Atividade 3 - Amostras confiáveis

1. Justificativa

Quando analisamos propriedades de um grupo de pessoas como, por exemplo, idade, estatura, escolaridade ou religião, podemos associar conceitos ou números a cada uma delas. Essas propriedades são chamadas de variáveis de pesquisa e normalmente classificadas em dois tipos: qualitativas e quantitativas .

As variáveis **qualitativas** representam um *atributo* do indivíduo e da população, e podem ser mensuradas apenas de uma forma conceitual. O *gênero* é feminino ou masculino, a *escolaridade* pode ser dividida em Ensino Fundamental, Ensino Médio, e Superior. Já as variáveis **quantitativas** permitem que associemos um valor à propriedade analisada, localizando-a inclusive de acordo com determinada escala. A estatura das pessoas, suas idades, o número de horas de sono de cada uma, e a renda familiar são exemplos de variáveis quantitativas.

Nesta atividade, trabalharemos apenas com variáveis do primeiro tipo, isto é, variáveis *qualitativas.*

O texto em anexo, extraído da Revista do Professor de Matemática, poderá ajudar o professor a compreender melhor a intenção da atividade, embora não seja de todo recomendável usá-lo diretamente com os alunos devido às equações matemáticas que provavelmente não são de seu conhecimento.

Como afirma o texto, é preciso ficar claro para o aluno que é necessário senso crítico e cuidado para escolher a amostra que posteriormente terá suas características *extrapoladas* para toda a população. Mesmo com toda técnica de escolha da amostra, não há garantias absolutas na transferência de conhecimento da amostra para o todo.

2. Descrição da atividade

Esta atividade pretende sensibilizar os alunos para o estabelecimento de critérios na formação de uma amostra de pesquisa. Partindo de uma proposta dara – *pesquisar a intenção de voto de uma população em dois possíveis candidatos* – será necessário que os alunos determinem as características da amostra de pesquisa, que

poderá ser formada com base em três critérios: escolaridade, sexo e classe social. Dependendo da composição da amostra, o sistema simulará um ou outro resultado para a pesquisa, além de confrontar o resultado obtido pelo aluno com o resultado real catalogado pelo sistema.

De início, o sistema fará simulações do resultado da pesquisa utilizando amostras divididas apenas em homens ou mulheres. Será importante o aluno perceber que os resultados da pesquisa serão diferentes dependendo do porcentual de cada gênero que compõe a amostra.

Na seqüência, os alunos poderão compor sua amostra de pesquisa com base também na escolaridade e na classe social de seus componentes. A atividade se encerra com a proposta de elaboração de um relatório a partir dos resultados obtidos na pesquisa.

3. Como conduzir a atividade

- Duração da atividade: uma aula de 50 minutos para a interação com o sistema e outra aula para discussão e elaboração do relatório.
- Organização: grupos de 2 ou de 3 alunos, a critério do professor.

Os alunos poderão iniciar a atividade sem qualquer explicação prévia, ou poderão participar de uma discussão, conduzida pelo professor, sobre acertos e erros de pesquisas de intenção de voto realizadas no Brasil, nos últimos tempos. Há uma grande curiosidade dos alunos em saber como é possível conseguir erros tão baixos em pesquisas que entrevistam porcentuais tão pequenos da população eleitoral do país. Nesse sentido, o texto da Revista do Professor de Matemática, citado anteriormente, pode dar elementos para o professor.

Além disso, poderá ser contada para os alunos a história da pesquisa realizada nos Estados Unidos, em 1945, na qual um grande instituto elaborou um plano de pesquisa via ligações telefônicas. O resultado anunciado pelo instituto era o da vitória do candidato Dewey sobre Harry Truman, com uma margem de vantagem de aproximadamente 10%. E o que ocorreu na realidade? Ocorreu a vitória de Truman. No dia seguinte à eleição, Harry Truman, que inclusive viria a se reeleger, deixou-se

fotografar com um jornal da véspera cuja manchete estampava em letras garrafais: “Dewey bate Truman”. Esse foi um dos episódios mais absurdos da história política americana.

Ao final desse relato, podemos perguntar aos alunos sobre o que eles avaliam ter sido o problema para o total fracasso da previsão. Sabemos que o problema ocorreu exatamente porque a pesquisa selecionou uma amostra completamente viciada, ao utilizar ligações telefônicas numa época em que só quem possuía telefone em casa eram pessoas de classe A ou B. Dessa forma, pessoas das classes menos favorecidas não foram pesquisadas, e a pesquisa serviu apenas para mostrar que o candidato que perdeu a eleição seria o mais votado nas classes A e B, mas não em toda a população de eleitores.

É importante, também, que o aluno acione o *link* e leia o texto que justifica o tamanho da amostra, “Por que 2400?” O texto é bastante simples e mostra um dos modos de se estabelecer um tamanho de amostra, que será usado na atividade.

Depois que a atividade virtual ocorrer, no momento de discutir a elaboração dos relatórios o professor poderá estabelecer uma boa conversa sobre pesquisas e como os institutos as realizam. Os institutos de pesquisa, como o IBGE[1], por exemplo, podem escolher suas amostras de forma simples, pelo sorteio, considerando as moradias. Partem do princípio de que todos moram em algum lugar e é lá que são feitas suas pesquisas. Todas as moradias têm igual chance de serem sorteadas para a pesquisa.

Há um tipo de amostragem, conhecida como *sistemática*, criada a partir de uma regra; pegar a lista telefônica de uma cidade e ir escolhendo os participantes de 10 em 10, segundo a lista, é um exemplo disso.

As amostras podem ser escolhidas depois que a população foi dividida em grupos mais homogêneos, ou *estratos,* sorteando ou sistematizando a escolha depois da população estratificada, que é o exemplo utilizado na atividade.

O professor pode discutir o exemplo que se segue:

Vamos supor que você vá fazer uma pesquisa em sua cidade e escolha uma amostra de 1000 pessoas. Para que ela corresponda à realidade da população, é necessário que todos os bairros sejam igualmente representados. Imagine que um bairro **A** *tenha uma população que corresponda a 8 % do total da cidade. Portanto,*

[1] Instituto Brasileiro de Geografia e Estatística

nessa amostra, deverão estar 80 pessoas do bairro **A** *, isto é, 8% do total da amostra, que poderão ser escolhidas por sorteio .*

Faz-se necessário deixar claro que, numa pesquisa real, a amostra bem escolhida pode evitar informações *tendenciosas* mas, ainda assim, é preciso alguns cuidados. Se você fizer uma pesquisa entre jovens que trabalham e só procurá-los no período diurno, não saberá o que pensam os que trabalham à noite. É preciso ter cuidado com o local, período, a formulação de questões, e até com quem faz a pesquisa. Há pesquisadores que inibem os pesquisados.

Por fim, os alunos poderão dispor de um certo tempo em sala de aula para a elaboração de seu relatório de pesquisa, cuja orientação é fornecida pelo próprio sistema. Nessa etapa, o professor poderá disponibilizar computadores para que os relatórios sejam elaborados com a ajuda de um editor de texto e de uma planilha eletrônica que os ajudaria a desenhar os gráficos.

Texto adaptado do original A ESTATÍSTICA E AS PESQUISAS ELEITORAIS , de *Flavio Wagner Rodrigues- IME – USP,* publicado na Revista do Professor de Matemática , número 40, 1999, da Sociedade Brasileira de Matemática.

Introdução

Neste artigo, serão discutidas algumas idéias intuitivas que estão por trás da Teoria Estatística da Estimação, que é a base teórica para a análise de pesquisas eleitorais. Serão apresentadas as principais fontes dos erros que podem ocorrer, discutindo-se também a possibilidade de que eles efetivamente ocorram.

Gostaríamos de deixar claro que nunca trabalhamos para nenhum instituto de pesquisa e nem temos nenhuma procuração para defendê-los. Temos, no entanto, duas fortes razões para acreditar que os poucos erros cometidos não foram intencionais. A primeira delas é a reputação dos institutos envolvidos, que têm uma longa história de seriedade e competência na realização de pesquisas. A segunda, mais pragmática, é que, embora as pesquisas eleitorais estejam longe de ser a

principal fonte de renda desses institutos, elas são um importante fator de prestígio, que contribui para que eles consigam projetos mais rendosos.

1. Universo e amostra

Serão consideradas apenas as pesquisas de intenção de voto, isto é, aquelas que são feitas antes da realização das eleições. As pesquisas de boca de urna (nas quais o eleitor que acabou de votar é entrevistado) e as pesquisas que se baseiam em contagens parciais já efetuadas não serão consideradas aqui.

Numa pesquisa de intenção de voto o conjunto de interesse (que os estatísticos chamam de universo) é formado por todos os eleitores aptos a votar naquela eleição. É claro que problemas de tempo e de custo tornam impraticável a consulta a todos os elementos desse conjunto. Temos que nos contentar em ouvir apenas uma pequena parcela dessa população e é esse conjunto de eleitores escolhidos para serem entrevistados que recebe o nome de amostra.

Para os estatísticos, uma boa amostra deve poder ser pensada como um retrato, em tamanho pequeno, do universo que está sendo considerado. Assim, por exemplo, nenhuma pessoa de bom-senso entrevistaria apenas moradores das mansões do Morumbi, em São Paulo, ou somente habitantes das favelas da periferia da cidade.

Os principais fatores utilizados para definir a composição da amostra são o nível sócio-econômico, grau de instrução, idade, etc. A escolha desses fatores é em grande parte determinada pela experiência passada, podendo em alguns casos refletir uma opinião pessoal do pesquisador que acredita que um determinado fator é importante para o problema considerado.

Resumindo, durante a realização de uma pesquisa existe uma proporção desconhecida de eleitores que pretendem votar num determinado candidato. Após a conclusão das entrevistas, obtemos a proporção de eleitores da amostra que manifestaram sua preferência por esse candidato.

2. Os problemas de interpretação da média

Provavelmente o conceito estatístico mais utilizado no dia-a-dia é a média. Expressões tais como renda média e vida média aparecem com freqüência nas nossas

conversas diárias, nos jornais e revistas e a televisão está sempre garantindo que 9 em cada 10 donas de casa preferem o sabão *X*.

Vamos recordar, através de um exemplo, a definição de média ou esperança matemática de uma distribuição de probabilidades. O lançamento de um dado perfeito admite como resultado qualquer um dos números 1, 2, 3, 4, 5 ou 6, a cada um deles sendo atribuída probabilidade $\frac{1}{6}$. A média dessa distribuição é definida como sendo a soma dos produtos de cada resultado possível pela probabilidade correspondente. Portanto:

$$M = \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = 3{,}5$$

Vamos considerar agora uma situação real na qual um dado perfeito é lançado 1000 vezes e calcula-se a média aritmética dos resultados obtidos. Essa média dificilmente será igual a 3,5, mas resultados bastante gerais nos permitem afirmar que a probabilidade de que ela se afaste muito de 3,5 é bastante pequena. Portanto, se a média teórica fosse desconhecida, esse experimento nos daria uma idéia sobre o seu valor. É importante observar que, ao contrário da média teórica, a média aritmética de 1000 observações não é constante, isto é, se alguém repetir esse experimento nas mesmas condições, irá, quase certamente, encontrar um valor diferente daquele que obtivemos.

É claro que o conhecimento apenas da média de uma distribuição não nos dá muita informação sobre ela. Assim, por exemplo, se em três faces de um dado perfeito for colocado o número 1 e nas outras três o número 6 (e portanto o 1 e o 6 irão aparecer com probabilidade $\frac{1}{2}$ cada um), a média dessa distribuição será também igual a 3,5, embora ela seja bastante diferente da distribuição associada a um dado comum. Como não poderia deixar de ser, a média nos dá apenas o centro da distribuição, não fornecendo nenhuma informação sobre como os demais valores se situam com relação ao centro. Para medir esse efeito, que os estatísticos chamam de variabilidade, a medida mais utilizada é a variância.

A variância de uma distribuição nunca é negativa e a determinação positiva da

raiz quadrada da variância recebe o nome de desvio padrão. É interessante observar que, embora existam infinitas distribuições com a mesma média e mesma variância, o conhecimento da média e da variância permite que se façam afirmações bastante gerais sobre os valores da distribuição. De fato, pode-se mostrar que o intervalo com centro na média e semi-amplitude igual a 2 desvios padrões contém, no mínimo, 75% dos valores da distribuição.

Quando dispomos de informações adicionais, essas estimativas podem ser bastante melhoradas. Assim, por exemplo, para variáveis contínuas com distribuição normal, esse mesmo intervalo conterá, no mínimo, 95% dos valores da distribuição. Esses resultados são bastante utilizados na clínica médica. São eles que possibilitam a construção das tabelas e dos gráficos que os pediatras utilizam para acompanhar o desenvolvimento das crianças com relação ao peso e à altura.

Os intervalos de normalidade para os resultados de exames laboratoriais são também determinados com base nessa teoria. Fica fácil agora explicar as brincadeiras que são feitas sobre a média. Dependendo do valor da variância é bastante provável que um rio cuja profundidade média é igual a um metro e meio tenha pontos onde a profundidade supere um metro e oitenta. Da mesma forma a variância da distribuição do tempo de vida do brasileiro mostra que não só é possível, como até bastante provável que alguém viva três ou quatro anos a mais. A única coisa a se lamentar é que também seja possível e até provável que muitos morram antes de atingir a idade média.

3. A determinação do intervalo de confiança

Nos meses que antecedem uma eleição encontramos com freqüência nos jornais informações que dizem que, de acordo com o instituto *X*, o candidato *A* tem 37% das intenções de voto e que a margem de erro da pesquisa é de dois pontos percentuais para mais ou para menos. Essa informação significa que, na amostra colhida pelo instituto, 37% dos entrevistados manifestaram sua preferência pelo candidato *A* e que, com uma probabilidade conhecida, que quase nunca é mencionada mas que geralmente vale 95%, o valor real da proporção de eleitores de *A* está compreendida entre 35 e 39%.

Para ver como esse intervalo é determinado, seja *p* a proporção de eleitores

que pretendem votar num candidato *A*. Vamos admitir que *p* é estritamente positiva e diferente de 1. Suponhamos que, numa amostra de *n* eleitores, *k* manifestem a intenção de votar em *A*. A proporção dos eleitores da amostra que pretendem votar em *A* será denotada por

$$p_1 = \frac{k}{n}$$

É claro que uma outra amostra de tamanho *n* irá, quase certamente, produzir um valor diferente para p_1. Utilizando a distribuição binomial de probabilidades podemos mostrar que a média de p_1 é igual a *p* e sua variância é igual a

$$\frac{p(1-p)}{n}$$

Um resultado teórico importante nos permite mostrar que, para valores grandes de *n*, p_1 tem uma distribuição aproximadamente normal. Uma consulta à tabela da normal mostra que, se *z* tem uma distribuição normal com média zero e variância 1, temos: $p(-1{,}96 < z < 1{,}96) = 95\%$

Segue-se que a probabilidade de que o intervalo $p_1 \pm 1{,}96\sqrt{\frac{p(1-p)}{n}}$ contenha o verdadeiro valor de *p* é aproximadamente igual a 95%. O problema que ainda resta é que os extremos desse intervalo dependem do valor desconhecido de *p*. Uma solução possível é aumentar o intervalo substituindo $p(1-p)$ pelo seu valor máximo, que é igual a $\frac{1}{4}$. Podemos então afirmar que a probabilidade de que o intervalo $p_1 \pm \frac{1{,}96}{2\sqrt{n}}$ contenha o verdadeiro valor de *p* é no mínimo igual a 95%.

Assim, por exemplo, se desejarmos uma confiança de 95% e uma margem de erro de dois pontos percentuais (para mais ou para menos), *n* deverá satisfazer:

$\frac{1{,}96}{2\sqrt{n}} = \frac{2}{100}$ e portanto *n* deverá ser igual a 2401.

Na determinação de um intervalo de confiança lidamos com três quantidades inter-relacionadas, que são as seguintes:

1. O tamanho da amostra *n*.
2. A precisão da estimativa que é definida pela amplitude do intervalo.

3. A confiança depositada no intervalo que é definida pela probabilidade de que o intervalo contenha o verdadeiro valor de p.

Assim, por exemplo, se o tamanho da amostra permanece fixo, um aumento da precisão implica necessariamente uma diminuição da confiança e reciprocamente. A única maneira de melhorar a precisão sem alterar a confiança é aumentar o tamanho da amostra. Analogamente, se estivermos dispostos a aceitar uma redução da confiança, a mesma precisão poderá ser obtida com uma amostra de tamanho menor. Se no exemplo anterior trabalharmos com uma confiança de 90% (o que corresponde a substituir o valor 1,96 por 1,64), o tamanho da amostra se reduzirá de 2401 para 1681.

Finalmente, é importante observar que a confiança e a precisão estão relacionadas com n e, assim, para manter a confiança e reduzir o intervalo à metade, nós vamos precisar de uma amostra quatro vezes maior. O preço a ser pago em termos de custos e do tempo necessário para obter as informações nem sempre compensa os ganhos de precisão.

4. A coleta da amostra ou onde mora o perigo

Nesta seção nós vamos discutir uma possível fonte de erro que muitas vezes não é sequer considerada pelos pesquisadores. Suponha que o número de elementos da amostra foi determinado, bem como os critérios que irão reger a sua composição. Resta definir o processo que será utilizado para selecionar os elementos que serão entrevistados. O saudoso professor José Severo de Camargo Pereira, que, entre os estatísticos que conheci, era o que mais sensibilidade tinha para os problemas dessa área, costumava contar uma história bastante ilustrativa sobre o que pode acontecer de errado no processo.

Um estudo foi realizado para determinar os gastos com alimentação de famílias de baixa renda na periferia da cidade de São Paulo. Pequenas vendas e mercados eram visitados, perguntando-se a pessoas escolhidas ao acaso o custo da compra que estavam fazendo no momento. Os valores encontrados na pesquisa foram significativamente maiores do que aqueles que eram esperados. Convidado para participar da análise dos resultados, o professor Severo descobriu que os pesquisadores entrevistavam a pessoa que se encontrava no caixa no momento em que eles chegavam à loja. A explicação para os valores mais altos estava no fato de

que quem gastava mais ficava mais tempo no caixa e tinha portanto uma probabilidade maior de ser incluído na amostra.

Nas pesquisas eleitorais, esse problema surge devido ao processo de seleção adotado pela maioria dos institutos, que consiste em entrevistar pessoas escolhidas entre as que passam pelos pontos mais movimentados das grandes cidades. É claro que, embora muita gente passe por esses pontos, existe um número maior de pessoas que raramente ou nunca passa por lá. Se por alguma razão esses dois grupos tiverem opiniões diferentes sobre a eleição, os resultados finais serão distorcidos. Infelizmente, no entanto, esse é um erro provável que é praticamente impossível de ser evitado. A adoção de um plano de amostragem por domicílios, que envolva a visita dos pesquisadores à casa do eleitor, teria um custo proibitivo e seria muito demorado em razão da rapidez que geralmente é exigida pelos patrocinadores das pesquisas eleitorais.