Estudo de Caso 01: estudantes da disciplina *Design and Analisys of Experiments* são bons estimadores para quantidade e valor de moedas colocadas em um copo?

Team 04

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

1- O experimento

Até que ponto a opinião de pessoas comuns, reunidas em grandes quantidades, podem revelar *"verdades"* acerca da natureza de determinado objeto ou fenômeno? Segundo Steiner [1], que realizou uma série de testes baseados no *best seller The Wisdom of Crouds* [2], o mais famooso experimento desta natrureza foi realizado pelo Cientista Vitoriano *Francis Galton*, em uma carta enviada à revista *Nature* [3], na qual analisa uma competição realizada em *Plymouth* (Inglaterra), em que diversas pessoas deveriam estimar a massa de um boi. Obviamente ninguém acertou exatamente o valor, mas a média das tentativas das quase 800 pessoas que participaram do concurso refletiu, com bastante proximidade, o real valor da medida procurada. E o que Steiner realizou foi testar a ideia utilizando-se de uma garrafa cheia de moedas, convidando pessoas que acessavam a internet a fazerem o mesmo, a partir de uma foto que mostrava a garrafa com as moedas. Mas seria isto verdade?

Da mesma forma podemos conjecturar que o experimento proposto pelo professor da disciplina *Design and Analisys of Experiments* foi inspirado nos mesmos experimentos. Com a diferença de que o material utilizado foram dois recipientes A e B, cheios de moedas, conforme descrito na referência [4] . O vigente estudo busca, portanto, investigar se as opiniões de 29 estudantes, isto é, o quanto a média das opiniões dos estudantes pode refletir o número e o valor real das moedas depositadas nos recipientes A e B?

2. Design Experimental

Como a média real não foi dada a conhecer pelo proponente do estudo, o time decidiu realizar uma montagem experimental semelhante (replicação do experimento), utilizando um recipiente de mesma natureza (copo plástico de $200ml$) para uma estimativa inicial do número de moedas no recipiente A, sabendo que era composto por moedas de natureza diferente (25 e 50 centavos e 1 real) e, no recipiente B, moedas de mesma natureza (5 centavos), utilizando contagem manual das moedas. O resultado seria utilizado como estimativa inicial para as médias: $Recipiente\,A$: 130 moedas; $Recipiente$ e
B : 9 reais e 10 centavos. Assim, formula-se a hipótese de que a média das estimativas dos estudantes é igual ou não ao "valor real":

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

e

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

O que forma um teste de hipóteses bilateral, considerando que os valores de $H_a e H_b$ são "reais", o que já é uma fonte de incertezas.

Por conveniência, julgou-se suficiente um nível de significância para o experimento de 5%, i.e., $\alpha = 0.05$, que implica em um grau de confiança $1 - \alpha$ de 95%. E a partir do conhecimento do time a respeito do problema em questão, adotou-se o menor efeito de significância prática como sendo $\delta_A^\star = 10$ moedas (para o caso A) e $\delta_B^\star = 0.50$ centavos (para o caso B). Além disso, o nível de potência estatística para o experimento escolhida foi (inicialmente) de $(1 - \beta) = 0.8$.

3. Teste de Hipóteses

O teste de hipóteses para os dois casos sugere rejeitar a hipótese nula, com $t_{0.05,28} = -4.51$ e $valor - p = 0.0001052 << \alpha$, no caso A, e $t_{0.05,28} = -15.15$ e $valor - p = 5.06e - 15 << \alpha$, para o caso B.

Em contrapartida, considerando-se o menor efeito de significância prática para os dois casos ($\delta_A^\star = 10$ moedas e $\delta_B^\star = 0.50$ centavos), realizou-se um teste de potência para determinar a sua sensibilidade em relação à ocorrência de erros do tipo II (falhar em rejeitar H_0 quando ela é falsa):

A potência do teste para o caso A obtida foi 0.18 e, para o caso B, foi de 0.37, considerando fixas o tamanho da amostra e o nível de significância. O teste demonstrou uma fraca sensibilidade para detectar erros do tipo II.

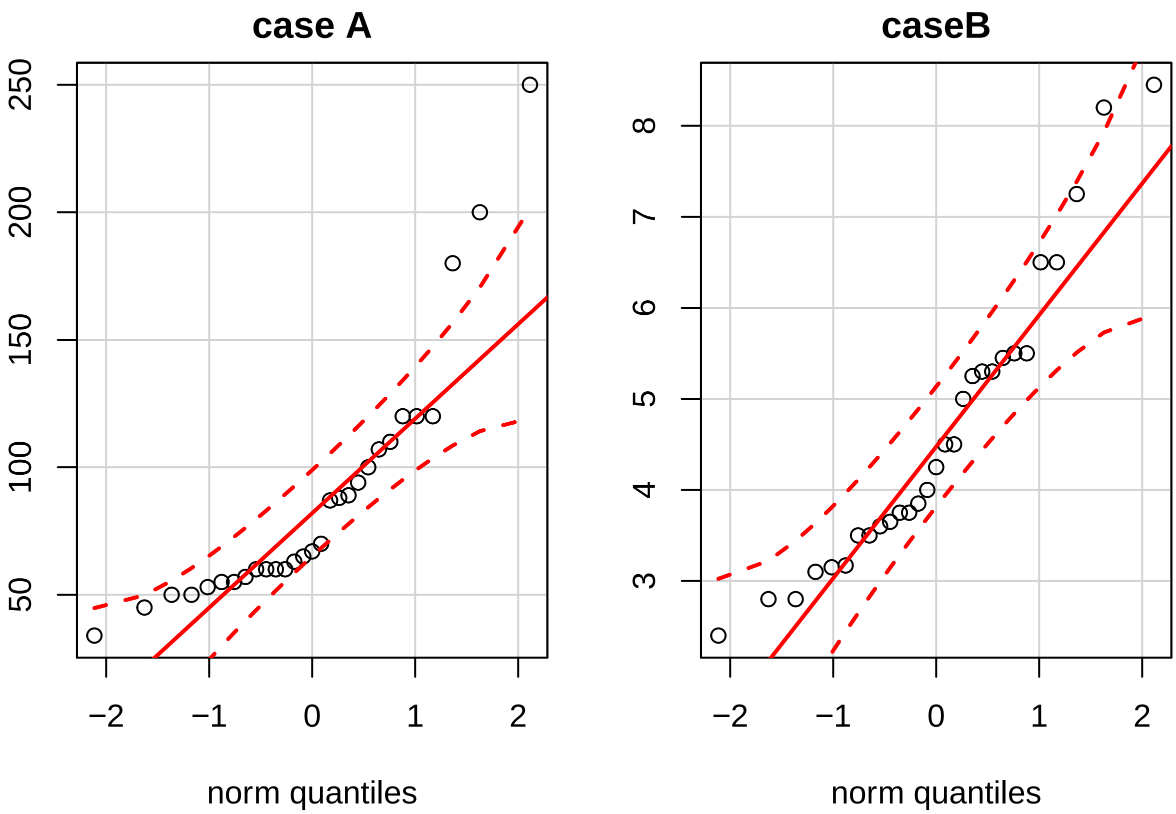
Como desejáva-se aumentar a potência do teste para o valor desejado de 0.8, observou-se, no caso A, que seria necessário um tamanho amostral de 194 e, para o caso B, 81, o que não é possível alcançar-se, já que a amostra única possui tamanho fixo. Logo, o aumento do número de observações seria desejável.

3. Validação

Como não se tem informações sobre a variância da população, o time escolheu adotar o teste de *t student*, assumindo as premissas: 1) As estimativas dos estudantes se distribuem em torno do valor real. 2) As observações são independentes. 3) A distribuição populacional das médias é normal.

Normalidade e independência são verificáveis por testes, porém, como a premissa de que a média dos *"chutes"* dos estudantes se aproxima do valor real não é facilmente testável *a priori*, acresenta-se mais um ponto de viéses para a análise.

A premissa de normalidade foi testada visualmente a partir dos gráficos *QQplots*:



o que demonstra que a população poderia não ser normal. Entretanto, dado o tamanho da amostra (29), não é possível descartar ainda a hipótese de normalidade, segundo Montgomery [6]. Nestes casos, apenas uma severa disparidade nos dados poderia ser um indicativo de não-normalidade. A recomendação seria a de se usar uma amostra um pouco maior. Assim, os supostos *outliers* não serão descartados e foi realizado o teste de *Shapiro-Wilk* para normalidade. Para o caso A, do número de moedas:

```
##
##  Shapiro-Wilk normality test
##
## data:  v_coins
## W = 0.79755, p-value = 7.551e-05
```

E para o caso B, para o valor:

```
##
##  Shapiro-Wilk normality test
##
## data:  v_value
## W = 0.92156, p-value = 0.0334
```

Para os dois casos, o valor de W_α e do valor p indicam que a população não segue uma distribuição normal.

Outro teste utilizado foi o de independência de Durbin-Watson:

```
dwtest(v_coins~1)
dwtest(v_value~1)
```

A hipótese nula desse teste afirma que a autocorrelação dos resíduos dos dados é zero. O p-valor alto obtido ($p >> \alpha$, nos dois casos) após o teste sugere que não há indícios para a rejeição da hipótese nula. Dessa forma, os autores concluem pela independência dos dados.

4. Análise descritiva e recomendações

A média dos valores obtida após a coleta das opiniões dos estudantes resultou em 88.5862069 moedas, com desvio padrão 49.4299773 no caso A e 4.6368966 reais, com desvio padrão de 1.5861051 no caso B, sendo que os valores "reais" eram, respectivamente, 130 e $9,10$.

O primeiro dado que, de certa forma, invalida o experimento é a observação do caráter de não normalidade da população, por meio dos testes. Neste caso, o teste de inferência de *t-student* não pode ser aplicado, pois não é possível aplicar-se o teorema do limite central.

No entanto, assumindo-se a premissa de normalidade, foi detectada uma baixa potência para o menor efeito de importância e nível de significância desejados e, por isso, uma repetiçao do experimento com um número maior de observações seria desejável.

A questão do cegamento dos participantes poderia ser repensada, pois acredita-se ter introduzido fontes de viéses na coleta dos dados. Uma sugestão seria a de se realizar uma replicação do experimento, utilizando grupos diferentes, altercando-se a ordem da coleta por grupo. Por exemplo, pergunta-se ao estudante 1 o número de moedas no recipiente A e o valor das moedas no recipiente B para o(a) estudante 2, e assim sucessivamente.

Outra sugestão seria a de diminuir o tempo de resposta de cada participante, de forma a reduzir diferenças entre simples "chutes" e possíves "cálculos mentais". Tais medidas poderiam, em princípio, sanar alguns do problemas em relação à normalidade da distribuição das médias.

Assim, a decisão final do time é a de que não há dados estatísticos suficientes para concluir que os alunos sejam bons estimadores para a quantidade ou valor das moedas depositadas nos recipientes A e B, nem garantir que, caso o experimento seja refeito segundo as recomendações, os objetivos serão, por isso, atingidos .

5. Considerações finais

Surowiecki, em seu estudo, lembra que a diferença não só contribui trazendo novas perspectivas para o ambiente, mas também ajuda os integrantes a expressarem mais livremente suas opiniões - sejam elas divergentes ou não [2, p. 38-39]. Isto revela o problema da coleta aberta no segundo caso, pois, não importa a magnitude do erro: mesmo que a "intuição" sugira o contrário, as pessoas dificilmente dariam respostas muito discrepantes da maioria. Gregory Berns, em seu *Iconoclast: A Neuroscientist Reveals How to Think Differently*[7] questiona inclusive a influência do grupo sobre a percepção das pessoas. Embora os estudantes garantirem terem dado a melhor resposta de acordo com suas observações, eles provavelmente questionavam suas convicções. Pode ser que alguns duvidassem daquilo que estavam vendo. Aparentemente as percepções permanecem intactas, mas a "fé" das pessoas nos seus sentidos, esta sim, parece ser moldada pela influência externa, alterando as decisões tomadas. E, no final, o que importa mesmo são as decisões. Vale ressaltar que o grupo experienciou um certo "alívio" ao saber que a experiência era por isso, de certo modo, uma pequena farsa.

Referências

[1] Steiner, E. B. *Turns Out the Internet Is Bad at Guessing How Many Coins Are in a Jar*. Wired Magazine: USA, 2017. Disponível em https://www.wired.com/2015/01/coin-jar-crowd-wisdom-experiment-results/

[2] Surowiecki, J. *The Wisdom of Crowds*. Anchor Books: New York, 2004.

[3] Galton, F. *Vox Populi*. Nature: England, mar. 1907.

[4] Campelo, F. *Estudo de caso 01.* Arquivo da disciplina Design and Analisys of Experiments. Disponível em https://goo.gl/b3IeAn.

[5] Ramirez, J.G. *Statistical Intervals: Confidence, Prediction, Enclosure.* Disponível em http://goo.gl/NJz7ot

[6] D.C. Montgomery, *Design and Analysis of Experiments*, 5th ed., John Wiley & Sons, 2001.

[7] Berns, G. *Iconoclast: A Neuroscientist Reveals How to Think Differently.* USA: Harvard Business press, 2008.

[8] Ramirez, J.G. *Statistical Intervals: Confidence, Prediction, Enclosure.* Disponível em http://goo.gl/NJz7ot

[9] D.C. Montgomery, *Design and Analysis of Experiments*, 5th ed., John Wiley & Sons, 2001.

[10] Felipe Campelo, *Lecture Notes on Design and Analysis of Experiments*, 2015.