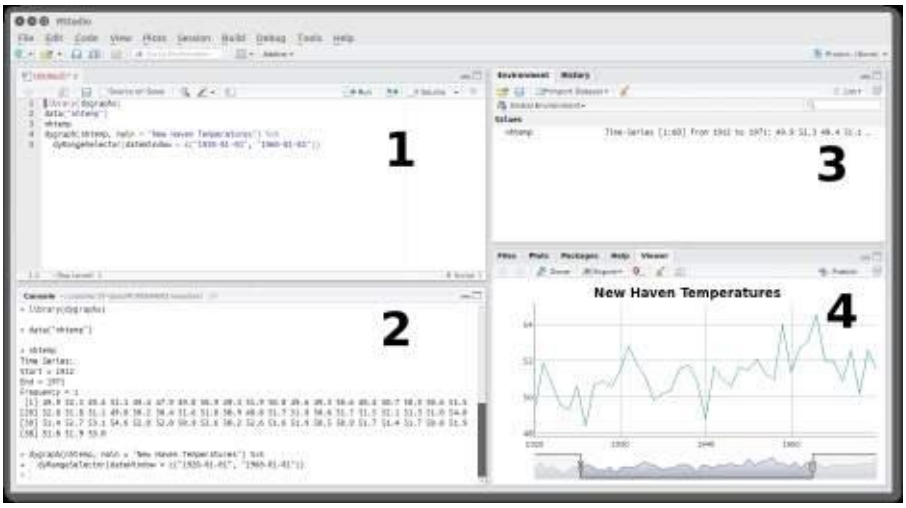
```
R version 4.3.0 (2023-04-21 ucrt) -- "Already Tomorrow"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R é um software livre e vem sem GARANTIA ALGUMA.
Você pode redistribuí-lo sob certas circunstâncias.
Digite 'license()' ou 'licence()' para detalhes de distribuição.

R é um projeto colaborativo com muitos contribuidores.
Digite 'contributors()' para obter mais informações e
'citation()' para saber como citar o R ou pacotes do R em publicações.

Digite 'demo()' para demonstrações, 'help()' para o sistema on-line de ajuda,
ou 'help.start()' para abrir o sistema de ajuda em HTML no seu navegador.
Digite 'q()' para sair do R.

[Área de trabalho anterior carregada]

> |
```

Source: Self-authorship.

When using R as a simple calculator, entering the desired command line directly in the R Console window (and pressing Enter to generate the result) is quite convenient. However, when desired outputs require the compilation of multiple command lines, it is appropriate to write them in a text file and then compile these lines all at once (or one by one). This type of file is called a script.

An facilitator in the creation and compilation of scripts for the R language is the RStudio program, which, like R, has versions for Windows, Linux, and Mac operating systems and can be downloaded for free at https://posit.co/download/rstudio-desktop/. Scripts created in RStudio are saved as files with the extension ".R". RStudio is a graphical interface with various features that enhance the use and learning of R, making daily work much more convenient (OLIVEIRA; GUERRA; MCDONELL, 2018, p.10). It should be noted that the installation and use of RStudio to compile scripts in the R language only makes sense after the prior installation of the R software. By default, the RStudio interface is divided into four windows, as illustrated in Figure 2. In the first window (code editor), you can write, edit, save, and load scripts. Below it is the Console, where script command lines are compiled by clicking on the Run options (which compile only the current line or selected lines) or Source (which compiles the entire script at once). On the right side, the upper window holds all objects created in the R session under the Environment tab, and the History tab creates a history of used commands (OLIVEIRA; GUERRA; MCDONELL, 2018, p.10). Finally, the last window has several tabs, with the most relevant one for the purposes of this paper being the Plots tab, where graphical outputs resulting from the compilation of command lines are displayed.

FIGURE 2: The default interface of the RStudio integrated development environment.



```
library(dygraphs)
data("nhtemp")
nhtemp
dygraph(nhtemp, main = "New Haven Temperatures") %>%
  dyRangeSelector(dateWindow = c("1920-01-01", "1960-01-01"))
```

```
> library(dygraphs)

> data("nhtemp")

> nhtemp
Time Series:
Start = 1912
End = 1971
Frequency = 1
[illegible]

> dygraph(nhtemp, main = "New Haven Temperatures") %>%
+   dyRangeSelector(dateWindow = c("1920-01-01", "1960-01-01"))
>
```

```
nhtemp    Time-Series [1:60] from 1912 to 1971: 49.9 52.3 49.4 51.1 ...
```

Source: Oliveira; Guerra; Mcdonell (2018, p.11).

Scripts to illustrate the frequentist interpretation of probability

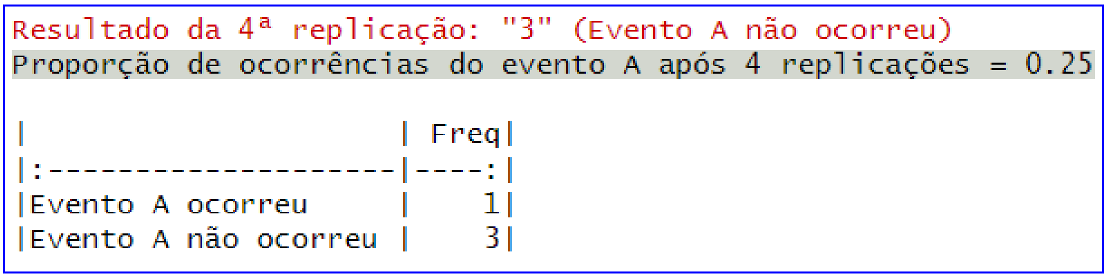
This section concerns to elucidate the two scripts created by the author in the R language to generate simulations that illustrate the frequentist interpretation of probability as a supplementary method in explaining this concept. Both, when compiled, produce frequency distribution tables in the Console window and graphical representations in the Plots tab of the lower right window in RStudio. These graphical representations, in particular, were inspired by and follow the same concept as discussed in PINHEIRO et al. (2012, p.7, Figure 1.1). To facilitate the understanding of the following paragraphs, here are the names given to these two script files:

- conceito_frequentista_2023-1.R;
- conceito_frequentista_sobrepostas_2023-1.R.

It is important to highlight that both scripts make use of functions from two R packages: crayon (CSÁRDI, 2022) and knitr (XIE, 2023). Therefore, before compiling either of these scripts, it is necessary to install these packages. A practical way to do this without the need for command line input is to access the menu Tools > Install Packages in RStudio. The "Install from" field should be filled with "Repository (CRAN)", and in the field below, the names of the packages to be installed should be entered, separated by commas or spaces. Just like the installations of R and RStudio, package installation is required only once.

When you open the file "conceito_frequentista_2023-1.R" in RStudio, you can see, in the mentioned "Window 1" in Figure 2, the definition of the number of replications of the random experiment, the sample space associated with this random experiment, the probabilities associated with each sample point, the event of interest and its probability of occurrence, encapsulated in the objects n, Omega, probs, A, and p, respectively. Subsequently, the graph's color is defined, and some objects that are part of the routine that concludes this script are created. This routine, after each of the n replications of the random experiment, generates: a table in the Console window, computing how many times event A occurred and how many times it did not occur after k replications, for k = 1, 2, ..., n, along with the result of the k-th replication; and a graph with the number of replications on the horizontal axis and the relative frequency of event A (up to the k-th replication) on the vertical axis. The k-th generated graph displays k points, each representing the total replications up to the moment considered on the horizontal axis and, on the vertical axis, the relative frequency of event A occurring up to the moment considered. To facilitate the visualization of the graph's evolution as the experiment is replicated, immediate neighboring points are connected by line segments. Throughout this paper, these graphs are frequently referred to as trajectories. To illustrate a simulation of this kind, consider the random experiment of rolling a biased die and observing the face that lands facing up. The probability of each face's occurrence is proportional to the number it represents: in other words, for i = 1, ..., 6, j = 1, ..., 6, and j > i, face j is j/i times more likely to occur than face i. In this scenario, the probability of face j occurring is equal to j/21, for j = 1, 2, 3, 4, 5, 6. But suppose this probabilistic structure of the die is unknown, and you want to simulate 100 rolls to obtain an approximation of the probability that the face obtained after a roll is an even number, using the script "conceito_frequentista_2023-1.R". Therefore, the event of interest is A = {2, 4, 6}. Assume that after the first 4 rolls, the results were: 3, 5, 6, and 3. Right after this fourth roll, a table is generated in the Console window, as shown in Figure 3, and a graph as in Figure 4 in the lower right window of RStudio. Since the first four rolls resulted in a sequence of two odd numbers, one even number, and one odd number, the relative frequencies of event A occurring after the 1st, 2nd, 3rd, and 4th roll are, respectively: 0/1 = 0, 0/2 = 0, 1/3 ≈ 0.3333 and 1/4 = 0.25. The four points in Figure 4 are the ordered pairs (1,0), (2,0), (3,1/3) and (4,1/4). In particular, the previously installed packages carry functions that focus on the formatting of the 100 tables generated in the Console window, as illustrated in Figure 3. The red font and gray background in the expressions above the table are due to the red and bgWhite functions (from the crayon package), respectively. The formatting of the table itself comes from the kable function of the knitr package."

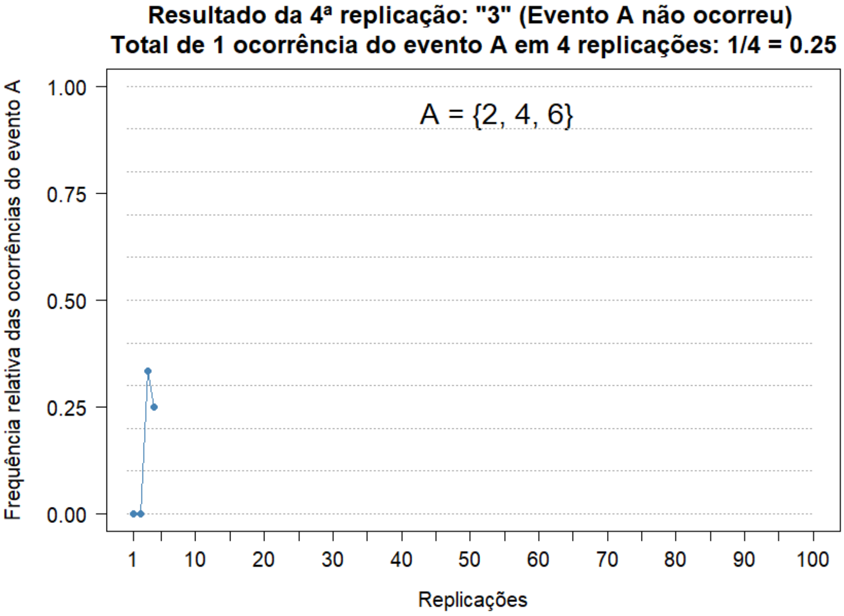
FIGURE 3: One of the tables generated from compiling the "conceito_frequentista_2023-1.R" script.



```
Resultado da 4ª replicação: "3" (Evento A não ocorreu)
Proporção de ocorrências do evento A após 4 replicações = 0.25

|                     | Freq|
|:--------------------|----:|
|Evento A ocorreu     |    1|
|Evento A não ocorreu |    3|
```

Source: Self-authorship.

FIGURE 4: One of the graphs generated from compiling the "conceito_frequentista_2023-1.R" script.

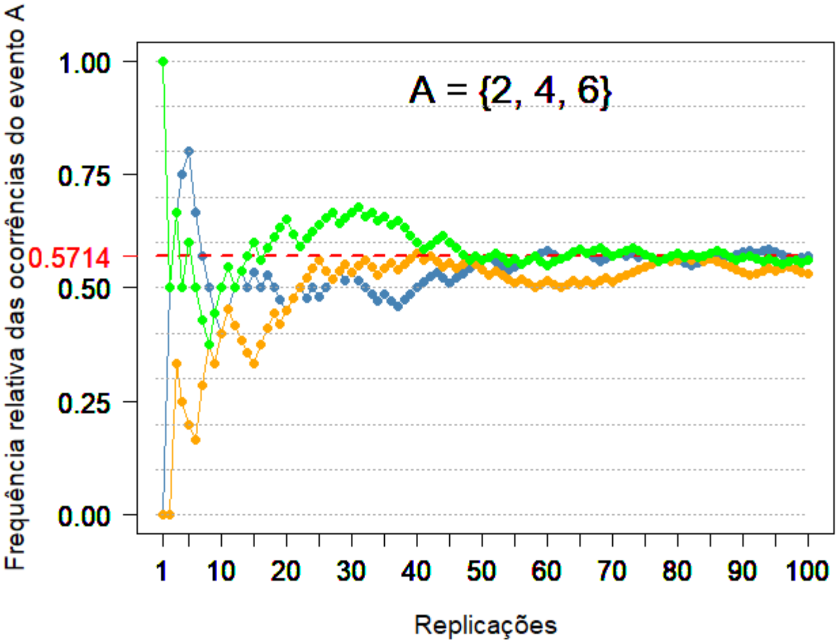


Source: Self-authorship.

Clearly, the primary interest lies in the relative frequency of event A occurring after 100 replications. When compiling the entire script, as explained in the previous section, the 100 graphs are stored in the lower right window of RStudio under the "Plots" tab. The last of these 100 graphs will be displayed, with the option to access any of the previous 99 graphs by clicking on the blue arrow "Previous plot" (or its keyboard shortcut Ctrl+Alt+F11). In particular, this final graph highlights the exact probability p of event A occurring, marked in red on the vertical axis, and a dashed line segment, also in red, from (0, p) to (n, p), reinforcing the idea of the convergence of fA to p – even though it is a convergence in probability. From a didactic perspective, a good teaching strategy for displaying these graphs to students is to return to the first generated graph (i.e., the one generated after the first replication) and use the blue arrow “Next plot” (or its keyboard shortcut Ctrl+Alt+F12) to show the evolution of the graph, replication after replication. In practice, emphasizing the graphs associated with the first and the last replications is the most important, in order not to consume too much class time and avoid making the activity tiresome or monotonous.

The script named “conceito_frequentista_sobrepostas_2023-1.R” closely resembles the previous one, but only the figure for the nth replication is generated. However, it is possible to execute a new sequence of n replications and represent it in the graph with a new color without removing the previous trajectory: simply modify the color in line 12 of the script, compile this line, and recompile all the lines below it. In this regard, it is possible to show that for each sequence of n replications of the same random experiment, the trajectories of the relative frequencies of the same event A occurring up to n replications are not the same, and at the end of the n replications, the values of f_A for the new n replications and the previous n replications are not necessarily equal. The application of this routine in the classroom should record the value obtained for f_A at the end of each simulation of n replications and verify that it may be above or below P(A) at times, and occasionally it may equal P(A). However, the arithmetic mean of the records of f_A at the end of each simulation of n replications approaches P(A) with high probability, considering a sufficiently large number of trajectories. For illustration, Figure 5 shows only three trajectories, with n = 100 replications each, and the same event A = {2, 4, 6} from the previous example, with a probability of P(A) = (2+4+6)/21 = 12/21 ≈ 0.5714. The observed values for f_A after 100 replications were 0.57 (blue trajectory), 0.53 (orange trajectory), and 0.56 (green trajectory), whose arithmetic mean returns approximately 0.5533, which is not necessarily expected to be close to 0.5714 due to the small number of simulations of 100 replications (only three replications).

FIGURE 5: Graph generated from the compilation of the "conceito_frequentista_sobrepostas_2023-1.R" script.



Source: Self-authorship.

Results: application of the activity

In the final minutes of the class that covered the classical and frequentist interpretations of probability, the professor of the probability course for the Production Engineering program at a university in Rio de Janeiro, represented by the author of this paper, presented and explored both scripts created by him. Initially, the "conceito_frequentista_2023-1.R" script was explored, simulating 100 rolls of a fair die with an interest in obtaining the relative frequency of event A = {2, 4, 6} occurring after these 100 rolls. According to the classical interpretation of probability, it is known that P(A) = 0.50. However, it is not guaranteed that f_A = 0.50 after 100 replications, but it is expected that this relative frequency will be close to 0.50. After several sequences of 100 replications, the students could realize that, in addition to f_A not necessarily being 0.50 after 100 replications, the trajectories from the 1st to the 100th replication are not the same, and the final value of f_A after 100 replications typically varies for each hundred replications.

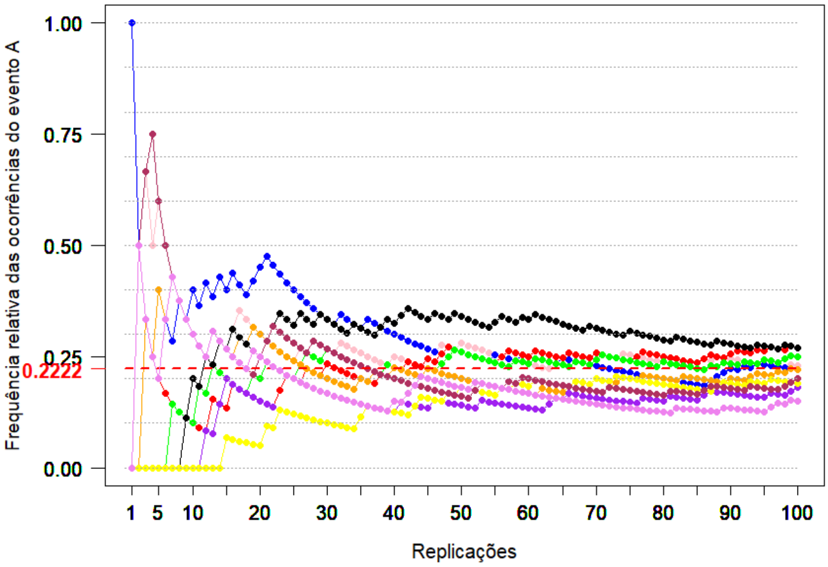
But the most interesting activity was yet to come. Based on the script "conceito_frequentista_sobrepostas_2023-1.R", the professor configured the random experiment as the selection of one of the 9 students present in the class, and the number of replications of this random experiment as n = 100. Therefore, the sample space was set as a set of nine elements, where each element is the name of one of the students. The event of interest considered was "student aged at least 24 years old", defined after a brief informal conversation between the professor and the students. This event contains only two elements, meaning exactly 2 of the 9 students present are aged 24 years or older. Denoting this event as A, it follows that P(A) = 2/9 ≈ 0.2222. Additionally, it should be noted that in this paper, the names in the sample space and the event of interest are omitted to avoid exposing the students' names.

First, the professor performed 100 replications of the random experiment, choosing the color blue for the trajectory of f_A from the first to the one hundredth replication. These 100 replications can be thought of as 100 draws with replacement of the students present in the class. In these 100 draws, exactly 22 of them resulted in one of the students aged at least 24 years old, and therefore, the observed value of f_A after 100 replications was 0.22. After presenting this result (which can be calculated from the last table generated in the Console window, as illustrated in Figure 3, by dividing the frequency of "Event A occurred" by 100), the professor wrote the result 0.22 on the board. Then, the professor called the student who was closest to him and asked him to choose a different color from the already used blue. The student chose the color red, and the professor changed the color in line 12 of the script (from blue to red) and executed another hundred replications, generating a red-colored trajectory that returned $f_A = 27/100 = 0.27$ after all replications. The professor returned to the board and wrote 0.27 below the 0.22 previously written, and at this point, we have a graph with two trajectories (one blue and one red), in addition to the values 0.22 and 0.27 written on the board by the professor. This procedure was replicated until all nine students present in the room participated in the activity, each choosing a color that had not yet been used. It is worth noting that the R language offers a native set of 657 different colors, whose names can be checked by executing the command line colors() or in freely available materials on the Internet, such as WEI (2021). Returning to the simulations, at the end of the activity, 10 overlapping trajectories were visualized (one with the color initially chosen by the professor and nine with the colors chosen by the students), as shown in Figure 6, in addition to 10 values of f_A (each after 100 replications) recorded in Table 1 in descending order of relative frequencies of event A = {student is at least 24 years old}.

To conclude the activity, the professor calculated the arithmetic mean of the 10 values displayed in the second column of Table 1, which returned 0.218. As expected, this result is close to $2/9 \approx 0.2222$, despite the small number of hundreds of replications – only 10 replications.

http://doi.org/10.14393/ER-v30a2023-47

FIGURE 6: Overlapping trajectories of 10 hundred replications of the random experiment, with the number of replications on the horizontal axis and the relative frequency of the event of interest up to that replication on the vertical axis.



Source: Self-authorship.

TABLE 1: Relative frequencies of the occurrence of the event of interest observed at the end of each hundred replications of the random experiment.

Color of the trajectory	f_A at the end of the trajectory
Black	27/100 = 0,27
Red	27/100 = 0,27
Green	25/100 = 0,25
Pink	23/100 = 0,23
Orange	22/100 = 0,22
Blue	22/100 = 0,22
Brown	20/100 = 0,20
Yellow	19/100 = 0,19
Purple	18/100 = 0,18
Violet	15/100 = 0,15

Source: Self-authorship.

Conclusion

Although the frequentist interpretation of probability was initially presented by the professor to his class through text and formulas, it is expected that the activity of simulating draws using the R software, with the accompanying monitoring of the relative frequency of the event of interest, along with the different trajectories with outcomes that typically change, has made the concept clearer to the students. According to MACHADO & WOJCICKOSKI (2017, p.11), playfulness can be a tool used by the educator to facilitate the teaching-learning process, but before using it, the educator should analyze and study the dynamics to be used. Therefore, the teaching strategy used was to compose the sample space as the set of students present in the classroom and the event of interest as something suggested by these students and easily interpretable for them. During the activity, moments of relaxation were experienced, bringing the essence of playfulness to the classroom in a light and natural way. In particular, for the ten final values of relative frequency for each trajectory (the same values present in Table 1 and discovered one by one), the expectation, led by the professor, for relative frequencies below 2/9 when most of the relative frequencies recorded on the board were above 2/9 was verified - an expectation that was not always fulfilled. At the end of the activity and consequently the end of the class, the professor agreed with his students to send them the figure generated through the virtual learning environment specially used for this course. Such a figure is Figure 6 of this paper. A way to share with the present (and absent) students a record of the activity carried out. In fact, this entire itinerary provided the students in the class with a better understanding of the frequentist concept of probability (compared to the initially presented text and formula), as well as a somewhat more refined understanding of the essence of randomness.

Several further reflections on the conducted activity are provided in this paragraph. Firstly, one might argue why the use of 100 replications for each trajectory. After all, is 100 a sufficiently large number? The choice of this value (instead of a larger value) was made to facilitate the visualization of the trajectories in a local manner and for computational time reasons in generating the results. Nevertheless, around the 5th or 6th generated trajectory, the professor's computer used in the classroom began to experience some slowdown until the completion of the result (a laptop with approximately 9 years of use, with a Core-i5 processor, 8GB of RAM, and a 480GB SSD). In addition, due to only 9 students being present in the class, only 10 trajectories were simulated. With a larger number of students in the class, more trajectories would be generated, and therefore, convergence in probability would be favored. For a positive ε close to zero and an integer value k greater than 10, the probability of the average of the outcomes of the new k trajectories being far from 2/9 by a value smaller than ε is greater than in the scenario with only 10 outcomes. However, the number of trajectories generated proved to be sufficient for the average of the outcomes to be close to 2/9: a difference of only 0.0042. With more class time, the activity could be restarted, generating new 10 trajectories and a new average value, in order to convince more skeptical students that the next 10 outcomes would follow with outcomes close to 2/9, and that it is expected that the average of these values continues to be close to the exact value of the probability of the event of interest. It is worth noting that the created scripts allow for changing the number of replications of the random experiment, as well as the sample space associated with this experiment, the event of interest, and the associated probabilities, making them highly versatile.

Finally, it is worth noting that the area of Data Science is of interest to many students in Engineering, Mathematics, and Computer Science programs as a professional career path. A solid foundation in probability, as well as statistics, is essential for Data Science, including an understanding of asymptotic results such as the Law of Large Numbers. The activity introduced in this work introduces this important probability result.

References

CSÁRDI, G. *crayon: Colored Terminal Output*. R package version 1.5.2, 2022. https://CRAN.R-project.org/package=crayon.

DEGROOT, M. H.; SCHERVISH, M. J. *Probability and Statistics*. Fourth Edition. Boston: Pearson, 2012.

MACHADO, B.; WOJCICKOSKI, V. S. O lúdico no ensino superior: uma proposta de inovação pedagógica. *Revista Múltiplo Saber*, v. 37, n. 1, mar. 2017. Disponível em: https://www.inesul.edu.br/site/revista_eletronica_volume.php?p=1&&vol=47 . Acesso em: 29 jun. 2023.

MEYER. P. L. *Probabilidade*: aplicações a estatística. Tradução Ruy de C. B. Lourenço Filho. Rio de Janeiro: Editora LTC, 1975.

OLIVEIRA, P. F.; GUERRA, S.; MCDONELL, R. *Ciência de dados com R*: Introdução. Brasília: Editora IBPAD, 2018. Disponível em: https://cdr.ibpad.com.br/#. Acesso em: 29 jun. 2023.

PINHEIRO, J. I. D. *et al. Probabilidade e estatística*: quantificando a incerteza. Rio de Janeiro: Elsevier, 2012.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. https://www.R-project.org/.

REIS, B. de O. F. B.; SILVA, R. A. da; DEMO, P. O lúdico e o ensino universitário combinam? *Políticas Públicas, Educação e Diversidade: Uma Compreensão Científica do Real*, v.1, n.1, p. 714-727, nov. 2020. DOI: https://doi.org/10.37885/200801058.

SONDEREGGER, D. L. *A Sufficient Introduction to R*. 2018. Disponível em https://dereksonderegger.github.io/570L/. Acesso em: 29 jun. 2023.

http://doi.org/10.14393/ER-v30a2023-47

WEI, Y. Colors in R. 2021. Disponível em http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf. Acesso em: 29 jun. 2023.

XIE, Y. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.42, 2023.

Received in August 2023.

Approved in November 2023.