Teste de Significância Totalmente Bayesiano para Hipóteses Precisas

Laboratório de Computação e Simulação

(MAP2212)

Alunos:	**Lucka de Godoy Gianvechio e Milton Leal Neto**
Curso:	Bacharelado de Matemática Aplicada Computacional
NUSP:	11352442 e 8973974
Professor:	Julio Stern

São Paulo, 20 de julho de 2021

Conteúdo

1 Apresentação

Este relatório apresenta uma solução para o sexto Exercício Programa (EP) proposto pelo professor Julio Stern no âmbito da disciplina de Laboratório de Computação e Simulação (MAP 2212) do Bacharelado de Matemática Aplicada Computacional (BMAC) do Instituto de Matemática e Estatística (IME) da Universidade de São Paulo (USP).

A título de registro, nos foi solicitado escrever um programa para computar o *e-valor* padronizado (SEV, na sigla em inglês) da hipótese de *Hardy-Weinberg* no modelo trinomial-Dirichlet para determinados conjuntos de observações (totalizando 72 conjuntos de dados) descritos no artigo original que apresentou este método, escrito por C.A.B Pereira e J.M. Stern em 1999.

A intenção do método é dar uma alternativa Bayesiana aos testes de significância (testes de hipóteses) ou, equivalentemente, à médida de evidência da estatística frequentista, o *p-valor*.

Para isso, um conjunto é definido no espaço paramétrico e a probabilidade a posteriori, sua credibilidade, é avaliada. Este conjunto é a "Região de Mais Alta Densidade a Posteriori"e é tangente ao conjunto que define a hipótese nula. A medida de evidência *e-valor* (EV) é o complemento da credibilidade na região tangente e sua padronização será explicada mais adiante.

Para cada conjunto de dados, computamos a função Verdade, $W(v)$, utilizada no EP 5, onde v é o argumento dado por s^*, que por sua vez é o $sup_{\theta \in H} s(\theta)$, ou seja, o valor de θ que maximiza a função surpresa $s(\theta)$ no espaço paramétrico H da hipótese nula.

Definimos a função surpresa como $s(\theta) = \frac{p_n(\theta)}{r(\theta)}$, onde $p_n(\theta)$ é a distribuição posteriori e $r(\theta)$ é a distribuição referencial, que neste EP é proporcional a 1. Temos que o *e-valor*, então, é dado por $ev(H|X) = W(s^*)$.

Vale a pena dizer que a hipótese de Hardy-Weinberg é definida como $H : \theta_3 = (1 - \sqrt{\theta}_1)^2, \theta_2 = 1 - \theta_1 - \theta_3$. Note que a dimensão do espaço paramétrico da hipótese nula é 1, enquanto que o espaço paramétrico da distribuição é 2. Como $dim(H) < dim(\Theta)$, temos uma hipótese precisa.

A partir do *e-valor* podemos chegar ao SEV, que é definido como $sev(H|X) = 1 - \overline{sev}(H|X)$, onde $\overline{sev}(H|X)$ é equivalente a $QQ(t, h, z)$, uma função definida como $QQ(t, h, z) \equiv Chi2(t - h, Chi2^{-1}(t, z))$, onde t é a dimensão do espaço paramétrico, h é a dimensão do espaço paramétrico da hipótese nula,z é o *e-valor* e Chi2 é a distribuição Chi-Quadrado.

Se H é verdadeira, isto é, se o valor do parâmetro θ^0 que gera as observações no modelo encontra-se em H, então, conforme o número de observações $n \to \infty$, a distribuição do $sev(H|X)$ converge para a distribuição uniforme em $[0,1]$, ou seja, $sev(H|X) \approx U_{0,1}$. Caso contrário, se $\theta^0 \notin H$, $sev(H|X) \to 0$. Dessa forma, $sev(H|X)$ passa a ter propriedades "similares"ao p-valor da estatística frequentista.

Após computar o SEV, nos foi solicitado decidir pela rejeição ou não rejeição da hipótese nula para cada um dos conjuntos de dados incluídos no artigo supracitado em dois casos diferentes em relação ao vetor de observações a priori, quando $y = [1,1,1]$, ou seja, quando existe algum conhecimento prévio sobre os dados e quando $y = [0,0,0]$, ou seja, quando não temos informação a priori.

Apenas para relembramos, definimos a função $W(v)$ como:

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

que representa a massa de probabilidade a posteriori no domínio $T(v)$, ou seja, a massa de probabilidade correspondente da função $f(\theta|x,y)$ que não ultrapassa um determinado nível v.

Também vale lembrar que a função f, que tem distribuição de probabilidades Dirichlet, dada por

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

representa o modelo estatístico m-dimensional Multinomial-Dirichlet e recebe como parâmetros um vetor de observações x, um vetor de informações a priori y e um vetor de probabilidades θ, sendo que $x, y \in N^m, \theta \in \Theta = S_m = \{\theta \in R_m^+ | \theta' 1 = 1\}$ e B representa a distribuição *Beta*. Aqui trabalhamos com $m = 3$.

Além disso, vale ressaltar que a região $T(v)$ foi definida como

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

2 Estratégia de resolução

Boa parte do que seria necessário ser feito neste EP 6 já havia sido construída nos EPs anteriores, como a função $U(v)$ que estima a função $W(v)$.Dessa forma, restou-nos definir apenas aspectos mais práticos para a implementação do método. Dentre estes, estava a maneira como iríamos realizar o processo de otimização no espaço paramétrico da hipótese nula. Em outras palavras, qual seria a estratégia para encontrar o $\theta \in H$ que maximasse a função supresa $s(\theta)$, que na prática neste EP pode ser enxergada como a função a posteriori da Dirichlet.

Optamos por utilizar uma função do pacote *Scipy* do *Python* chamada *Opitimize Minimize Scalar* restringida ao intervalo $[0, 1]$. Além dessa otimização, construímos a função que calcula o SEV a partir das fórmulas apresentadas no enunciado do EP.

3 Definindo n

Assim como no EP4 e EP5, também utilizamos o Teorema Central do Limite para estimar o valor ótimo de n que nos levasse à acurácia desejada. Para tanto, para cada um dos 72 conjuntos de dados, utilizamos a fórmula

$$n_{final} = \frac{\Phi^{-1}(1 - \frac{\delta}{2})^2 \cdot \hat{\sigma}^2}{\epsilon^2},$$

na qual $\Phi^{-1}(1-\frac{\delta}{2})$ corresponde ao percentil de uma distribuição $Normal(0, 1)$, $\hat{\sigma}^2$ corresponde à variância da amostra piloto e ϵ corresponde ao erro máximo suportado, no caso 0.0005. Neste trabalho, decidimos considerar $\delta = 95\%$.

Para estimarmos a variância, criamos 100 experimentos aleatórios nos quais geramos 1000 pontos aleatórios de uma Dirichlet, avaliamos a f nestes pontos, ordenamos os resultados e consideramos a média entre os dois menores valores obtidos em cada simulação como uma estimativa para a posição daquele que poderia ser considerado o v_1, ou seja, o primeiro nível de corte a ser feito na função.

Dessa forma, obtivemos 100 valores de v_1 e calculamos a variância destes valores e a utilizamos como variância amostral do nosso experimento. Assim como no EP4, obtivémos valor final final de n em torno de 20 a 30 mil pontos.

4 Resultados e Decisão do teste

Começamos com resultados do vetor de observações a priori $y = [1, 1, 1]$.

X1	X3	Y	EV	SEV	Decisão (alfa = 5%)
01	02	[1, 1, 1]	0.0032	0.0007	Rejeita H0
01	03	[1, 1, 1]	0.0133	0.0033	Rejeita H0
01	04	[1, 1, 1]	0.0384	0.0107	Rejeita H0
01	05	[1, 1, 1]	0.0910	0.0286	Rejeita H0
01	06	[1, 1, 1]	0.1786	0.0634	Não Rejeita H0
01	07	[1, 1, 1]	0.3075	0.1246	Não Rejeita H0
01	08	[1, 1, 1]	0.4772	0.2238	Não Rejeita H0
01	09	[1, 1, 1]	0.6638	0.3653	Não Rejeita H0
01	10	[1, 1, 1]	0.8338	0.5465	Não Rejeita H0
01	11	[1, 1, 1]	0.9538	0.7584	Não Rejeita H0
01	12	[1, 1, 1]	0.9998	0.9849	Não Rejeita H0
01	13	[1, 1, 1]	0.9615	0.7792	Não Rejeita H0
01	14	[1, 1, 1]	0.8437	0.5598	Não Rejeita H0
01	15	[1, 1, 1]	0.6682	0.3692	Não Rejeita H0
01	16	[1, 1, 1]	0.4665	0.2168	Não Rejeita H0
01	17	[1, 1, 1]	0.2760	0.1086	Não Rejeita H0
01	18	[1, 1, 1]	0.1308	0.0437	Rejeita H0
05	00	[1, 1, 1]	0.0135	0.0034	Rejeita H0
05	01	[1, 1, 1]	0.0920	0.0289	Rejeita H0
05	02	[1, 1, 1]	0.2917	0.1165	Não Rejeita H0
05	03	[1, 1, 1]	0.6045	0.3157	Não Rejeita H0
05	04	[1, 1, 1]	0.8893	0.6281	Não Rejeita H0
05	05	[1, 1, 1]	1.0000	1.0000	Não Rejeita H0
05	06	[1, 1, 1]	0.8993	0.6450	Não Rejeita H0
05	07	[1, 1, 1]	0.6640	0.3655	Não Rejeita H0
05	08	[1, 1, 1]	0.4028	0.1775	Não Rejeita H0
05	09	[1, 1, 1]	0.2006	0.0731	Não Rejeita H0
05	10	[1, 1, 1]	0.0843	0.0261	Rejeita H0
09	00	[1, 1, 1]	0.1937	0.0700	Não Rejeita H0
09	01	[1, 1, 1]	0.6613	0.3631	Não Rejeita H0
09	02	[1, 1, 1]	0.9930	0.9060	Não Rejeita H0
09	03	[1, 1, 1]	0.8530	0.5729	Não Rejeita H0
09	04	[1, 1, 1]	0.4906	0.2327	Não Rejeita H0
09	05	[1, 1, 1]	0.2037	0.0745	Não Rejeita H0
09	06	[1, 1, 1]	0.0587	0.0172	Rejeita H0
09	07	[1, 1, 1]	0.0139	0.0034	Rejeita H0

Vejamos agora resultados do vetor de observações a priori $y = [0, 0, 0]$.

X1	X3	Y	EV	SEV	Decisão (alfa = 5%)
01	02	[0, 0, 0]	0.0004	0.0001	Rejeita H0
01	03	[0, 0, 0]	0.0021	0.0004	Rejeita H0
01	04	[0, 0, 0]	0.0075	0.0017	Rejeita H0
01	05	[0, 0, 0]	0.0221	0.0057	Rejeita H0
01	06	[0, 0, 0]	0.0536	0.0155	Rejeita H0
01	07	[0, 0, 0]	0.1087	0.0351	Rejeita H0
01	08	[0, 0, 0]	0.1912	0.0689	Não Rejeita H0
01	09	[0, 0, 0]	0.2991	0.1203	Não Rejeita H0
01	10	[0, 0, 0]	0.4287	0.1931	Não Rejeita H0
01	11	[0, 0, 0]	0.5647	0.2850	Não Rejeita H0
01	12	[0, 0, 0]	0.6950	0.3937	Não Rejeita H0
01	13	[0, 0, 0]	0.8104	0.5168	Não Rejeita H0
01	14	[0, 0, 0]	0.8999	0.6461	Não Rejeita H0
01	15	[0, 0, 0]	0.9586	0.7713	Não Rejeita H0
01	16	[0, 0, 0]	0.9875	0.8737	Não Rejeita H0
01	17	[0, 0, 0]	0.9983	0.9529	Não Rejeita H0
01	18	[0, 0, 0]	0.9968	0.9361	Não Rejeita H0
05	00	[0, 0, 0]	NA	NA	NA
05	01	[0, 0, 0]	0.0221	0.0058	Rejeita H0
05	02	[0, 0, 0]	0.1295	0.0432	Rejeita H0
05	03	[0, 0, 0]	0.3843	0.1667	Não Rejeita H0
05	04	[0, 0, 0]	0.7280	0.4256	Não Rejeita H0
05	05	[0, 0, 0]	0.9665	0.7941	Não Rejeita H0
05	06	[0, 0, 0]	0.9710	0.8082	Não Rejeita H0
05	07	[0, 0, 0]	0.7663	0.4656	Não Rejeita H0
05	08	[0, 0, 0]	0.4896	0.2320	Não Rejeita H0
05	09	[0, 0, 0]	0.2496	0.0957	Não Rejeita H0
05	10	[0, 0, 0]	0.1001	0.0319	Rejeita H0
09	00	[0, 0, 0]	NA	NA	NA
09	01	[0, 0, 0]	0.2956	0.1185	Não Rejeita H0
09	02	[0, 0, 0]	0.8426	0.5583	Não Rejeita H0
09	03	[0, 0, 0]	0.9676	0.7976	Não Rejeita H0
09	04	[0, 0, 0]	0.6090	0.3193	Não Rejeita H0
09	05	[0, 0, 0]	0.2478	0.0948	Não Rejeita H0
09	06	[0, 0, 0]	0.0695	0.0209	Rejeita H0
09	07	[0, 0, 0]	0.0134	0.0033	Rejeita H0

Antes de compararmos os resultados, vamos esclarecer o critério utilizado para a tomada de decisão em relação ao teste.

De acordo com o artigo "The Rules of Logic Composition for Bayesian Epistemic e-Values", escrito por J.M. Stern e Wagner Borges, que lançou as bases do *e-valor* padronizado (SEV), esta medida possui propriedades assintóticas muito similares ao *p-valor* da estatística frequentista.

Por isso, decidimos fixar um nível de significância $alfa = 5\%$ e rejeitar toda hipótese que apresentasse valor de SEV inferior a 0.05.

Dessa forma, para o conjunto de dados a priori $y = [1, 1, 1]$, obtivemos 10 rejeições da hipótese nula e 26 não rejeições.

No caso do conjunto de dados a prior $y = [0, 0, 0]$, obtivemos 11 rejeições, 23 não rejeições e dois casos anômalos que denominamos por NA, por possuírem uma das componentes $x_i + y_i = 0$.

Podemos observar que no geral as decisões sobre as hipóteses nulas não se alteraram muito quando trocamos o vetor de observações a priori $y = [1, 1, 1]$ por $y = [0, 0, 0]$.

5 Estrutura do programa

O programa escrito em *Python* está estruturado em seis funções:

1) calcula_n_final():

Realiza o experimento para estimar a variância amostral e calcula o n_{final} a ser utilizado no programa.

2) calcula_f_():

Computa a constante de normalização, a função f e ordena os resultados obtidos.

3) estima_W():

Verifica quantos pontos existem abaixo do nível de corte desejado e retorna a proporção de pontos em relação ao total de pontos gerados como resultado da função $U(v)$, que estima a $W(v)$.

4) dirichlet():

Função auxiliar que calcula a função f quando da realização da otimização.

5) sev():

Calcula o *e-valor* padronizado.

6) main():

Chamada principal do programa. Inclui as linhas de código imprimem os resultados na tela e consistências do programa.

6 Conclusão

Podemos concluir que o Teste de Hipótese Totalmente Bayesiano para Hipóteses Precisas é uma excelente alternativa aos métodos frequentistas.

A metodologia de cálculo de evidências baseada em conjuntos de credibilidade apresentada neste trabalho é calculada no espaço paramétrico, considera apenas a amostra observada e traz consigo propriedades similares ao *p-valor*.

O método usado neste trabalho leva em consideração apenas a localização da máxima verossimilhança sob a hipótese nula. Esta metodologia também é independente da parametrização da hipótese nula.

Esta independência de parametrização dá à metodologia uma caracterização geométrica e está em nítido contraste com alguns procedimentos bem conhecidos, como o teste de hipóteses proposto por Ronaldo Fisher na década de 20 do século passado.

7 Referências

[1] C.A.B.Pereira, J.M.Stern, (1999). Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. Entropy Journal, 1, 69-80.

[2] Wagner de Souza Borges, Julio Michael Stern (2007). The Rules of Logic Composition for the Bayesian Epistemic e-Values. Logic Journal of the IGPL, 15, 5-6, 401-420.