Instituto de Matemática, Estatística e Computação Científica – IMECC
Universidade Estadual de Campinas – UNICAMP
Monografia PIBIC/CNPq

PROCEDIMENTOS BAYESIANOS ÓTIMOS PARA DETERMINAÇÃO DE TAMANHO DE AMOSTRA

Luís Guilherme de Oliveira Dantas RA: 147096

Orientadora: Prof.ª Dra. Laura Letícia Ramos Rifo

Julho de 2016

SUMÁRIO

Introdução....3

1. Testes de hipóteses e tamanho de amostra....5
   - 1.1. Teste de significância clássico....5
   - 1.2. Critério da ponderação dos erros....8
   - 1.3. Critério da ponderação dos erros sob a ótica bayesiana....10
2. Construção de funções para determinação de tamanho de amostra....13
   - 2.1. Hipóteses simples....15
   - 2.2. Hipóteses simples vs. bilateral....18
   - 2.3. Informação de Lindley....19
3. Dados composicionais....22
   - 3.1. Espaço simplex....23
4. Fundamentos da Geometria de Aitchison....25
   - 4.1. Princípios de Aitchison....25
   - 4.2. Razão logarítmica de proporções....26
   - 4.3. Distância de Aitchison....27
5. Ensaios clínicos....29
6. Amostragem intencional....31
   - 6.1. Alocação....32
7. Amostragem sequencial....36
   - 7.1. Teste sequencial da razão de probabilidades de Wald....36
   - 7.2. Retas de decisão....37
   - 7.3. Abordagem Bayesiana para o teste sequencial da razão de probabilidades de Wald....39

Conclusão....42

INTRODUÇÃO

Em qualquer análise estatística é de extrema importância que o estatístico utilize, durante todo seu estudo, as ferramentas mais práticas e eficientes para que seja possível encontrar os melhores caminhos que o levem à solução do problema. Deter conhecimentos de inferência estatística é imprescindível em inúmeros trabalhos, pois a partir de uma amostra buscamos extrair todas as informações necessárias para inferir a respeito de toda a população. Sendo assim, a amostra é o alicerce de todos estes tipos de estudos, uma vez que toda a análise é construída através dela. Para isso, é essencial considerar também que para uma análise ser proveitosa, ou seja, que consiga bons resultados, deve-se partir de uma amostra que tenha sido bem planejada e coletada, contendo as características do comportamento da população de estudo.

Uma das maiores dificuldades que está presente na prática estatística é a determinação do tamanho de amostra a ser estudada. Em ensaios clínicos, por exemplo, que serão discutidos posteriormente, trata-se de um problema rotineiro em seus planejamentos, pois os custos éticos e humanos bem como a disponibilidade de recursos para o estudo devem ser considerados, respeitando os limites que foram propostos inicialmente pelo pesquisador responsável do estudo. Em linhas gerais, podemos dizer que qualquer que seja o ponto de vista adotado, a solução para determinarmos o tamanho amostral exige um equilíbrio entre o custo experimental e a qualidade final da análise.

Podemos utilizar abordagens frequentistas e bayesianas para encontrarmos um tamanho de amostra conveniente. Uma das mais utilizadas é a aplicação do instrumental frequentista decorrente dos procedimentos de teste de hipótese em que podemos escolher o menor tamanho de amostra que permita atingir poder e nível de significância dados, por exemplo. Uma alternativa bayesiana para solucionar o mesmo problema pode ser escolher o menor tamanho de amostra possível para obter determinada probabilidade para a correta classificação das hipóteses consideradas. Neste trabalho, estaremos particularmente interessados em procedimentos bayesianos ótimos para determinarmos o tamanho da amostra.

Primeiramente apresentaremos uma das dificuldades que pode ser encontrada ao procurar o tamanho de amostra através de procedimentos de teste baseados no lema de Neyman-Pearson e como podemos contornar este problema, através de uma metodologia de minimização dos erros cometidos em testes de hipóteses que, na prática, pode ser aplicada sob o ponto de vista frequentista e bayesiano. Sequencialmente, apresentaremos alguns métodos de construção de funções para determinação de tamanho de amostra que podem ser utilizados tanto em uma análise frequentista como bayesiana. Introduziremos alguns conceitos básicos da análise de dados composicionais e da geometria de Aitchison e discutiremos um método de amostragem intencional para determinação do tamanho amostral aplicado em ensaios clínicos. Por fim, abordaremos o plano de amostragem sequencial no qual as unidades amostrais são inspecionadas uma a uma e, com base nos resultados obtidos, tomamos uma decisão a respeito das hipóteses do parâmetro de interesse do experimento ou inspecionamos uma nova unidade até que seja possível chegar a uma conclusão ou até quando o custo de coleta não fique extremamente dispendioso.

1) Testes de hipóteses e tamanho de amostra

1.1) Teste de significância clássico:

Quando precisamos tomar alguma decisão utilizando técnicas estatísticas, uma das soluções mais buscadas e utilizadas são os testes de hipóteses. Antes de prosseguirmos com a discussão, é preciso defini-los e ter em mente as definições dos erros presentes nestes tipos de teste.

Em diversos experimentos que necessitam de análise estatística o maior interesse dos pesquisadores é saber se os resultados observados oriundos de uma amostra contrariam, ou não, uma determinada afirmação (hipótese) que foi estabelecida inicialmente a respeito da população. Esta hipótese estatística é o que denominamos *hipótese nula* (H_0) e é usada para verificar se alguma hipótese estabelecida inicialmente poderá ser rejeitada ou não. Em geral, estas hipóteses são afirmações ou conjecturas a respeito do(s) parâmetro(s) da população ou até mesmo ao tipo ou natureza da população. A hipótese que utilizamos como alternativa à hipótese nula, ou seja, a que conjecturamos quando H_0 é rejeitada é a que chamamos de *hipótese alternativa* (H_1).

Num procedimento de teste, se a decisão for H_0, dizemos que: "a amostra observada não entrega evidências para rejeitar a hipótese nula". Mas se decidir por H_1, dizemos que: "a amostra observada entrega evidências para rejeitar a hipótese nula".

Definimos dois tipos de erros em um procedimento:

Erro do tipo I: rejeitar a hipótese nula H_0 quando ela é de fato verdadeira.

Erro do tipo II: não rejeitar a hipótese nula H_0 quando a hipótese alternativa H_1 é verdadeira.

As probabilidades associadas às ocorrências destes tipos de erros são usualmente denotadas por α e β, respectivamente.

Em planejamento e análise de experimentos, que tem grande aplicação em processos industriais, ensaios clínicos e pesquisas científicas, utiliza-se em grande escala o procedimento de teste baseado no lema de Neyman-Pearson, no qual se

estabelece um nível de significância α e a partir da informação extraída da amostra é calculada a estatística de teste de modo que β seja mínimo. Um dos mecanismos para controlar a probabilidade do erro do tipo I e minimizar o erro do tipo II é desenhar um experimento que seja capaz de encontrar um tamanho de amostra que respeite tais restrições. A seguir descrevemos um exemplo de teste de hipótese proposto em [4], no qual é determinado um tamanho amostral seguindo tal procedimento.

Exemplo 1: Suponha que $X_1, \dots, X_n$ sejam variáveis aleatórias normais independentes e identicamente distribuídas com desvio padrão $\sigma = 3$ e estamos interessados em comparar duas possíveis médias:

$$H_0: \theta = \theta_0 = -1 \; vs \; H_1: \theta = \theta_1 = 1.$$

É desejado planejar um experimento tal que a probabilidade do erro do erro do tipo I seja 0,05 e que a do tipo II seja no máximo 0,1.

Aplicando o procedimento de teste de Neyman-Pearson teremos o seguinte critério ótimo baseado na razão de verossimilhanças:

$$rejeita\ H_0\ se\ \frac{\bar{x} - (-1)}{3/\sqrt{n}} \geq k_\alpha$$

Em que $\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$ *, n o tamanho da amostra e* k_α *o quantil escolhido de forma que* $P\left(\frac{\bar{X} - (-1)}{3/\sqrt{n}} \geq k_\alpha | \, H_0\right) = \alpha.$

Sob a hipótese nula temos que $Z_0 = \frac{\bar{X} - (-1)}{3/\sqrt{n}} \sim N(0,1)$ e consequentemente $k_\alpha \cong 1{,}645$.

Agora definiremos o tamanho da amostra a partir do valor de β proposto no início do problema.

$$P\left(\frac{\bar{X} - (-1)}{3/\sqrt{n}} < 1{,}645 | \, H_1\right) = \beta$$

$$P\left(\frac{\bar{X} - 1 + 1 + 1}{3/\sqrt{n}} < 1{,}645 | \, \theta = 1\right) = 0{,}1$$

Sob a hipótese alternativa temos que $Z_1 = \frac{\bar{X}-1}{3/\sqrt{n}} \sim N(0,1)$, assim:

$$P\left(Z_1 < 1{,}645 - \frac{2\sqrt{n}}{3} \mid \theta = 1\right) = 0{,}1$$

$$P(Z_1 < z_\alpha \mid \theta = 1) = 0{,}1$$

Calculando o quantil z_α chegamos que $1{,}645 - \frac{2\sqrt{n}}{3} = -1{,}28$ e finalmente é possível encontrar que $n \cong 19{,}25 \rightarrow \boxed{n = 20}$

Portanto, ao utilizar o tamanho de amostra acima, note que a região crítica será rejeitar H_0 se $\bar{x} > 0{,}1$ e que $\beta = 0{,}091$. Veja também que nesta situação $\frac{\alpha}{\beta} = 0{,}55$, ou seja, temos que a probabilidade do erro do tipo II é aproximadamente duas vezes maior que a do tipo I.

Suponha agora que ao entregar este tamanho de amostra encontrado para os pesquisadores, eles dizem que podem analisar 100 unidades amostrais ao invés de 20, uma vez que o custo de coleta para ambos os tamanhos seria o mesmo. Vejamos como esta maior quantidade de dados impacta nos resultados da metodologia tradicional que foi empregada.

Como o nível de significância $\alpha = 0{,}05$ foi pré-estabelecido anteriormente, mantemos este valor. Porém ao substituirmos $n = 100$ para encontrar o valor da probabilidade do erro do tipo II teremos que $\beta = 2{,}6 \times 10^{-7}$. Consequentemente a razão $\frac{\alpha}{\beta}$ será aproximadamente igual a 195 que é extremamente superior ao que foi calculado com o tamanho de amostra $n = 20$, ou seja, apesar de mantermos o valor de α não temos controle sobre o de β. Além disso, temos a seguinte região de rejeição de H_0:

$$\frac{\bar{x} - (-1)}{3/\sqrt{100}} \geq 1{,}645 \Longrightarrow \bar{x} \geq -0{,}51$$

Note que se observarmos $\bar{x} = -0{,}5$ que é um valor mais próximo da que está sendo testada para o parâmetro populacional em $H_0: \theta_0 = -1$, rejeitaremos esta hipótese em favor da hipótese alternativa ao utilizar o critério acima.

Portanto, precisamos buscar uma abordagem que seja capaz de lidar efetivamente com estes problemas, independente da quantidade de informação disponível no estudo.

Nas duas seções seguintes apresentaremos alguns tópicos abordados em [10] com respeito a este paradigma de teste de hipótese clássico. Como alternativa a esta metodologia, eles descrevem uma mudança que foi proposta por DeGroot em [4], que basicamente procura minimizar uma combinação linear das probabilidades dos erros cometidos em testes de hipótese.

1.2) Critério da Ponderação dos Erros

Quando realizamos um teste de hipóteses queremos construir um teste que tenha o melhor poder possível, ou seja, com a menor probabilidade possível de cometer um erro de tipo I ou de tipo II. Uma das propostas mais utilizadas é limitar o tamanho do teste, no qual fixamos um nível de significância α e minimizamos o valor de β, como no exemplo anterior.

Ao invés de fixar a probabilidade do erro do tipo I e minimizar a do tipo II, uma abordagem conveniente de teste de hipótese é escolher o teste que minimize a soma ponderada das probabilidades dos erros:

$$min_{\delta}\left[a \cdot \alpha_{\theta_0}(\delta) + \ b \cdot \beta_{\theta_1}(\delta)\right] \qquad (1)$$

Em que δ denota o teste: $\delta(\boldsymbol{x}) = I_R(\boldsymbol{x})$, Ré a região de rejeição de H_0, I função indicadora (igual a 1 se $x \in R$ e 0 caso contrário).

Para testes com hipóteses simples, ou seja, $\Theta_0 = \{\theta_0\}$ *e* $\Theta_1 = \{\theta_1\}$, DeGroot prova que o seguinte teste é ótimo:

Teorema 1: Seja δ^*o teste que:

- Não rejeita H_0 se $a \cdot f(\boldsymbol{x}|\theta_0) > \ b \cdot f(\boldsymbol{x}|\theta_1)$;
- Rejeita H_0 se $a \cdot f(\boldsymbol{x}|\theta_0) < \ b \cdot f(\boldsymbol{x}|\theta_1)$;
- É indiferente entre H_0 *e* H_1 se $a \cdot f(\boldsymbol{x}|\theta_0) = \ b \cdot f(\boldsymbol{x}|\theta_1)$;

  Então, para qualquer outro procedimento de teste δ:

$$a \cdot \alpha(\delta) + b \cdot \beta(\delta) \geq a \cdot \alpha(\delta^*) + b \cdot \beta(\delta^*)$$

em que $f(\boldsymbol{x}|\theta_i), i = 0{,}1$ *denota as funções de verossimilhança e* a *e* b *são constantes.*

Os valores de a e b refletem a gravidade relativa entre os erros de tipo I e tipo II. Por exemplo, se ambos os erros tiverem o mesmo custo para o pesquisador então tomamos a = b; neste caso, o teste δ^* rejeita H_0 se e somente se $f(\boldsymbol{x}|\theta_0) < f(\boldsymbol{x}|\theta_1)$, ou seja, decidimos a favor da hipótese mais verossímil. Mas se o erro tipo I for mais grave que o erro tipo II, tomamos a > b, por exemplo, a/b = 2; neste caso, o teste δ^* rejeita H_0 se $2 \cdot f(\boldsymbol{x}|\theta_0) < f(\boldsymbol{x}|\theta_1)$.

Quanto à determinação da quantidade de unidades experimentais que devem ser observadas, que é o foco deste estudo, esta abordagem consegue lidar efetivamente com *qualquer* tamanho de amostra. Para visualizarmos isto, voltemos à situação apresentada no exemplo anterior:

Inicialmente, foi escolhido um tamanho de amostra $n = 20$, para um nível de significância fixo $\alpha = 0{,}05$ e $\beta = 0{,}091$. Utilizando a região de rejeição pelo critério da ponderação dos erros e algumas manipulações algébricas, temos que:

$$rejeite\ H_0\ se\ exp\left(\frac{n}{\sigma_0^2}(\theta_1 - \theta_0)\left[\frac{(\theta_0 + \theta_1)}{2} - \bar{x}\right]\right) < \frac{b}{a}$$

Neste caso, se substituirmos na expressão acima o valor de n que foi proposto no início do planejamento do experimento e também o valor crítico de $\bar{x}$ = 0,1 que havia sido encontrado na construção do teste usual, obteremos implicitamente que $\frac{b}{a} \cong 0{,}63$. Assim, se tomarmos $a = 1$, teremos que $b = 0{,}63$. Consequentemente, ao considerarmos a região acima temos que para qualquer tamanho amostral n:

$$A\ região\ de\ rejeição\ ótima\ R\ é: \bar{x} > \frac{2{,}07}{n}$$

Além disso, a razão entre as probabilidades dos erros de tipo I e tipo II pode ser expressa como função de n. A partir da figura 1 podemos enxergar que esta razão é bem estável conforme aumentamos o tamanho da amostra, variando de

0,55 em $n = 20$ a 0,61 em $n = 100$, resultado bastante diferente da razão de 195 encontrada na abordagem clássica.

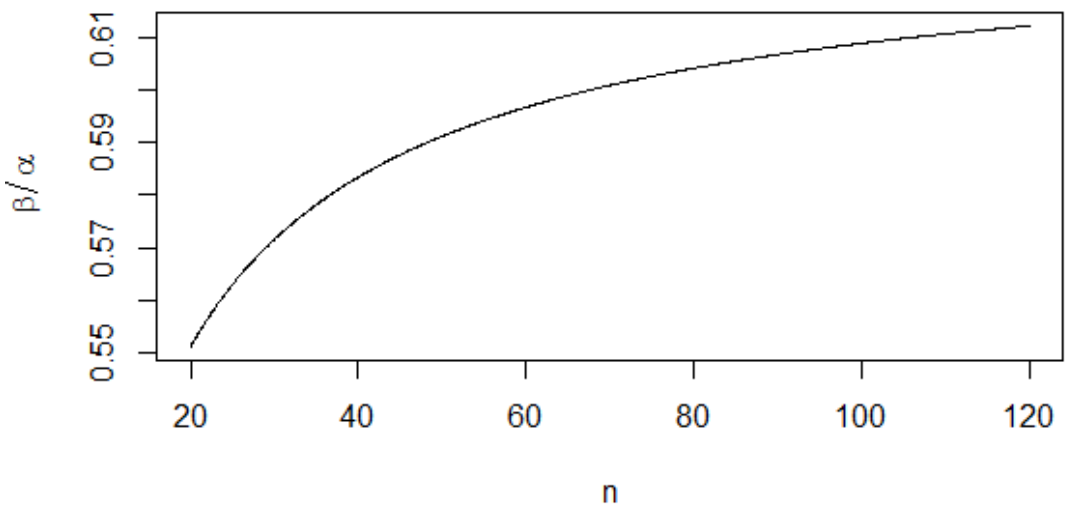

Figura 1: Gráfico da razão entre as probabilidades dos erros em função do tamanho amostral

1.3) Critério da ponderação dos erros sob a ótica bayesiana:

Suponha que queremos testar:

$$H_0: \theta \in \Theta_0 \; vs \; H_1: \theta \in \Theta_1$$

As medidas de evidência (verossimilhanças ponderadas) para os dados **y** sob cada hipótese são definidas como:

$$m_0(\boldsymbol{y}) = \int_{\Theta_0} f(\boldsymbol{y}|\theta)\pi_0(\theta)d\theta$$

$$m_1(\boldsymbol{y}) = \int_{\Theta_1} f(\boldsymbol{y}|\theta)\pi_1(\theta)d\theta$$

em que $\pi_j(\theta) \geq 0 \; \forall \; j$ são as densidades a priori dos valores dos parâmetros condicionados em cada hipótese.

A seguir apresentamos um procedimento de teste δ que minimiza, para valores de a > 0 e b > 0, a combinação linear (1).

Lema 1: Seja δ^* o teste definido como:

- Não Rejeita H_0 se $\frac{m_0(\boldsymbol{y})}{m_1(\boldsymbol{y})} > \frac{b}{a}$,
- Rejeita H_0 se $\frac{m_0(\boldsymbol{y})}{m_1(\boldsymbol{y})} < \frac{b}{a}$,
- Indiferente entre as duas hipóteses se $a \cdot m_0(\boldsymbol{y}) = b \cdot m_1(\boldsymbol{y})$

Então para qualquer procedimento de teste δ:

$$a \cdot \alpha(\delta) + b \cdot \beta(\delta) \geq a \cdot \alpha(\delta^*) + b \cdot \beta(\delta^*)$$

Neste caso, como podemos determinar o valor da razão $r = \frac{b}{a}$? Uma das abordagens mais utilizadas na inferência bayesiana é considerar a função de perda $a_0 - a_1$.

Definição 1: Definimos a função de perda $a_0 - a_1$ como:

$$L(\theta, \delta) = \begin{cases} 0 \; se \; \delta \neq I_{\Theta_0}(\theta) \\ a_0 \; se \; \theta \in \Theta_0; \delta = 1 \; (rejeita \; H_0) \\ a_1 \; se \; \theta \in \Theta_1; \delta = 0 \; (não \; rejeita \; H_0) \end{cases}$$

Sendo assim, podemos determinar o valor de r utilizando a razão da probabilidade a priori de H_1 vezes a perda sofrida pela falsa rejeição de H_1 sobre o produto entre a probabilidade a priori de H_0 e a perda sofrida pela falsa rejeição de H_0, ou seja:

$$r = \frac{b}{a} = \frac{P(H_1) \cdot a_1}{P(H_0) \cdot a_0}$$

Antes de apresentarmos o último resultado desta seção é preciso ter em mente um dos conceitos mais importantes da inferência estatística, o princípio da verossimilhança. Em linhas gerais, este princípio postula que em inferências sobre uma quantidade de interesse θ toda informação experimental relevante está contida na função de verossimilhança da amostra observada. Além disso, duas funções de verossimilhança contém a mesma informação sobre θ se elas forem proporcionais entre si. A seguir apresentamos este conceito sob a ótica bayesiana:

Princípio da Verossimilhança (bayesiano):

Se x_1 e x_2 são duas observações dependendo do parâmetro θ tais que $l_1(x_1|\theta) = c \cdot l_2(x_2|\theta) \; \forall \; \theta$ em que l_i é a verossimilhança da observação x_i e c constante, então $\pi(\theta|x_1) = \pi(\theta|x_2)$, em que $\pi(\theta|x_i)$ denota a probabilidade a posteriori de θ considerando a verossimilhança l_i.

Corolário 1: O teste pela minimização da soma ponderada dos erros automaticamente obedece o princípio da verossimilhança.

PROVA: Pelo Lema 1 temos que o teste ótimo é

$$Rejeitar\ H_0\ se\ \frac{m_0(\boldsymbol{y})}{m_1(\boldsymbol{y})} = \frac{\int_{\Theta_0} f(\boldsymbol{y}|\theta)\pi_0(\theta)d\theta}{\int_{\Theta_1} f(\boldsymbol{y}|\theta)\pi_1(\theta)d\theta} < \frac{b}{a}$$

Na razão $\frac{\int_{\Theta_0} f(\boldsymbol{y}|\theta)\pi_0(\theta)d\theta}{\int_{\Theta_1} f(\boldsymbol{y}|\theta)\pi_1(\theta)d\theta}$ a constante da função de verossimilhança se cancela.

2) Construção de funções para determinação de tamanho de amostra

Abordaremos nesta seção alguns procedimentos para determinação de tamanho de amostra apresentados em [4] no qual definimos uma função objetivo, frequentista ou bayesiana, para encontrar um tamanho de amostra que atenda às especificações desejadas pelo pesquisador do experimento. Ademais, discutiremos algumas relações que podem ser observadas circunstancialmente entre as abordagens frequentistas e bayesianas.

Podemos formalizar o objetivo de escolha de tamanho de amostra definindo uma função objetivo G que seja capaz de nos auxiliar na determinação do número de observações que deverão ser coletadas durante o experimento. Como exemplo desta função pode citar o poder do teste, a quantidade de informação, o tamanho do intervalo de confiança ou de credibilidade, erros de classificação, etc. Assim, esta função depende do tamanho de amostra n e também de outros elementos que dependem do cenário que está sendo utilizado na análise do experimento, frequentista ou bayesiano.

Denotamos como $G_F(n, \mathbf{u})$ a função objetivo frequentista em que $\mathbf{u}$ é um vetor cujos elementos são típicos de uma análise frequentista (nível de significância, comprimento desejado de intervalo de confiança, etc.); e $G_B(n, \mathbf{v})$ a função objetivo bayesiana em que $\mathbf{v}$ é um vetor cujos elementos são típicos de uma análise bayesiana (distribuição a priori das hipóteses ou do parâmetro de interesse, perda sofrida pela classificação incorreta, etc.). Assim, é de interesse construir um critério no qual seja encontrado o menor tamanho de amostra n tal que:

$$G_F(n, \mathbf{u}) \geq G_F^*, \qquad frequentista \quad (1)$$

$$G_B(n, \mathbf{v}) \geq G_B^*, \qquad bayesiano \quad (2)$$

em que G_F^* e G_B^* denotam, respectivamente, os níveis desejados das funções objetivo frequentista e bayesiana.

O problema de minimização do tamanho de amostra sob as restrições (1) ou (2) pode ser resolvido quando G_F e G_B são funções monótonas em n e as soluções podem ser obtidas através das expressões abaixo:

$$n_F^* = \min\{n \in N : G_F(n, \mathbf{u}) \geq G_F^*\}$$
$$= F(\mathbf{u}) \quad \textit{frequentista} \qquad (3)$$

$$n_B^* = \min\{n \in N : G_B(n, \mathbf{v}) \geq G_B^*\}$$
$$= F(\mathbf{v}) \quad \textit{bayesiano} \qquad (4)$$

em que N denota o conjunto dos número naturais e as funções F e B representam funções para determinação do tamanho amostral e dependem de **u** *e* **v**, *respectivamente.*

Para tamanhos amostrais iguais, as funções F e B definem, implicitamente, dois subconjuntos de valores de **u** *e* **v** tais que $n_F^* = n_B^* = n^*$. Observe o exemplo a seguir:

Exemplo 2: Suponha que dois estatísticos estavam interessados em testar $H_0 : \theta = \theta_0$ contra $H_1 : \theta = \theta_1$. Um estatístico optou por inferência frequentista e para tal decidiu utilizar o poder para determinar $n_F^*(\theta_1)$, enquanto que o outro, por inferência bayesiana, usou a informação esperada para calcular $n_B^*(\theta_1)$. O tamanho amostral encontrado por ambos estatísticos foi exatamente o mesmo e na prática isto não é algo incomum de acontecer, pois muitos planos amostrais independentemente se forem frequentista ou bayesiano, acabam determinando um mesmo tamanho de amostra. Entretanto, se θ_1 mudar, consequentemente $n_F^*(\theta_1)$ e $n_B^*(\theta_1)$ se alteram e é esperado que sejam diferentes. Mas, suponha agora que $n_F^*(\theta_1) = n_B^*(\theta_1)$ para todos os possíveis valores de θ_1. Isto evidenciaria uma forte equivalência entre as abordagens que utilizam o poder ou a informação esperada para determinar o tamanho de amostra.

A seguir, apresentaremos alguns casos de determinação de tamanho amostral em diferentes cenários de testes de hipótese para compararmos e identificarmos relações entre as abordagens frequentista e bayesiana. Iremos

considerar que $x_1, \dots, x_n$ é uma amostra aleatória $N(\theta, \sigma^2)$, com média θ desconhecida e variância σ^2 conhecida. Além disso, utilizaremos a notação usual para as probabilidades de cometer os erros de tipo I e tipo II, α e β, respectivamente, e assumiremos perda $0 - 1 - K$ para o caso bayesiano.

Tabela 1: Função de perda $0 - 1 - K$

	H_0 é verdadeiro	H_1 é verdadeiro
Não rejeitar H_0	0	1
Rejeitar H_0	K	0

Note que a perda sofrida é igual a 0 se tomar a decisão correta, 1 se não rejeitar H_0 quando H_1 é verdadeira e se rejeitar H_0 quando ela de fato era verdadeira a perda sofrida será igual a K. Vale ressaltar que na prática o valor de K é definido pelo pesquisador do experimento de modo que reflita a gravidade ao cometer tal erro.

2.1) Hipóteses simples

Suponha que queremos testar as hipóteses $H_0: \theta = \theta_0$ contra $H_1: \theta = \theta_1$, em que $\theta_1 > \theta_0$. Como vimos na seção anterior, sob a abordagem frequentista, fixamos um nível de significância e construímos um teste que minimize a probabilidade de erro tipo II. Considerando que $\bar{x}$ seja uma estatística suficiente, ao utilizar o lema de Neyman-Pearson a região crítica obtida é $\bar{x} \geq \theta_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$, de modo que $\Phi(z_\alpha) = P(Z \leq z_\alpha) = \alpha$.

Como o poder do teste $(1 - \beta)$ foi definido como a função objetivo frequentista G_F, poderemos encontrar, após algumas manipulações algébricas, que o tamanho amostral será:

$$n_F = \left(z_\alpha + z_\beta\right)^2 \left(\frac{\sigma}{\delta}\right)^2, \qquad (5)$$

$$em\ que\ \delta = \theta_1 - \theta_0$$

Assim, em termos do que foi definido anteriormente, o vetor de entrada frequentista para determinação do tamanho de amostra é $\mathbf{u} = (\alpha, \sigma, \delta)$.

Sob a metodologia de testes de hipóteses bayesiana, a incerteza sobre θ é representada por sua distribuição a priori e, neste caso, assuma que $P(\theta_0) = 1 - P(\theta_1) = \pi$. Uma decisão bayesiana entre H_0 e H_1 é baseada em suas probabilidades a posteriori.

A perda esperada a posteriori é definida por:

$E^{\pi}(L(\theta, \varphi)|x) = \int_{\Theta} L\big(\theta, \varphi(x)\big)\pi(\theta|x)d\theta$

em que $\varphi(x) = 1$ *é a decisão de rejeitar* H_0 *e* $\varphi(x) = 0$ *não rejeitar* H_0.

Assim, sob perda $0 - 1 - K$ temos que:

$E^{\pi}(L(\theta, \varphi)|x) = \int_{\Theta_0} L\big(\theta, \varphi(x)\big)\pi(\theta|x)d\theta + \int_{\Theta_1} L\big(\theta, \varphi(x)\big)\pi(\theta|x)d\theta$

x: $\varphi(x) = 1$ *(rejeitar* H_0*):*

$E^{\pi}(L(\theta, \varphi)|x) = KP(\theta \in \Theta_0|x)$;

x: $\varphi(x) = 0$ *(não rejeitar* H_0*):*

$E^{\pi}(L(\theta, \varphi)|x) = P(\theta \in \Theta_1|x) = 1 - P(\theta \in \Theta_0|x)$;

Minimizando a perda esperada resulta que:

$\varphi^* = 0 \Leftrightarrow 1 - P(\theta \in \Theta_0|x) < KP(\theta \in \Theta_0|x) \Leftrightarrow P(\theta \in \Theta_0|x) > \frac{1}{1+K}$

$\varphi^* = 1 \Leftrightarrow KP(\theta \in \Theta_0|x) < 1 - P(\theta \in \Theta_0|x) \Leftrightarrow P(\theta \in \Theta_0|x) < \frac{1}{1+K}$

Podemos determinar um limitante para a probabilidade a posteriori de H_0 que seja consistente com o resultado da minimização encontrado acima, ou seja, não rejeitar a hipótese nula se o valor de sua probabilidade a posteriori for pelo menos $\frac{1}{1+K}$. Considerando essa regra de decisão, podemos escrever que H_0 não é rejeitada se:

$$\bar{x} \leq \frac{\sigma^2 \log\left(K\frac{\pi}{1-\pi}\right)}{n\delta} + \frac{\theta_1 + \theta_0}{2}.$$

Neste exemplo, a função objetivo bayesiana escolhida é a taxa de classificação correta das hipóteses e o vetor de entrada bayesiana é $\mathbf{v} = (\pi, K, \delta, \sigma)$. Então:

$G_B(n, \mathbf{v}) = KP(H_0)P(decisão\ correta|H_0) + P(H_1)P(decisão\ correta|H_1)$

$$= K\pi P_{\theta_0}\left(\bar{x} \leq \frac{\sigma^2 \log\left(K\frac{\pi}{1-\pi}\right)}{n\delta} + \frac{\theta_1+\theta_0}{2}\right)$$

$$+(1-\pi)P_{\theta_1}\left(\bar{x} > \frac{\sigma^2 \log\left(K\frac{\pi}{1-\pi}\right)}{n\delta} + \frac{\theta_1+\theta_0}{2}\right)$$

$$= K\pi P\left(Z \leq \frac{\sigma \log\left(K\frac{\pi}{1-\pi}\right)}{\sqrt{n}\delta} + \frac{\delta\sqrt{n}}{2\sigma}\right)$$

$$+(1-\pi)P\left(Z \geq \frac{\sigma \log\left(K\frac{\pi}{1-\pi}\right)}{\sqrt{n}\delta} - \frac{\delta\sqrt{n}}{2\sigma}\right) \qquad (6)$$

em que $Z \sim N(0,1)$.

Note que esta função objetivo pondera a probabilidade em fazer uma decisão correta pelas probabilidades a priori em alcançar cada decisão. Além disso, a utilidade de não rejeitar H_0 corretamente não é necessariamente a mesma que rejeitá-la corretamente, uma vez que depende do valor de K.

Podemos observar que há elementos comuns entre esta abordagem com a frequentista na qual se controla o erro de tipo II fixado o de tipo I. Entretanto, ao invés de considerar as probabilidades α e β, a função objetivo bayesiana utiliza uma probabilidade a priori para a hipótese nula, uma função de perda e um valor que assegure uma taxa mínima (r^*) desejada de classificação correta.

Para exemplificarmos uma determinação do tamanho amostral, suponha que $\sigma^2 = 1$ e que $\delta = 0{,}10$. Assumindo que $\alpha = 0{,}05$ e $\beta = 0{,}10$, um frequentista obtém $n_F = 857$; um bayesiano ao considerar que $\pi = 0{,}5$, $K = 1$ e $r^* = 0{,}9283$ encontrou o mesmo tamanho, $n_B = 857$ (neste caso, n_B pode ser obtido computacionalmente pela equação (6), calculando um valor de n tal que $G_B^* = r^*$). Neste caso, uma equivalência entre as duas abordagens pode ser visualizada ao substituirmos $n = n_F$ da equação (5) em (6), obtendo:

$$G_B(n = n_F, \mathbf{v}) = K\pi P\left(Z \leq \frac{\log\left(K\frac{\pi}{1-\pi}\right)}{|z_\alpha + z_\beta|} + \frac{|z_\alpha + z_\beta|}{2}\right)$$

$$+(1-\pi)P\left(Z \geq \frac{\log\left(K\frac{\pi}{1-\pi}\right)}{|z_\alpha + z_\beta|} - \frac{|z_\alpha + z_\beta|}{2}\right), \qquad (7)$$

Portanto, $G_B(n = n_F, \mathbf{v})$ é algebricamente constante como uma função de δ, ou seja, a igualdade $n_F = n_B$ ocorre para todos os valores de δ. Este é um exemplo de uma correspondência formal entre as entradas frequentista e bayesiana.

2.2) Hipóteses simples vs. bilateral

Considere agora a hipótese nula $H_0: \theta = \theta_0$ vs a hipótese bilateral $H_1: \theta \neq \theta_0$. Adotando o poder do teste como função objetivo frequentista, o tamanho mínimo da amostra para detectar uma diferença δ com probabilidades fixas de erros tipo I e II é ligeiramente diferente da equação (5):

$$n_F = \left(z_{\alpha/2} + z_\beta\right)^2 \left(\frac{\sigma}{\delta}\right)^2, \qquad (8)$$

Em análise Bayesiana é comum adotar prioris que são simétricas em torno de θ_0, atribuindo assim a mesma probabilidade prévia para as regiões nula e alternativa. Se assumirmos que θ é uma priori $N(\theta_0, \tau^2)$ e que $P(H_0) = 1 - P(H_1) = \pi$ facilitaremos a ilustração da relação entre as duas abordagens. O método bayesiano não rejeita a hipótese nula quando a probabilidade posteriori é de pelo menos $\frac{1}{1+K}$. Desenvolvendo os cálculos necessários, temos que H_0 não será rejeitada quando $\bar{x} \in I = (-y_n + \theta_0, y_n + \theta_0)$ em que:

$$y_n = \frac{\sigma}{\sqrt{n}}\sqrt{2\left(1 + \frac{\sigma^2}{n\tau^2}\right)\left(\log\left(K\frac{\pi}{1-\pi}\right) + \frac{1}{2}\log\left(1 + \frac{n\tau^2}{\sigma^2}\right)\right)}.$$

A taxa de classificação correta é

$$G_B = (n, \boldsymbol{v}) = K\pi P_{\theta_0}(\bar{x} \in I) + (1-\pi)\int_{-\infty}^{\infty} P_\theta(\bar{x} \notin I)\pi(\theta)d\theta. \qquad (9)$$

Pelo Apêndice A.3 em [4] temos que:

$$G_B(n, \boldsymbol{v}) = [1 + (K-1)\pi] - 2K\pi\Phi\left(-\frac{y_n\sqrt{n}}{\sigma}\right)$$

$$+(1-\pi)\int_{-\infty}^{\infty}\left[\Phi\left(\frac{(-y_n-u)\sqrt{n}}{\sigma}\right) - \Phi\left(\frac{(y_n-u)\sqrt{n}}{\sigma}\right)\right]\frac{1}{\sqrt{2\pi}\tau}e^{-\frac{u^2}{2\tau^2}}du \quad (10)$$

O conjunto de entrada frequentista é novamente $\boldsymbol{u} = (\alpha, \delta, \sigma)$. Na bayesiana a busca pelo tamanho mínimo da amostra com taxa r^* de decisões corretas requer, a partir de (10), a entrada $\boldsymbol{v} = (\pi, \tau, K, \sigma)$. Ao considerarmos o caso em que $\tau^2 = C\delta^2$ (C uma constante positiva) com $n = n_F$ na equação (8), temos que:

$$G_B(n = n_F, \boldsymbol{v}) = [1 + (K-1)\pi] - 2K\pi\Phi(-A)$$
$$+ (1-\pi)\int_{-\infty}^{\infty}[\Phi(-A - tB) - \Phi(A - tB)]\frac{1}{\sqrt{2\pi C}}e^{\frac{t^2}{2C}}dt, \quad (11)$$

em que

$$A = \sqrt{2\left(1 + \frac{1}{CB^2}\right)\left(\log\left(K\frac{\pi}{1-\pi}\right) + \frac{1}{2}\log(1 + CB^2)\right)}$$

$e\ B = \left|z_{\alpha/2} + z_{\beta}\right|$ (veja Apêndice A.3 em [4] para mais detalhes).

Novamente, essa quantidade é constante como uma função de δ e depende de $z_{\alpha/2}$ e z_{β}.

2.3) Informação de Lindley

Considere uma observação X com distribuição $p(x|\theta)$. Lindley (1956) propôs medir a quantidade de informação sobre θ fornecida pelo resultado do experimento usando a quantidade

$$\mathcal{I} = E\{\log\pi(\theta|X)\} - E\{\log\pi(\theta)\} = E\left\{\log\left(\frac{\pi(\theta|X)}{\pi(\theta)}\right)\right\}, \quad (12)$$

em que as esperanças são calculadas com respeito à distribuição conjunta de θ e X. Esta quantidade quantifica a diferença entre o conhecimento disponível sobre θ antes e depois do experimento, representada pela distribuição a priori $\pi(\theta)$ e pela distribuição a posteriori $\pi(\theta|X)$, respectivamente.

Um dos conceitos primordiais de teoria da informação é chamado de entropia que pode ser entendido como uma medida de incerteza de uma variável aleatória. A entropia $H(X)$ de uma variável X é definida por

$$H(X) = E\{-\log[p(x)]\}$$

Deste modo, a interpretação para $\mathcal{I}$ seria a diferença das entropias entre as distribuições priori e posteriori. Em linhas gerais podemos dizer que $\mathcal{I}$ é um quantificador da informação fornecida por um experimento.

Considerando o plano de amostragem sequencial que será discutido com mais detalhes na seção 8, Lindley sugere que um experimento deva continuar até a obtenção de informações suficientes $\mathcal{I}$ sobre o parâmetro de interesse. Utilizaremos as informações para investigar paralelos entre as abordagens bayesiana e frequentista para determinar o tamanho de amostra sob diferentes funções objetivo. A função objetivo frequentista $G_F$continua sendo o poder e a função objetivo bayesiana G_B é a informação de Lindley.

Quando uma amostra aleatória de tamanho n é obtido a partir de uma distribuição $N(\theta, \sigma^2)$ e uma distribuição priori $\theta \sim N(\mu, \tau^2)$, a informação esperada para o experimento (Lindley 1956) é:

$$\mathcal{I} = \frac{1}{2}\log\left(1 + n\frac{\tau^2}{\sigma^2}\right) \qquad (13)$$

A partir da equação (13), o tamanho amostral bayesiano para atingir uma quantidade fixa da informação esperada $\mathcal{I}^*$ é

$$n_B = (\exp(2\mathcal{I}^*) - 1)\left(\frac{\sigma}{\tau}\right)^2, \quad (14)$$

e a entrada Bayesiana é $\boldsymbol{v} = (\tau, \sigma)$.

Note que a informação esperada $\mathcal{I}$ não é constante em δ quando escolhemos $n = n_F$ na equação (13), em que n_F é o tamanho de amostra baseado no poder frequentista com a entrada $\boldsymbol{u} = (\alpha, \delta, \sigma)$. Quando $n = n_F$, assumindo que $\tau^2 = C\delta^2$ (para $C > 0$) tornamos $\mathcal{I}$ constante em δ e pelas equações (5) e (14), $n_B = n_F$ se e somente se $(\exp(2\mathcal{I}^*) - 1) = C\left(z_\alpha + z_\beta\right)^2$, o que implica:

$$\mathcal{I}^* = \frac{1}{2}\log\left(1 + C\left(z_\alpha + z_\beta\right)^2\right).$$

3) Dados Composicionais

Uma maneira simples de definir dados composicionais é dizer que são dados que se referem, especificamente, às proporções ou frações de um todo. Através deles, conseguimos obter informações relacionadas às frequências relativas das categorias das variáveis que são de interesse ao estudo, permitindo-nos ferramentas para analisar o comportamento destas variáveis ao longo de suas componentes.

Definição 2: *Uma composição* $\boldsymbol{x}$ *de D partes é um vetor de dimensão D em que cada componente* $x_1, x_2, \dots, x_D$ *é um número real estritamente positivo e a soma total é 1.*

Os dados composicionais são facilmente encontrados e aplicados nas mais variadas áreas de conhecimento como geologia (composição mineral de rochas), biologia (dietas e genética populacional), demografia, economia, saúde (ensaios clínicos) e psicologia. Quando trabalhamos com este tipo de dados devemos ter em mente que o objetivo do estudo é encontrar respostas acerca do comportamento das proporções, uma vez que eles não nos entregam nenhuma informação relacionada às frequências absolutas de cada categoria. A seguir apresentamos um exemplo que ilustra tal limitação de interpretação:

Exemplo 3: Suponha que tenha sido coletado um vetor de proporções $\mathbf{x} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ de uma variável com três categorias de um mesmo indivíduo. Em um instante posterior, foi coletado do mesmo indivíduo o vetor $\mathbf{x} = \left(\frac{1}{6}, \frac{1}{6}, \frac{2}{3}\right)$. Note que neste caso não podemos concluir se frequência absoluta da terceira categoria aumentou, diminuiu ou manteve-se o mesmo, uma vez que também não sabemos o que houve com as demais categorias. Neste caso, só podemos afirmar que a proporção da terceira categoria foi maior no instante seguinte.

O espaço amostral associado aos dados composicionais é o simplex e será definido a seguir.

3.1) Espaço simplex

Definição 3: *O espaço d-simplex é definido como o conjunto:*

$$S^D = \left\{ (X^1, \dots, X^D) \in \mathbb{R}^D : \sum_{j=1}^{D} X^j = 1, X^j > 0 \; para \; j = 1, \dots, D \right\}$$

em que $D = d + 1$.

É importante notarmos pela definição que uma composição de D partes pertence a um espaço (D-1)-dimensional. Se conhecermos as proporções de D-1 componentes, conseguimos determinar a proporção associada à última componente devido à restrição do espaço simplex em que a soma delas deve ser igual a 1.

Exemplo 4:

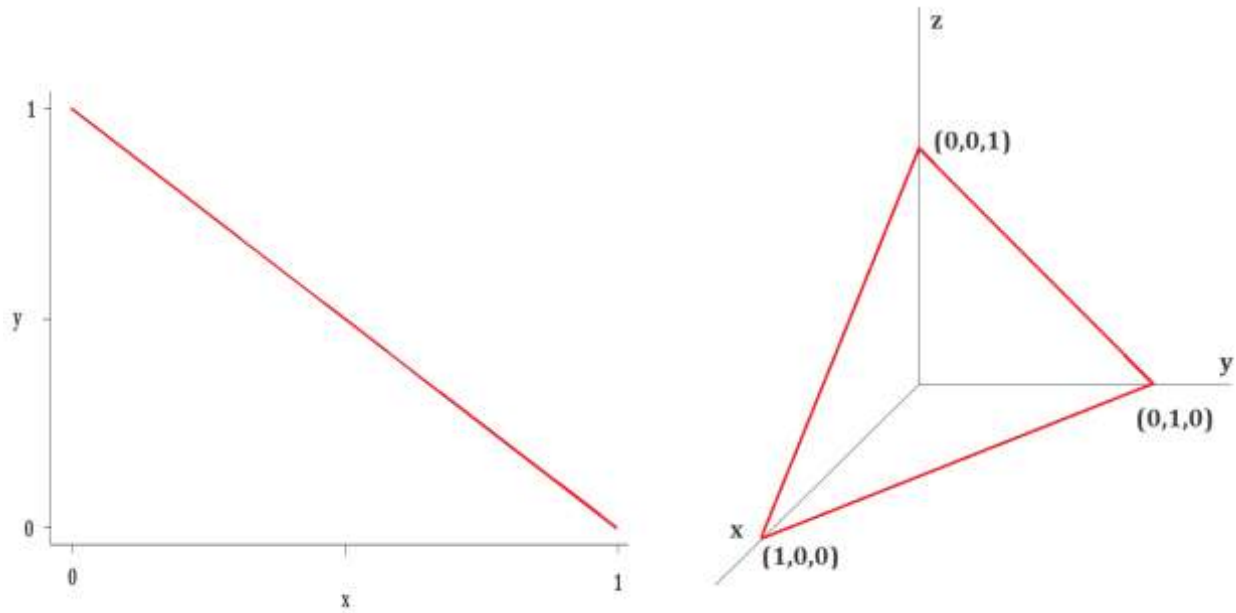

Figura 2: Espaço 1-simplex (esquerda) e espaço 2-simplex (direita)

Algumas referências que caracterizam este tipo de dados sugerem buscar outras formas de representações gráficas quando a dimensão ultrapassar D = 3. Um exemplo de fácil visualização comparativa é o gráfico de radar (*spiderplot*), em que os valores de cada categoria são colocados ao longo de um eixo separado que inicia no centro do gráfico e termina no anel externo.

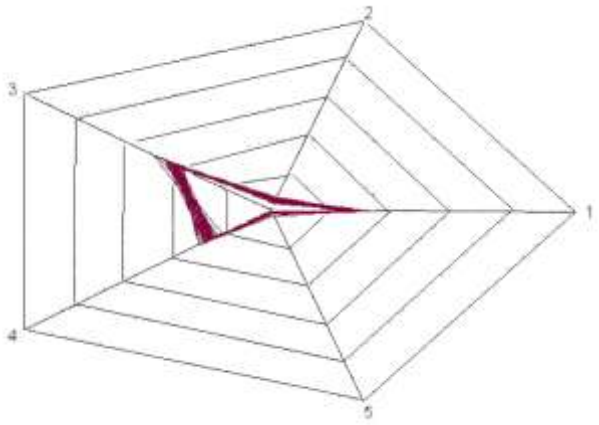

Figura 3: Gráfico de radar para uma variável com cinco componentes

Na figura 3 temos um exemplo no caso em que D = 5. As curvas coloridas se referem a vários indivíduos. Note através do gráfico que as proporções das categorias 1,3 e 4 são relativamente próximas entre si, enquanto que as proporções das demais categorias são muito pequenas (próximas do eixo central). Além disso, podemos depreender também que a variabilidade destas observações é muito baixa, fato facilmente observado pela sobreposição das curvas.

Um ponto fundamental para qualquer análise envolvendo dados composicionais é ter ciência de alguns problemas ao utilizar os métodos usuais de inferência frequentista. Como cada vetor $\mathbf{x}$ de proporções está sujeito a uma restrição de soma, algumas limitações não desejadas na matriz de variância-covariância acabam sendo impostas, invalidando técnicas baseadas em regressão e análise multivariada. Algumas das alternativas mais utilizadas para sanar este tipo de problema são as transformações aplicadas nos dados que removem a restrição a fim de aplicar em seguida os métodos usuais de inferência, além de outras propriedades consequentes de alguns princípios apresentados por John Aitchison, uma das maiores referências sobre a análise de dados composicionais.

4) Fundamentos da geometria de Aitchison

John Aitchison propôs alguns princípios fundamentais que regem toda a análise de dados composicionais. Ao utilizar as transformações e demais ferramentas sugeridas por ele, devemos ter em mente cada um destes princípios a fim de nos permitir maior conhecimento e facilidade de manipulação deste tipo de dados.

4.1) Princípios de Aitchison

a) invariância por escala: inferências estatísticas aplicadas em dados composicionais não dependem da escala que está sendo utilizada. Como exemplo desta propriedade, se tivermos os valores de cada categoria em porcentagens (total igual a 100%) ou em proporções (total igual a 1), os resultados obtidos pelas análises devem ser semelhantes.

b) coerência subcomposicional: inferências estatísticas aplicadas em um subconjunto particular de componentes dependem somente dos dados cujas componentes se encontram neste subconjunto. Podemos dizer que tal propriedade garante que os resultados das comparações serão as mesmas se tivermos acesso a todas as proporções, concentrando-nos apenas na comparação de um subconjunto de componentes ou até mesmo quando tivermos somente os dados correspondentes às proporções deste subconjunto.

c) invariância por permutação: os resultados das comparações não dependem da ordem (marcação) das componentes. Assim, não importa qual categoria nós escolhermos para ser a última ou a primeira do nosso vetor **x** de proporções.

4.2) Razão logarítmica de proporções

Como vimos anteriormente, qualquer vetor de proporções está sujeito à restrição de que a soma das componentes devem ser iguais a um. Consequentemente, cada elemento do vetor só pode assumir um valor entre zero e um. Logo a suposição de distribuição normal multivariada não é razoável para os dados composicionais, podendo acarretar equívocos em nossa análise.

Para o caso em que temos um vetor de proporções com D = 2, Aitchison propôs uma transformação que contorna a situação acima, permitindo-nos utilizar técnicas de inferência baseadas na distribuição normal. Ele assume que a razão logarítmica, $log\left(\frac{X_2}{X_1}\right)$, tenha distribuição normal, utilizando como um dos argumentos o fato de esta transformação conseguir mudar o suporte da composição, podendo agora assumir qualquer número real, o que é pré-requisito essencial para a suposição de normalidade.

No caso em que D > 2, utilizamos uma generalização da transformação acima, denotada por **alr**, definida como segue.

Definição 4: *Transformação **alr** para dados composicionais com D≥2 categorias:*

$$\emptyset_j(\mathbf{X}) = \left(log\left(\frac{X_1}{X_j}\right), \dots, log\left(\frac{X_{j-1}}{X_j}\right), log\left(\frac{X_{j+1}}{X_j}\right), \dots, log\left(\frac{X_D}{X_j}\right) \right)$$

É interessante destacar da definição acima, a influência do princípio da invariância por permutação definida anteriormente, uma vez que as inferências estatísticas nos entregarão os mesmos resultados, independentemente da variável X_j que for tomada para o denominador das razões-logarítmicas. Portanto, como a transformação **alr** é invariante por permutação, a escolha do denominador pode ser feita de modo arbitrário.

Utilizando os mesmos argumentos apresentados anteriormente, Aitchison assume que as $D-j$ razões logarítmicas tenham conjuntamente distribuição normal multivariada. Assim, basta utilizar os métodos de análise multivariada para fazer inferências a respeito do vetor de médias $\boldsymbol{\mu}_{D-j}$ e da matriz de variância-covariância $\boldsymbol{\Sigma}_{D-j\ x\ D-j}$.

4.3) Distância de Aitchison

Em nosso estudo, utilizaremos a distância de Aitchison por esta ser compatível com as operações básicas no simplex e com a natureza particular dos dados composicionais.

Definição 5: *Dadas duas composições* $\mathbf{X}^{(1)} = \left(x_1^{(1)}, \dots, x_D^{(1)}\right)$ *e* $\mathbf{X}^{(2)} = \left(x_1^{(2)}, \dots, x_D^{(2)}\right)$*, definimos como a distância de Aitchison entre elas como:*

$$\Delta\left[\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right] = \sqrt{\sum_{j=1}^{k}\left[ln(r_j) - L\right]^2}$$

em que $r_j = \left(\frac{x_j^{(1)}}{x_j^{(2)}}\right)$ *é a razão entre as frequências relativas de mesma componente e* $L = \frac{1}{D}\sum_{j=1}^{D} ln(r_j)$ *a média aritmética do logaritmo destas razões.*

Exemplo 5: Considere os vetores $\mathbf{X}^{(1)} = (0{,}05; 0{,}65; 0{,}30)$, $\mathbf{X}^{(2)} = (0{,}10; 0{,}60; 0{,}30)$, $\mathbf{X}^{(3)} = (0{,}50; 0{,}20; 0{,}30)$ e $\mathbf{X}^{(4)} = (0{,}55; 0{,}15; 0{,}30)$.

Então: $\Delta\left[\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right] = 0{,}60$ e $\Delta\left[\mathbf{X}^{(3)}, \mathbf{X}^{(4)}\right] = 0{,}28$.

Calculemos agora a distância Euclidiana entre eles:

$$\left|\left|\mathbf{X}^{(1)} - \mathbf{X}^{(2)}\right|\right| = \sqrt{(-0.05)^2 + (0.05)^2} = 0{,}07 = \left|\left|\mathbf{X}^{(3)} - \mathbf{X}^{(4)}\right|\right|$$

Note que a primeira componente do vetor $\mathbf{X}^{(1)}$ teve um aumento de 100% se considerarmos $\mathbf{X}^{(2)}$ uma observação seguinte, enquanto que a primeira componente do vetor $\mathbf{X}^{(3)}$ teve um aumento relativo de apenas 10% se considerarmos $\mathbf{X}^{(4)}$ como a seguinte. Assim, é esperado que a distância entre $\mathbf{X}^{(1)}$ e $\mathbf{X}^{(2)}$ não seja a mesma que entre $\mathbf{X}^{(3)}$ e $\mathbf{X}^{(4)}$. De fato, pelo resultado obtido, concluímos que a distância de Aitchison é maior entre o primeiro par de vetores. Por outro lado, a distância Euclidiana foi a mesma nos dois casos, sendo razoável acreditar que não seja uma ferramenta adequada para estudar este tipo de dados, uma vez que não consegue captar e descrever a variabilidade composicional de modo adequado como a de Aitchison.

Exemplo 6: Para efeito de compreensão do comportamento da função de distância de Aitchison, foi feito um exercício computacional considerando o espaço 1-simplex. Fixado o ponto (0,5; 0,5) plotamos as funções correspondentes aos valores das distâncias de Aitchison e das distâncias Euclidianas entre este ponto com todos os outros pertencentes à reta y = 1 - x na direção de (0,5; 0,5) até (1;0). Vale ressaltar que as distâncias têm mesmo comportamento se o sentido for de (0,5; 0,5) para (0; 1).

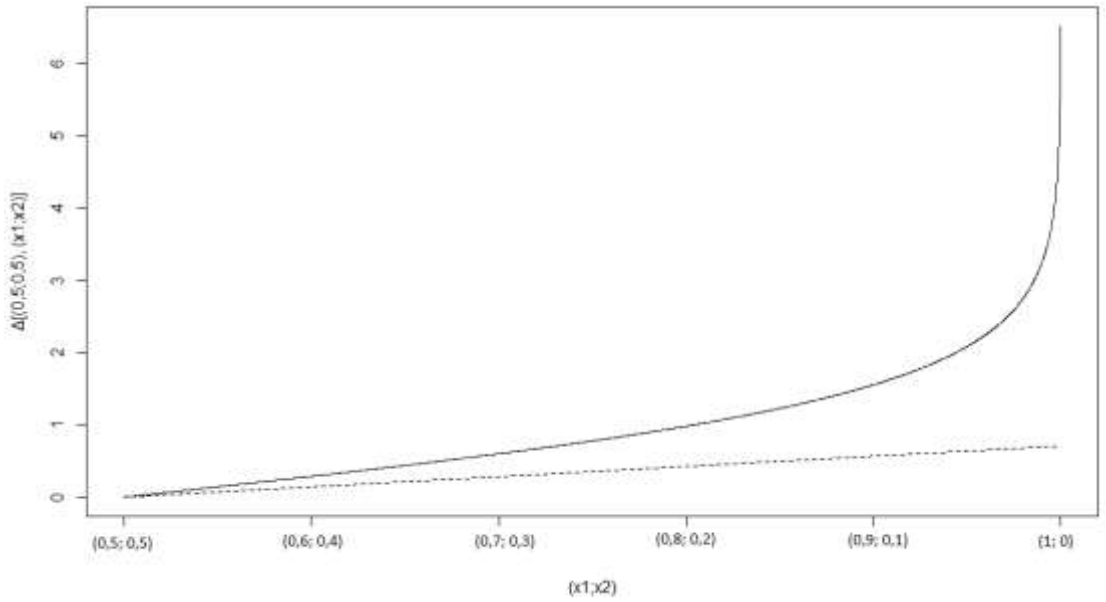

Figura 4: Gráfico da distância de Aitchison (linha contínua) e da distância Euclidiana (linha tracejada) correspondente à descrição do exemplo 6.

Podemos obervar pela figura 4 que a distância de Aitchison de todos os pontos entre o que foi fixado até (0,8; 0,2) cresce de modo aproximadamente linear e lento se comparado com o aumento significativo a partir de (0,9; 0,1), uma vez que os valores das distâncias concentravam-se somente entre 0 e 1 e depois cresceram muito rápido. Notamos também que além do crescimento da distância Euclidiana ser muito demorado, não há diferenças significativas entre seus valores ao longo dos pontos. Logo, podemos enxergar que a distância de Aitchison de fato parece ser adequada ao estudo de composições, já que ela consegue captar pequenas e grandes diferenças entre as proporções, descrevendo isto numericamente.

5) Ensaios clínicos

Os ensaios clínicos correspondem a estudos que visam avaliar e testar a eficácia e segurança de um novo medicamento ou tratamento médico. Um ensaio clínico bem planejado e estruturado permite que o desenvolvimento de novos tratamentos seja feito de forma rápida, segura e eficiente. Para tal, ainda deve-se ter em mente todas as questões éticas e os riscos que podem impactar na saúde das unidades experimentais envolvidas, elaborando estratégias criteriosamente avaliadas. Os participantes do estudo devem ser previamente informados a respeito da metodologia, de todos os testes que serão realizados, dos princípios da voluntariedade e da possibilidade de interromperem suas participações caso queiram abandonar o estudo.

Podem ser classificados em quatro fases de acordo com o nível de complexidade e objetivo da avaliação. Resumidamente iremos descrevê-las:

Fase I – são realizados sob supervisão médica, envolvendo um número limitado de voluntários sadios, e tem como objetivo principal avaliar a velocidade de reação e a toxicidade do produto no caso de ser um medicamento.

Fase II – realizam-se os estudos em um pequeno número de pacientes a fim de encontrar a dosagem ótima do produto, além de verificar a necessidade de estudos adicionais. Os testes desta fase requerem um grupo de controle para efeitos de comparação entre o medicamento/tratamento e um placebo.

Fase III – o principal objetivo é o de demonstrar a eficácia a curto e longo prazo do medicamento/tratamento, envolvendo um número maior de participantes.

Fase IV – testes que são realizados após a aprovação do medicamento/tratamento a fim de identificar possíveis reações adversas e inesperadas.

Na fase II e III ocorrem os estudos de intervenção comparativa que incluem pelo menos dois grupos de estudos: o de intervenção, cujos participantes receberão efetivamente o remédio/tratamento que está sendo testado e o de controle em que uma substância sem atividade (placebo) é aplicada. Nestas fases a

aleatorização dos pacientes entre os grupos podem ser cruciais para o delineamento do experimento. Além de evitar vícios na atribuição de tratamentos, a aleatorização é fundamental para os princípios de inferências estatísticas no plano frequentista.

Na próxima seção discutiremos um planejamento de ensaio clínico em que o tratamento não é atribuído aleatoriamente para um número fixo de pacientes. O planejamento sequencial permite que o monitoramento dos dados seja feito de modo contínuo e/ou periódico em que o tamanho da amostra (número total de pacientes) é determinado durante o estudo, respeitando as necessidades e o desenvolvimento do experimento.

6) Amostragem intencional

Nesta seção discutiremos brevemente o artigo [6] que descreve um planejamento bayesiano aplicado em um ensaio clínico em psiquiatria envolvendo indivíduos com transtorno obsessivo compulsivo (TOC).

Na primeira fase do experimento, é de interesse comparar as respostas dos pacientes para dois braços de estudo. Um grupo de n_1 pacientes será submetido a um tratamento farmacológico, enquanto que n_2 a um psicoterapêutico. Os participantes serão alocados sequencialmente, de acordo com a ordem de chegada à clínica e cada um será colocado em somente um dos braços.

Baseando-se em estudos prévios, foram consideradas seis variáveis explicativas para maior compreensão dos efeitos de cada um dos tratamentos entre os pacientes bem como dos fatores que podem influenciar o curso e a evolução da doença. São elas: idade (contendo três categorias: menor que 30, entre 31 e 45 e acima de 45 anos); severidade do sintoma de TOC (obsessão e compulsão são classificadas como de nível baixo (L), moderado (M) e alto (A), acarretando um total de oito combinações possíveis: (L;L), (L;M), (L;A), (M;L), ..., (A;A). A categoria (L;L) porém não é considerada, uma vez que não indicaria sintomas de TOC); histórico de tratamento (contendo três categorias: não fez tratamento, já fez um tratamento e não obteve resultados e fez dois ou mais tratamentos e não obteve resultados); nível de escolaridade (contendo quatro categorias: sem escolaridade, até oito anos de escolaridade, entre 9 e 12 anos de escolaridade e nível superior); estado civil (contendo duas categorias: casado ou divorciado e solteiro que nunca casou) e gênero (masculino e feminino). Vale ressaltar que algumas destas variáveis não foram consideradas nesta fase inicial de apenas dois braços de estudo, mas sim em uma fase seguinte em que foram comparadas três diferentes estratégias farmacológicas. Como o procedimento utilizado nas duas fases é análogo, iremos descrever aqui apenas a primeira para facilitar a compreensão da metodologia empregada.

O objetivo principal da alocação é o de permitir um equilíbrio das proporções de cada uma destas variáveis entre os braços do estudo, a fim de evitar viés de comparação e outros problemas na interpretação dos resultados. Como as

respostas aos tratamentos podem variar entre pacientes com características diferentes, como por exemplo, pessoas idosas reagirem mais lentamente que os mais novos, é de extrema importância procurar uma estratégia que permita que a distribuição dos envolvidos entre os tratamentos seja feita da forma mais homogênea possível.

Foi desenvolvido um método de alocação que utiliza a distância de Aitchison e todas as propriedades dos dados composicionais que foram descritas nas seções anteriores. Como as variáveis explicativas seriam melhores classificadas como vetores contendo as proporções de cada uma de suas categorias, decidiram utilizar a distância de Aitchison por ser coerente com a natureza dos dados composicionais.

6.1) Alocação:

Inicialmente iremos descrever o processo de alocação para o caso mais simples, contendo apenas uma variável explicativa: idade. Considere a situação contida na fase um do experimento que foi descrita anteriormente e as seguintes notações:

- A variável idade "**X**" é divida em três categorias: x_1 se $x \leq 30$; x_2 se $30 < x \leq 45$; e x_3 se $x > 45$;
- $n_1 = (n_{11}, n_{12}, n_{13})$ e $n_2 = (n_{21}, n_{22}, n_{23})$ em que para $i = 1\ ou\ 2$ e $j = 1{,}2\ ou\ 3$, n_{ij} é o número de pacientes de idade correspondente a categoria x_j no braço i.
- $\boldsymbol{X_1} = (x_{11}, x_{12}, x_{13})$ e $\boldsymbol{X_2} = (x_{21}, x_{22}, x_{23})$ em que $x_{ij} = \frac{n_{ij}}{n_i}$ representa a frequência relativa de pacientes com idade na categoria x_j no braço i.

Assim que um paciente chega à clínica, os passos para decidir em qual braço o mesmo será alocado consistem em:

1) Suponha que aloquemos o novo paciente no primeiro braço. Assim, atualizamos $\boldsymbol{X_1}$ em uma nova composição $\boldsymbol{X_1^*}$ e calculamos a distância de Aitchison entre $\boldsymbol{X_1^*}$ e $\boldsymbol{X_2}$ e denotamos $d_1 = \Delta[\boldsymbol{X_1^*}; \boldsymbol{X_2}]$.

2) Considere o mesmo procedimento em (1) só que agora supondo que o aloquemos no segundo braço e denotemos $d_2 = \Delta[\boldsymbol{X_1}; \mathbf{X_2^*}]$.
3) Comparamos os valores de d_1 e d_2. Se $d_1 < d_2$, nós alocamos o novo paciente no braço 1; se $d_1 > d_2$, nós alocamos o novo paciente no braço 2; se de $d_1 = d_2$, o paciente pode ser alocado aleatoriamente em qualquer um dos braços.

Exemplo 7: Suponha que em um estágio do processo nós encontramos os seguintes vetores de frequências absolutas para a variável idade: $n_1 = (3;\ 7;\ 5)$ e $n_1 = (5;\ 6;\ 6)$. Um novo paciente entre 30 e 45 anos chega à clínica e deve-se decidir qual tratamento será aplicado. Assim, teremos: $\boldsymbol{X_1} = (0{,}200;\ 0{,}467;\ 0{,}333)$, $\boldsymbol{X_1} = (0{,}294;\ 0{,}353;\ 0{,}353)$ e consequentemente $\boldsymbol{X_1^*} = (0{,}187;\ 0{,}500;\ 0{,}313)$ e $\boldsymbol{X_2^*} = (0{,}278;\ 0{,}389;\ 0{,}333)$. Em seguida, calculamos as distâncias de Aitchison para que seja feita a decisão:

Se alocarmos o novo paciente no primeiro braço, temos $d_1 = 0{,}47$. Caso a alocação seja feita no outro braço, temos $d_2 = 0{,}37$. Portanto, como $d_1 > d_2$, o paciente deveria ser alocado no braço 2.

Considere agora a situação real do problema em que temos mais de uma variável explicativa. Além de tentar equilibrar as proporções das categorias destas variáveis em cada um dos braços, procura-se também controlar o tamanho amostral dos mesmos na tentativa de evitar uma diferença discrepante entre o número de pacientes alocados em cada tratamento. Assim, é acrescida a seguinte notação:

- O tamanho amostral é denotado por s tal que $s = s_1 + s_2$, em que s_1 corresponde ao número de pacientes alocados no primeiro braço e s_2 no segundo braço.
- $\boldsymbol{S_1} = (\frac{s_1}{s}, \frac{s_2}{s})$ e $\boldsymbol{S_2} = (\frac{s_2}{s}, \frac{s_1}{s})$ são os vetores composicionais correspondentes aos tamanhos amostrais do primeiro e segundo braço, respectivamente. Assim como descrito no processo de alocação acima, $\boldsymbol{S_1^*}$ e $\boldsymbol{S_2^*}$ indicam que um novo paciente entrou em um braço específico.

Como estamos interessados em controlar as proporções das variáveis e o tamanho amostral entre os braços, a estratégia proposta para resolvermos isso é dar "pesos de importância" para cada uma das variáveis, ou seja, utilizar uma média ponderada das distâncias de Aitchison. Tais pesos devem ser decididos em acordo com os pesquisadores e investigadores da área da qual está sendo realizado o ensaio clínico. O mecanismo será exemplificado a seguir:

Exemplo 8: Considere a mesma situação proposta pelo exemplo 6 com a restrição de que queremos controlar o tamanho amostral de cada braço. A fim de respeitar a importância de cada variável para o estudo, os pesquisadores da área concordaram em dar peso um para o tamanho amostral e dois para a idade. Deve-se decidir em qual dos tratamentos o mesmo paciente deverá ser alocado.

Dado que antes da chegada do paciente tínhamos 15 no primeiro braço e 17 no outro, teremos $\boldsymbol{S_1} = (0{,}47; 0{,}53)$ e $\boldsymbol{S_2} = (0{,}53;\ 0{,}47)$ e consequentemente $\boldsymbol{S_1^*} = \left(\frac{15+1}{32+1}; \frac{17}{32+1}\right) = (0{,}48; 0{,}52)$ e $\boldsymbol{S_2^*} = \left(\frac{17+1}{32+1}; \frac{15}{32+1}\right) = (0{,}55;\ 0{,}45)$. Em seguida, calculamos as distâncias de Aitchison $\Delta[\boldsymbol{S_1^*}; \boldsymbol{S_2}] = 0{,}13$ e $\Delta[\boldsymbol{S_1}; \boldsymbol{S_2^*}] = 0{,}22$ e utilizamos as médias ponderadas D_1 e D_2 para guiar nossa decisão:

Se alocarmos o novo paciente no primeiro braço, temos $D_1 = \{2\Delta[\boldsymbol{X_1^*}; \boldsymbol{X_2}] + \Delta[\boldsymbol{S_1^*};\ \boldsymbol{S_2}]\} \div 3 \cong 0{,}42$. Caso a alocação seja feita no outro braço, temos $D_2 = \{2\Delta[\boldsymbol{X_1}; \boldsymbol{X_2^*}] + \Delta[\boldsymbol{S_1}; \boldsymbol{S_2^*}]\} \div 3 \cong 0{,}32$. Portanto, como $D_1 > D_2$, o paciente deveria ser alocado no braço 2 para que os critérios propostos fossem respeitados.

Um problema que pode ser recorrente neste tipo de metodologia é a presença de uma frequência igual à zero em alguma categoria de uma variável que pode acarretar indeterminação no cálculo das razões das frequências relativas da distância de Aitchison. Para corrigir isto, sugerem que uma correção seja utilizada, consistindo em adicionar para cada categoria um pequeno número às frequências absolutas observadas. Como uma analogia à análise bayesiana de dados categorizados, uma mesma constante é adicionada em cada categoria e a recomendação é que este valor seja igual a $\frac{1}{k}$ em que k corresponde ao número de categorias da variável.

Por fim, quando as proporções de cada uma das categorias das variáveis entre os tratamentos estiverem satisfatoriamente equilibradas e um número de pacientes que seja conveniente para os pesquisadores tenham sido atendidos sem extrapolar os recursos disponíveis do estudo, o processo se encerra e o tamanho total da amostra é finalmente determinado.

7) Amostragem sequencial

Na amostragem sequencial as unidades experimentais são examinadas continuamente até que os resultados acumulados possam nos entregar uma decisão acerca das hipóteses que estão sendo testadas ou até quando o custo de coleta das observações não extrapolem os limites propostos inicialmente. Logo, o tamanho da amostra neste caso é variável uma vez que não é determinado logo na fase preliminar de um experimento.

Usualmente estamos interessados em testar uma hipótese a respeito dos parâmetros populacionais e em muitos casos utilizamos um procedimento de teste que necessita estimar estes parâmetros para chegar a uma decisão, seja através do cálculo de uma estatística de teste, p-valor, etc. Em um teste sequencial podemos testar estas hipóteses sem a necessidade de estimar as quantidades de interesse e segundo [12] podemos chegar a uma decisão com um número menor de unidades amostrais do que um processo de amostragem usual.

Com base na primeira observação que foi coletada, após examiná-la podemos chegar a uma das três possíveis decisões:

1ª) não rejeitar H_0;

2ª) rejeitar H_0;

3ª) não é possível decidir entre H_0 e H_1, então continua a amostragem;

Caso os resultados nos levem a escolher a terceira opção, outra unidade amostral é coletada e com base nos resultados acumulados com as observações anteriores decidimos novamente entre as opções acima, e assim sucessivamente.

A seguir apresentamos o teste sequencial da razão de probabilidades de Wald proposto em [12].

7.1) Teste sequencial da razão de probabilidades de Wald

Suponha que queremos testar as seguintes hipóteses acerca do parâmetro de interesse θ:

$$H_0: \theta = \theta_0 \; vs \; H_1: \theta = \theta_1$$

Sejam $L(\theta_0; x_1, \dots, x_n)$ e $L(\theta_1; x_1, \dots, x_n)$ as verossimilhanças consideradas na etapa que estamos com n observações sob H_0 e H_1, respectivamente. O procedimento de Wald consiste em calcular o logaritmo da razão destas verossimilhanças e compará-lo com os limites propostos A e B. Estas constantes são definidas de modo que $0 < A < B$ e que o teste sequencial tenha os valores desejados para as probabilidades de cometer o erro do tipo I (α) e tipo II (β).

O teste se baseia na seguinte regra de decisão:

- Se $\Lambda > B$, rejeite H_0 e encerre a amostragem;
- Se $\Lambda < A$, não rejeite H_0 e encerre a amostragem;
- Se $A < \Lambda < B$, não decida ainda em nenhuma das hipóteses, colete uma nova observação para obter Λ_{n+1} e verifique todos estes passos novamente.

  Em que: $\Lambda_n = \frac{L(\theta_1; x_1, \dots, x_n)}{L(\theta_0; x_1, \dots, x_n)}$, $A = \ln\left(\frac{\beta}{1-\alpha}\right)$ e $B = \ln\left(\frac{1-\beta}{\alpha}\right)$.

Wald afirma em [12] que o teste sequencial de razão de probabilidades pode ser considerado ótimo no sentido que minimiza o tamanho amostral esperado antes de uma decisão ser tomada dentre todos os testes sequenciais que não têm probabilidades de erros maiores que ele.

A seguir apresentamos outra metodologia baseada no teste sequencial de Wald no qual buscamos tomar nossa decisão a partir de um procedimento gráfico bastante intuitivo.

7.2) Retas de decisão

Considere as seguintes funções:

i. $R_0(n) = h_0 + S \cdot n$

ii. $R_1(n) = h_1 + S \cdot n$

Em que h_0 e h_1 são os coeficientes lineares, S o coeficiente angular e n o tamanho da amostra no estágio considerado.

As retas de decisão i e ii são traçadas num gráfico de modo que na abscissa temos a escala do tamanho amostral n e no eixo das ordenadas a estatística $\sum_{i=1}^{n} y_i$,

que pode indicar, por exemplo, o número de unidades amostrais que contenha alguma característica específica entre as n unidades observadas (dados de classificação) , número acumulado de ocorrências de interesse (dados de contagem), etc.

Nos casos das distribuições de probabilidade mais conhecidas como os modelos de Poisson, Binomial, Normal e Binomial Negativa, por exemplo, o procedimento para determinar as fórmulas para os coeficientes das retas, consiste em isolar na regra de decisão apresentada em 6.1 o termo $\sum_{i=1}^{n} y_i$, de modo que seja escrito como uma função linear do tamanho amostral da amostra, ou seja, $\sum_{i=1}^{n} y_i \leq h_0 + S \cdot n$ e $\sum_{i=1}^{n} y_i \geq h_1 + S \cdot n$. Em [5] contém para todos os modelos citados acima, os coeficientes lineares e angulares obtidos ao seguir este procedimento.

Exemplo 9: Para o modelo probabilístico Binomial, os coeficientes das retas de decisão para testar as hipóteses de 6.1 são dados por:

$$h_0 = \frac{\ln\left(\frac{\beta}{1-\alpha}\right)}{\ln\left[\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right]}, \qquad h_1 = \frac{\ln\left(\frac{1-\beta}{\alpha}\right)}{\ln\left[\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right]}$$

$$S = \frac{\ln\left[\frac{(1-\theta_0)}{(1-\theta_1)}\right]}{\ln\left[\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right]}$$

Considerando as hipóteses de 6.1, como é feita a decisão ao construir as retas? Suponha que estejamos com n unidades amostradas e que tenhamos calculado o valor de $\sum_{i=1}^{n} y_i$. Se o ponto $(n, \sum_{i=1}^{n} y_i)$ pertencer à reta R_0 ou estiver abaixo dela, a amostragem é terminada e decidimos por não rejeitar H_0. Se o ponto pertencer à reta R_1 ou estiver acima dela, paramos de coletar novas observações e rejeitamos H_0. Entretanto, se o ponto estiver entre R_0 e R_1 uma nova unidade amostral é observada e o processo continua até que consiga custear as limitações estipuladas no experimento. Na figura 5 veja como este processo pode ser representado graficamente.

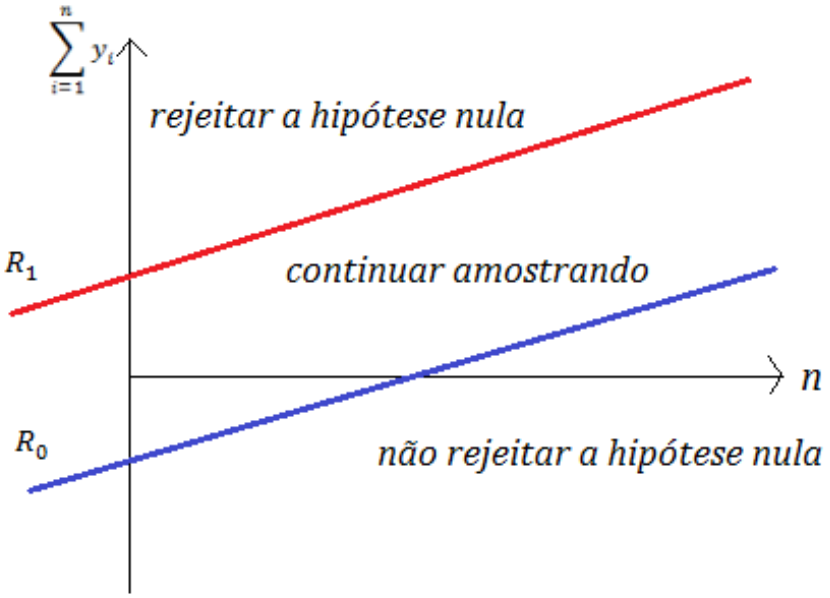

Figura 5: Exemplificação da representação gráfica das retas de decisão. Note que $\boldsymbol{R_0}$ e $\boldsymbol{R_1}$são paralelas, uma vez que o coeficiente angular $\boldsymbol{S}$ é o mesmo para ambas as retas.

7.3) Abordagem Bayesiana para o teste sequencial da razão de probabilidades de Wald

Nesta seção apresentaremos a versão bayesiana para o teste sequencial de Wald sob as mesmas condições que foram introduzidas em 6.1.

Consideremos a existência de uma probabilidade a priori de H_0 ser verdadeiro e que seu valor seja conhecido, $P(H_0) = Q_0$. Consequentemente, a probabilidade a priori de H_1 ser verdadeiro é denotada por $P(H_1) = Q_1 = 1 - Q_0$.

Sob o ponto de vista da inferência bayesiana, a informação acerca do parâmetro de interesse é atualizada ao ter acesso aos dados que foram observados. No caso do teste de hipóteses ganharemos informações que afetarão as probabilidades de H_0 e H_1. Assim, sejam $Q_{0|x}$ e $Q_{1|x}$ as probabilidades a posteriori de H_0 e H_1 serem verdadeiros após n unidades amostrais terem sido observadas, respectivamente. Pelo teorema de Bayes temos que:

$$Q_{0|x} = \frac{Q_0 f_0(x_1, \dots, x_n|\theta_0)}{Q_0 f_0(x_1, \dots, x_n|\theta_0) + Q_1 f_1(x_1, \dots, x_n|\theta_1)}$$

$$Q_{1|x} = \frac{Q_1 f_1(x_1, \dots, x_n|\theta_1)}{Q_0 f_0(x_1, \dots, x_n|\theta_0) + Q_1 f_1(x_1, \dots, x_n|\theta_1)}$$

Em que $f_i(x_1, \dots, x_n|\theta_i)$ *é a função de verossimilhança da amostra* $\boldsymbol{x}$ *condicionada a* θ_i.

A regra de decisão para a $n-$ésima etapa desta versão do teste, proposta por Wald em [12], consiste em computar os valores de $Q_{0|x}$ e $Q_{1|x}$ e com base nos resultados escolher uma das três opções abaixo:

- Se $Q_{1|x} \geq d_1$, rejeite a hipótese nula;
- Se $Q_{0|x} \geq d_0$, não rejeite a hipótese nula;
- Se $Q_{1|x} < d_1$ e $Q_{0|x} < d_0$, colete uma nova observação e repita o procedimento.

Neste caso, d_0 e d_1 são dois números positivos menores que 1 e maiores que $1/2$ definidos pelo pesquisador com base em seus critérios para decidir acerca das hipóteses de interesse. As restrições acerca dos valores destas constantes podem ser explicadas pela seguinte situação.

Note que $Q_{1|x} \geq d_1$e $Q_{0|x} \geq d_0$ não podem ser satisfeitas simultaneamente. Suponha que ambas as relações sejam satisfeitas. Então $Q_{1|x} + Q_{0|x} \geq d_1 + d_0$. Mas $Q_{1|x} + Q_{0|x} = 1$. Consequentemente $d_1 + d_0 \leq 1$, o que é uma contradição, pois definimos que $d_i > 1/2$. Portanto, os eventos $Q_{1|x} \geq d_1$ e $Q_{0|x} \geq d_0$ são mutuamente exclusivos.

É importante destacar que em algumas situações práticas, especificamente quando θ_0 estiver muito distante do real valor do parâmetro θ que queremos testar, a amostragem sequencial permite chegar a uma decisão com um número menor de unidades amostrais do que a amostragem probabilística usual. Por outro lado, quando θ_0 for muito próximo de θ o procedimento de Wald exige maior número de unidades amostrais para chegar a uma decisão. Sob o ponto de vista frequentista, Wald propõe uma nova regra de decisão para as hipóteses caso o processo sequencial exceda o tamanho amostral encontrado através do procedimento de teste de Neyman-Pearson. Sob o ponto de vista bayesiano, o processo continua enquanto o custo de coleta das observações não exceda os limites propostos pelo experimento. Caso o processo fique extremamente

dispendioso e não consiga chegar a uma decisão acerca das hipóteses, a amostragem é encerrada e o teste é inconclusivo.

CONCLUSÃO

No contexto de determinação de tamanho de amostra, um dos papeis fundamentais do estatístico é a escolha de um método que seja apropriado para o estudo que será conduzido. Ao examinar rigorosamente todas as condições do problema, desde o custo de coleta, formas de condução e coleta das observações e demais limitações impostas pelo experimentador, definimos quais procedimentos são mais adequados para determinar o tamanho de amostra.

Através deste trabalho conhecemos alguns procedimentos de determinação de tamanho de amostra, tanto do ponto de vista frequentista e bayesiano, para diversas situações em que desejamos testar hipóteses a cerca de um parâmetro populacional de interesse. Vimos que a maioria destes procedimentos utilizam funções específicas para determinar o tamanho de amostra ótimo, de modo que os erros inerentes aos testes de hipóteses possam ser controlados e que o tamanho obtido atenda aos preceitos do pesquisador. Apresentamos alguns casos específicos nos quais os tamanhos de amostra encontrados pelas abordagens bayesiana ou frequentista podem ser iguais, mesmo que estas metodologias estejam baseadas em critérios e entradas completamente diferentes. Além disso, através do estudo de procedimentos de amostragem intencional e sequencial vimos que existem alternativas adequadas e eficientes ao plano de amostragem aleatória usual, permitindo encontrar tamanhos de amostra menores e que sejam capazes de respeitar as necessidades e o desenvolvimento do experimento.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Aitchison, J. (1986). *"The Statistical Analysis of Compositional Data"*, Chapman and Hall, New York.

[2] Andrade, A. L. (1996). *"Métodos de Investigação Epidemiológica em Doenças Transmissíveis – Volume I – Ensaios Clínicos e Exercícios"*, Universidade Federal de Goiás.

[3] Butler, A., Bierman, S. e Marion, G. (2005). *"Statistical methods for environmetal risk assessment: Compositional data module"*, University of Edinburgh.

[4] DeGroot, M.H. and Schervish, M.J., (2012). *"Probability and Statistics"*, 4th ed., Addison Wesley.

[5] Estefanel, V. e Barbin, D. (1979). *"Amostragem Sequencial baseada no teste sequencial da razão de probabilidades e seu uso na determinação da época de controle das lagartas da soja no estado do Rio Grande do Sul"*, Rev. Centro Ciência Rurais, 29-48.

[6] Fossaluza, V., Diniz, J. B., Pereira, B.B., Miguel, E.C., Pereira C.A.B. (2009). *"Sequence allocation to balance prognostic factors in a psychiatric clinical trial"*, Clinics 64 (6) 511-8.

[7] Inoue, L.Y.T., Berry, D.A., Parmigiani, G. (2005) *"Relationship between Bayesian and frequentist sample size determination"*. The American Statistician 59 (1) 79–87.

[8] Martin-Fernández, J. A., Barceló-Vidal, C. e Pawlowsky-Glahn, V. (2006). *"Measures of Difference for Compositional Data and Hierarchial Clustering Methods"*, Universitat de Girona and Universitat Politécnica de Catalunya.

[9] Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R. (2007). *"Lecture Notes on Compositional Data Analysis"*, Catedrática de Universidad.aruzza, L. (2008). *"Métodos Estatísticos para a análise de bibliotecas digitais de expressão genética"*, Tese apresentada ao Programa Interunidades de Bioinformática da Universidade de São Paulo.

[10] Pericchi, L. e Pereira, C. (2016). *"Adaptative significance levels using optimal decision rules: Balancing by weighting the error probabilities"*, Brazilian Journal of Probability and Statistics, Vol. 30, No. 1, 70-90.

[11] Robert, C. (2007). *"The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation".* 2th ed., Springer.

[12] Wald, A. (1945). *"Sequential Tests of Statistical Hypotheses".* Columbia University.