Estudo de Caso 01: A Sabedoria das Massas - Estimativa para Moedas em um Copo

Coordenador: Gustavo Vieira
Relator: Danny Tonidandel
Verificador: Alessandro Dias
Monitor: Bernardo Marques

1- Descrição do Problema

No livro *The Wisdom of Crowds* [1], são apresentados anedotas e estudos de caso que ilustram um argumento interessante: o julgamento médio de um conjunto diverso de indivíduos independentes tende a convergir para soluções corretas em alguns tipos de problemas. Essa idéia não é tão recente: em 1907, Francis Galton [2] relata na revista *Nature* um experimento no qual 800 pessoas tentaram estimar a massa de um boi. Ele notou que, apesar de palpites diversos, a média de todos ele refletiu com bastante proximidade valor real.

No entanto, a sabedoria das multidões não é infalível. Surowiecki expõe também diversas situações na qual o julgamento das massas errou e discute aspectos que influênciam a decisão. Segundo ele, dentre os fatores que podem prejudicar esse processo estão incluídos homogeneidade no grupo e ausência de independência entre as opiniões dos diversos indivíduos. Outros estudos subsequentes [3] reforçam a ideia de que a influência da opinião de alguns indivíduos sobre os demais pode prejudicar o resultado final.

A fim de explorar as ideias apresentadas, dois experimentos foram realizados [4]. No primeiro, foi apresentado a 29 alunos da disciplina de Planejamento e Análise de Experimentos um copo de 200ml contendo moedas diversas. Os alunos foram então instruídos a estimar a quantidade de moedas ali após 30 segundos de deliberação. Os palpites deveriam ser escritos secretamente em um papel e não deveria haver comunicação entre os indivíduos.

O segundo experimento foi similar. Dado um mesmo copo de 200ml preenchido apenas com moedas de 5 centavos, os alunos foram instruídos a estimar o valor total das moedas nele. Nesse experimento, no entanto, os palpites foram realizados em sequência e em voz alta.

A partir dos dados coletados, duas questões devem ser discutidas:

1) Qual a quantidade real de moedas no primeiro experimento?
2) Qual o valor real das moedas no segundo experimento?

2. Projeto Experimental

Para responder às perguntas propostas a partir dos dados obtidos, é necessário assumir como verdadeira a premissa de que o julgamento da multidão converge para o valor verdadeiro. Aplicada ao caso específico, espera-se que as estimativas dos estudantes estejam distribuídas em torno do número real de moedas no recipiente (experimento 1) e do valor total das moedas (experimento 2). Nesse caso, a média de todos os palpites refletiria a melhor estimativa para os valores reais desejados

Além disso, assumindo que a distribuição dos palpites indivuais seja normal, é possível determinar intervalos de confiança para os valores estimados, analisando o desvio padrão da amostra.

Conforme discutido anteriormente, existem outros fatores interessantes a serem considerados nos experimentos. A partir das diferenças entre os experimentos 1 e 2, seria interessante investigar diferenças decorrentes da influência entre indivíduos, presente no segundo experimento e não no primeiro.

No entanto, o fato do valor real ser desconhecido limita o alcance dos experimentos como forma de avaliar fatores que influenciam a estimativa de grupos. Além disso, não há um valor esperado para que as amostras possam ser avaliadas com testes de hipóteses. Para contornar essa limitação, foi realizado um procedimento que permite estabelecer uma quantidade esperada para os experimentos 1 e 2.

O grupo preencheu um copo de 200ml com moedas de valores diversos (5, 10, 25, 50 centavos e 1 real) e realizou a contagem das moedas. Em seguida, foram utilizadas apenas moedas de 5 centavos para encher outro copo de 200ml, e as mesmas foram também contadas. Os valores obtidos foram 130 moedas no primeiro caso e 182 moedas (equivalentes ao valor de R$9.10) no segundo.

Dessa forma, torna-se possível realizar teste de hipótese para os experimentos realizados, tomando como hipótese nula as estimativas obtidas através dos dos procedimentos descritos. Tem-se então:

$$\begin{cases} H_0 : \mu = 130, \\ H_1 : \mu \neq 130. \end{cases}$$

para o experimento 1 e

$$\begin{cases} H_0 : \mu = 9.10, \\ H_1 : \mu \neq 9.10. \end{cases}$$

para o experimento 2.

Nos testes, foi estabelecido o nível de significância desejado em 5% (i.e. $\alpha = 0.05$), que resulta em grau de confiança de 95%. Como menor efeito de significância prática, foi adotado $\delta_1^\star = 10$ para o experimento 1 e $\delta_2^\star = 0.50$ para o experimento 2.

É importante notar que os procedimentos realizados carregam incertezas. Não se sabe a distribuição de moedas utilizadas no primeiro experimento, de forma que a aleatorização foi utilizada para tentar bloquear esse fator. Além disso, há variabilidade na forma como as moedas são posicionadas e não exatidão na definição de "um copo cheio". Como não foi possível controlar esses fatores ou obter essas informações a respeito do experimento original, a confiabilidade das estimativas obtidas com o procedimento é prejudicada.

3. Análise Estatística

3.1 Análise Descritiva

No experimento 1, foi obtida uma média de aproximadamente 88.586, com desvio padrão 49.43. A partir daí, *assumindo as premissas discutidas como verdadeiras*, é possível afirmar que o intervalo [69.7840491, 107.3883647] possui 95% de chance de conter a quantidade real de moedas no experimento 1.

Com os dados do experimento 2, a média obtida foi 4.637 e o desvio padrão 1.586. Analogamente, afirma-se que o intervalo [4.0335744, 5.2402187] possui 95% de chance de conter o valor real em dinheiro no recipiente do segundo experimento.

3.2 Teste de Hipóteses

Para o experimento 1 com hipótese nula $H_0 = 130$, o teste de hipótese apresenta $p = 1.0520784 \times 10^{-4} << \alpha$, de forma que a hipótese nula é rejeitada. No experimento 2, $H_0 = 9.10$ e $p = 5.061044 \times 10^{-15} << \alpha$, o que também indica rejeição da hipótese nula. Para ambos os experimentos, portanto, é possível afirmar que as estimativas para hipóteses nula obtidas através dos procedimentos realizados não são confirmadas pelos dados obtidos.

Considerando os valores definidos $\delta_1^\star = 10$ e $\delta_2^\star = 0.50$, a potência $(1 - \beta)$ de ambos os testes calculada é 0.1820489 e 0.374285, respectivamente. A baixa potência implica em uma probabilidade alta de erros do Tipo II, isto é, falha em rejeitar uma hipótese nula falsa. Embora isso não seja um problema nos testes atuais (uma vez que a hipótese nula já foi rejeitada), seria desejável uma potência maior. Para obter potência de pelo menos 0.8 nos testes para os mesmos valores de $\delta_1^\star$ e $\delta_2^\star$, seriam necessários tamanhos amostrais de 194 e 81, respectivamente (assumindo que os desvios padrão nas amostras permanecessem os mesmos).

3.3 Validação das Premissas

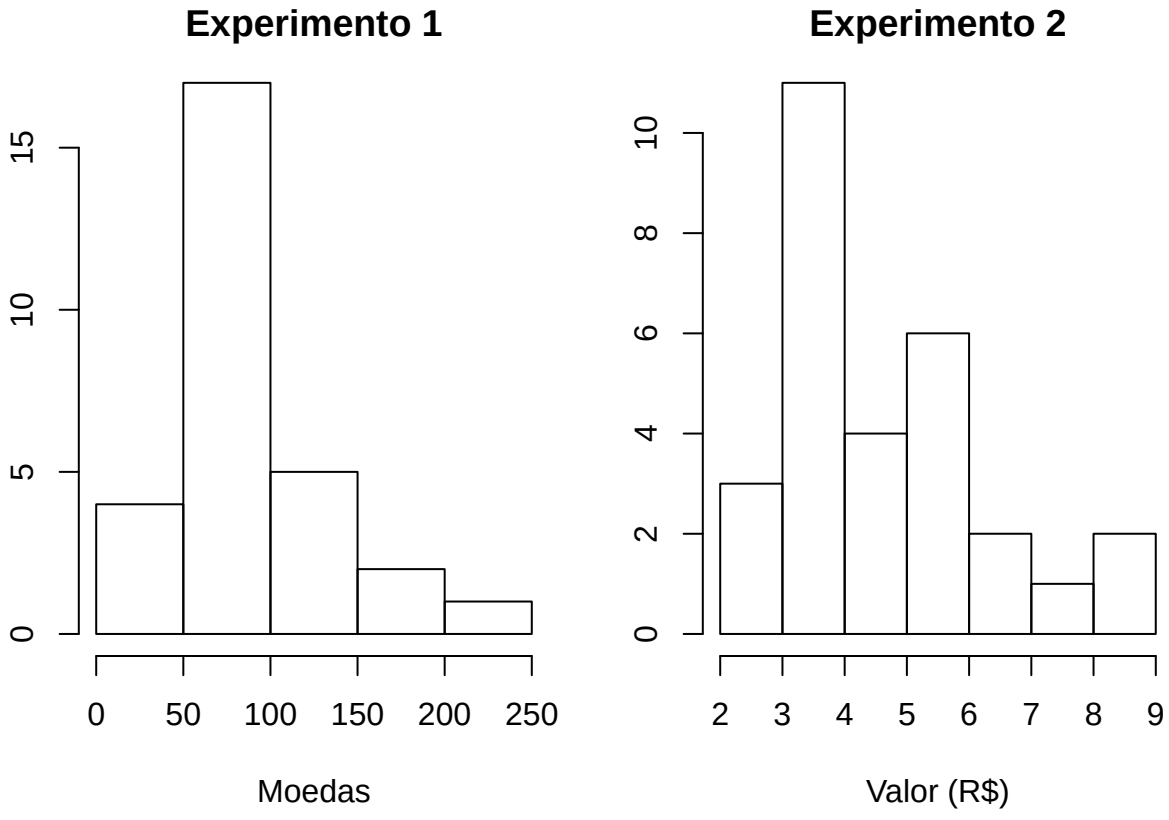
Todas as análises feitas até então assumem que as premissas estabelecidas são verdadeiras, isto é:

1) observações são distribuídas em torno do valor real considerado;
2) observações possuem distribuição normal;
3) observações são independentes entre si.

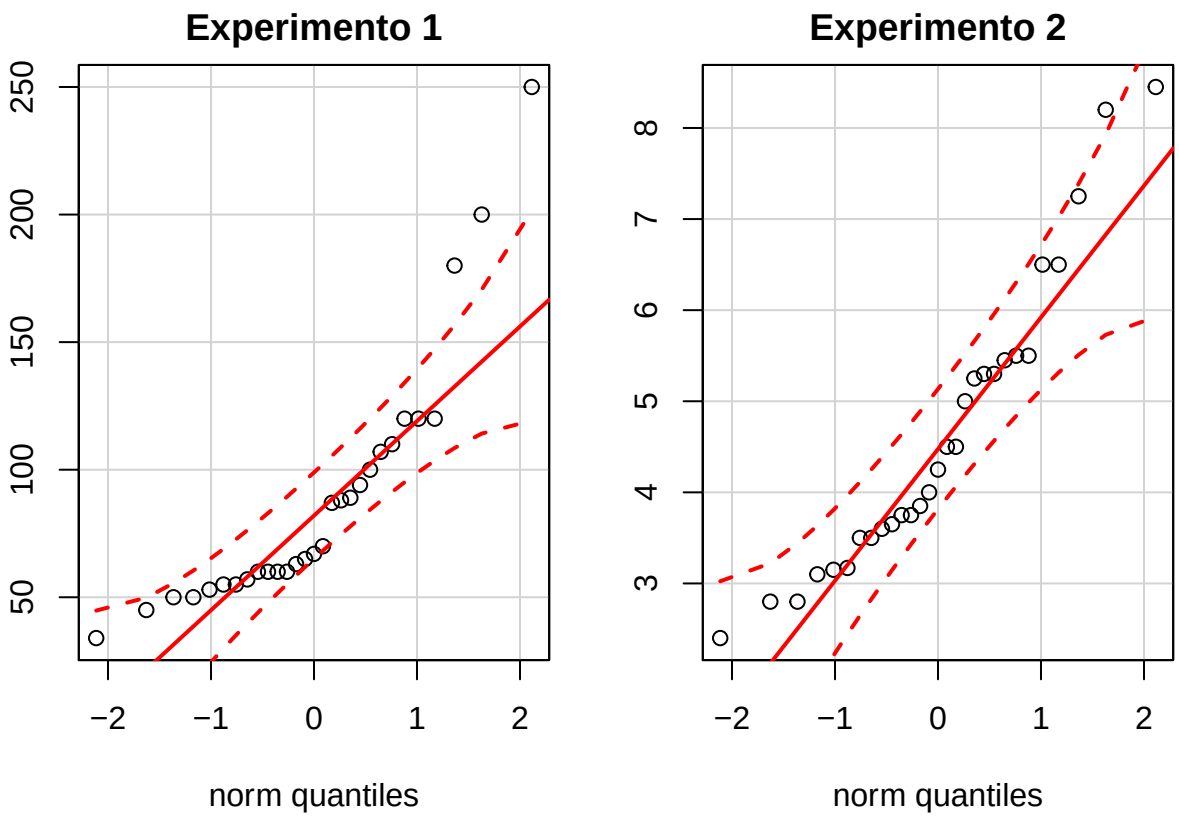
Não é possível verificar a premissa 1 sem conhecer os valores reais considerados (moedas no copo do experimento 1 e valor no copo do experimento 2). No entanto, existem testes que permitem avaliar a normalidade e interdependência em amostras

Normalidade

Os histogramas abaixo mostram a distribuição das observações nos experimentos e oferecem um primeiro indício da sua não normalidade:



Os QQplots abaixo comparam a distribuição das amostras em relaçõ à distribuição normal esperada (linha vermelha):



A análise dos gráficos indica que as distribuição das observações pode não ser normal. No entanto, essa análise ainda é inconclusiva, especialmente para o tamanho amostral considerado ($n = 29$) [5]. Para confirmar a não normalidade, é realizado o teste Shapiro-Wilk. Os resultados são $p = 7.551158 \times 10^{-5}$ e $W = 0.7975494$ para o experimento 1 e $p = 0.0333998$ e $W = 0.9215585$ para o experimento 2. Considerando $\alpha = 0.05$, o tamanho amostral $n = 29$ e a tabela de referência para o teste em [6], é possível rejeitar a hipótese nula (população tem distribuição normal) em ambos os experimentos, o que invalida uma das premissas assumidas.

Independência

Embora não haja um teste específico para independência das observações, é possível avaliar a autocorrelação serial através do teste de Durbin-Watson. No teste, foram obtidos valores $p = 0.9295731$ e $p = 0.2966482$ para os experimentos 1 e 2, respectivamente.

Esses resultados não rejeitam a hipótese de que as observações são independentes. No entanto, os procedimentos utilizados no experimento dois claramente violam o princípio de cegamento entre diferentes observações e permitem que influência entre amostras subsequêntes, o que é um indício de possibilidade de violação da independência. Analisando o desvio padrão relativo nos experimentos, tem-se respectivamente 0.5579873 e 0.3420618. O menor desvio padrão observado no segundo experimento pode ser consequência da influência entre observações, que tende a diminuir a variância e aumentar o viés [3].

Dessa forma, embora não seja possível provar a violação da premissa de independência, ela é enfraquecida para o experimento 2.

4. Discussão e Conclusões

A análise descritiva dos dados permite estabelecer intervalos de confiança para os valores desejados. Considerando as observações realizadas e assumindo as premissas estabelecidas como verdadeiras, é razoável assumir que os valores desejados estão entre os intervalos obtidos.

Para realizar o teste de hipótese, foram realizados procedimentos experimentais para determinar valores razoáveis para as hipóteses nulas. No entanto, os valores obtidos foram rejeitados pelos dados obtidos. Dessa forma, há uma contradição entre os valores esperados e os valores indicados pelos dados. Daí, duas conclusões são possíveis:

1) Os procedimentos experimentais realizados não são estimativas representativas dos valores reais utilizados nos experimentos realizados em sala e/ou
2) As premissas estabelecidas são falsas e, portanto, as análises realizadas não permitem determinar corretamente os valores considerados.

Os testes das premissas reforçam a segunda conclusão, uma vez que foi verificada a não normalidade das observações obtidas em sala. A análise dos procedimentos experimentais e dos desvios relativos também enfraquece a premissa de independência das observações no experimento 2, embora ela não seja possível provar sua violação.

Por outro lado, há também diversas fontes de incerteza nos procedimentos realizados para determinação da hipótese nula. A distribuição de moedas do experimento original é desconhecida, o método para determinar a quantidade que representa um copo cheio é incerta e a forma como as moedas são colocadas influencia muito o resultado. Dessa forma, não é possível confiar nos valores estimados também.

Dessa forma, a recomendação mais adequada seria que o experimento fosse refeito, com algumas melhorias procedimentais. Em primeiro lugar, é recomendado realizar cegamento dos indvíduos participantes, de maneira a evitar que uma observação afete as demais. Um maior tamanho amostral também é necessário para obtenção de potência maior nos testes, conforme discutido na seção 3.1. A replicação do experimento seria um fator positivo para identificar e eliminar erros experimentais e obter resultados mais confiáveis. Além disso, grupos mais heterogêneos são recomendados em [1]. Finalmente, a aleatorização na ordem dos experimentos poderia ajudar a eliminar possíveis bias entre estimativas subsequentes realizadas pelo mesmo indivíduo.

Em um experimento semelhante é descrito em [7], observa-se um outro fator que possivelmente interfere nos resultados. Os participantes que pensaram por mais tempo no problema e utilizaram matemática para estimar os valores erraram por uma margem maior que aqueles que apenas chutaram. Dessa forma, a variação no método utilizado para fazer a estimativa pode ser prejudicial. Para tentar bloquear esse efeito, sugere-se reduzir drasticamente o tempo disponível para dar um palpite para no máximo 10 segundos.

Mesmo com os procedimentos experimentais reformulados, ainda não é possível garantir que grupos de alunos seriam bons estimadores para a quantidade de moedas. Embora resultados da literatura sugiram que sim, podem haver fatores específicos desse problema que impedem a sabedoria das massas de prevalecer. Um pressuposto básico para o sucesso, por exemplo, é o conhecimento básico e habilidade dos indivíduos de chegar em conclusões corretas. Caso o problema seja simplesmente muito difícil, é esperado que o consenso do grupo falhe. Dessa forma, é necessário que o valor real das moedas seja registrado ao longo dos experimentos. Assim seria possível avaliar criticamente o método de estimativa a partir de grupos.

V. Atividades Desepenhadas

Todos os membros do grupo participaram da análise e planejamento do experimento e dos procedimentos para estimativa da hipótese nula. O código em R para testes foi majoritariamente desenvolvido pelo coordenador do grupo. O relatório foi escrito pelo relator, aprimorado pelo verificador e finalizado pelo coordenador.

Referências

[1] J. Surowiecki, *The wisdom of crowds*. Anchor, 2005.

[2] F. Galton, "Vox populi (the wisdom of crowds)," *Nature*, vol. 75, no. 7, pp. 450–451, 1907.

[3] J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing, "How social influence can undermine the wisdom of crowd effect," *Proceedings of the National Academy of Sciences*, vol. 108, no. 22, pp. 9020–9025, 2011.

[4] F. Campelo, "Estudo de caso 01," *Arquivo da disciplina Design and Analisys of Experiments*. 2017.

[5] D. C. Montgomery and D. C. Montgomery, *Design and analysis of experiments*, vol. 7. Wiley New York, 1984.

[6] "Teste de shapiro-wilk," *Portalaction.com.br*. 2017.

[7] E. B. Steiner, "Turns out the internet is bad at guessing how many coins are in a jar," *Wired Magazine*. 2015.