Atividade 2 - Amostras e populações

1. Justificativa

Esta atividade introduz o aluno no estudo de alguns elementos de estatística, especialmente medidas de posição, análise de gráficos e princípios de amostragens. Com esse objetivo, o sistema convidará os alunos a escolherem uma característica a ser estudada em uma cidade, de sua escolha, e pedirá que estabeleçam o tamanho da amostra que desejam pesquisar para quantificar percentualmente as condições da característica escolhida. O sistema simulará pesquisas e apresentará resultados que deverão ser analisados pelos alunos, e validados ou não a partir de índices de erros percentuais pré-estabelecidos.

Os conhecimentos sobre probabilidades que, espera-se, tenham sido adquiridos na atividade anterior poderão agora ser contextualizados em uma situação muito conhecida. Vale observar, entretanto, que os algoritmos de cálculo seguidos pelo sistema em suas simulações são próximos dos algoritmos estatísticos tradicionais, e que simplificações foram feitas com o objetivo de facilitar o primeiro contato dos alunos com esse tipo de simulação.

2. Descrição da atividade

Ao entrar na atividade, o aluno será convidado a selecionar uma dentre 4 cidades fictícias. O que difere uma cidade da outra, além do nome, é a quantidade de habitantes de cada uma e a distribuição porcentual das características que serão analisadas. Dessa forma, o aluno poderá fixar uma característica e analisar sua distribuição nas 4 cidades, da mesma forma que poderá, escolhendo uma cidade, analisar 3 características disponíveis pelo sistema: classe social, intenção de voto e escolaridade.

Escolhida a cidade e a característica a ser estudada, o aluno será, na seqüência, convidado a selecionar o tamanho da amostra de pesquisa que deseja estabelecer para estudar a característica selecionada. Por exemplo, se optar por analisar a escolaridade, o sistema simulará um sorteio que fornecerá a composição porcentual dos habitantes da cidade distribuídos em 5 classes: ensino fundamental incompleto, ensino fundamental completo, ensino médio incompleto, ensino médio completo e ensino superior, completo ou não. Ocorre, entretanto, que dependendo do tamanho da amostra selecionada e também do número de sorteios realizados o processo terá um determinado custo, e o sistema gerará um ou outro resultado, calculado de acordo com algoritmo estatístico aproximado. Dessa forma, o aluno perceberá que o grau de certeza dos resultados de uma pesquisa depende da correta dimensão dos parâmetros envolvidos, e também dos custos.

Após a simulação da pesquisa pelo sistema, o aluno será convidado a optar por uma medida de posição - média, mediana ou moda – para fazer, a partir dela, uma estimativa da distribuição porcentual da população de sua cidade, de acordo com a característica escolhida. Nessa opção, o aluno aprenderá, se ainda não sabia, a calcular cada uma dessas medidas e também a avaliar a confiabilidade de cada uma na análise de toda a população. O sistema não pede a análise da dispersão do conjunto de dados, mas seria importante que isso pudesse ser feito, de acordo com a prioridade dos conteúdos programados, a fim de enriquecer ainda mais a atividade.

3. Como conduzir a atividade?

- Duração da atividade – 50 minutos
- Organização – grupos de 2 alunos

Para introduzir a atividade, o professor poderá confeccionar um objeto de sorteio que consiste em um determinado número de bolinhas, 150 por exemplo, divididas em 2 ou 3 cores. Colocando as bolinhas em um saco, o professor poderá sortear um determinado número delas e contar quantas de cada cor saíram na amostra. Com isso, poderá fazer uma estimativa, ainda que precária, do número total de bolinhas de cada cor no saco. Por exemplo, vamos imaginar que no saquinho haja apenas bolinhas de duas cores: 70 azuis e 80 amarelas. Se o aluno sortear uma amostra de 20 bolinhas e, dentre elas, saírem 8 azuis e 12 amarelas, poderá fazer:

Azuis: 8 em 20, quantas em 150?

Resposta: 60

Amarelas: 12 em 20, quantas em 150?

Resposta: 90

Esses valores, 60 e 90, são estimativas das quantidades das cores dentro do total. Evidentemente, esses valores contêm erros que podem, até, ser além dos esperados, e caberá ao professor discutir com seus alunos como fazer para aumentar a confiabilidade da estimativa. A possibilidade recomendada, nesse caso, para que as estimativas tenham maior confiabilidade, será a de repetir outras vezes as amostragens de 20 bolinhas e calcular a média aritmética de todos os resultados obtidos. Sabemos que a média terá mais significado quanto maior for o número de amostras sorteadas. Dividindo a classe em vários grupos, de modo que cada grupo repita o experimento algumas vezes, o professor poderá, ao final, calcular a média geral de todas as estimativas e mostrar aos alunos que os valores estimados, a partir dessa média, se aproximam bastante das quantidades reais de bolinhas no saco.

Numa outra possibilidade, o professor poderá argumentar com seus alunos sobre a eficácia do aumento do número de elementos em cada amostra sorteada, isto é, aumentando, por exemplo, de 20 para 30 a cada sorteio. Nesse caso, é importante considerar que, apesar desse processo mostrar alguma eficiência, especialmente no caso de um pequeno número total de bolinhas como 150, em um caso real em que estiverem diante de uma população muito numerosa e com alguma característica a ser estimada, tal conduta poderá encarecer por demais o processo, e não permitir resultados tão precisos quanto permitiria a opção pela outra possibilidade, de aumentar o número de sorteios.

Essas duas opções – aumentar o tamanho da amostra ou aumentar o número de sorteios – serão colocadas para os alunos na atividade simulada. Caberá a ele fazer a escolha e analisar os resultados obtidos em cada caso.

É preciso que o professor esteja preparado para, algumas vezes, se surpreender com os resultados obtidos por um ou outro aluno. Por exemplo, um aluno poderá optar por fazer apenas 4 sorteios de amostras com o menor número de elementos (0,1% da população), e mesmo assim obter um melhor resultado em sua estimativa do que um outro aluno que tenha optado pelo mesmo tamanho de amostra e que tenha efetuado 8 sorteios.

Vale lembrar que estamos trabalhando com estatística e não haverá, em nenhum caso, 100% de certeza nas previsões. Mas vale considerar, também, que a repetição do experimento mostrará que, na maior parte das vezes, a estimativa será mais eficiente, na atividade, se forem escolhidas amostras de poucos elementos e repetido o sorteio o maior número de vezes possível. Caberá, assim, ao professor ficar atento para efetuar um levantamento dos resultados obtidos pelos vários grupos e discutir com eles os resultados obtidos. Outra possibilidade de introdução do trabalho, também interessante, e costuma dar ótimos resultados e envolver toda a classe, consiste na simulação de uma previsão sobre o número de corredores de uma prova de atletismo. Veja a descrição na seqüência.

Adivinhe quantos corredores são

A atividade mostra a eficácia de inferência feita a partir da média aritmética de elementos numéricos em uma amostra.

O professor deve providenciar um saco com números escritos em bolinhas de isopor bem pequenas, todas do mesmo tamanho, numeradas de 1 a 172. O material escolhido para ser numerado pode ser outro, mas deve-se garantir minimamente que a chance de tirar um número é igual para todos os números a serem sorteados. Os alunos não têm conhecimento do número de peças.

A atividade inicia-se com o professor contando a seguinte história: *Você está parado na rua assistindo a uma prova de atletismo, como uma maratona, por exemplo. Por você passam alguns corredores que carregam o número de inscrição preso ao peito. Esses corredores, que na inscrição foram numerados de 1 a* ***n,*** *passam agora todos misturados, de maneira que se você observar a numeração de alguns deles perceberá que não existe regularidade alguma. Como poderíamos fazer para avaliar o número total de corredores apenas observando a numeração de um punhado deles*?

Nesse momento, o professor procura extrair da classe que o interessado deve analisar amostras de corredores. Em seguida, o professor mostra o saco cheio de peças e diz : *eis aqui todos os corredores. Estimem quantos são eles*. De início, os alunos poderão fazer suas estimativas apenas utilizando o toque nas peças, sem retirá-las do saco.

É interessante escrever o número de peças do saco em um papel e escondê-lo, sem mostrar aos alunos o valor escrito. Depois, anotar num canto da lousa alguns valores estimados pelos alunos.

Para continuar, o professor escolhe o tamanho de sua amostra: por exemplo, 11 peças. Retira-as e anota os números no quadro.Em seguida, pergunta à classe: o *que devo fazer com esses números*? Certamente, após vários palpites, aparecerá alguém para dizer : *calcule a média .* O cálculo deve ser feito e os alunos devem dizer o que fazer com a média obtida. Isso trará novamente um valor estimado pelos alunos, e que deve ser valorizado. No entanto, o professor poderá propor um outro método cuja experiência mostrou ser o mais eficiente, acarretando, na maioria das vezes, erro de, no máximo, 5%. Esse método, que será descrito a seguir, permite em cerca de 80% das vezes resultados mais fiéis do que os demais métodos tentados, como demonstraram simulações realizadas em computador.

1) Escrever os valores sorteados na primeira vez e colocá-los em ordem crescente, como por exemplo:

3 17 21 34 45 67 70 89 93 104 121

2) Calcular as diferenças entre os termos sucessivos da seqüência:

3 17 21 34 45 67 70 89 93 104 121

14 4 13 11 22 3 19 4 11 17

3) O próximo passo é calcular a média das diferenças obtidas. A idéia é imaginar que os elementos sorteados estão em progressão aritmética de razão igual à média das diferenças, que é 11,8. Dessa forma, se adicionarmos 11,8 a 121, podemos pensar que o número de peças é aproximadamente 133 peças.

O professor diz aos alunos que esse procedimento pode trazer um erro muito grande, caso o experimento seja realizado apenas uma vez. Entretanto, se o experimento for repetido pelo menos 8 vezes e, ao final, for tirada a média dos valores obtidos, o erro comparado não será grande. Naturalmente, cada vez que o experimento for repetido, todos as peças devem ser colocadas novamente no saco e misturadas. O professor deve fazer os cálculos junto com seus alunos.

Ao final da atividade, a classe deve ser dividida em grupos de 4 alunos com o objetivo de criar um método que eles consideram melhor do que o proposto pelo professor.

Se for possível cada grupo ter o seu saco de amostras, todos poderão verificar a validade de seu método.

Depois de preparada a classe com uma ou duas das discussões aqui propostas, os alunos poderão interagir com a atividade virtual praticamente sem a participação do professor, o qual, no entanto, como já foi dito, precisará estar atento para fazer um levantamento das estimativas que os grupos conseguiram e discutir sobre a eficiência da escolha das amostras em cada caso.

Após o término da atividade virtual, o professor, em sala de aula, poderá formalizar os conceitos de “Medidas de posição”, elaborando alguns exemplos de conjunto de dados para que os alunos calculem média, mediana e moda, exemplos esses que poderão ser recolhidos de livros didáticos de ensino médio. Caso o professor se sinta seguro e seu planejamento permita, ele poderá também introduzir uma discussão sobre “Medidas de dispersão” de um conjunto de dados, mostrando para os alunos que há casos em que média e mediana coincidem em um conjunto de dados mas, mesmo assim, tais resultados são sem significância se a dispersão do conjunto de valores for alta. Por exemplo, vamos imaginar que em uma das séries de 8 sorteios realizados na atividade foram obtidos os seguintes valores para uma das características analisadas:

32 18 27 23 29 16 24 31

média = 25 mediana = 25,5

Valerá a pena comentar o fato de que os limites desse conjunto de valores são 16 e 32, e a diferença entre eles, portanto a *Amplitude*, é 16, que é um valor muito alto quando comparado com a média de 25. Para melhorar ainda mais a análise, o professor poderá pedir para os alunos calcularem algum desvio, sendo recomendado, nesse caso, o *desvio médio*.

$$\text{Desvio médio} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

x = valores do conjunto; $\bar{x}$ *= média; n = número de elementos do conjunto*

O desvio médio desse conjunto de valores será igual a 4,7, o que é também um valor alto quando comparado com a média de 25 - quase 20%.

Será possível comentar, ainda, que o desvio médio não é a medida de dispersão mais indicada, e o melhor seria calcular o *desvio padrão*. O planejamento e a motivação dos alunos determinarão a viabilidade de se introduzir também o cálculo do desvio padrão.