Teste da hipótese Hardy-Weinberg utilizando a função dirichlet - MAP2212

Antonio Gabriel Freitas da Silva - 13687290
Guilherme Vaz das Neves Hummel - 13733732
Marco Antonio Soares de Campos - 13686469

Maio 2023

1 Introdução

Este trabalho tem como objetivo testar, para diversos valores inciais de x e y, a hipótese de Hardy-Weinberg por meio do cálculo do e-valor utilizando diversas dirichlet tridimensionais. Para cada valor inicial, será necessário calcular a função dirichlet, encontrar o θ^* mais próximo da hipótese, achar o seu valor na função verdade e por fim calcular o sev desse valor.

2 Cálculo da dirichlet

Queremos calcular a função definida por $W(v)$ aproximada por uma função condensada de massa de probabilidade $U(v)$, sendo aquela dada como:

$$W(v) = \int_{T(v)} f(\theta|x,y)d\theta \tag{1}$$

Além disso, o domínio da função será dada por um vetor definido por:

$$T(v) = \{\theta \in \Theta \,|\, f(\theta|x,y) \leq v\} \tag{2}$$

A função $f(\theta|x,y)$, que denotaremos por potencial, por fins práticos, é definida por uma função de Dirichlet de parâmetros x, y e θ dados por:

$$f(\theta|x,y) = \frac{1}{B(x+y)} \prod_{i=1}^{m} \theta_i^{x_i+y_i-1} \tag{3}$$

Sendo B uma distribuição Beta e $x, y \in \aleph^m, \theta \in \Theta = S_m = \{\theta \in \Re_m^+ \,|f(\theta|x,y) \leq v\,\}$, e θ um vetor de probabilidades que neste trabalho terá uma dimensão definida por $m = 3$ e será gerado a partir de um valor inicial θ_1,que ocupará o primeiro valor do vetor, $\theta_3 = \left(1 - \sqrt{\theta_1}\right)_2$ e $\theta_2 = 1 - \theta_1 - \theta_3$

2.1 Amostragem desejada e cálculo do número de bins

Foi utilizada uma aproximação assintótica através de uma distribuição Bernoulli com variância máxima de 0.25 e sua normalização em 95% de confiança e um erro $\varepsilon = 0.05\%$, a quatidade de bins será dada a partir de k, que reduz o erro que será parametrizado como:

$$W(v_j) - W(v_{j-1}) \approx \tfrac{1}{k} \leq \varepsilon \tag{4}$$

E pelo resultado do Teorema Central do Limite, teremos que:

$$n = (\tfrac{\sigma \cdot Z_{\alpha/2}}{\varepsilon})^2 \tag{5}$$

Dados que $\sigma^2 = 0.25$, $Z_{\alpha/2} = 1.96$ para a nossa amostragem geral, que torna:

$$n = \tfrac{3.8416 \cdot 0.25}{0.0005^2} = 3.841.600$$

Logo, precisamos de 3.841.600 de pontos para conseguirmos a precisão desejada.
Ao considerarmos apenas a igualdade, o cálculo dos bins foi feito da seguinte forma:

$$\tfrac{1}{k} = \varepsilon \implies k = \tfrac{1}{\varepsilon} \implies k = 2000$$

Nós utilizaremos o valor mínimo de bins (2000) para realizarmos a simulação.

3 Teste da hipótese

3.1 E-valor

Para iniciar o teste da hipótese temos que calcular o e-valor do x e y utilizados. Para fazer isso, calcularemos a dirichlet desses dados iniciais e com tamanho n e depois encontramos o θ que maximiza a hipótese, que chamaremos de θ^* . Por fim, o e-valor é obtido pelo seguinte cálculo;:

$$e - valor = 1 - W(v) = 1 - \int_{T(v)} f(\theta^x | x, y) d\theta$$

3.2 Calculo do Sev e aceitação do teste

Para aceitar ou recusar o teste devemos calcular o sev, e-valor padronizado, e para fazer isso utilizamos a seguinte fórmula:

$$sev\,(H|X) = 1 - QQ\left(t - h, QQ^{-1}\,(e - valor, z)\right)$$

Sendo QQ a cumulada de uma função χ^2, QQ^{-1} a função de ponto percentual da χ^2, t a dimensão de Θ que nesse trabalho será 2, h a dimensão do teste que será 1 e z o grau de liberdade da função percentual, que novamente será 1.

Após obtermos o sev, verificamos o seu valor e rejeitamos o teste se o valor for menor do que 0.05 e aceitamos se for maior do que 0.05.

4 Programa

Inicialmente no programa é pedido uma seed para a geração da dirichlet e são armazenados todos os valores obtidos pela tabela de testes de x e y em dois vetores bidimensionais, contudo para os valores de x é necessário calcular o x_2 pela seguinte fórmula:

$$x_2 = 20 - x_1 - x_3$$

Após isso, por meio de dois for, são selecionados todas as combinações entre um vetor de y com um vetor de x para fazer o teste. Inicialmente é chamado a função theta star que retorna o potencial do θ^*,que será chamado de máximo, que é obtido utilizando a função minimize da biblioteca optimize que testa valores delimitados em seu bound em uma certa função a fim de encontrar o valor inicial que retorna o maior valor possível, neste trabalho a função dada para o minimize gera um θ, obedecendo as regras colocadas nesse trabalho, a partir do valor obtido no bound e calcula o seu potencial.

Depois, é gerada a dirichlet utilizando a função Cal T com o x e y da iteração com n pontos e é calculado o e-valor por meio da função Cal v ,utilizando como argumentos o vetor de pontos da dirichlet e o valor do máximo, que calcula quantos pontos anteriores ao máximo existem no vetor e divide pelo tamanha do vetor . Por fim, é calculado o sev pela função sev cal utilizando a função chi2 da scipy.stats para calcular a cumulada e percentual da χ^2 e checa-se o seu valor, alterando a mensagem impressa dependo do valor e resetando os valores do vetor de pontos da dirichlet no final da iteração.

Um problema encontrado durante o desenvolvimento do programa foi observado quando os valores de y e x_3 resultavam em 0, fazendo que a função minimize para-se de funcionar. Por essa razão, certas iterações do programa resultam no estado que chamamos de anomalia e para evitar a quebra do programa essa iteração é pulada quando é encontrada.

5 Conclusão

Executando o programa utilizando a seed 13 obtemos que com $y = [0, 0, 0]$,o caso de não ter observação prévia, obtemos que 23 testes foram rejeitados,11 foram aceitos e houveram 2 anomalias. Já com $y = [1, 1, 1]$ foram rejeitados 26 testes, 10 foram aceitos e não houve nenhuma anomalia. Com esses dados concluímos que houve uma mudança significativa entre os testes já que houve uma mudança no número de testes aceitos e rejeitados, porém a principal mudança foi o número de anomalias que desapareceu no segundo teste , fazendo com que nenhuma testagem tenha sido perdida, obtendo assim uma análise mais completa.