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

Explicando a interpretação frequentista da probabilidade com simulações no *software* R

Felipe Rafael Ribeiro Melo[1]

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.

[1] Doutor em Estatística. Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, RJ, Brasil. Orcid: https://orcid.org/0000-0002-1482-8533. *E-mail*: felipe.ribeiro@uniriotec.br.

Explaining the frequentist interpretation of probability with simulations in R software

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.

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.

* * *

O homem não é nada em si mesmo. Não passa de uma probabilidade infinita. Mas ele é o responsável infinito dessa probabilidade.
Albert Camus

Introdução

Desde as primeiras civilizações, o ser humano lida com os conceitos de acaso e incerteza. O uso da probabilidade para medir a incerteza e a variabilidade remonta a centenas de anos. Segundo DEGROOT & SCHERVISH (2012, p.1), a teoria da probabilidade tem se desenvolvido constantemente desde o século XVII e acredita-se que foi iniciada pelos matemáticos franceses Blaise Pascal (1623-1662) e Pierre Fermat (1601-1665), quando conseguiram obter probabilidades exatas para certos problemas de jogos envolvendo dados. Todavia, a prática de jogos de azar data de muitos séculos atrás.

> "Por volta do ano 3500 a.C., jogos de azar jogados com objetos de osso, que poderiam ser considerados precursores de dados, estavam aparentemente bem desenvolvidos no Egito e em outros lugares. Dados cúbicos com marcações praticamente idênticas às dos dados modernos foram encontrados em túmulos egípcios datados de 2000 a.C. Sabemos que jogar com dados tem sido popular desde aquela época e desempenhou um papel importante no desenvolvimento inicial da teoria da probabilidade." (DEGROOT & SCHERVISH, 2012, p.1).

Hoje, a teoria da probabilidade é uma ferramenta importante na maioria das áreas de engenharia, ciência e administração. Por conta disto, diversos cursos universitários contam com uma disciplina de probabilidade (ou probabilidade e estatística em uma única disciplina), nas quais são explorados conceitos como cálculo de probabilidades, variáveis aleatórias e distribuições de probabilidade. Este artigo, em particular, volta-se à interpretação frequentista da probabilidade, trazendo um relato de experiência em sala de aula de uma atividade baseada em simulações de sorteios no *software* R (R CORE TEAM, 2023) para ilustrar tal interpretação, mostrando saídas numéricas e gráficas, em uma turma do curso do Bacharelado em Engenharia de Produção de uma universidade no estado do Rio de Janeiro. Uma forma lúdica e, por alguns momentos, divertida, de explorar a interpretação frequentista da probabilidade por meio de simulações de sorteios com advento do *software* R. A saber, a interpretação frequentista da probabilidade estabelece que, para um "número grande" de replicações de um experimento aleatório, a probabilidade de um evento é razoavelmente bem aproximada pela frequência relativa de ocorrência do evento em questão nestas muitas replicações.

Muito se estuda sobre a utilização do lúdico no aprendizado infantil, porém, pouco se discute sobre sua utilização no processo ensino-aprendizado em universidades (REIS; SILVA; DEMO, 2020, p.717). O aprendizado por meio de apenas textos e fórmulas no ensino superior pode se tornar um processo demorado, sobretudo para alunos com maior dificuldade em matemática, a qual costuma vir desde o ensino médio ou mesmo desde o ensino fundamental. Uma atividade de simulação mostrada pelo professor, por meio de um *software* livre e gratuito, e com participação ativa dos alunos, como atividade suplementar, tende a tornar o entendimento do conceito algo mais rápido e prazeroso.

Além desta Introdução, este artigo está dividido em mais seis seções. A primeira delas traz alguns conceitos importantes que devem ser abordados em uma disciplina de probabilidade antes de abordar a interpretação frequentista da probabilidade, tais como experimento aleatório, espaço amostral e eventos. Na sequência, uma seção abordando as interpretações clássica e frequentista da probabilidade, com uma atenção maior à última e uma breve discussão sobre convergência em probabilidade. As duas seções seguintes protagonizam a linguagem de programação R: primeiramente, uma seção explicando o que é o *software* R, o ambiente integrado de desenvolvimento Rstudio e o conceito de *scripts* a serem compilados na linguagem R, seguida de uma seção dedicada aos dois *scripts* criados pelo autor para a atividade em sala de aula. A seção de resultados conta toda a aplicação da atividade em sala de aula, incluindo a saída gráfica obtida e tabela com frequências relativas de interesse obtidas via simulação. Por fim, a seção de Conclusões levanta a discussão sobre a atividade executada, seus pontos positivos e reflexões pertinentes a respeito da interpretação frequentista da probabilidade e da geração de simulações via linguagem de programação.

Conceitos preliminares em uma disciplina de probabilidade

Espera-se, do ponto de vista didático, que uma disciplina de probabilidade, a nível de graduação, apresente os conceitos de experimento aleatório, espaço amostral e eventos antes de abordar e definir formalmente a probabilidade, uma vez que esta é uma função aplicada a eventos, os quais são subconjuntos de um espaço amostral, e este último está associado a um experimento aleatório.

Um experimento cujo resultado não pode ser previsto com exatidão é chamado de experimento aleatório. O arremesso de uma moeda, o lançamento de um dado, o sorteio de um elemento (dentro de um grupo de elementos), o placar do jogo do seu time de coração na próxima rodada e a temperatura máxima que será registrada amanhã na sua cidade são alguns

exemplos de experimentos aleatórios. Apesar da impossibilidade de prever com exatidão o resultado de um experimento aleatório, é possível definir o conjunto de todos os seus desfechos possíveis. Este conjunto é chamado de espaço amostral deste experimento aleatório e é frequentemente denotado por Ω, com seus elementos sendo frequentemente referenciados como pontos amostrais. Por fim, qualquer subconjunto de um espaço amostral é um evento desde espaço amostral. Eventos são denotados por letras maiúsculas do alfabeto latino, comumente do início do alfabeto. Em particular, se Ω possui n elementos (isto é, n pontos amostrais), então há 2^n eventos diferentes neste espaço amostral, incluindo o conjunto vazio (o que representa um evento impossível) e o próprio Ω (chamado de evento certo).

Para ilustrar os conceitos abordados no parágrafo acima, seja o experimento (aleatório) de arremessar um dado e verificar qual face cairá voltada para cima. Tem-se, como espaço amostral, Ω = {1, 2, 3, 4, 5, 6}, e seguem alguns exemplos de eventos: A = {2, 4, 6}, B = {5, 6} e C = {3}. Enquanto o evento A representa "face voltada para cima é um número par", o evento B consiste em "face voltada para cima é maior que 4", e o evento C, "face voltada para cima é 3". É dito que um evento ocorre se e somente se o resultado do experimento aleatório está contido no evento. No exemplo acima, quando é dito que o evento A ocorreu, então o resultado do experimento aleatório foi 2, 4 ou 6. E a recíproca é válida: se o resultado foi 2, 4 ou 6, diz-se que o evento A ocorreu.

A partir da definição de eventos, torna-se razoável começar a apresentar a medida de probabilidade em suas diferentes abordagens, além de sua definição axiomática. De fato, do ponto de vista da probabilidade, a principal motivação da definição de eventos se dá em obter suas respectivas probabilidades de ocorrência. Para um evento A qualquer em um espaço amostral Ω, a probabilidade de ocorrer o evento A será denotada por P(A). Ou seja: a probabilidade é uma função P aplicada a eventos, com contradomínio real e imagem no intervalo [0 ; 1].

A interpretação frequentista da probabilidade

Antes de abordar a interpretação frequentista da probabilidade, é comum a abordagem da interpretação clássica da probabilidade. Bastante popular por ser de fácil aplicação, a interpretação clássica só é válida quando se assume que todos os pontos amostrais são igualmente prováveis. Sob esta condição, para um evento A qualquer num espaço amostral Ω, a probabilidade de ocorrência do evento A é dada pela razão do número de elementos de A pelo número de elementos de Ω, ou seja:

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

O numerador desta fração é comumente referenciado por "número de casos favoráveis ao evento", ao passo que o denominador é referenciado como "número de casos possíveis".

Em que pese a praticidade da interpretação clássica da probabilidade, ela não pode ser expandida para qualquer espaço amostral. Neste sentido, a chamada interpretação frequentista da probabilidade é mais interessante, uma vez que vale para qualquer espaço amostral. Nesse cenário mais geral, suponha que são realizadas n replicações (sob condições semelhantes) de um experimento aleatório – com espaço amostral Ω – e o interesse reside em obter a probabilidade de ocorrência de um evento A em Ω. Denotando por n_A o número de vezes que o evento A ocorreu nestas n replicações (ou seja, a frequência absoluta de ocorrência do evento A) e por $f_A = n_A/n$ a frequência relativa de ocorrência do evento A nestas n replicações, a interpretação frequentista da probabilidade estabelece que, se n for suficientemente grande, f_A é uma boa aproximação para P(A), isto é,

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

Em outras palavras: f_A converge, em algum sentido probabilístico, para P(A) quando $n \rightarrow \infty$.

Há algumas críticas em relação a esse tipo de interpretação. Primeiramente, ela fornece apenas uma aproximação. Ainda, aborda que o número de replicações deve ser "suficientemente grande", mas não há indicação definitiva de um número real que seria considerado grande o suficiente. Além disso, as condições semelhantes sob as quais os experimentos devem ser replicados não são claras. Considerando um experimento aleatório de lançamento de uma moeda (que pode não ser honesta),

> "...afirma-se que a moeda deve ser lançada a cada vez sob 'condições semelhantes', mas essas condições não são descritas com precisão. As condições sob as quais a moeda é lançada não devem ser completamente idênticas para cada jogada, porque os resultados seriam os mesmos e haveria todos cara ou tudo coroa. De fato, uma pessoa habilidosa pode lançar uma moeda para o ar repetidamente e pegá-la de tal forma que uma cara seja obtida em quase todos os lançamentos. Portanto, os lançamentos não devem ser totalmente controlados, mas devem ter algumas características 'aleatórias'. Além disso, afirma-se que a frequência relativa de caras deve ser 'aproximadamente 1/2', mas nenhum limite é especificado para a variação permitida de 1/2. Se uma moeda fosse lançada 1.000.000 vezes, não esperaríamos obter exatamente 500.000 caras." (DEGROOT & SCHERVISH, 2012, p.3).

É importante esclarecer sobre o comportamento assintótico que estabelece a convergência de f_A para P(A) quando $n \rightarrow \infty$. Tal convergência se dá em sentido probabilístico, e não é um processo de convergência como comumente verificado em cursos de Cálculo Diferencial e Integral para funções determinísticas. Neste último tipo de convergência, a partir de um determinado ponto k, existe um intervalo [a ; b], que depende de k e

contém o limite da função, tal que para todo n > k, a função está sempre neste intervalo [a ; b]. Todavia, não é possível garantir, com probabilidade 100%, que f_A sempre pertença a um mesmo intervalo [a ; b] para n > k. O que podemos garantir é que, para qualquer ε > 0, quanto maior n, mais provável a diferença entre f_A e P(A) ser menor que ε. Ou seja: f_A converge em probabilidade para P(A) quando n → ∞. Para maiores detalhes, consultar MEYER (1975, p.285).

Ainda no âmbito da convergência em probabilidade, considere a realização de n replicações de um mesmo experimento aleatório e o consequente cálculo de f_A não apenas uma vez, mas m vezes, independentes umas das outras. Sejam $n_{A,j}$ e $f_{A,j}$ as frequências absoluta e relativa do evento A ao término da j-ésima sequência de n replicações (desconsiderando as sequências anteriores). Conforme a Lei dos Grandes Números estabelece, a expressão

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

converge em probabilidade para P(A) quando m → ∞, uma vez que os termos da sequência $\{n_{A,j}\}$ são variáveis aleatórias independentes e identicamente distribuídas seguindo distribuição binomial com parâmetros n e p = P(A), e portanto o valor esperado de cada $f_{A,j}$ é a probabilidade de ocorrência do evento A:

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

A título de ilustração, suponha uma turma de 50 alunos, cada qual com um dado (equilibrado). Cada um destes alunos arremessará seu dado 100 vezes e registrará quantas vezes obteve a face 6, para que seja possível calcular a frequência relativa do resultado "face 6" nos 100 arremessos para cada um dos 50 estudantes. Posto isto, a média aritmética das 50 frequências relativas observadas tende a ser próxima de 1/6 ≈ 0,1667.

O *software* R e a IDE Rstudio

Os próximos parágrafos consistem em alguns esclarecimentos sobre o *software*/linguagem de programação R, o ambiente de desenvolvimento integrado Rstudio e o que são, de fato, *scripts* do ponto de vista da linguagem R. Em particular, esta seção volta-se principalmente ao leitor interessado nesta atividade de simulações para explicação da interpretação frequentista da probabilidade, mas que tem pouco ou nenhum conhecimento sobre R, Rstudio e criação/compilação de *scripts* em alguma linguagem de programação.

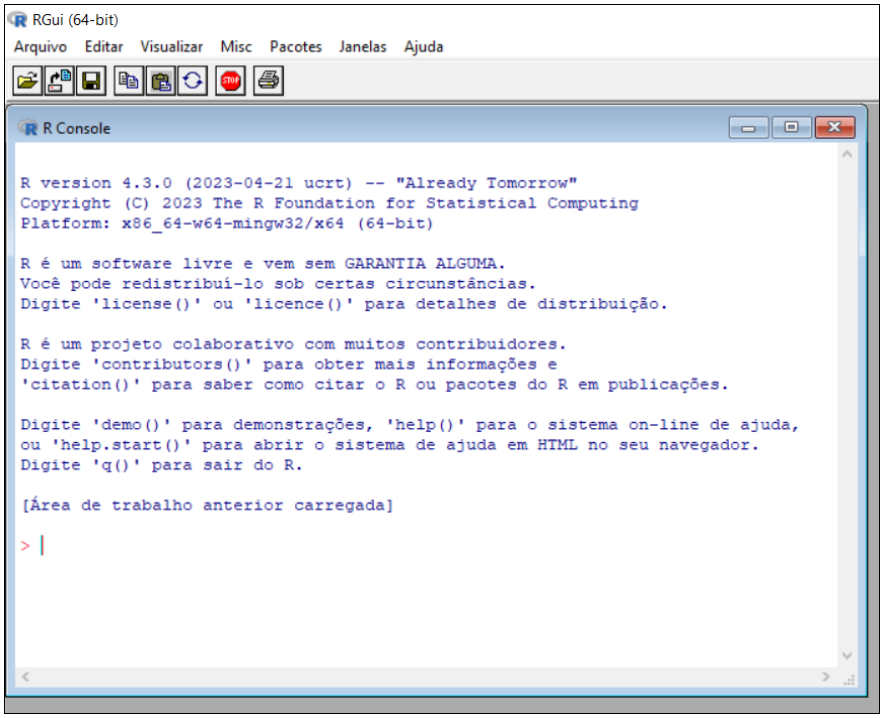
O *software* R é um programa de código aberto comumente usado para tratamento de dados e análises estatísticas, mas não apenas para isto! Também pode ser pensado como uma linguagem de programação, em particular uma linguagem orientada a objetos. A linguagem R é baseada em *script*, portanto sua interface padrão está apta a receber apenas linhas de comando, em que pese a existência de alguns pacotes que fornecem interface de "apontar e clicar", com certas limitações (SONDEREGGER, 2018, p.9). Cada vez mais difundido no meio acadêmico, o *software* R carrega um histórico de mais de 20 anos e, apesar da sua interface pouco atraente, traz diversos pontos positivos, uma vez que se trata de um programa gratuito, potente e estável, disponível para Windows, Linux e Mac, apoiado por uma grande equipe de desenvolvedores em todo o mundo e com uma grande quantidade de pacotes disponíveis que fornecem funcionalidades específicas. Além disso, metodologias de ponta são desenvolvidas primeiramente em R (e disponibilizadas em forma de pacotes) e há diversos materiais, tutoriais e fóruns de discussão disponíveis gratuitamente na Internet.

O download do *software* R pode ser realizado gratuitamente em https://cloud.r-project.org/. Para o sistema operacional Windows, uma forma mais rápida é acessar https://cloud.r-project.org/bin/windows/base/. Após feito o download e concluído o processo de instalação, basta abrir o novo

programa instalado para visualizar sua interface apta a receber linhas de comando, conforme mostra a Figura 1: uma janela denominada *R Console*, dentro da janela de nome *R Gui*.

Ainda na Figura 1, note o sinal “>” na cor vermelha. Ele é chamado *prompt* de comando. Ele indica que o R está apto a receber uma linha de comando (que pode ser algo tão simples como 2+3 ou uma linha de comando bastante extensa e complexa). Nesta janela *R Console*, basta apertar *Enter* após concluir a digitação da linha de comando para que ela seja compilada e seu resultado seja exibido. Faça um teste digitando 2+3, aperte *Enter* e “ignore” o número 1 entre colchetes do lado esquerdo do resultado.

FIGURA 1: Interface padrão do *software* R.



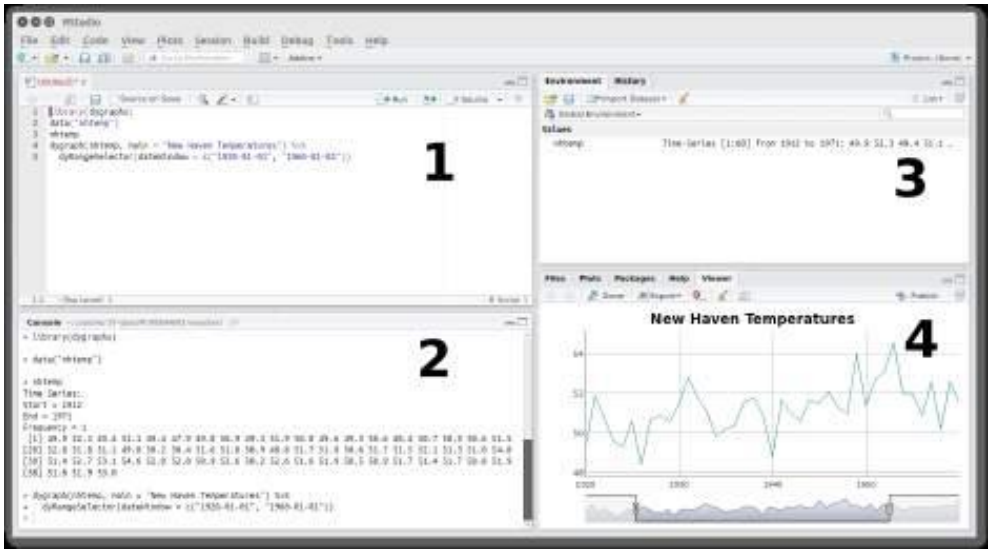
Fonte: Autoria própria.

Para usar o R como uma mera calculadora, digitar a linha de comando desejada diretamente na janela *R Console* (e apertar *Enter* para gerar o resultado) é bastante conveniente. Todavia, quando as saídas desejadas requerem a compilação de várias linhas de comando, é adequado escrevê-las

em um arquivo de texto, para em seguida compilar tais linhas de uma vez só (ou uma a uma). Este tipo de arquivo é denominado *script*.

Um facilitador na criação e na compilação de *scripts* para a linguagem R é o programa Rstudio, que, assim como o R, possui versões para os sistemas operacionais Windows, Linux e Mac e pode ser baixado gratuitamente em https://posit.co/download/rstudio-desktop/. *Scripts* criados no Rstudio são salvos como arquivos na extensão ".R". O Rstudio é uma interface gráfica com diversas funcionalidades que melhoram ainda mais o uso e aprendizado do R que, na prática, facilita muito o dia a dia de trabalho (OLIVEIRA; GUERRA; MCDONELL, 2018, p.10). Cabe ressaltar que a instalação e uso do Rstudio para compilar *scripts* na linguagem R só faz sentido mediante instalação prévia do *software* R. Por padrão, a interface do Rstudio é desmembrada em quatro janelas, conforme ilustrado na Figura 2. Na primeira janela (editor de código), é possível escrever, editar, salvar e carregar os *scripts*. Abaixo dela, segue o *Console*, onde são compiladas as linhas de comando do *script* ao clicar nas opções *Run* (que compila apenas a linha corente ou as linhas selecionadas) ou *Source* (que compila todo o *script* de uma só vez). No lado direito, a janela superior guarda, na aba *Environment*, todos os objetos criados na sessão do R, e a aba *History* cria um histórico de comandos utilizados (OLIVEIRA; GUERRA; MCDONELL, 2018, p.10). Por fim, a última das janelas possui várias abas, sendo a mais relevante, para os propósitos deste artigo, a aba *Plots*, na qual saídas gráficas, fruto de compilações de linhas de comando, são exibidas.

FIGURA 2: Interface padrão do ambiente de desenvolvimento integrado Rstudio.



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

Scripts para ilustrar a interpretação frequentista da probabilidade

Esta seção é voltada a elucidar sobre os dois *scripts* que foram criados pelo próprio autor, na linguagem R, para gerar simulações que ilustram a interpretação frequentista da probabilidade como método suplementar na explicação deste conceito. Ambos, após compilados, geram tabelas de distribuição de frequências na janela *Console* e representações gráficas na aba *Plots* da janela inferior direita do Rstudio. Estas representações gráficas, em particular, foram inspiradas e seguem a mesma ideia abordada em PINHEIRO *et al.* (2012, p.7, Figura 1.1). De forma a facilitar a compreensão dos próximos parágrafos, seguem os nomes dados a estes dois arquivos de *script*:

- conceito_frequentista_2023-1.R;

- conceito_frequentista_sobrepostas_2023-1.R.

É importante ressaltar que ambos os *scripts* utilizam funções de dois pacotes do R: *crayon* (CSÁRDI, 2022) e *knitr* (XIE, 2023). Portanto, antes de compilar qualquer um destes *scripts*, se faz necessária a instalação destes

pacotes. Uma maneira prática de se fazer isto sem a necessidade de digitação de linhas de comando é acessar, no Rstudio, o menu *Tools* > *Install Packages*. O campo *Install from* deve estar preenchido com *Repository (CRAN)* e, no campo abaixo, devem ser digitados os nomes dos pacotes a serem instalados, separados por vírgula ou por espaço. Tal como as instalações do R e do Rstudio, instalação de pacote é necessária apenas uma vez.

Ao abrir o arquivo "conceito_frequentista_2023-1.R" no Rstudio, é possível visualizar, na referenciada "Janela 1" da Figura 2: a definição do número de replicações do experimento aleatório, o espaço amostral associado a este experimento aleatório, as probabilidades associadas a cada ponto amostral, o evento de interesse e sua probabilidade de ocorrência, encapsulados nos objetos n, Omega, probs, A e p, respectivamente. Em seguida, é definida a cor do gráfico e são criados alguns objetos que fazem parte da rotina que encerra este *script*. Tal rotina gera, após cada uma das n replicações do experimento aleatório: uma tabela, na janela *Console*, computando quantas vezes o evento A ocorreu e quantas vezes este evento não ocorreu após k replicações, para k = 1, 2, ..., n, além do resultado da k-ésima replicação; e um gráfico com o número de replicações no eixo horizontal e a frequência relativa do evento A (até a k-ésima replicação) no eixo vertical. O k-ésimo gráfico gerado mostra k pontos, onde cada um deles representa o total de replicações até o instante considerado no eixo horizontal e, no eixo vertical, a frequência relativa de ocorrência do evento A até o instante considerado. Para facilitar a visualização da evolução no gráfico conforme o experimento é replicado, pontos vizinhos imediatos são conectados por segmentos de reta. No decorrer deste artigo, estes gráficos são frequentemente referenciados como trajetórias. Para ilustrar uma simulação do tipo, seja o experimento aleatório de arremessar um dado viciado e verificar a face voltada para cima, onde a probabilidade de ocorrência de cada face é proporcional ao número que representa a face: ou seja, para i = 1, ..., 6, j = 1, ..., 6 e j > i, face j é j/i vezes mais provável de ocorrer que a face i. Nesse cenário, a probabilidade de ocorrer face j é igual a

j/21, para j = 1, 2, 3, 4, 5, 6. Mas suponha que esta estrutura probabilística do dado é desconhecida e deseja-se simular 100 arremessos para obter uma aproximação da probabilidade da face obtida após um lançamento ser um número par, por meio do *script* "conceito_frequentista_2023-1.R". Portanto, o evento de interesse é A = {2, 4, 6}. Admita que, após os 4 primeiros arremessos, os resultados foram: 3, 5, 6 e 3. Logo após este quarto arremesso, é gerada uma tabela, na janela *Console*, conforme a Figura 3 e um gráfico conforme a Figura 4 na janela inferior direita do Rstudio. Uma vez que os quatro arremessos resultaram numa sequência de dois números ímpares, um número par e um número ímpar, as frequências relativas de ocorrência do evento A após 1º, 2º, 3º e 4º arremesso são, respectivamente: 0/1 = 0, 0/2 = 0, 1/3 ≈ 0,3333 e 1/4 = 0,25. Os quatro pontos na Figura 4 são os pares ordenados (1,0), (2,0), (3,1/3) e (4,1/4). Em particular, os pacotes previamente instalados carregam funções que focam na formatação das 100 tabelas geradas na janela *Console*, tal como ilustrado na Figura 3. A fonte vermelha e o fundo cinza nas expressões acima da tabela se devem às funções (do pacote *crayon*) *red* e *bgWhite*, respectivamente. A formatação da tabela em si provém da função *kable*, do pacote *knitr*.

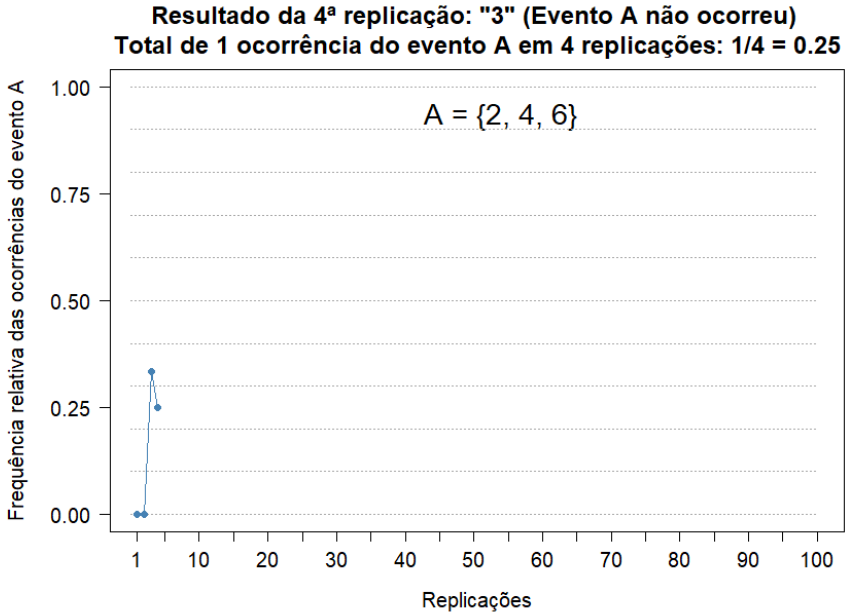
FIGURA 3: Uma das tabelas provenientes de compilação do *script* conceito_frequentista_2023-1.R.

```
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|
```

Fonte: Autoria própria.

FIGURA 4: Um dos gráficos provenientes de compilação do *script* conceito_frequentista_2023-1.R.



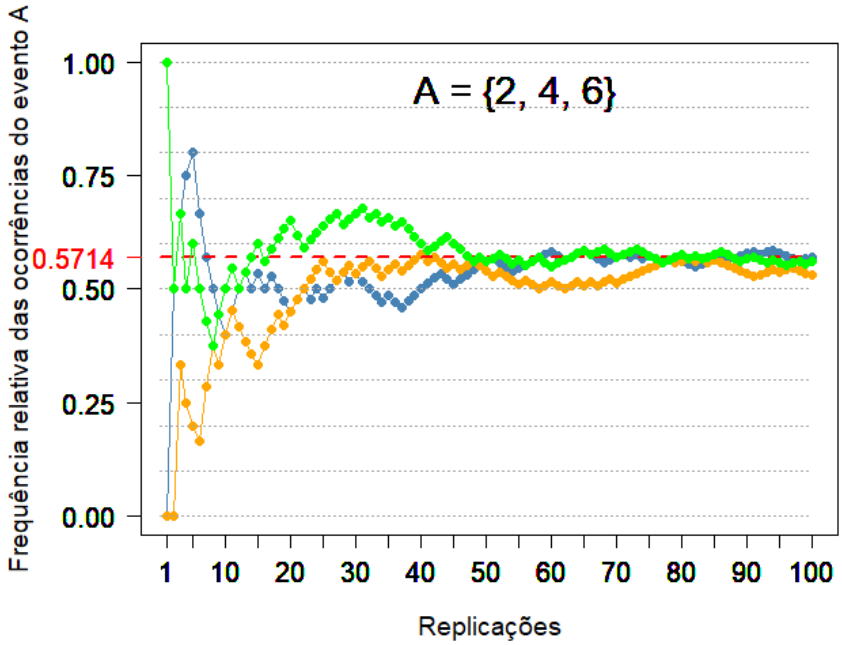
Fonte: Autoria própria.

Claramente, o interesse maior se dá na frequência relativa de ocorrência do evento A após as 100 replicações. Ao compilar todo o *script* – conforme procedimento explicado na seção anterior – os 100 gráficos são armazenados na janela inferior direita do Rstudio, na aba *Plots*. O último destes 100 gráficos será exibido, com a possibilidade de acessar qualquer um dos 99 gráficos anteriores clicando na seta azul *Previous plot* (ou seu atalho no teclado Ctrl+Alt+F11). Em particular, este último gráfico gerado traz, em destaque, a probabilidade exata p de ocorrência do evento A marcada de vermelho no eixo vertical e um segmento de reta tracejado, também na cor vermelha, de (0, p) até (n, p), reforçando a ideia de convergência de f_A para p – ainda que seja convergência em probabilidade. Do ponto de vista didático, uma boa estratégia docente para exibir estes gráficos aos seus alunos é retornar ao primeiro gráfico gerado (isto é, aquele gerado após a primeira replicação) e utilizar a seta azul *Next plot* (ou seu atalho no teclado Ctrl+Alt+F12) para mostrar a evolução do gráfico, replicação após replicação. Na prática, enfatizar os gráficos associados às primeiras e às últimas replicações é o

mais importante, de forma a não perder muito tempo de aula e não tornar a atividade cansativa ou monótona.

O *script* "conceito_frequentista_sobrepostas_2023-1.R" assemelha-se bastante ao anterior, mas apenas a figura relativa a n-ésima replicação é gerada. Contudo, é possível executar uma nova sequência de n replicações e representá-la no gráfico com uma nova cor e sem remover a trajetória anterior: basta modificar a cor na linha 12 do *script*, compilar esta linha e recompilar todas as linhas abaixo dela. Neste sentido, é possível mostrar que, a cada sequência de n replicações realizadas de um mesmo experimento aleatório, as trajetórias das frequências relativas de ocorrências de um mesmo evento A até n replicações não são as mesmas e, ao final das n replicações, os valores de f_A das novas n replicações e das n replicações anteriores não necessariamente são iguais. A aplicação desta rotina em sala de aula deve registrar o valor obtido para f_A ao final de cada simulação de n replicações e verificar que ora ele está acima de P(A), ora abaixo de P(A) e eventualmente pode se igualar a P(A). Contudo, a média aritmética dos registros de f_A ao final de cada simulação de n replicações se aproxima de P(A) com alta probabilidade, considerando um número suficientemente grande de trajetórias. A título de ilustração, a Figura 5 exibe apenas três trajetórias, com n = 100 replicações cada e o mesmo evento A = {2, 4, 6} do exemplo anterior, cuja probabilidade é P(A) = (2+4+6)/21 = 12/21 ≈ 0,5714. Os valores observados para f_A após 100 replicações foram 0,57 (trajetória azul), 0,53 (trajetória laranja) e 0,56 (trajetória verde), cuja média aritmética retorna aproximadamente 0,5533, o que não necessariamente espera-se estar próximo de 0,5714 em virtude do número pequeno de simulações de 100 replicações (apenas três).

FIGURA 5: Gráfico proveniente de compilação do *script* conceito_frequentista_sobrepostas _2023-1.R.



Fonte: Autoria própria.

Resultados: aplicação da atividade

Nos minutos finais da aula na qual foram abordadas as interpretações clássica e frequentista da probabilidade, o professor da disciplina de probabilidade para o curso de Engenharia de Produção de uma universidade do Rio de Janeiro, na figura do autor deste artigo, apresentou e explorou ambos os *scripts* por ele criados. Primeiramente, foi explorado o *script* "conceito_frequentista_2023-1.R", simulando 100 arremessos de um dado equilibrado e com interesse na obtenção da frequência relativa de ocorrências do evento A = {2, 4, 6} após estes 100 arremessos. Sabe-se, pela interpretação clássica da probabilidade, que P(A) = 0,50. Contudo, não é garantido que f_A = 0,50 após 100 replicações, mas espera-se que esta frequência relativa esteja próxima de 0,50. Após várias sequências de 100 replicações, os discentes puderam perceber que, além de f_A não ser necessariamente 0,50 após 100 replicações, as trajetórias da 1ª até a 100ª

replicação não são as mesmas e que o valor final de f_A após as 100 replicações tipicamente varia para cada centena de replicações.

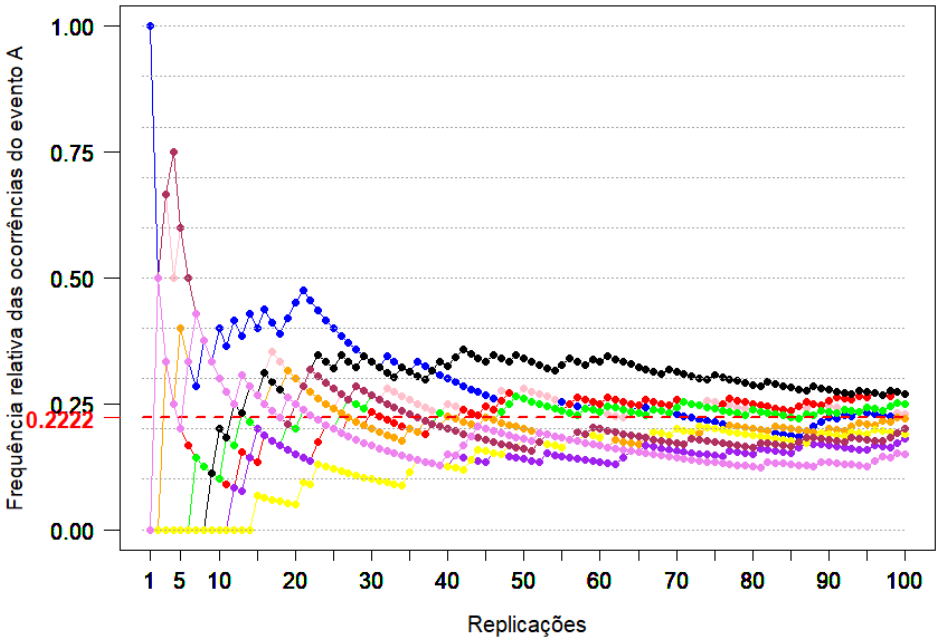
Mas a atividade mais interessante ainda estava por vir. Tomando por base o *script* "conceito_frequentista_sobrepostas_2023-1.R", o docente configurou o experimento aleatório como o sorteio de um dos 9 estudantes presentes na turma e o número de replicações deste experimento aleatório como n = 100. Portanto, o espaço amostral foi configurado como um conjunto de nove elementos no qual cada elemento é o nome de um dos alunos. O evento de interesse considerado foi "estudante com pelo menos 24 anos", definido após rápida conversa informal entre docente e discentes. Tal evento contém apenas dois elementos, ou seja, exatamente 2 dos 9 alunos presentes possuem idade igual ou maior que 24 anos. Denotando tal evento por A, tem-se $P(A) = 2/9 \approx 0{,}2222$. Além disso, cabe ressaltar que, neste artigo, os nomes presentes no espaço amostral e no evento de interesse são omitidos, de forma a não expor os nomes dos alunos.

Primeiramente, o docente realizou 100 replicações do experimento aleatório, escolhendo a cor azul para o traçado da trajetória de f_A da primeira até a centésima replicação. Estas 100 replicações podem ser pensadas como 100 sorteios com reposição dos alunos presentes em aula. Nestes 100 sorteios, exatamente 22 deles resultaram em um dos estudantes com pelo menos 24 anos e, portanto, o valor de f_A observado após 100 replicações foi de 0,22. Após exposição deste resultado (que pode ser calculado por meio da última tabela gerada na janela *Console*, tal qual ilustrado na Figura 3, dividindo a frequência de "Evento A ocorreu" por 100), o professor escreveu o resultado 0,22 na lousa. Em seguida, chamou o aluno que estava mais próximo dele e pediu para que escolhesse uma cor diferente do azul que já fora utilizado. O aluno escolheu a cor vermelha e, assim, o professor mudou a cor na linha 12 do *script* (de *blue* para *red*) e executou uma nova centena de replicações, gerando uma trajetória na cor vermelha que retornou $f_A = 27/100 = 0{,}27$ após todas as replicações. O docente voltou para a lousa e escreveu 0,27 abaixo do 0,22 previamente escrito e, a esta altura, temos um gráfico com duas trajetórias (uma azul e uma vermelha), além

dos valores 0,22 e 0,27 escritos na lousa pelo docente. Tal procedimento foi replicado até que todos os nove estudantes presentes na sala participassem da atividade, cada qual escolhendo uma cor que ainda não tivesse sido utilizada. Aqui, cabe destacar que a linguagem R oferece, de forma nativa, um conjunto de 657 cores diferentes, cujos nomes podem ser conferidos compilando a linha de comando *colors()* ou em materiais gratuitamente disponibilizados na Internet, como em WEI (2021). Voltando às simulações, visualizou-se, ao final da atividade, 10 trajetórias sobrepostas (uma com a cor escolhida inicialmente pelo docente e mais nove com as cores escolhidas pelos discentes), conforme mostra a Figura 6, além de 10 valores de f_A (cada um após 100 replicações) que seguem registrados na Tabela 1 em ordem decrescente de frequências relativas de ocorrência do evento A = {estudante com pelo menos 24 anos}.

Para encerrar a atividade, o docente calculou a média aritmética dos 10 valores dispostos na segunda coluna da Tabela 1, que retornou 0,218. Conforme esperado, tal resultado está próximo de $2/9 \approx 0{,}2222$, em que pese uma pequena quantidade realizada de centenas de replicações – apenas 10 replicações.

FIGURA 6: Trajetórias sobrepostas de 10 centenas de replicações do experimento aleatório com o número de replicações no eixo horizontal e a frequência relativa de evento de interesse até tal replicação no eixo vertical.



Fonte: Autoria própria.

TABELA 1: Frequências relativas da ocorrência do evento de interesse observadas ao final de cada centena de replicações do experimento aleatório.

Cor da trajetória	f_A ao final da trajetória
Preta	27/100 = 0,27
Vermelha	27/100 = 0,27
Verde	25/100 = 0,25
Rosa	23/100 = 0,23
Laranja	22/100 = 0,22
Azul	22/100 = 0,22
Marrom	20/100 = 0,20
Amarela	19/100 = 0,19
Roxa	18/100 = 0,18
Violeta	15/100 = 0,15

Fonte: Autoria própria.

Conclusão

Apesar da interpretação frequentista da probabilidade ter sido apresentada inicialmente pelo docente à sua turma por meio de texto e fórmulas, espera-se que a atividade das simulações de sorteios pelo *software* R, com consequente acompanhamento da frequência relativa do evento de interesse, além das diferentes trajetórias com desfechos que tipicamente mudam, tenha tornado o conceito mais claro para os discentes. Segundo MACHADO & WOJCICKOSKI (2017, p.11), o lúdico pode ser um instrumento a ser utilizado pelo educador para facilitar o processo de ensino-aprendizagem, porém este, antes de utilizá-lo deve analisar e estudar a dinâmica a ser usada. Sendo assim, a estratégia docente utilizada foi compor o espaço amostral como o conjunto de alunos presentes em sala de aula e o evento de interesse como algo sugerido por estes alunos e de fácil interpretação para estes. Durante a atividade, momentos de descontração foram vivenciados, trazendo a essência do lúdico para a sala de aula de forma leve e natural. Em particular, para os dez valores finais de frequência relativa para cada trajetória (os mesmos valores presentes na Tabela 1 e descobertos um a um), verificou-se a expectativa, capitaneada pelo docente, por frequências relativas inferiores a 2/9 quando a maior parte das frequências relativas registradas no quadro situavam-se acima de 2/9 – expectativa esta que nem sempre foi satisfeita. Ao final da atividade e consequente final da aula, o docente combinou com os seus alunos de enviá-los a figura gerada, por meio do ambiente virtual de aprendizagem utilizado especialmente para esta disciplina. Tal figura é a Figura 6 deste artigo. Uma forma de compartilhar com os alunos presentes (e ausentes) um registro da atividade realizada. De fato, todo este roteiro executado propiciou aos discentes da turma uma melhor compreensão do conceito frequentista da probabilidade (em relação ao texto e à fórmula incialmente apresentados), tal como um entendimento um pouco mais refinado da essência do aleatório.

Algumas outras reflexões a respeito da atividade realizada seguem neste parágrafo. Primeiramente, pode-se argumentar por que o uso de 100 replicações

para cada trajetória. Afinal, 100 é um número suficientemente grande? A escolha por esse valor (em vez de um valor maior) se deu por facilitar a visualização das trajetórias de maneira local e por questões de tempo computacional para a geração dos resultados. Ainda assim, por volta da 5ª ou 6ª trajetória gerada, o computador do professor usado em sala de aula passou a apresentar alguma lentidão até a finalização do resultado (a saber, um *notebook* com cerca de 9 anos de uso, com processador *Core-i5*, 8Gb de memória RAM e SSD de 480Gb). Além disso, por conta de apenas 9 alunos estarem presentes na aula, apenas 10 trajetórias foram simuladas. Com um número maior de alunos na turma, mais trajetórias seriam geradas e, portanto, a convergência em probabilidade seria favorecida. Para um valor positivo ε próximo de zero e um valor inteiro k maior que 10, a probabilidade da média dos desfechos da novas k trajetórias estar distante de 2/9 por um valor menor que ε é maior do que no cenário com apenas 10 desfechos. Entretanto, o número de trajetórias geradas mostrou-se suficiente para que a média dos desfechos se localiza-se próxima de 2/9: uma diferença de apenas 0,0042. Com um tempo maior de aula, a atividade poderia ser recomeçada, gerando novas 10 trajetórias e um novo valor médio, de forma a convencer alunos mais incrédulos que os próximos 10 desfechos seguiriam com desfechos próximos de 2/9 e que se espera que a média destes valores continue próxima do valor exato da probabilidade do evento de interesse. Vale ressaltar que os *scripts* criados permitem alterar o número de replicações do experimento aleatório, assim como o espaço amostral associado a este experimento, o evento de interesse e as probabilidades associadas, tornando-os bastante versáteis.

Por fim, vale ressaltar que a área de Ciência de Dados mostra-se interessante para diversos alunos de cursos de Engenharia, Matemática e Informática como área de atuação profissional. Uma boa base em probabilidade, tal como em estatística, é fundamental para esta carreira, incluindo a compreensão de resultados assintóticos, como a Lei dos Grandes Números. A atividade realizada introduz este importante resultado da probabilidade.

Referências

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.

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.

Recebido em agosto de 2023.

Aprovado em novembro de 2023.