Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Inferência Bayesiana e Clássica sobre o Coeficiente de Variação

Ana Paula Rodrigues Berçot

Orientador: Prof. Dr. Eduardo Yoshio Nakano

Brasília

2017

Ana Paula Rodrigues Berçot

Inferência Bayesiana e Clássica sobre o Coeficiente de Variação

Monografia apresentada para obtenção de título de Bacharel em Estatística, Instituto de Exatas, Universidade de Brasília

Orientador: Prof. Dr. Eduardo Yoshio Nakano

Brasília

2017

Agradecimentos

Agradeço primeiramente à minha família, principalmente aos meus pais e ao meu irmão, por toda a paciência que tiveram comigo nos momentos de irritação, pela força e motivação que sempre buscavam me dar, além de todo apoio afetivo. Sem vocês eu nada seria. Agradeço também ao Pitoco, que sempre esteve comigo e me apoiando, independente da situação.

Sou grata aos professores que me acompanharam durante a graduação, em especial ao Prof. Nakano, responsável pela realização deste estudo. Obrigada pela orientação, dedicação e inestimável auxílio com este trabalho. Ter seus conhecimentos e ideias compartilhados comigo foi uma honra.

Agradeço também a toda equipe do Departamento de Estatística da UnB, principalmente o pessoal da secretaria, da segurança e da limpeza, os quais estavam sempre dispostos a dar-nos o melhor.

Agradeço à ESTAT Consultoria, tanto pelas experiências profissionais e empresariais, como pelos desafios, crescimento profissional e pessoal.

A todos aqueles que, de alguma forma, estiveram e estão próximos de mim, fazendo esta vida valer cada vez mais a pena.

Aos meus amigos Bruno Matos e Thayanne Sales, que estiveram ao meu lado do início ao fim do curso, ajudando em todos os momentos de desespero e compartilhando aqueles de alegria. Vocês deixaram marcas inesquecíveis nessa fase da minha vida. Obrigada também aos demais amigos que me ajudaram e apoiaram ao longo dessa jornada, principalmente aqueles com os quais me reunia para jogar Catan (Alexandre, Alfredo, Ana Carolina, Cadu, Gongora, Maroneze, Pedro e Tuler) e todos aqueles que que conheci na ESTAT: com vocês, as pausas entre um parágrafo e outro de produção melhora tudo o que tenho produzido na vida. Tenho certeza que sem o apoio, ajuda e companhia de vocês tudo teria sido muito mais difícil e sem graça.

"PER ASPERA AD ASTRA"

Resumo

O Coeficiente de Variação (CV) é uma medida com ampla aplicabilidade, contanto que sejam obedecidos os pré-requisitos para utilização do mesmo. Este trabalho visa mostrar a estimação dessa medida via estimativa pontual e intervalo de confiança, bem como a aplicação do Teste de Significância Genuinamente Bayesiano (FBST). Toda a metodologia será desenvolvida com base na visão Bayesiana e considerando apenas distribuições que possuem suporte positivo visto ser essa uma das restrições para utilização do CV. Além disso, os procedimentos estudados nesse trabalho foram ilustrados por meio de dados simulados e dados reais obtidos através da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2015, sendo que no último caso o objetivo era analisar a desigualdade econômica entre os estados do Amazonas e de São Paulo. Por fim, o presente trabalho disponibiliza rotinas no Software R (R CORE TEAM, 2017) que permitem a realização de tais procedimentos.

Palavras-chave: Coeficiente de variação, Inferência Bayesiana, Desigualdade econômica, FBST, HPD.

Abstract

The Coefficient of Variation (CV) is a measure with a wide applicability, provided that the prerequisites for its use are obeyed. This study aims to show the estimation of this measure using point estimate and confidence interval, as well as the application of the Full Bayesian Significance Test (FBST). Methodology will be developed on Bayesian view basis and considering only distributions with positive support seeing that this is one of the CV's use restrictions. Futhermore, the procedures studied in this work were illustrated by simulated data and real data obtained through the National Household Sample Survey (PNAD) of 2015, aiming, in the latter case, to analyze the economic inequality between Amazonas and São Paulo states. Lastly, the present work provides routines in Software R (R CORE TEAM, 2017) allowing the accomplishment of such procedures.

Keywords: Coefficient of Variation, Coefficient, Bayesian Inference, Economic Inequalit, FBST, HPD.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

O Coeficiente de Variação (CV) é uma medida de variação relativa que elimina o efeito da magnitude dos dados, exprimindo sua variabilidade em relação à média. Essa característica permite que se compare variáveis com unidades de medidas diferentes, por exemplo, a variabilidade do peso de indivíduos (em kg) com a sua altura (em cm). Na saúde, o CV pode ser utilizado como medida de precisão (ou reprodutibilidade) de um teste de laboratório. Quanto menor o CV, mais preciso é o teste. Além disso, visto que o CV elimina o efeito da magnitude das medidas, ele permite comparar a precisão de dois instrumentos.

O coeficiente de variação é recomendado para variáveis quantitativas com escala de mensuração do tipo razão, isto é, variáveis que apresentam um zero absoluto (definidas para valores maiores ou iguais a zero). Por exemplo, calcular o CV da temperatura medida em graus Fahrenheit não tem significado, pois ao transformar essa temperatura em graus Celsius, essa transformação não altera a temperatura (só foi modificada a escala de medida), mas muda o valor do CV.

Em geral, inferências como testes de hipóteses ou intervalos de confiança do CV são realizadas sob a suposição de normalidade dos dados. No entanto essa suposição pode não ser razoável, visto que só faz sentido calcular o CV para variáveis não negativas.

O presente projeto utiliza procedimentos alternativos para inferências sobre o Coeficiente de Variação que independam da suposição de normalidade dos dados. Inicialmente realizou-se uma revisão bibliográfica das principais metodologias para realização de inferências do CV, apresentou-se os principais modelos probabilísticos de variáveis aleatórias não negativas e suas propriedades, e o teste de Kolmogorov-Smirnov foi utilizado para a verificação da suposição que os dados seguem algum desses modelos propostos.

Inferências do CV foram realizadas segundo a suposição que os dados seguem cada um dos modelos propostos. A inferência bayesiana baseou-se na média da distribuição à posteriori do CV, seu respectivo intervalo de credibilidade HPD (*highest posteriori density*) e o FBST (*Full Bayesian Significance Test*).

A metodologia proposta foi ilustrada em conjuntos de dados artificiais e de dados de rendimento domiciliar fornecidos pela PNAD 2015 (Pesquisa Nasional por Amostra de Domicílios). Todas as análises foram realizadas pelo software livre R (R CORE TEAM, 2017).

2 Coeficiente de Variação (CV)

O coeficiente de variação (CV) é uma medida estatística usada para analisar a dispersão dos dados (consistência dos dados) excluindo a influência da ordem de grandeza da variável, ou seja, o CV é definido como o desvio padrão em porcentagem da média sem importar qual a dimensão da unidade da variável. Essa medida permite a comparação da precisão entre experimentos, sem a necessidade de igualdade de unidades, podendo comparar dois bancos de dados medidos em unidades diferentes. Por ser uma medida relativa seu valor pode ser dado em forma de porcentagem. Sua definição matemática é dada por:

Definição 1 *O coeficiente de variação consiste na razão entra o desvio padrão e a média, ou seja:*

$$CV = \frac{\sigma}{\mu} \times 100\%$$

na qual:

- μ *é a média populacional;*
- σ *é o desvio padrão populacional.*

Esta medida representa a dispersão dos dados (homogeneidade) em relação à média, sendo seu resultado apresentado em forma de porcentagem. Sua interpretação é "o desvio padrão equivale a K% da média", k é o valor obtido para o coeficiente de variação.

Alguns pontos importantes sobre o CV são:

- Quanto menor o CV mais homogêneo é o conjunto de dados;
- É uma medida adimensional, ou seja, um número puro, assumindo valores positivos ou, no mínimo, igual a zero (ocorre quando não há variabilidade entre os dados, $\sigma = 0$);
- Indica a porcentagem do desvio padrão em relação à media. Quando $CV < 100\%$, o desvio padrão é menor que a média; quando $CV > 100\%$, a média é menor que o desvio padrão.

Por ser uma medida relativa, o CV possuirá valores muito semelhantes em um grande grupo de experimentos se, em cada um desses, o desvio-padrão for diretamente proporcional à média de cada grupo [15]. Estes mesmos autores lembram, entretanto, que o pesquisador deve ter cuidado, pois o CV é uma medida sem sentido no caso em que as observações experimentais envolvam valores positivos e negativos. Vale ressaltar que o coeficiente de variação deve ser

usado apenas nos casos em que a variável de interesse é positiva e do tipo razão, sendo seu significado pouco válido ou inexistente em variáveis medidas em escala intervalar, como por exemplo, graus Celsius e Fahrenheit; tendo em contraposição a escala Kelvin, que não pode assumir valores negativos e possui o zero como absoluto.

Em igualdade de condições, é mais preciso o experimento com menor coeficiente de variação [9]. De acordo com Snedecor e Cochran (1980), a distribuição do CV possibilita estabelecer faixas de valores que orientem os pesquisadores sobre a validade de seus experimentos, mas a definição de grandeza do coeficiente de variação (baixo, médio, alto e outros) é algo muito relativo e depende do estudo que está sendo realizado, sendo necessária familiaridade com o material que é objeto de pesquisa. Pimentel Gomes (1990) [19], estudando os coeficientes de variação obtidos nos ensaios agrícolas (atributos de solo), classifica-os da seguinte forma:

- CV baixo: inferiores a 10%
- CV médio: entre 10 e 20%
- CV alto: entre 20 e 30%
- CV muito alto: valores acima de 30%.

Em adição, Pimentel Gomes (2000) [20] classifica o coeficiente de variação de forma diferente, agora para experimentos de campo:

- Baixo: menor ou igual a 15%
- Médio: entre 15 e 30%
- Alto: maior que 30%

Isso mostra que, de fato, a classificação dessa medida requer conhecimento quanto ao assunto e depende fortemente do que está sendo estudado. Sendo assim, para cada situação deve-se procurar referências que estejam de acordo com o que será avaliado no estudo.

Outro exemplo de classificação do Coeficiente de Variação é o definido por Amaral et al. (1997) [1] e Judice et al. (1999) [13]. Estes sugerem verificar a normalidade da distribuição dos coeficientes de variação para encontrar as faixas de variabilidade. Contudo, Costa et al. (2002) [5], apresentaram um novo método de classificação dos coeficientes de variação que pode ser aplicado independentemente da distribuição dos valores de CV. Este método baseou-se no uso da mediana (Md) e do pseudo-sigma (PS), medidas mais resistentes que a média e o desvio-padrão, segundo Costa et al. (2002) [5]. Quando os dados não têm distribuição normal, o uso do pseudo-sigma e da mediana será mais resistente que o desvio padrão e a média. Se os dados têm distribuição aproximadamente normal, o pseudo-sigma produz uma estimativa

próxima de s, que é o desvio-padrão da amostra [2, 11]. Nesse caso, o coeficiente de variação é dado pela fórmula a seguir:

Definição 2 *O coeficiente de variação, quando considera-se uma distribuição não normal, é dado por:*

$$CV = \frac{PS}{Md} \times 100$$

onde:

$$Md = \frac{Q1 + Q3}{2}$$

é a mediana dos coeficientes de variação para $Q1$ e $Q3$ o primeiro e o terceiro quartil, respectivamente, e delimitam 25% de cada extremidade da distribuição.

O pseudo-sigma é uma medida equivalente ao desvio padrão, porém, difere pelo fato de que esse é resistente a ruídos ou outliers, enquanto o desvio padrão não. Sua fórmula é dada por:

$$PS = \frac{IQR}{1,35}$$

onde IQR é a amplitude interquartílica (Q3-Q1), medida resistente que indica o quanto os dados estão distantes da mediana.

Costa et al. (2002) [5] definem as faixas de classificação do coeficiente de variação dado pelo pseudo sigma e pela mediana como:

- Baixo: $CV \leq (Md - 1PS)$
- Médio: $(Md - 1PS) < CV \leq (Md + 1PS)$
- Alto: $(Md + 1PS) < CV \leq (Md + 2PS)$
- Muito alto: $CV > (Md + 2PS)$

Costa et al (2002) [5] argumenta que o pseudo-sigma corresponde ao desvio padrão que uma distribuição normal precisa ter para produzir a mesma distância interquartílica com os dados utilizados, o que justifica o fator 1,35 presente na fórmula do IQR. Essa definição é melhor explicada a seguir:

$$\begin{cases} Q1 = \mu - 0,675\,\sigma \\ Q3 = \mu + 0,675\,\sigma \end{cases}$$

Assim,

$$Q3 - Q1 = (\mu + 0,675\,\sigma) - (\mu - 0,675\,\sigma) = 0,675 + 0,675 = 1,35$$

Nesse trabalho essa metodologia não será utilizada, servindo apenas como exemplo de faixas de classificação do Coeficiente de Variação.

Uma outra aplicação do coeficiente de variação é na questão de desigualdade social: este preenche as exigências para uma medida de desigualdade econômica. Nesse caso, é importante ressaltar que:

- Seja x_i o salário recebido pelo *i-ésimo* indivíduo. Se todos $x_i's$ forem iguais, então $V(x_i) = 0$ e, consequentemente, CV=0;
- CV é uma medida invariante à escala: $c_v(x) = c_v(\alpha x)$, em que α é um número real;
- CV é não invariante à translação: se for adicionada (ou subtraída) a mesma quantidade de todos os salários originais, o valor do CV irá mudar;
- Anonimato: CV é independente do ordenamento da lista de salários;
- Satisfaz o princípio da transferência de Pingu-Dalton: a redistribuição da riqueza de uma pessoa mais rica para uma menos rica diminui o valor da desigualdade, visto que o denominador permanece igual, enquanto o numerador diminui. Esse decréscimo na desigualdade (menor valor do CV) será maior quando a redistribuição da renda envolver extremos (mais ricos e mais pobres).

3 Inferência Bayesiana

A teoria estatística clássica consiste em pensar no processo aleatório que produziu os dados observados, como por exemplo o lançamento de um dado: há seis possíveis resultados, representados pelos seis números das faces dos dados. Aqui, vamos considerar como sucesso o caso em que se obtém a face com número 3 (três). O estatístico frequentista acredita no processo aleatório do lançamento dos dados sendo repetido infinitas vezes, tendo por base a experiência de estabilidade da frequência relativa de ocorrência dos eventos quando esses são realizados muitas vezes. A frequência relativa é dada por:

$$f_i = \frac{n_i}{N} :$$

- f_i representa a frequência com que se obteve o *i-ésimo* possível resultado, $i = 1, 2, ..., 6$;
- n_i representa o número de sucessos do *i-ésimo* resultado possível, $i = 1, 2, ..., 6$;
- N representa o número total de vezes que o experimento (lançar o dado) foi repetido.

No caso do lançamento de dados, as observações consistem na quantidade de vezes que se obteve a face de número 3, dentre todos os lançamentos independentes realizados. Sendo assim, temos que a frequência relativa da faces de número 3 é $f_3 = \frac{n_3}{N}$. Essa frequência equivale à quantidade de sucessos no exemplo em questão.

Os frequentistas estimam a probabilidade θ de obter a face de número 3 como o valor de f_3, calculada com N tendendo ao infinito. Sendo assim, o parâmetro θ pode ser estimado pelos dados observados.

A repetição do processo aleatório gerador dos dados (ou amostragem repetida), é fundamental na visão clássica, a qual considera o parâmetro como fixo, não sendo aceitável colocar distribuições de probabilidade nos parâmetros. Aqui, apenas os dados são aleatórios e, assim, somente esses podem ter uma distribuição de probabilidade associada. Sendo assim, na visão clássica calcula-se $P(x|\theta)$, sendo x o vetor dos dados.

Em adição, os clássicos também consideram duas estatísticas amplamente utilizadas: valor-p e intervalo de confiança, as quais são definidas abaixo para uma confiança de $(1-\alpha) \times 100\%$ [6]:

- O valor-p (nível descritivo do teste) corresponde à probabilidade de observar valores mais extremos que o observado para a estatística do teste (considerada uma variável aleatória), dado que a hipótese nula (H_0) é verdadeira. Rejeita-se a hipótese nula quando o valor-p é menor ou igual ao nível de significância do teste (α);

- Um intervalo de confiança (*IC*) fornece um intervalo de estimativas prováveis para o valor do parâmetro de interesse (no caso do dado, a probabilidade de obter a face de número 3), ao invés de dar apenas uma estimativa pontual. O nível de confiança desse intervalo é determinado pelo coeficiente de confiança $(1-\alpha)$, $\alpha \in (0,1)$ e corresponde ao nível de significância do teste. Estes servem para identificar a confiabilidade de uma estimativa: quanto menor o *IC*, mais confiável é o intervalo, pois isso ocorre principalmente quanto menor for a variância do parâmetro a ser estimado, já que a amplitude desse intervalo está associada à incerteza que temos a respeito do parâmetro. Uma interpretação para o intervalo de confiança é que ele contém os valores plausíveis que θ pode assumir. Resumidamente, se o processo aleatório gerador dos dados for realizado um número grande de vezes e fosse construído um IC com confiança de $(1-\alpha)\times 100\%$ para cada uma dessas repetições, espera-se que $(1-\alpha)\times 100\%$ dos intervalos contenham o real valor do parâmetro desejado e, por consequência, $(\alpha)\times 100\%$ deles não conterão o valor real do parâmetro θ.

As definições acima descritas dependem fortemente do conceito de múltiplas repetições do processo aleatório, o qual é o gerador dos dados. Contudo, seriam mais naturais definições que contam somente com os dados observados em mãos. Dessa maneira, por meio da visão Bayesiana, as definições ideais seriam [6]:

- valor-*p* é a probabilidade estimada da hipótese nula ser verdadeira, dadas as observações;
- Intervalo de confiança (*IC*) de $(1-\alpha)\times 100\%$ é um intervalo que contém o valor real do parâmetro desejado com uma probabilidade de $(1-\alpha)$.

3.1 Paradigma Bayesiano

A inferência clássica baseia-se no princípio da repetibilidade, como mencionado anteriormente. Uma vez determinado o modelo estatístico, a estimação de parâmetros e testes de hipóteses sobre os parâmetros são realizados considerando-se a variabilidade inerente à amostra observada. Sendo assim, a inferência clássica consiste em considerar a variabilidade passível de ser observada na estimação do parâmetro θ por $\hat{\theta}$ caso um mesmo experimento fosse repetido uma grande quantidade de vezes, sob as mesmas condições, evidenciando o fato de que diferentes amostras levarão à diferentes valores de $\hat{\theta}$.

A teoria de verossimilhança tem papel de destaque na inferência clássica, pois, dado um vetor de observações x, para qualquer $\theta \in \Theta$, $f(x|\theta)$ é uma probabilidade e, dessa maneira, há vários possíveis valores para $\hat{\theta}$, sendo o melhor valor aquele que seja mais provável (verossímil) para a amostra observada, dado pela função de verossimilhança definida abaixo [6]:

$$\hat{\theta} = \arg max_{\theta\in\Theta} L(\theta;x) = \arg max_{\theta\in\Theta} f(x|\theta)$$

Vale ressaltar que $L(\theta;x)$ é condicional nos dados observados (informação dos dados é utilizada no processo inferencial) e deve ser vista como função de θ. O valor que maximiza essa função, $\hat{\theta}$ é conhecido como estimador de máxima verossimilhança de θ, o parâmetro de interesse.

No caso da inferência Bayesiana, adota-se uma postura subjetivista com o uso explícito de probabilidades para quantificar o grau de incerteza acerca de quantidades de interesse não observadas. Sendo assim, essa vertente busca combinar a informação subjetiva (inerente, a priori) referente a um problema, com a informação obtida através dos dados observados, fazendo uso de declarações probabilísticas por meio do teorema de Bayes. No caso da inferência Bayesiana, inicialmente especifica-se um modelo probabilístico completo, o qual contém uma distribuição conjunta para todas as quantidades observadas ou não do problema em estudo. O segundo passo consiste na obtenção de uma distribuição a posteriori: distribuição condicional das quantidades não observadas dado as observadas. Por fim, verifica-se a adequação do modelo ajustado aos dados.

O parâmetro θ, $\theta \in \Theta$, no modelo clássico, é um escalar ou vetor desconhecido, mas fixo e igual ao valor particular que melhor representa os dados observados através da realização do experimento. No modelo Bayesiano, θ, $\theta \in \Theta$, é um escalar ou vetor aleatório (não observável). Sendo assim, a filosofia Bayesiana é definida como "aquilo que é desconhecido também é incerto, e toda a incerteza deve ser quantificada em termos de probabilidade". Dessa maneira, passa-se a tentar reduzir este desconhecimento.

Suponha que o interesse do estudo é estimar o parâmetro θ, do qual tem-se uma informação inicial, anterior ou externa em relação à experiência e que seja relevante [16]. A intensidade de incerteza a respeito de θ pode assumir diferentes graus, os quais são representados por modelos probabilísticos para θ, $\pi(\theta)$ no paradigma Bayesiano [7], chamados de distribuição à priori de θ. Da mesma forma que na visão clássica, toda informação gerada pelos dados observados está contida na função de verossimilhança $L(\theta;x) = f(x|\theta)$. Assim, a distribuição a priori é atualizada através da informação contida em $L(\theta;x)$, o que aumenta a informação sobre o parâmetro. O teorema de Bayes é a regra e atualização utilizada para quantificar o aumento da informação [6].

$$\pi(\theta|x) = \frac{L(\theta;x)\pi(\theta)}{\int_\Theta L(\theta;x)\pi(\theta)\mathrm{d}\theta} = \frac{f(x|\theta)\pi(\theta)}{\pi(x)} \tag{3.1}$$

A Equação (3.1) fornece a probabilidade a posteriori de θ como função da verossimilhança, $L(x;\theta)$, e da distribuição a priori de θ, $\pi(\theta)$. Vale ressaltar que o teorema de Bayes define a posteriori como uma função proporcional à verossimilhança vezes a priori, ou seja:

$$\pi(\theta|x) \propto L(\theta;x)\pi(\theta) \tag{3.2}$$

Uma tradução prática é que o novo estado do conhecimento é proporcional ao produto

da informação dos dados novos e o estado atual do conhecimento. Essa relação é possível pelo fato do denominador da fórmula de Bayes não depender de θ, sendo este apenas uma constante normalizadora da distribuição a posteriori.

É fácil notar que a função de verossimilhança é de suma importância para a inferência Bayesiana, pois, como mostrado, representa o meio pelo qual os dados x obtidos pelo experimento transformam o conhecimento a priori sobre θ.

Resumidamente, a distribuição a posteriori, $\pi(\theta|x)$, é responsável por, através do teorema de Bayes, incorporar toda a informação disponível sobre o parâmetro (priori e verossimilhança), sendo todos os procedimentos Bayesianos baseados em $\pi(\theta|x)$. É importante lembrar que, após atualizada, a distribuição a posteriori torna-se uma nova priori, podendo essa ser usada em estudos futuros. Ou seja, o conceito de priori e posteriori são relativos à observação que está sendo considerada no momento: $\pi(\theta|x)$ é a posteriori de θ em relação aos dados observados, mas a priori de θ em relação a uma nova observação y.

3.2 Prioris

Defina θ como o parâmetro do qual de deseja fazer inferências. A informação que se tem disponível sobre este parâmetro, mesmo que algumas vezes seja mais subjetiva, possibilita associar a ele uma distribuição de probabilidade que será capaz de traduzir esse conhecimento para a forma probabilística. Nos casos em que não existe informação alguma a priori ou que o conhecimento é pouco significativo, utiliza-se prioris minimamente informativas, também conhecidas como distribuições não informativas.

Para melhor entender o que é uma priori, tome como exemplo o lançamento de uma moeda, onde sucesso ocorre quando retira-se cara, $P(sucesso) = \theta$. Inicialmente imagina-se uma moeda honesta, onde a probabilidade de obter cara é a mesma de coroa e igual a $0,5$. Mas, supondo que não seja uma moeda honesta, qual informação prévia se tem acerca da moeda? Primeiramente pode-se imaginar uma distribuição Uniforme (priori não informativa) ou Beta (1;1), que associa igual probabilidade à todas as porcentagens. Nesse caso, $P(\theta) \propto c$, $0 < \theta < 1$, c é uma constante e, consequentemente, $\pi(\theta|x) \propto L(\theta;x)$. Sendo assim, ao fazer uso de uma priori não informativa, gera-se uma posteriori completamente dependente dos dados observados. Aqui, os resultados numéricos da análise Bayesiana e frequentista tornam-se muito semelhantes.

Contudo, se o pesquisador já possui algum conhecimento e sabe que, ao lançar a moeda repetidas vezes, tende-se a obter mais caras do que coroas, é possível definir, por exemplo, uma Beta(45,5), a qual concentra mais valores na cauda direita da distribuição, que equivale à valores mais elevados para θ. E assim sucessivamente, escolhendo-se a distribuição que melhor se adapta às informações prévias relacionadas ao lançamento da moeda.

A escolha de prioris deve ser feita de forma cautelosa, dado que pode trazer dificuldade

técnicas como, por exemplo, se o intervalo de variação de θ for ilimitado, então a distribuição a priori constante é imprópria [7].

A classe de prioris não informativas proposta por Jeffreys [12] é invariante a transformações 1 a 1, embora em geral seja imprópria [7]. Faz-se necessário definir a Informação de Fisher, antes de especificar o que é a priori de Jeffreys.

Definição 3 *Considere uma única observação x com densidade de probabilidade* $P(x|\theta)$. *A informação de Fisher é definida como:*

$$I(\theta) = E\Big[-\frac{\partial^2 log(P(x|\theta))}{\partial\theta^2}\Big] \tag{3.3}$$

A informação aqui é associada a uma espécie de curvatura da função de verossimilhança: quanto maior for a curvatura, maior será o valor de $I(\theta)$ e mais precisa é a informação contida na verossimilhança.

Definição 4 *Seja X uma observação com função de probabilidade* $P(x|\theta)$. *A priori não informativa de Jeffrey tem densidade dada por [7]:*

$$\pi(\theta) \propto [I(\theta)]^{\frac{1}{2}} \tag{3.4}$$

Se θ *for um vetor paramétrico, então:*

$$\pi(\theta) \propto |I(\theta)|^{\frac{1}{2}} \tag{3.5}$$

3.3 Estimativas Pontuais

Uma estimativa pontual é um valor (ou ponto) único usado para aproximar um parâmetro populacional, obtido através dos dados observados. A distribuição a posteriori dos parâmetros de interesse θ contém toda a informação a respeito do parâmetro (priori e verossimilhança). Contudo, ocorre de algumas vezes fazer-se necessário resumir toda essa informação em um valor numérico, ao qual se dá o nome de estimativa pontual. Para cada parâmetro tem-se uma estimativa.

A estatística Clássica resolve o problema da estimação por meio do uso dos estimadores de máxima verossimilhança. Já a Bayesiana opta pelo conceito da função perda para auxílio na escolha do estimador [7]:

Definição 5 *A cada decisão* δ *e a cada possível valor do parâmetro* θ *podemos associar uma perda* $l(\delta, \theta)$*, que assume valores positivos. Definimos assim uma função de perda* $l(\delta,\theta)$*, que*

representa a perda sofrida ao se tomar a decisão dado o real estado $\theta \in \Theta$. *A esta função perda está relacionado um risco, o qual corresponde à perda esperada a posteriori.*

$$R(\delta) = E_{\theta|x}[L(\delta, \theta)] \tag{3.6}$$

Uma regra de decisão ótima, δ^* *é aquela que possui risco mínimo, chamada de regra de Bayes, associada ao risco de Bayes.*

Se considerarmos uma amostra aleatória $X_1, ..., X_n$ de uma função de densidade de probabilidade $p(x|\theta)$, com θ desconhecido. Nesse caso, a estimação desse parâmetro depende dos valores observados na amostra, sendo que, se $\theta \in \Theta$, então os possíveis valores do estimador $\delta(x)$ também devem pertencer ao espaço δ. Um bom estimador é aquele para o qual o erro $\delta(x) - \theta$ estará próximo de zero, com alta probabilidade [7]. Associando uma perda $L(a^*, \theta)$ para cada possível valor de θ e cada possível estimativa $a^* \in \Theta$, de maneira que quanto menor a distância entre os dois parâmetros, menor o valor da perda. Assim, aqui o valor da perda esperada a posteriori é dada por:

$$E_{\theta|x}[l(a^*, \theta)] = \int l(a^*, \theta)\pi(\theta|x)d\theta \tag{3.7}$$

e a regra de Bayes consiste em escolher a^* de forma que minimize (3.7)

Algumas escolhas usuais para a função perda são [7]:

- Função Perda Quadrática: $L(a^*, \theta) = (a^* - \theta)^2$
- Função Perda Absoluta : $L(a^*, \theta) = |a^* - \theta|$
- Função "Perda Tudo ou Nada":

$$L(a^*, \theta) = \begin{cases} 1 & , \text{se } |a^* - \theta| > \varepsilon \\ 0 & , \text{se } |a^* - \theta| < \varepsilon \end{cases}$$

para algum $\varepsilon > 0$.

Os estimadores de Bayes para as três funções perda acima são, respectivamente, média a posteriori, mediana a posteriori e moda a posteriori [7].

3.4 Estimação por Intervalos

A posteriori contem toda a informação que se tem acerca do parâmetro, e a estimativa pontual do parâmetro resume a informação contida na posteriori em único valor (ou em k valores,

k representa o número de parâmetros). Também é importante associar uma medida sobre o quão precisa é essa estimação ou especificação desse valor, sendo esse o papel do intervalo de credibilidade Bayesiano, definido a seguir:

Definição 6 *C é um intervalo de credibilidade de* $100 \times (1-\alpha)\%$*, ou nível de credibilidade* $(1-\alpha)$*, para* θ*, se* $P(\theta \in C) \geq 1-\alpha$*.*

Esse intervalo fornece a probabilidade de um parâmetro (θ) pertencer ao intervalo, de modo que, sabendo a distribuição da quantidade de interesse, torna-se possível calcular os limites que estabelecem essa probabilidade. Quanto menor for o intervalo, mais concentrada é a distribuição do parâmetro, ou seja, o tamanho do intervalo informa sobre a dispersão do parâmetro [7]. Além disso, a exigência de que a probabilidade possa ser maior que a credibilidade exigida é meramente técnica, pois para distribuições contínuas é possível calcular essa probabilidade exata mas, já para as discretas, nem sempre é possível satisfazer a igualdade (optando por intervalos com a menor credibilidade possível, mas acima da exigida).

Com a Definição 6 é possível construir diversos intervalos, porém, o interesse está baseado em apenas aquele que possui a menor amplitude (comprimento) possível, os quais são obtidos através de valores de θ com a maior densidade a posteriori. A esse intervalo se dá o nome de HPD (*Highest Posterior Density)*, ou MPD (Maior densidade a posteriori), em português, e sua definição é dada a seguir:

Definição 7 *Um intervalo de credibilidade C de* $(1-\alpha) \times 100\%$ *para* θ *é dito HPD se* $C = \{\theta \in \Theta : \pi(\theta|x) \geq k(\alpha)\}$*, sendo* $k(\alpha)$ *a maior constante tal que* $P(\theta \in C) \geq 1-\alpha$*.*

Sendo assim, qualquer ponto contido nesse intervalo terá densidade maior que de pontos fora do intervalo. Vale ressaltar que os intervalos *HPD* não são invariantes para transformações 1 a 1 que não sejam transformações lineares.

3.5 Teste de Hipótese

Os testes de hipóteses são usados para determinar se existe evidência suficiente nos dados gerados pelo experimento aleatório para se inferir que uma determinada condição é verdadeira para toda a população. Um teste analisa duas hipóteses, nula (H_0) e alternativa (H_a).

Considerando θ como o parâmetro de interesse (ou vetor de parâmetros), e Θ o espaço paramétrico, o problema de testar $H_0 : \theta \in \Theta_0$ *vs* $H_a : \theta \in \Theta_1 = \Theta - \Theta_0$ na inferência Bayesiana consiste apenas em calcular as probabilidades à posteriori:

- $P(H_0|X) = P(\theta \in \Theta_0|X)$ e

- $P(H_a|X) = P(\theta \in \Theta_1|X) = 1 - P(H_0|X)$

Ademais, pode-se optar por uma das hipóteses em função de uma grandeza relativa. Nesse caso, calcula-se o *Odds* (Chance) à posteriori a favor de H_0 (ou de H_0 sobre H_a) [16]:

$$O(H_0, H_a|x) = \frac{P(H_0|X)}{P(H_a|X)} = \frac{P(H_0|X)}{1 - P(H_0|X)} \tag{3.8}$$

A evidência a favor de H_0 antes da observação dos dados pode ser calculada de forma análoga, denominada *Odds* à priori:

$$O(H_0, H_a) = \frac{P(H_0)}{P(H_a)} = \frac{P(H_0)}{1 - P(H_0)} \tag{3.9}$$

Para avaliar a influência dos dados observados X na alteração da credibilidade relativa de H_0 e H_a, faz-se a a contraposição da razão a posteriori com a razão das probabilidades a priori, gerando a *Odds Ratio* ou Razão de Chances:

$$B_0(x) = \frac{O(H_0, H_a|x)}{O(H_0, H_a)} = \frac{\frac{P(H_0|X)}{1-P(H_0|X)}}{\frac{P(H_0)}{1-P(H_0)}} \tag{3.10}$$

Aqui $B_0(x)$ é denominado como Fator de Bayes a favor de H_0 (ou contra H_a). Aplicando o logaritmo tem-se uma relação de adição entre os termos:

$$ln\,(B_0(x)) = ln\,(O(H_0, H_a|x)) - ln\,(O(H_0, H_a)) \tag{3.11}$$

Esses termos são chamados de pesos de evidência. Assim, o Fator de Bayes ou o logaritmo dele representam o peso relativo da evidência contida nos dados a favor de uma ou outra hipótese em confronto. O logaritmo do fator de Bayes é visto como o peso da evidência a posteriori descontado do correspondente peso da evidência a priori [16], onde valores muito grandes ou muito pequenos representam uma tendência forte nos dados a favor de uma hipótese contra a outra, ou seja, uma hipótese é muito mais ou menos provável do que era a priori. A Tabela 1 apresenta a escala de evidências de Jeffreys para o fator de Bayes [12]

Tabela 1 – Escala de evidências de Jeffreys para o Fator de Bayes

$B_0(X)$	Evidência à favor de H_0
<1	negativa (evidencia a favor de H1)
1 a 3	insignificante
3 a 10	moderada
10 a 30	forte
30 a 100	muito forte
> 100	decisiva

3.6 Hipóteses precisas

O problema de se testar uma hipótese do tipo $H_0 : \theta = \theta_0$ *vs* $H_a : \theta \neq \theta_0$, $\theta_0 \in \Theta$, se torna impossível quando Θ segue uma distribuição contínua, já que é inviável calcular o Fator de Bayes, visto que $P(\theta = \theta_0) = 0, \forall\, \theta_0 \in \Theta$.

Consolidou-se na literatura a abordagem de Jeffreys, Savage e Good, os quais defendem o fato de θ_0 possuir, a priori, uma ordem de importância diferente da que é atribuída aos demais valores de θ devido a própria formulação da hipótese nula. Assim, essa ordem de importância deve ser formalizada e integrada na distribuição a priori, passando essa a ter uma natureza mista e com massa pontual concentrada em θ_0, $p_0 = P(H_0)$, e uma distribuição contínua da massa restante, $P\{H1\} = 1 - p_0$, em $\theta \neq \theta_0$. Ou seja, sendo $h(\theta)$ a priori para θ, tem-se:

$$h(\theta) = \begin{cases} p_0 & \text{, se } \theta = \theta_0 \\ (1 - p_0)\, P_1(\theta) & \text{, se } \theta \neq \theta_0 \end{cases}$$

na qual $P_1(\theta)$ é a distribuição a priori dos valores de $\theta \neq \theta_0$. Assim, a distribuição a posteriori mista de θ é definida como:

$$P(\theta|x) = \begin{cases} \frac{p_0 L(\theta_0;x)}{p(x)} & \text{, se } \theta = \theta_0 \\ \frac{p_1 (1-p_0) L(\theta|x)}{p(x)} & \text{, se } \theta \neq \theta_0 \end{cases}$$

$p(x) = p_0 L(\theta_0|x) + (1 - p_0) = \int_{\theta \neq \theta_0} p_1(\theta) L(\theta; x) \mathrm{d}\theta$.

Neste caso, o fator de Bayes a favor de H_0 é dado por:

$$B_0(x) = \frac{O(H_0, H_a|x)}{O(H_0, H_a)} = \frac{\frac{\frac{p_0 L(\theta_0|x)}{p(x)}}{1 - \frac{p_0 L(\theta_0;x)}{p(x)}}}{\frac{p_0}{1-p_0}}$$

3.7 Teste de significância Genuinamente Bayesiano (FBST)

Jeffrey [12] aponta que uma hipótese nula pode ser rejeitada utilizando-se teste de significância clássicos, enquanto a probabilidade de não rejeitar no campo Bayesiano é alta, baseando-se em uma priori pouco provável para a hipótese nula e uma distribuição difusa de probabilidade remanescente para hipóteses alternativas [22]. O Paradoxo de Lindley, como ficou conhecido esse conflito entre as abordagens estatísticas, está baseado em grandes amostras e no pressuposto de que θ é possível de se igualar a θ_0. A evidência contra uma hipótese nula (medida pela probabilidade posteriori, fator de Bayes ou verossimilhança comparativa) pode diferir de forma significativa em relação ao valor-p. Concluiu-se, com Berger e Selke (1987) [3], que os valores-p podem ser, em algumas situações, medidas enganosas quanto à evidência trazida pelos dados contra a hipótese nula precisa.

As abordagens geram desentendimentos devido a diversos fatores. Primeiramente, os clássicos não consideram a hipótese alternativa no cálculo do valor-p [17]; enquanto os Bayesianos consideram ambas hipóteses. Em adição, o fato de atribuir uma probabilidade a priori para a hipótese nula faz com que a probabilidade remanescente, distribuída de forma difusa para a hipótese alternativa, pode resultar em uma probabilidade posteriori pequena para a mesma, o que favorece a hipótese nula. Sendo assim, segundo Lindley (1997) [14], a utilização do fator de Bayes para testes de significância com hipóteses precisas é controverso. Dessa forma, levando em consideração o que fora citado e todo o desconforto de se trabalhar com distribuições mistas, Pereira e Stern (1999) [18] propôs a medida de evidência genuinamente Bayesiana, a qual pode ser obtida através do procedimento *"Full Bayesian Significance Test (FBST)"*.

O FBST recebe o termo "genuinamente Bayesiano"porque a medida de evidência baseia-se apenas na distribuição posteriori e pelo fato da mesma ser caracterizada pela teoria da decisão. A medida de evidência a favor de uma hipótese nula precisa é definida a seguir:

Definição 8 *Sendo X uma variável aleatória associada a um modelo estatístico paramétrico. Considere que após a observação de x obtém-se a densidade a posteriori de* θ *(parâmetro de interesse),* $\pi(\theta|x)$*, restringindo a uma função de densidade de probabilidade. Seja* T_ϕ *o conjunto do espaço paramétrico onde a densidade a posteriori é maior que* ϕ*:*

$$T_\phi = \{\theta \in \Theta : \pi(\theta|x) \geq \phi\}$$

A probabilidade a posteriori $k = \int_{T_\phi} \pi(\theta|x)d\theta$ *é a credibilidade de* T_ϕ*. Além disso, o máximo da densidade posteriori sob a hipótese nula (*H_0*),* P^**, é obtido por* θ^**:*

$$\theta^* \in sup_{\theta \in \Theta_0} \pi(\theta|x), \pi^* = \pi(\theta^*|x)$$

O conjunto "tangente"à hipótese nula com credibilidade k^* *é dado por* $T^*(x) = \{\theta \in \Theta : \pi(\theta|x) > \pi^*\}$*. A medida de evidência de Pereira-Stern em favor de* H_0*, valor-e, é a probabilidade complementar do conjunto* T^**, isto é:*

$$valor\text{-}e = 1 - k^* = 1 - P(\theta \in T^*(x)|x)$$

O procedimento FBST consiste em aceitar H_0 *sempre que o valor-e for grande.*

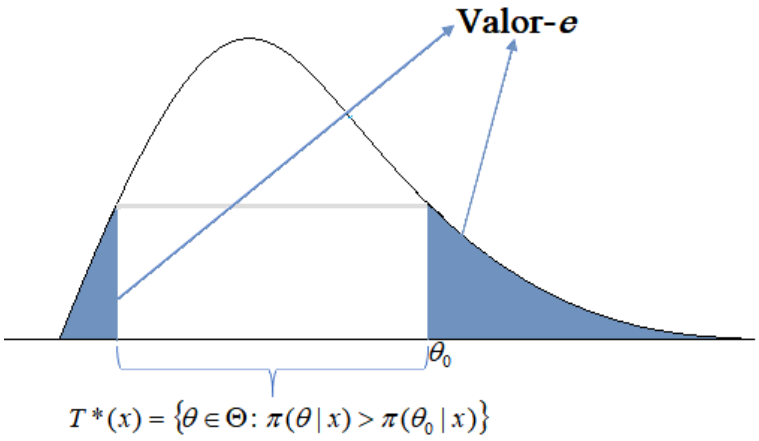

Figura 1 – Representação geométrica do cálculo do valor-e do FBST

Sendo assim, a medida de evidência proposta em favor de H_0 considera os os pontos do espaço paramétrico com densidade posteriori no máximo tão grande quanto o supremo em Θ_0. Sendo assim, um grande valor de valor-e significa que o subconjunto Θ_0 cai em uma região do espaço paramétrico com alta probabilidade, fazendo com que $T^*(x)$ tenha probabilidade posteriori pequena e, assim, os dados "suportam"H_0. Em oposição, um pequeno valor$e$ leva à rejeição da hipótese nula [17].

Vale ressaltar que, embora as definições apresentadas aqui sejam para hipóteses com dimensão menor que a do espaço paramétrico, nada impede que seja usado para o caso em que $dim(\Theta_0) = dim(\Theta)$.

A realização do FBST consiste em:

1. Encontrar $\theta^* = \underset{\theta \in \Theta_0}{\arg max}\, \pi(\theta|x)$;
2. Obter o valor-e da densidade a posteriori no ponto θ^*, $\pi^* = \pi(\theta^*|x)$;
3. Obter a região tangente à hipótese nula $T^* = \{\theta \in \Theta : \pi(\theta|x) > \pi^*\}$;
4. Calcular a credibilidade $k^* = P(\theta \in T^*(x)|X)$;
5. Calcular o valor da evidência valor-e = $P(\theta \notin T^*(x)|X) = 1 - k^*$

4 Variável aleatória

Considere um experimento , com Ω sendo o espaço amostral associado a esse experimento. Uma função X, que associa a cada elemento $\omega \in \Omega$ um número real, $X(\omega)$, é denominada variável aleatória (v.a.). Ou seja, variável aleatória é uma função que a cada acontecimento do espaço de resultados faz corresponder um valor real. Estas podem ser classificadas em:

- Discreta: se o número de valores possíveis de X for enumerável (finito ou infinito). Isto é, os possíveis valores de X podem ser postos em lista como $x_1, x_2, \ldots$. No caso finito, a lista possui um valor final x_n, e no caso infinito, a lista continua indefinidamente. Exemplos: número de filhos, quantidade de peças defeituosas em uma produção de uma fábrica, número de pontos em um jogo de basquete e outros.

- Contínua: se o número de valores possíveis de X for não enumerável. Isto é, a variável pode assumir qualquer valor dentro de um intervalo. Exemplos: altura, peso, concentração de CO_2 na água e outros.

Uma variável aleatória é uma variável que tem um valor único (determinado aleatoriamente) para cada resultado de um experimento. A palavra aleatória indica que em geral só conhecemos aquele valor depois do experimento ser realizado [23].

4.1 Distribuição de probabilidade

Uma distribuição de probabilidade é um modelo matemático que relaciona um certo valor da variável aleatória em estudo com a sua probabilidade de ocorrência. Isto é, a distribuição de probabilidade associa uma probabilidade a cada resultado numérico de um experimento, dando a probabilidade de cada um dos valores da variável aleatória. Há dois tipos de distribuições, definidos conforme o tipo de variável em estudo:

- Discreta: quando a variável que está sendo medida só assume valores enumeráveis, ou seja, a variável é discreta;

- Contínuas: quando a variável que está sendo medida é expressa em uma escala contínua, ou seja, o estudo ocorre em relação à uma variável contínua.

Definição 9 *Seja X uma variável aleatória discreta. Dá-se o nome de função de probabilidade da variável aleatória X à função que a cada x_i associa sua probabilidade de ocorrência, ou seja:*

$$p(x_i) \;=\; P(X = x_i),\; i = 1,2,3,...$$

tal que:

- *Soma de todos os valores de uma distribuição de probabilidade, $P(x)$, x assume todos os valores possíveis da variável aleatória, deve ser igual a 1:*

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

- *A probabilidade de ocorrência de um dado evento deve ser um número entre 0 e 1, ou seja:*

$$0 \leq p(x_i) \leq 1, i = 1,2,...$$

Definição 10 *Seja X uma variável aleatória contínua. Dá-se o nome de função de densidade de probabilidade (FDP) da variável aleatória X à função $f(x)$ que satisfaça:*

- $f(x) \geq 0$, $\forall x \in \mathbb{R}$
- $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$
- *Para quaisquer a e b, com $-\infty < a < b < +\infty$, tem-se: $P(a \leq X \leq b) = \int_a^b f(x)\mathrm{d}x$*

Observações:

1. *$P(a < X < b)$ representa a área sob a curva da função densidade de probabilidade entre a e b.*
2. *Para qualquer valor específico de X, x_0 por exemplo, tem-se que $P(X = x_0) = 0$, pois $P(X = x_0) = \int_{x_0}^{x_0} f(x)\mathrm{d}x = 0$*
3. *Dado que a probabilidade de X assumir valores em pontos isolados é nula, tem-se que: $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$*

É possível calcular algumas características para ambos os tipos de distribuições, tais como:

- Média (ou esperança): indica o valor médio que esperaríamos ter se pudéssemos repetir as provas infinitamente;

- Desvio padrão: obtido através da raiz quadrada da variância, essa medida indica quanto a distribuição de probabilidades se dispersa em torno da média. Um grande desvio padrão reflete dispersão considerável, enquanto que um desvio padrão menor traduz menor variabilidade, com valores relativamente mais próximos da média.

- Coeficiente de variação: definido desvio padrão em porcentagem da média e serve para analisar a consistência dos dados excluindo-se a unidade de medição da variável.

As fórmulas para o cálculo de cada uma dessas medidas são definidas abaixo, separadas de acordo com o tipo da variável aleatória:

- Distribuições discretas:
  - Média:
$$\mu(x) = \sum_{i=1}^{\infty} x_i \, p(x_i)$$
  - Variância:
$$\sigma^2(x) = \sum_{i=1}^{\infty} (x_i - \mu)^2 \, p(x_i)$$
  - Desvio padrão:
$$\sigma(x) = \sqrt{\sigma^2(x)}$$

- Distribuições contínuas:
  - Média:
$$\mu(x) = \int_{-\infty}^{+\infty} x \, f(x) \mathrm{d}x$$
  - Variância:
$$\sigma^2(x) = \int_{-\infty}^{+\infty} (x - \mu)^2 \, f(x) \mathrm{d}x$$
  - Desvio padrão:
$$\sigma(x) = \sqrt{\sigma^2(x)}$$

Tendo como base os conceitos acima definidos, as próximas seções apresentam as principais distribuições a serem discutidas nesse documento.

4.1.1 Distribuições discretas

4.1.1.1 Binomial

A distribuição Binomial é uma repetição de n ensaios independentes de Bernoulli, sendo adequada para descrever situações nas quais os resultados da variável aleatória podem ser agrupadas em apenas dois resultados possíveis. Estes devem ser mutuamente excludentes, de

forma que não haja dúvida na classificação do resultado; e coletivamente exaustivos, de maneira que não seja possível nenhum outro resultado diferente. Em geral, as categorias são denominadas como "sucesso"ou "fracasso", com $P(sucesso) + P(fracasso) = 1$.

Condições para aplicação:

1. São feitas n repetições do experimento de Bernoulli;
2. Há apenas dois possíveis resultados no experimento;
3. A probabilidade de sucesso, p, e de fracasso, $(1-p)$, permanecem constantes em todas as n repetições;
4. As repetições são independentes, ou seja, o resultado de uma repetição não é influenciado pelos outros resultados.

Em uma sequência de n observações independentes com probabilidade de sucesso constante e igual a p, a distribuição do número de sucessos (x) seguirá um modelo Binomial$(n;\ p)$:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, 2..., n$$

$\binom{n}{x}$ representa o número de combinações de n objetos tomados x a x, calculado como:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

A média, variância e coeficiente de variação da distribuição Binomia são dados por, respectivamente:

$$\mu(X) = np,$$
$$\sigma^2(X) = np(1-p) \text{ e}$$
$$CV = \sqrt{\frac{1-p}{np}}.$$

4.1.1.2 Geométrica

Considere que ensaios de Bernoulli independentes sejam realizados sucessivamente. Uma variável X que fornece o número de falhas (fracassos) até que se obtenha o primeiro sucesso, com probabilidade de sucesso igual a p, segue uma distribuição Geométrica. Sendo assim, diz-se que $X \sim Geom(p)$ com função de probabilidade dada por:

$$P(X = x) = p(1-p)^x, x = 0, 1, 2, ...$$

p é a probabilidade de sucesso. A média, variância e coeficiente de variação da distribuição Geométrica são dados, respectivamente, por:

$$\mu(X) = \frac{1-p}{p},$$
$$\sigma^2(X) = \frac{1-p}{p^2} \text{ e}$$
$$CV = \frac{1}{\sqrt{1-p}}.$$

4.1.1.3 Binomial Negativa

Uma variável aleatória X segue uma distribuição Binomial Negativa se X conta o número de fracassos necessários para se obter k ($k \geq 1$) sucessos em ensaios de Bernoulli independentes, com probabilidade de sucesso igual a p em cada ensaio. Sendo assim, o último ensaio (repetição) ocorre quando obtém-se o *k-ésimo* sucesso.

O termo "binomial negativa"vem da inversão do interesse da análise: na binomial tem-se n repetições (n fixo) e deseja-se verificar quantos sucessos se obtém; na binomial negativa deseja-se obter k sucessos (k fixo) e verifica-se quantas observações são necessárias para isso.

A função de probabilidade de $X \sim BN(p;\, k)$, é dada por:

$$P(X = x) = \binom{x+k-1}{x} p^k (1-p)^x,\ x = 0, 1, 2, ...$$

A média, variância e coeficiente de variação da distribuição Binomial Negativa são dados por, respectivamente:

$$\mu(X) = \frac{k(1-p)}{p},$$
$$\sigma^2(X) = \frac{k(1-p)}{p^2} \text{ e}$$
$$CV = \frac{1}{\sqrt{k(1-p)}}.$$

4.1.2 Poisson

A distribuição de Poisson expressa a probabilidade de ocorrência de uma serie de eventos em um certo período de tempo (ou área ou volume), sendo estes eventos independentes de quando ocorreu o último evento. A função de probabilidade de $X \sim Poisson(\lambda)$, sendo $\lambda \geq 0$ a taxa média de ocorrência do evento, é dada por:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!},\ x = 0, 1, 2, ...$$

A média, variância e coeficiente de variação da distribuição de Poisson são dados, respectivamente, por:

$$\mu(X) = \lambda,$$
$$\sigma^2(X) = \lambda \text{ e}$$
$$CV = \frac{1}{\sqrt{\lambda}}.$$

4.1.3 Distribuições Contínuas

4.1.3.1 Gama

A distribuição Gama tem como principal aplicação a análise do tempo de vida de produtos, sendo comumente usada em estudos de sobrevivência e confiabilidade. É uma das distribuições mais gerais, pois muitas outras são um caso particular desta, como por exemplo a exponencial.

A função de densidade de uma variável X que segue uma distribuição Gama ($x \sim Gama(\alpha; \beta)$) é dada por:

$$f(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, x > 0$$

na qual $\alpha > 0$ (conhecido como parâmetro de forma) e $\beta > 0$ (conhecido como parâmetro de taxa).

A média, variância e coeficiente de variação da distribuição Gama são dados, respectivamente, por:

$$\mu(X) = \frac{\alpha}{\beta},$$
$$\sigma^2(X) = \frac{\alpha}{\beta^2} \text{ e}$$
$$CV = \frac{1}{\sqrt{\alpha}}.$$

4.1.3.2 Exponencial

Na distribuição exponencial a variável aleatória é definida como o tempo entre duas ocorrências, sendo a média de tempo entre ocorrências igual a $\frac{1}{\lambda}$, $\lambda > 0$ é a taxa de ocorrências.

A função densidade de probabilidade ($X \sim Exp(\lambda)$) é dada por:

$$f(x) = \lambda e^{-\lambda x}, x \geq 0$$

A média, variância e coeficiente de variação da distribuição Exponencial são dados por, respectivamente:

$$\mu(X) = \frac{1}{\lambda},$$
$$\sigma^2(X) = \frac{1}{\lambda^2} \text{ e}$$
$$CV = 1.$$

4.1.3.3 Weibull

A distribuição de Weibull é muito flexível e pode assumir uma variedade de formas. Primeiramente proposta por W. Weibull (1954) em estudos relacionados ao tempo de falha devido a fadiga de metais, essa distribuição tem sido usada para modelar tempos de processo ou tempo até a falha de componentes, sejam elétricos, mecânicos, estruturais e outros. Uma propriedade da distribuição de Weibull é que, mesmo assumindo uma grande variedade de formas, sua função de taxa de falha é monótona, ou seja, ou ela é crescente, ou decrescente ou constante.

A função de densidade de uma variável X que segue uma distribuição Weibull com parâmetros α e β ($X \sim Weibull(\alpha;\, \beta)$) é dada por:

$$f(x) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-(\frac{x}{\beta})^\alpha},\ \alpha > 0,\ \beta > 0,\ \forall x > 0$$

A média, variância e coeficiente de variação da distribuição Weibull são dados por, respectivamente:

$$\mu(X) = \beta\, \Gamma\Big(1 + \frac{1}{\alpha}\Big),$$
$$\sigma^2(X) = \beta^2 \Big\{ \Gamma\Big[1 + \frac{2}{\alpha}\Big] - \Big[\Gamma\Big(1 + \Big(\frac{1}{\alpha}\Big)\Big)^2\Big] \text{ e}$$
$$CV = \frac{\sqrt{\Gamma\Big(1+\frac{2}{\alpha}\Big) - \Big(\Gamma\Big(1+\frac{1}{\alpha}\Big)\Big)^2}}{\Gamma\Big(1+\frac{1}{\alpha}\Big)}$$

na qual $\Gamma(.)$ é a Função Gama, definida por $\Gamma(a) = \int_0^\infty u^{a-1}\, e^{-u}\, du$.

4.1.3.4 Log-normal

Assim como a distribuição de Weibull, a Log-Normal é muito usada para caracterizar tempo de vida de materiais e produtos.

A função de densidade de uma variável X que segue uma distribuição Log-normal com parâmetros μ e σ, $-\infty < \mu < \infty$ e $\sigma > 0$, ($X \sim Log - normal(\mu;\, \sigma^2)$) é dada por:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} exp\Big\{\frac{-(log(x)-\mu)^2}{2\sigma^2}\Big\},\ x > 0$$

A média, variância e coeficiente de variação da distribuição Log-Normal são dados por, respectivamente:

$$\mu(X) = exp\Big\{\mu + \frac{\sigma^2}{2}\Big\},$$
$$\sigma^2(X) = exp\{2\mu+\sigma^2\}(exp\{\sigma^2\}-1) \text{ e}$$
$$CV = \sqrt{exp\{\sigma^2\}-1}.$$

4.2 Teste de Kolmogorov-Smirnov

Em Estatística, grande parte dos problemas assumem que os dados são retirados de uma população com uma distribuição de probabilidade específica, sendo o formato da distribuição um possível objetivo do estudo. Os testes paramétricos para média, variância e outras medidas, muitas vezes baseiam-se na suposição de que os dados seguem uma distribuição normal, por exemplo, fazendo com que seja necessário testar se de fato isso ocorre. Para dar suporte a esta suposição e auxiliar nas análises, consideramos, dentre outros, o teste de Kolmogorov-Smirnov [4], definido a seguir:

Definição 11 *O teste de Kolmogorov-Smirnov é baseado na diferença entre a função de distribuição acumulada teórica ($F_0(x)$) e a função de distribuição acumulada empírica, $S_n(x)$, da amostra X, na qual $S_n(x)$ é definida como a proporção das observações da amostra que são menores ou iguais a x, ou seja:*

$$S_n(x) = \frac{número\ de\ elementos\ na\ amostra \leq x}{tamanho\ da\ amostra}$$

Nota: para que o teste possa ser usado, é necessário que os dados sejam pelo menos ordinais.

Se a hipótese nula é verdadeira, espera-se que as diferenças entre $S_n(x)$ e $F_0(x)$ sejam pequenas e estejam dentro dos limites dos erros aleatórios. O teste de Kolmogorov-Smirnov usa a maior dessa diferença como estatística do teste.

Assim, as hipóteses a serem testadas e as respectivas estatísticas do teste são:

$$\begin{cases} H_0 : X \text{ segue o modelo proposto} \\ H_a : X \text{ segue outro modelo} \end{cases}$$

As hipóteses acima podem ser reescritas como:

$$\begin{cases} H_0 : S_n(X) = F_0(X) \ \forall \ x \\ H_a : S_n(X) \neq F_0(X) \text{ para pelo menos um x} \end{cases}$$

Estatística do teste:

$$D = \max_x |S_n(x) - F_0(x)|, \forall x.$$

Rejeita-se a hipótese nula, sob um nível de significância α, se a estatística D for maior que o quantil α do teste de Kolmogorov-Smirnov, o qual pode ser obtido no anexo A

5 Inferência Bayesiana do Coeficiente de Variação: aplicação em dados simulados

Neste capítulo será apresentada a inferência Bayesiana para o Coeficiente de Variação segundo as distribuições propostas anteriormente.

5.1 Binomial

Seja $x_1, x_2, x_3, ..., x_n$ uma amostra que, dado θ, segue uma distribuição Binomial (m; θ). Assim, a função de verossimilhança é dada por:

$$L(\theta; X) = \prod_{i=1}^{n} \binom{m}{x_i} \theta^{x_i} (1-\theta)^{m-x_i}$$

Considerando a priori que $\theta \sim Beta(a;\, b)$, em que a e b são hiper-parâmetros positivos conhecidos, temos que a distribuição a posteriori é dada por [7]:

$$\theta|x \sim Beta(a + \sum_{i=1}^{n} x_i;\; b + \sum_{i=1}^{n} (m - x_i))$$

Como visto no Capítulo 4, o Coeficiente de Variação da Binomial é dado por:

$$\phi = \sqrt{\frac{1-\theta}{m\,\theta}} \tag{5.1}$$

Pela dificuldade de obter analiticamente a distribuição de ϕ, a mesma pode ser obtida numericamente por meio da geração de valores da posteriori de $\theta|x$. Abaixo segue o algoritmo para obtenção da distribuição à posteriori de $\phi|x$, bem como para geração dos dados necessários para as inferências:

1. Gerar M valores de $\theta|x \sim Beta(A;\, B)$, $A = a + \sum_{i=1}^{n} x_i$ e $B = b + \sum_{i=1}^{n} (m - x_i)$;
2. Obter o gráfico da densidade a posteriori de θ e de ϕ (fórmula 5.1);
3. Obter a estimativa intervalar e pontual de θ e ϕ;
4. Obter o valor-*e* resultante do teste a ser realizado;
5. Analisar os resultados gerados.

5.1.1 Exemplo 1

Um inspetor de qualidade extrai, durante 30 dias, uma amostra de tamanho 10 de uma máquina que produz energéticos com intuito de verificar a acurácia dessa máquina. Suponha que a probabilidade de uma lata ter algum defeito é de 30%. O número de latas defeituosas obtidas em cada uma das amostras são dados a seguir:

2	4	3	5	5	1	3	5	3	3	6	3	4	3	1
5	2	1	2	6	5	4	3	7	4	4	3	3	2	1

Assim, o número de latas defeituosas observadas em um dia segue uma distribuição $Binomial(10;\ \theta)$. Considerando a priori que $\theta \sim Beta(1;\ 1)$, tem-se que, a posteriori, $\theta|x \sim Beta(83;\ 219)$. O gestor da fábrica, após alguns estudos, determinou que a máquina está desbalanceada quando o coeficiente de variação é maior que 0,3.

As Figuras 2 e 3 apresentam a densidade a posteriori de θ e ϕ, respectivamente.

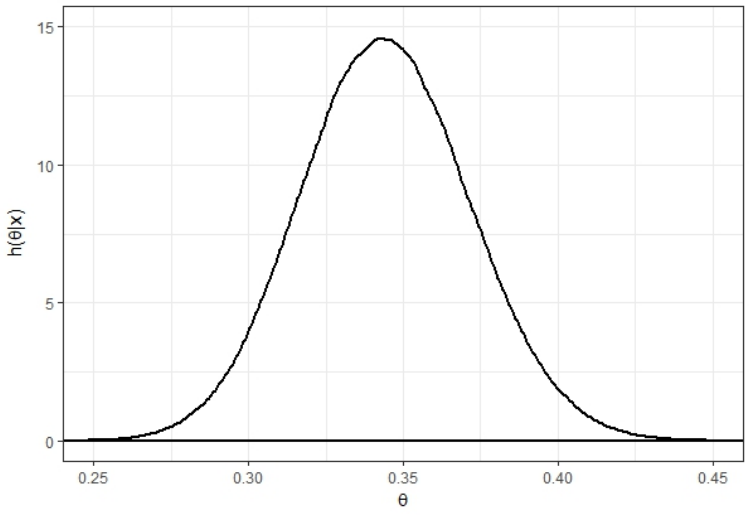

Figura 2 – Densidade a posteriori de θ - Binomial

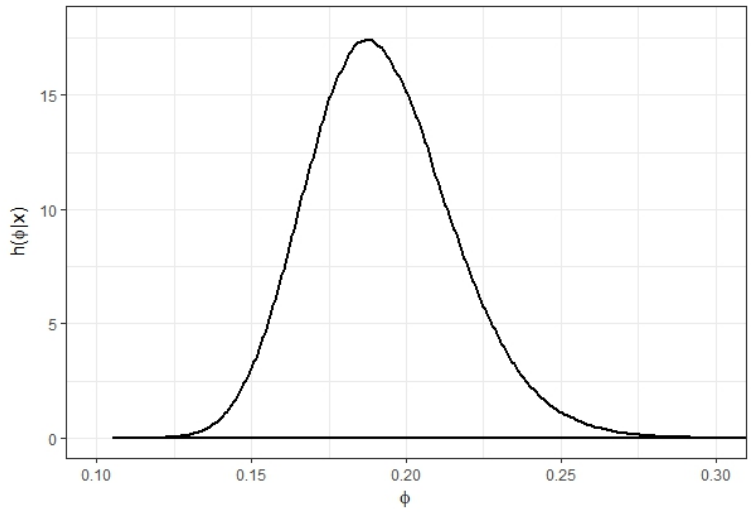

Figura 3 – Densidade a posteriori de ϕ - Binomial

A Tabela 2 apresenta as inferências dos parâmetros θ, probabilidade de uma lata ser defeituoso, e ϕ, o coeficiente de variação.

Tabela 2 – Inferência Bayesiana dos parâmetros θ e ϕ da distribuição Binomial

Parâmetro	Média a posteriori	IC *HPD* 95%
θ	0,3444	[0,2913; 0,3982]
ϕ	0,1922	[0,1483; 0,2393]

Analisando os resultados, nota-se que a estimativa de θ (a probabilidade de uma lata defeituosa) é de aproximadamente $0,35$ (*IC HPD 95%*: 0,29; 0,39). Além disso, pela distribuição a posteriori de ϕ tem-se que $P(\phi > 0,3|x) = 0,0002$, probabilidade a posteriori de se ter uma máquina desbalanceada. Desta forma, conclui-se que a máquina está balanceada, visto que a probabilidade de estar desbalanceada é muito baixa. Para este exemplo, tem-se a priori que $P(\phi > 0,3) = 0,2496$, resultando em um Fator de Bayes $B_0(X) = 0,0008$, o que implica em $B_0^{-1} = 1256,99$, isto é, uma evidência decisiva da máquina estar desregulada.

5.1.2 Exemplo 2: Comparação do CV de duas populações Binomiais independentes

Agora considere, por exemplo, que o dono da mesma fábrica deseja adquirir uma nova máquina. Contudo, ele não sabe se esta é melhor calibrada em relação a que ele já possui em sua indústria. Sendo assim, o inspetor de qualidade retira uma amostra de tamanho 10 da máquina antiga (MA) e uma de tamanho 15 da nova (MN). Sabe-se que, dado θ_1, $MA \sim Binomial(10;\ \theta_1)$ e, dado θ_2, $MN \sim Binomial(15;\ \theta_2)$. Abaixo, seguem as amostras obtidas:

Considerando a priori que θ_1 e $\theta_2 \sim Beta(1;\ 1)$, tem-se que, a posteriori, $\theta_1|x \sim Beta(35;\ 67)$ e $\theta_2|x \sim Beta(80;\ 72)$. As Figuras 4 e 5 apresentam as densidades a posteriori de

Amostra pop 1: 2 4 3 5 5 1 3 5 3 3
Amostra pop 2: 3 6 5 6 8 4 7 9 7 3 4 5 5 2 5

θ_1, θ_2, ϕ_1 e ϕ_2, respectivamente, para ambas as populações.

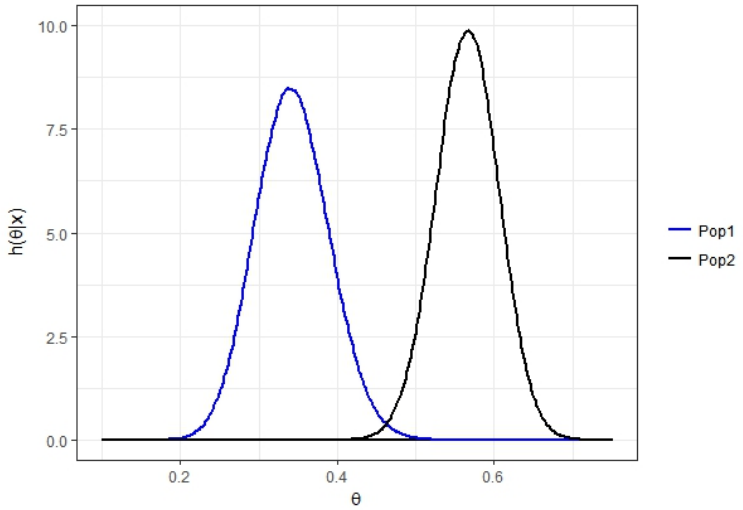

Figura 4 – Densidades marginais a posteriori de θ_1 e θ_2, populações independentes - Binomial

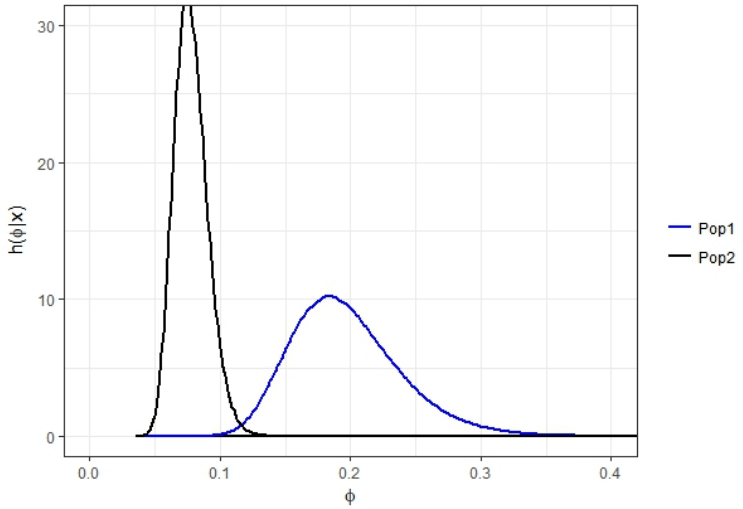

Figura 5 – Densidades marginais a posteriori de ϕ_1 e ϕ_2, populações independentes - Binomial

A Tabela 3 apresenta as inferências dos parâmetros $\theta_1, \theta_2, \phi_1$ e ϕ_2 para ambas as amostras, na qual ϕ representa o coeficiente de variação.

Tabela 3 – Inferência Bayesiana dos parâmetros $\theta_1, \theta_2, \phi_1$ e ϕ_2 das duas populações de distribuição Binomial

Parâmetros	Máquina antiga - População 1		Máquina nova - População 2	
	Média a posteriori	IC *HPD* 95%	Média a posteriori	IC *HPD* 95%
θ	0,3432	[0,2532; 0,4359]	0,5263	[0,4474; 0,6054]
ϕ	0,1971	[0,1213; 0,2811]	0,0912	[0,0634; 0,1211]

A densidade a posteriori de θ para a máquina nova e a máquina antiga se assemelham bastante quanto ao formato. No entanto, a máquina nova assume valores maiores de θ em relação à máquina antiga. A análise da densidade de ϕ permite verificar que a máquina antiga assume valores de ϕ maiores que a máquina nova.

Analisando os resultados dispostos na Tabela 3 observa-se que o coeficiente de variação obtido é menor para a máquina nova em relação à máquina antiga. Com intuito de comparar os parâmetros θ_1 e θ_2 e, consequentemente, o coeficiente de variação, aplica-se o Teste de Significância Genuinamente Bayesiano, ou *FBST*:

$$\begin{cases} H_0 : \phi_1 = \phi_2 & \Rightarrow \quad H_0 : \theta_1 = \theta_2 \\ H_a : \phi_1 \neq \phi_2 & \qquad\quad H_a : \theta_1 \neq \theta_2 \end{cases}$$

Pressupondo a priori que $\theta_1 \sim Beta(1;\ 1)$ e $\theta_2 \sim Beta(1;\ 1)$ (θ_1 e θ_2 independentes a priori), tem-se que, a posteriori, $\theta_1|x_1 \sim Beta(35;\ 67)$ e $\theta_2|x_2 \sim Beta(80;\ 72)$. Dessa maneira, a posteriori conjunta de $\theta = (\theta_1;\ \theta_2)$ é dada por:

$$h(\theta|x_1,x_2) = \frac{1}{B(35,67)}\ \frac{1}{B(80,72)}\ \theta_1^{34}\ (1-\theta_1)^{66}\ \theta_2^{79}(1-\theta_2)^{71}$$

A posteriori sob H_0 é dada por:

$$h_0(\theta_1|x_1,x_2) = \frac{1}{B(35;67)}\ \frac{1}{B(80;72)}\ \theta_1^{113}\ (1-\theta_1)^{137}$$

Nesse caso, $B(.)$ é a Função Beta, definida por $B(a,b) = \int_0^1 u^{a-1}\ (1-u)^{b-1}\ du$.

O valor de θ_1 que maximiza $h_0(\theta_1|x_1,x_2)$ é dado por:

$$\theta_1^* = \frac{A_1 + A_2 - 2}{A_1 + A_2 + B_1 + B_2 - 4} = \frac{113}{250}$$

Portanto, o valor máximo da posteriori sob H_0 é dado por:

$$h_0^*(\theta|x_1,x_2) = h_0(\theta_1^*|x_1,x_2) = 1,1652$$

Assim, a região tangente à hipótese H_0 é dada pelos valores de $\theta = (\theta_1,\theta_2)$ cuja posteriori é maior do que $h_0^*(\theta_1|x_1,x_2) = 1,1652$, isto é:

$$T^*(x) = \{\theta = (\theta_1,\theta_2) \in (0,1) : h(\theta|x_1,x_2) > 1,1652\}$$

Essa região é descrita pela Figura 6:

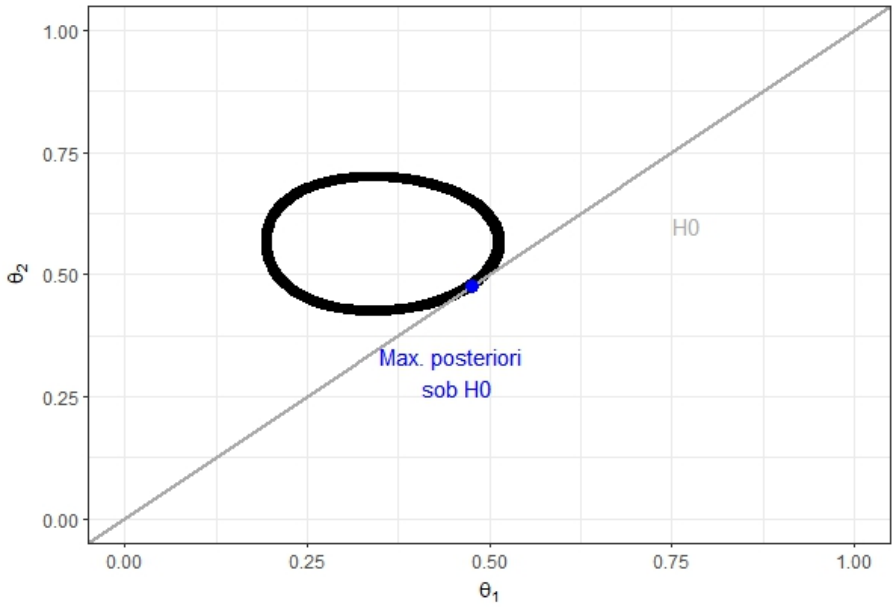

Figura 6 – Região tangente à hipótese $H_0: \theta_1 = \theta_2$ - Binomial

Assim, o valor-e do *FBST* é dado por:

$$\text{valor-}e = 1 - P(\theta \in T^*(x)|X) = 0,0133$$

Sendo assim, de acordo com os resultados obtidos, há evidência estatística suficiente para rejeitar a hipótese de que $\phi_1 = \phi_2$, ou seja, a máquina antiga e a nova são estatisticamente diferente. Pelos resultados obtidos na Tabela 3 é possível concluir que a máquina nova é mais balanceada do que a antiga (mesmo apresentando mais latas defeituosas).

5.2 Binomial Negativa

Seja $x_1, x_2, x_3, ..., x_n$ uma amostra que, dado θ, segue uma distribuição Binomial Negativa $(\theta;\ k)$. Assumindo k conhecido, a função de verossimilhança é dada por:

$$L(\theta;X) = \prod_{i=1}^{n} \binom{x_i + k - 1}{x_i} \theta^k (1-\theta)^{x_i}$$

Considerando a priori que $\theta \sim Beta(a;\ b)$, em que a e b são hiper-parâmetros positivos conhecidos, temos que a distribuição a posteriori é dada por:

$$\theta|x \sim Beta(a + nk;\ b + \sum_{i=1}^{n} x_i)$$

O Coeficiente de Variação da Binomial Negativa foi definido no Capítulo 4 como:

$$\phi = \frac{1}{\sqrt{k(1-\theta)}} \tag{5.2}$$

Assim como destacado anteriormente, devido à dificuldade de obtenção da distribuição de ϕ de maneira analítica, a mesma pode ser obtida numericamente por meio da geração de valores da posteriori de $\theta|x$. Abaixo segue o algoritmo para obtenção da distribuição à posteriori de $\phi|x$, bem como para geração dos dados necessários para as inferências:

1. Gerar M valores de $\theta|x \sim Beta(A;\ B)$, sendo $A = a + nk$ e $B = b + \sum_{i=1}^{n} x_i$;
2. Obter o gráfico da densidade a posteriori de θ e de $\phi = CV$ (fórmula 5.2);
3. Obter a estimativa intervalar e pontual de θ e ϕ;
4. Obter o valor-e resultante do teste a ser realizado;
5. Analisar os resultados gerados.

5.2.1 Exemplo 1

Considere uma amostra retirada de uma população que, dado θ, segue uma distribuição $BN(\theta;\ 10)$:

15	8	15	4	14	31	11	8	7	10
18	6	16	11	16	19	15	8	17	16
9	12	29	11	5	15	14	7	17	18

Se, a priori, $\theta \sim Beta(2;\ 1)$, tem-se que, a posteriori, $\theta|x \sim Beta(302;\ 403)$. As Figuras 7 e 8 apresentam a densidade a posteriori de θ e ϕ, respectivamente.

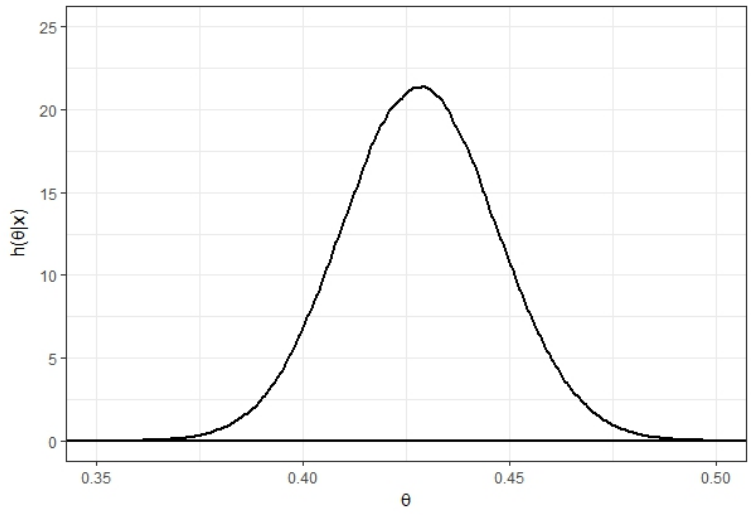

Figura 7 – Densidade a posteriori de θ - Binomial Negativa

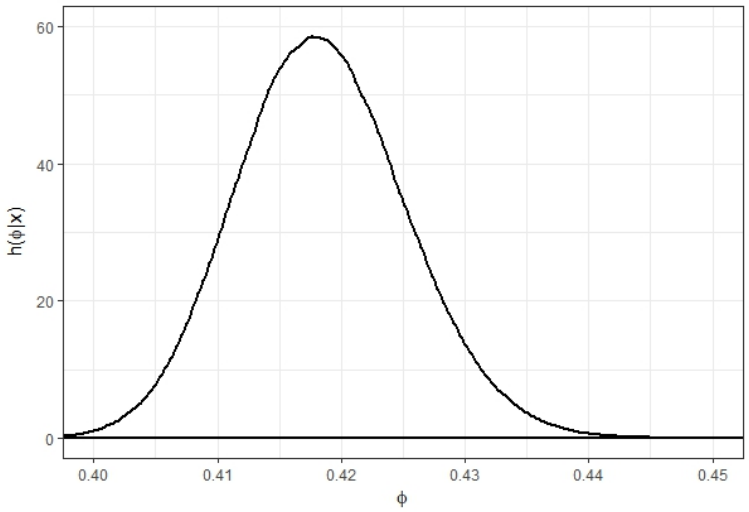

Figura 8 – Densidade a posteriori de ϕ - Binomial Negativa

A Tabela 4 apresenta as inferências dos parâmetros θ e ϕ, sendo ϕ o coeficiente de variação.

Tabela 4 – Inferência Bayesiana dos parâmetros θ e ϕ da distribuição Binomial Negativa

Parâmetro	Média a posteriori	IC *HPD* 95%
θ	0,4284	[0,3917; 0,4647]
ϕ	0,4184	[0,4052; 0,4319]

Analisando a Tabela 4 e as Figuras 7 e 8 é possível observar que a estimativa de θ é de aproximadamente 0,43. Além disso, a distribuição a posteriori de ϕ mostra que $P(\phi > 0,45|x) < 0,0001$. Logo, a probabilidade de obter um coeficiente de variação maior que 0,45 é muito baixa. Para esse exemplo, a priori $P(\phi > 0,45) = 0,0334$, resultando em um Fator de Bayes $B_0(X) = 0,0006$, o que implica em $B_0^{-1} = 1698,60$: uma evidência decisiva a favor de H_0.

5.2.2 Exemplo 2: Comparação do CV de duas populações Binomiais Negativas independentes

Considere uma amostra de tamanho 10 de uma população que, dado θ_1, segue uma distribuição *Binomial Negativa*$(\theta_1;\ 10)$; e uma amostra de tamanho 15 de uma outra população (independente da primeira, a priori) que, dado θ_2, segue uma distribuição *Binomial Negativa*$(\theta_2;\ 10)$. Abaixo, seguem as amostras obtidas:

Amostra pop 1:	2	2	1	2	3	3	2	3	0	1					
Amostra pop 2:	6	6	5	8	7	7	7	10	3	3	5	4	3	2	1

Considerando a priori que θ_1 e $\theta_2 \sim$ *Beta*$(3;\ 2)$, tem-se que, a posteriori, $\theta_1|x \sim$ *Beta*$(103;\ 21)$ e $\theta_2|x \sim$ *Beta*$(153;\ 79)$. As Figuras 9 e 10 apresentam as densidades a posteriori marginais de θ_1, θ_2, ϕ_1 e ϕ_2, respectivamente, para ambas as populações.

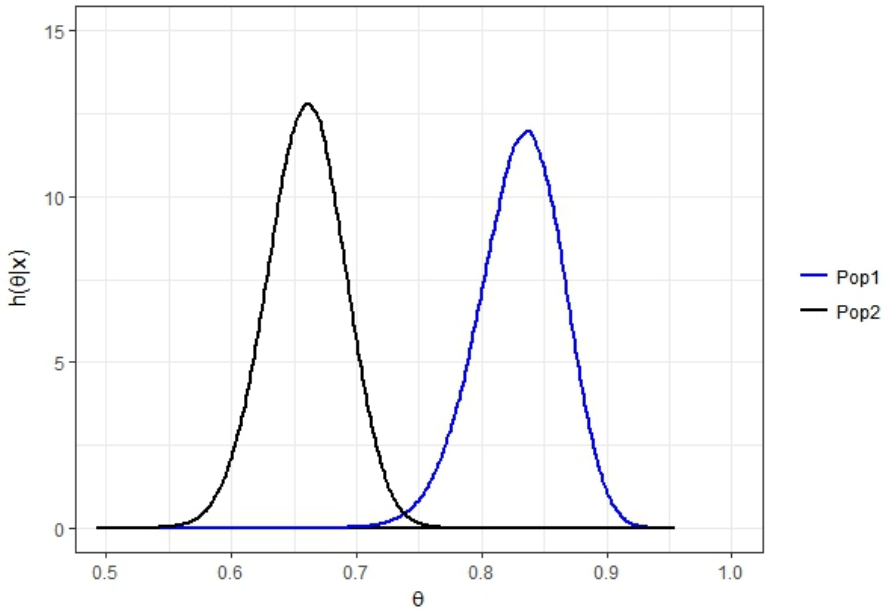

Figura 9 – Densidades marginais a posteriori de θ_1 e θ_2, populações independentes - Binomial Negativa

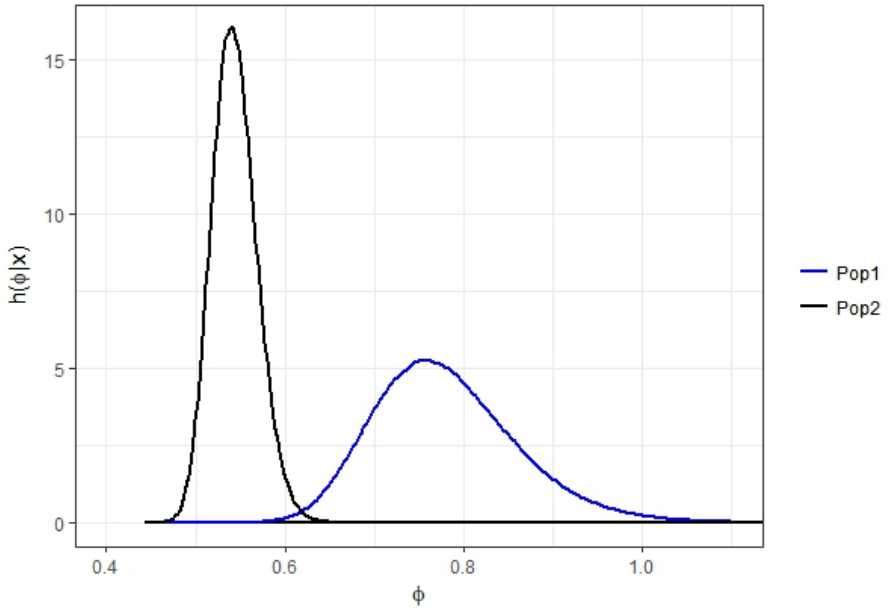

Figura 10 – Densidades marginais a posteriori de ϕ_1 e ϕ_2, populações independentes - Binomial Negativa

A Tabela 5 apresenta as inferências dos parâmetros θ e ϕ (coeficiente de variação) para ambas as amostras.

Tabela 5 – Inferência Bayesiana de θ e ϕ das duas populações de distribuição Binomial Negativa

Parâmetros	População 1		População 2	
	Média a posteriori	IC *HPD* 95%	Média a posteriori	IC *HPD* 95%
θ	0,8301	[0,7639; 0,8942]	0,6595	[0,5984; 0,7199]
ϕ	0,7801	[0,6338; 0,9407]	0,5436	[0,4959; 0,5935]

Analisando os resultados da Tabela 5 e os gráficos 9 e 10 nota-se que, aqui, tanto ϕ quanto θ assumem valores maiores para a amostra da população 1. Em média, ϕ assume valores maiores na população 1, visto que o intervalo HPD desta possui limite inferior maior que o limite superior da população 2 ($LI_{pop1} = 0,6338 > 0,5935 = LS_{pop_2}$).

Para comparar os parâmetros θ_1 e θ_2 e, consequentemente o coeficiente de variação, utilizou-se o Teste de Significância Genuinamente Bayesiano, ou *FBST*:

$$\begin{cases} H_0 : \phi_1 = \phi_2 & \Rightarrow \quad H_0 : \theta_1 = \theta_2 \\ H_a : \phi_1 \neq \phi_2 & \qquad\quad H_0 : \theta_1 \neq \theta_2 \end{cases}$$

Pressupondo a priori que $\theta_1 \sim Beta(3;\ 2)$ e $\theta_2 \sim Beta(3;\ 2)$, tem-se que, a posteriori, $\theta_1|x \sim Beta(103;\ 21)$ e $\theta_2|x \sim Beta(153;\ 79)$. Assim, a posteriori de $\theta = (\theta_1;\ \theta_2)$ é dada por:

$$h(\theta|x_1,x_2) = \frac{1}{B(103;21)}\ \frac{1}{B(153;79)}\ \theta_1^{102}\ (1-\theta_1)^{20}\ \theta_2^{152}\ (1-\theta_2)^{78}$$

E a posteriori sob H_0 é dada por:

$$h_0(\theta_1|x_1,x_2) = \frac{1}{B(103;21)}\ \frac{1}{B(153;79)}\ \theta_1^{254}\ (1-\theta_1)^{98}$$

O valor de θ_1 que maximiza $h_0(\theta_1|x_1,x_2)$ é dado por:

$$\theta_1^* = \frac{A_1 + A_2 - 2}{A_1 + A_2 + B_1 + B_2 - 4} = \frac{254}{352}$$

Portanto, o valor máximo da posteriori sob H_0 é dado por:

$$h_0^*(\theta_1|x_1,x_2) = h_0(\theta_1^*|x_1,x_2) = 0,2409$$

Assim, a região tangente à hipótese H_0 é dada pelos valores de $\theta = (\theta_1,\theta_2)$ cuja posteriori é maior do que $h_0^*(\theta_1|x_1,x_2) = 0,2409$, isto é:

$$T^*(x) = \{\theta = (\theta_1,\theta_2) \in (0,1) : h(\theta|x_1,x_2) > 0,2409\}$$

Essa região é descrita pela Figura 11:

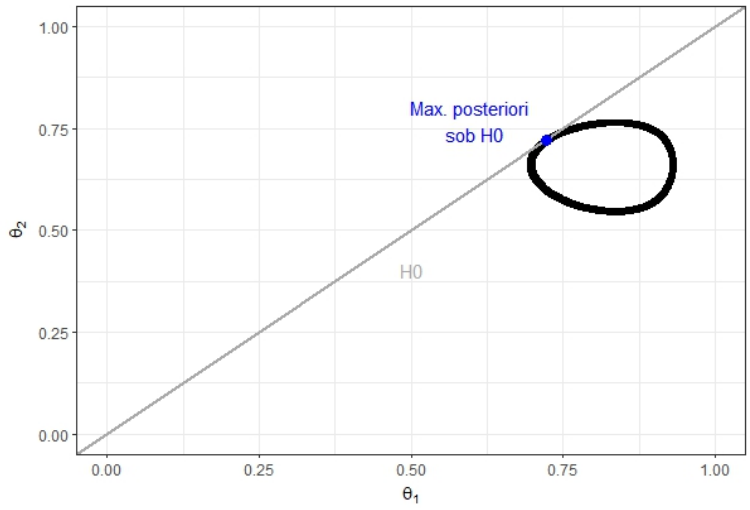

Figura 11 – Região tangente à hipótese $H_0: \theta_1 = \theta_2$ - Binomial Negativa

Assim, o valor-e do *FBST* é dado por:

$$\text{valor-}e = 1 - P(\theta \in T^*(x)|X) = 0,0154$$

Concluí-se, dessa maneira, que há evidências suficientes para rejeitar a hipótese de $\theta_1 = \theta_2$ e, consequentemente, $\phi_1 \neq \phi_2$. Analisando a Tabela 3 pode-se observar que $\phi_1 > \phi_2$ e, consequentemente, a amostra da população 2 é mais homogênea que a da população 1.

5.3 Poisson

Seja $x_1, x_2, x_3, ..., x_n$ uma amostra que, dado θ, segue uma distribuição Poisson (θ). Assim, a função de verossimilhança é dada por:

$$L(\theta; X) = \prod_{i=1}^{n} \frac{\theta^{x_i}\, e^{-\theta}}{x_i\,!}$$

Considerando a priori que $\theta \sim Gama(\alpha;\, \beta)$, em que α e β são hiper-parâmetros positivos conhecidos, temos que a distribuição a posteriori é dada por:

$$\theta|x \sim Gama(\alpha + \sum_{i=1}^{n} x_i;\, \beta + n)$$

O Coeficiente de Variação da Poisson foi definido no Capítulo 4 como:

$$\phi = \frac{1}{\sqrt{\theta}} \tag{5.3}$$

A obtenção da distribuição de $\phi|x$ de maneira analítica é muito complicada. Sendo assim, a mesma pode ser obtida numericamente por meio da geração de $\theta|x$. Abaixo segue o algoritmo para obtenção da distribuição à posteriori de $\phi|x$, bem como para geração dos dados necessários para as inferências:

1. Gerar M valores de $\theta|x \sim Gama(\alpha + \sum_{i=1}^{n} x_i;\ \beta + n)$
2. Obter o gráfico da densidade a posteriori de θ e de $\phi = CV$ (fórmula 5.3);
3. Obter a estimativa intervalar e pontual de θ e ϕ;
4. Obter o valor-*e* resultante do teste a ser realizado;
5. Analisar os resultados gerados.

5.3.1 Exemplo 1

Considere uma amostra de tamanho $n = 20$ de uma distribuição $Poisson(\theta)$:

0	1	0	2	2	0	1	2	1	0
2	0	1	1	0	2	0	0	0	2

Se, a priori, $\theta \sim Gama(4;\ 2)$, tem-se que, a posteriori, $\theta|x \sim Gama(21;\ 22)$. As Figuras 12 e 13 apresentam a densidade a posteriori de θ e ϕ, respectivamente. Além disso, considere também que, para ser considerado um baixo coeficiente de variação, nesse caso a estimativa do mesmo deve ser menor ou igual a 0,8.

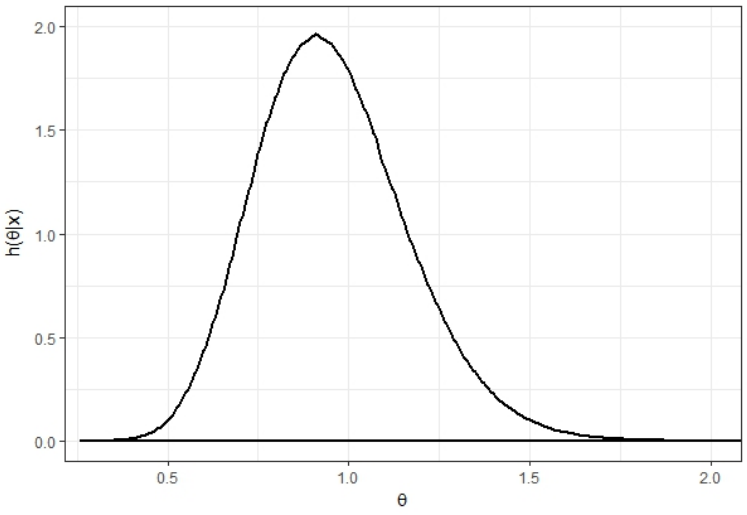

Figura 12 – Densidade a posteriori de θ - Poisson

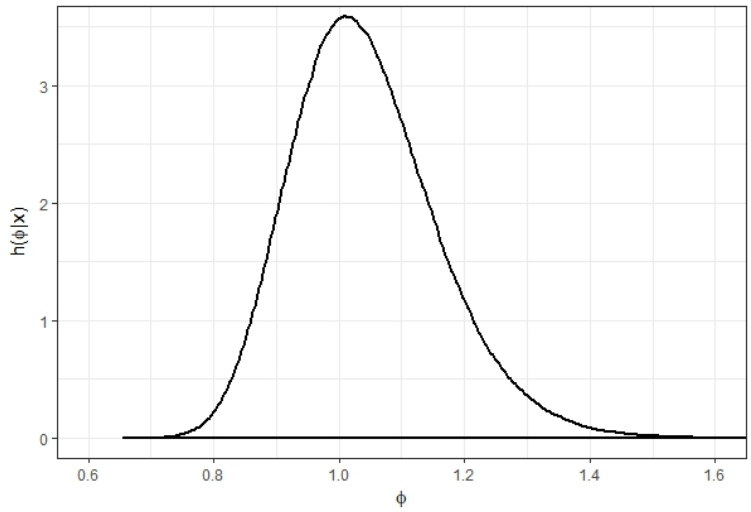

Figura 13 – Densidade a posteriori de ϕ - Poisson

A Tabela 6 apresenta as inferências dos parâmetros θ e ϕ, na qual ϕ representa o coeficiente de variação.

Tabela 6 – Inferência Bayesiana dos parâmetros θ e ϕ da distribuição Poisson

Parâmetro	Média a posteriori	IC *HPD* 95%
θ	0,9544	[0,5649; 1,3692]
ϕ	1,0424	[0,8278; 1,2773]

Analisando as Figuras 12 e 13 e a Tabela 6 nota-se que a estimativa de θ é de aproximadamente 0,95 (*IC HPD 95%*: 0,57; 1,37). Pela distribuição a posteriori de ϕ, a estimativa do mesmo é de 1,04 aproximadamente. Por fim, a probabilidade de obtenção de um baixo coeficiente de variação é dada por $P(\phi \leq 0,8|x) =$ 0056. Aqui, tem-se a priori que $P(\phi < 0,8) = 0,6196$. Assim, o Fator de Bayes obtido foi $B_0(X) = 0,0034$, resultando em $B_0^{-1} = 292,1204$, o que representa uma evidência decisiva a favor de H_0.

5.3.2 Exemplo 2: Comparação do CV de duas populações Poisson independentes

Considere uma amostra de tamanho 13 de uma população que, dado θ_1, segue uma distribuição $Poisson(\theta_1)$; e uma amostra de tamanho 22 de uma outra população (independente da primeira, a priori) que, dado θ_2, segue uma distribuição $Poisson(\theta_2)$. Abaixo, seguem as amostras obtidas:

Amostra pop 1: 0 1 0 2 2 0
0 2 1 0 2 0 1

Amostra pop 2: 0 1 0 2 2 0 1 2 1 1 3
1 1 1 0 2 0 0 0 3 2 1

Considerando a priori que θ_1 e $\theta_2 \sim Gama(4;\ 2)$, tem-se que, a posteriori, $\theta_1|x \sim Gama(14;\ 15)$ e $\theta_2|x \sim Gama(27;\ 24)$. As Figuras 14 e 15 apresentam as densidades marginais a posteriori de θ_1, θ_2, ϕ_1 e ϕ_2, respectivamente, para ambas as populações e a Tabela 7 apresenta as inferências dos parâmetros θ e ϕ (coeficiente de variação) para ambas as amostras.

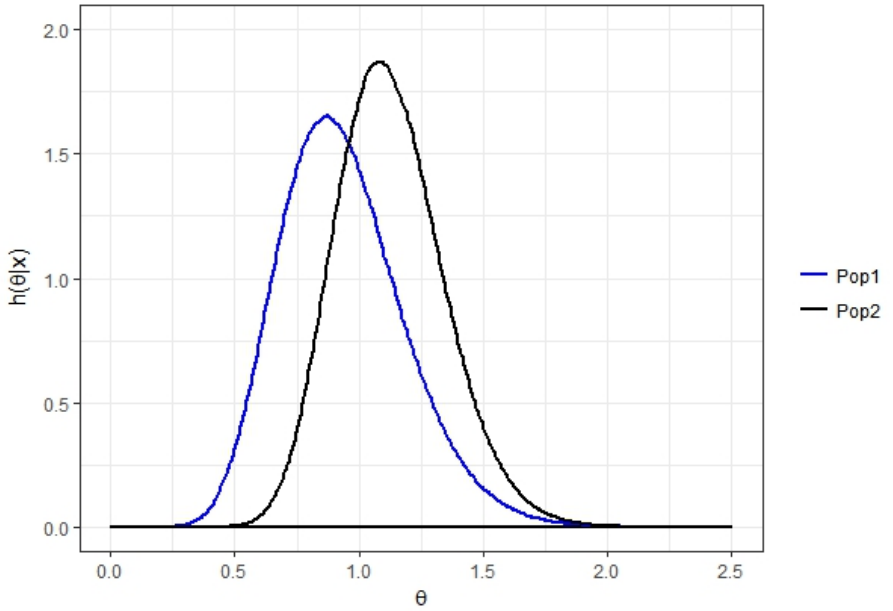

Figura 14 – Densidades marginais a posteriori de θ_1 e θ_2, populações independentes - Poisson

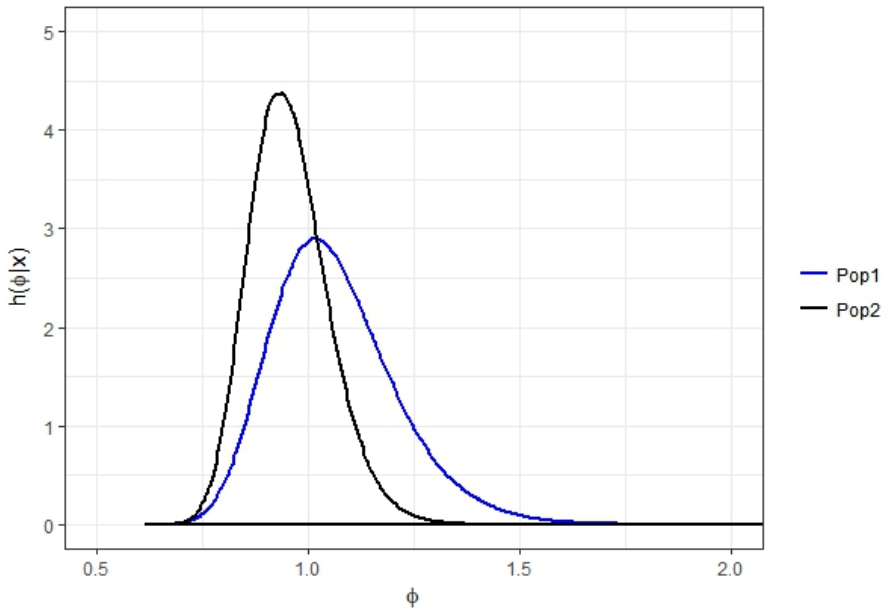

Figura 15 – Densidades marginais a posteriori de ϕ_1 e ϕ_2, populações independentes - Poisson

Tabela 7 – Inferência Bayesiana de θ_1, θ_2, ϕ_1 e ϕ_2 das duas populações de distribuição Poisson

Parâmetros	População 1		População 2	
	Média a posteriori	IC *HPD* 95%	Média a posteriori	IC *HPD* 95%
θ	0,9334	[0,4764 ; 1,4309]	1,1251	[0,7164 ; 1,5548]
ϕ	1,0638	[0.7977 ; 1,3615]	0,9561	[0,7413 ; 1,1435]

De acordo com os resultados apresentados na Tabela 7 e nas Figuras 14 e 15 é possível observar que, em média, $\theta_1 < \theta_2$ e, com confiabilidade de 95%, θ_1 assume sempre valores menores que θ_2, visto que o limite superior do intervalo HPD do primeiro é menor que do segundo. Além disso, em média, $\phi_1 > \phi_2$, sugerindo que a amostra da população 1 seja mais heterogênea que da população 2.

Para comparar os parâmetros θ_1 e θ_2 e, consequentemente o coeficiente de variação, utilizou-se o Teste de Significância Genuinamente Bayesiano, ou *FBST*:

$$\begin{cases} H_0 : \phi_1 = \phi_2 & \Rightarrow \quad H_0 : \theta_1 = \theta_2 \\ H_a : \phi_1 \neq \phi_2 & \qquad\quad H_a : \theta_1 \neq \theta_2 \end{cases}$$

Pressupondo a priori que $\theta_1 \, e \theta_2 \sim Gama(4;\, 2)$, tem-se que, a posteriori, $\theta_1|x \sim Gama(14;\, 15)$ e $\theta_2|x \sim Gama(27;\, 24)$. A posteriori conjunta de $\theta = (\theta_1;\theta_2)$ é dada por:

$$h(\theta|x_1,x_2) = \frac{15^{14}}{\Gamma(14)} \, \frac{24^{27}}{\Gamma(27)} \, \theta_1^{13} \, e^{-15(\theta_1)} \, \theta_2^{26} \, e^{-24(\theta_2)}$$

na qual $\Gamma(.)$ é a Função Gama, definida como $\Gamma(a) = \int_0^\infty u^{a-1} \, e^{-u} \, du$.

Assim, a distribuição a posteriori sob H_0 é dada por:

$$h_0(\theta_1|x_1,x_2) = \frac{15^{14}}{\Gamma(14)} \, \frac{24^{27}}{\Gamma(27)} \, \theta_1^{39} \, e^{-39(\theta_1)}$$

O valor de θ_1 que maximiza $h_0(\theta|x_1,x_2)$ é dado por:

$$\theta_1^* = \frac{\alpha_1 + \alpha_2 - 2}{\beta_1 \, \beta_2} = \frac{39}{39} = 1$$

Portanto, o valor máximo da posteriori sob H_0 é dado por:

$$h_0^*(\theta_1|x_1,x_2) = h_0(\theta_1^*|x_1,x_2) = 2,4751$$

Dessa maneira, a região tangente à hipótese H_0 é dada pelos valores de $\theta = (\theta_1, \theta_2)$ cuja posteriori é maior do que $h_0^*(\theta|x_1,x_2) = 2,4751$, isto é:

$$T^*(x) = \{\theta = (\theta_1, \theta_2) \in (0,1) : h(\theta|x_1,x_2) > 2,4751\}$$

Essa região é descrita pela Figura 16:

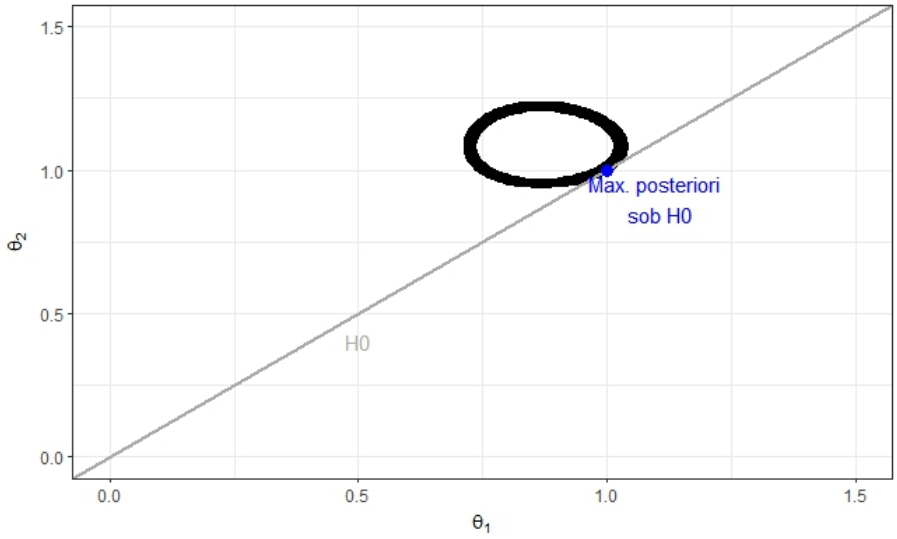

Figura 16 – Região tangente à hipótese H_0 - Poisson

Assim, o valor-*e* do *FBST* é dado por:

$$\text{valor-}e = 1 - P(\theta \in T(x)|x) = 0,8040$$

Pode-se concluir que não há evidências suficientes para rejeitar a hipótese de $\theta_1 = \theta_2$ e, consequentemente, não há evidências suficientes para afirmar que $\phi_1 \neq \phi_2$, visto que *valor-E*$> 0,05$.

5.4 Gama

Seja $x_1, x_2, x_3, ..., x_n$ uma amostra que, dado α e β, segue uma distribuição Gama (α; β). Assim, a função de verossimilhança é dada por:

$$L(\theta; X) = \prod_{i=1}^{n} \frac{\beta^{\alpha}\ x_i^{\alpha-1}\ e^{-\beta x_i}}{\Gamma(\alpha)}$$

Considerando a priori que $\alpha \sim Gama(a;\ b)$ e $\beta \sim Gama(c;\ d)$, em que a, b, c, d são hiper-parâmetros positivos conhecidos e, além disso, α e β são independentes, a priori, temos que o *log* da distribuição a posteriori é dado por:

$$\begin{aligned} log(\pi(\alpha,\beta|x)) = &\left[n\alpha\log(\beta)\ -\ n\log(\Gamma(\alpha))\ +\ (\alpha-1)\sum_{i=1}^{n}\log(x_i)\ -\ \beta\sum_{i=1}^{n}x_i\right] + \\ &log\left(\frac{a^b\ \alpha^{a-1}\ e^{-b\alpha}}{\Gamma(a)}\right) + log\left(\frac{c^d\ \beta^{c-1}\ e^{-d\beta}}{\Gamma(c)}\right) \end{aligned} \tag{5.4}$$

Seja, agora, $x_{11}, x_{12}, x_{13}, ..., x_{1n_1}$ uma amostra que, dado α_1 e β_1, segue uma distribuição $Gama(\alpha_1;\ \beta_1)$ e $x_{21}, x_{22}, x_{23}, ..., x_{2n_2}$ uma segunda amostra que, dado α_2 e β_2, segue uma distribuição $Gama(\alpha_2;\ \beta_2)$. Considere a priori que $\alpha_1 \sim Gama(a_1, b_1)$, $\beta_1 \sim Gama(c_1, d_1)$, $\alpha_2 \sim Gama(a_2, b_2)$ e $\beta_2 \sim Gama(c_2, d_2)$, em que $a_1, b_1, c_1, d_1, a_2, b_2, c_2, d_2$ são hiper-parâmetros positivos conhecidos e, além disso, α_1, β_1, α_2, β_2 são independentes, a priori. Assim, temos que o *log* da distribuição a posteriori é dado por:

$$
\begin{aligned}
log(\pi(\alpha_1, \beta_1, \alpha_2, \beta_2 | x_1, x_2)) = & \left\{ \left[n_1 \alpha_1 \log(\beta_1) \ - \ n_1 \log(\Gamma(\alpha_1)) \ + \ (\alpha_1 - 1) \sum_{i=1}^{n_1} \log(x_{1i}) \ - \ \beta_1 \sum_{i=1}^{n_1} x_{1i} \right] + \right. \\
& \left. log\left(\frac{a_1^{b_1} \ \alpha_1^{a_1 - 1} \ e^{-b_1 \alpha_1}}{\Gamma(a_1)} \right) + log\left(\frac{c_1^{d_1} \ \beta_1^{c_1 - 1} \ e^{-d_1 \beta_1}}{\Gamma(c_1)} \right) \right\} + \\
& \left\{ \left[n_2 \alpha_2 \log(\beta_2) \ - \ n_2 \log(\Gamma(\alpha_2)) \ + \ (\alpha_2 - 1) \sum_{i=1}^{n_2} \log(x_{2i}) \ - \ \beta_2 \sum_{i=1}^{n_2} x_{2i} \right] + \right. \\
& \left. log\left(\frac{a_2^{b_2} \ \alpha_2^{a_2 - 1} \ e^{-b_2 \alpha_2}}{\Gamma(a_2)} \right) + log\left(\frac{c_2^{d_2} \ \beta_2^{c_2 - 1} \ e^{-d_2 \beta_2}}{\Gamma(c_2)} \right) \right\}
\end{aligned}
\tag{5.5}
$$

Nota-se que, em ambos os casos, a posteriori não assemelha-se a nenhuma distribuição conhecida, sendo necessário, nesse caso, a utilização do método MCMC (Markov Chain Monte Carlo) [8].

O Coeficiente de Variação da distribuição Gama foi definido no Capítulo 4 como:

$$
\phi = \frac{1}{\sqrt{\alpha}} \tag{5.6}
$$

A obtenção da distribuição de ϕ de maneira analítica é muito complicada. Sendo assim, a mesma pode ser obtida numericamente por meio da geração de valores da posteriori conjunta de $(\alpha,\ \beta|x)$. Abaixo segue o algoritmo para obtenção da distribuição à posteriori conjunta de $\phi|x$, bem como para geração dos dados necessários para as inferências:

1. Gerar M valores de $(\alpha, \beta|x)$ segundo a distribuição a posteriori obtida;
2. Obter o gráfico da densidade a posteriori de $\phi = CV$ (fórmula 5.6);
3. Obter a estimativa intervalar e pontual de $\alpha,\ \beta$ e ϕ;
4. Obter o valor-*e* resultante do teste a ser realizado;
5. Analisar os resultados gerados.

5.4.1 Exemplo 1

Considere uma amostra de tamanho $n = 15$ de uma variável aleatória X que, dado α e β, segue uma $Gama(\alpha;\ \beta)$. Além disso, considere que, hipoteticamente, $\phi \geq 0,71$ é considerado um coeficiente de variação alto.

0,8920936	3,3118474	0,1443750	1,6625233	4,3358790
2,1176157	0,3507177	0,1304001	3,3737820	1,9730466
2,0309972	1,6386208	0,8964763	3,5048576	2,6814547

Se, a priori, $\alpha \sim Gama(\frac{1}{1000};\ \frac{1}{1000})$, uma priori difusa, e $\beta \sim Gama(\frac{1}{1000};\ \frac{1}{1000})$, tem-se que, a posteriori, $(\alpha,\beta)|x$ segue a distribuição definida em (5.4) com $a = b = c = d = \frac{1}{1000}$, $n = 15$ e:

$$\sum_{i=1}^{n} x_i = 29,0447 \text{ e}$$
$$\sum_{i=1}^{n} log(x_i) = 4,0178$$

A Figura 17 apresenta a densidade a posteriori de ϕ.

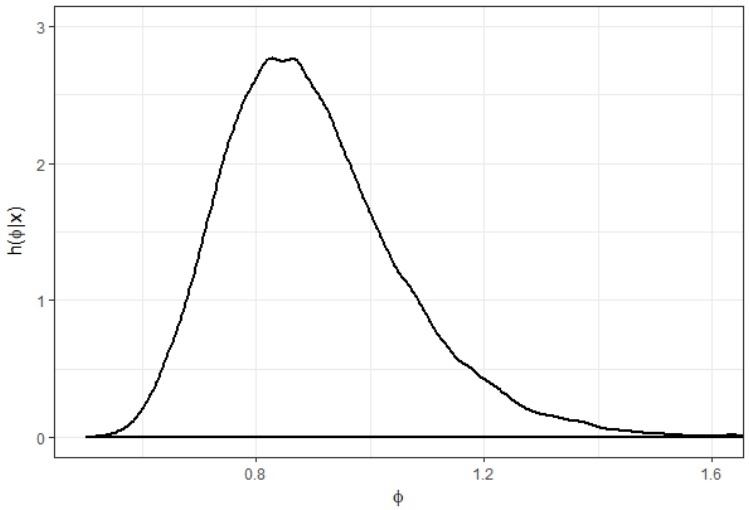

Figura 17 – Densidade a posteriori de ϕ - Gama

A Tabela 8 apresenta as inferências dos parâmetros α, β e ϕ, sendo ϕ o coeficiente de variação

Tabela 8 – Inferência Bayesiana dos parâmetros α, β e ϕ da distribuição Gama

Parâmetro	Média a posteriori	IC *HPD* 95%
α	2,0809	[1,9159; 2,2511]
β	1,1072	[1,0067; 1,2093]
ϕ	0,6937	[0,6663; 0,7222]

Ao analisar a Tabela 8 nota-se que, em média, $\phi = 0,6937$, variando entre $0,6663$ e $0,7222$ com 95% de credibilidade. A probabilidade de obter um alto coeficiente, nesse caso, é de $P(\phi \geq 0,71|x) = 0,3240$.

5.4.2 Exemplo 2: Comparação do CV de duas populações Gama independentes

Considere uma amostra de uma população que, dado α_1 e β_1, segue uma distribuição $Gamma(\alpha_1;\ \beta_1)$; e uma amostra de uma outra população (independente da primeira, a priori) que, dado α_2, β_2, segue uma distribuição $Gama(\alpha_2;\ \beta_2)$.

População 1	0,8921	3,3118	0,1444	1,6625	4,3359	2,1176
	0,3507	0,1304	3,3738	1,9731	2,0309	1,6386
	5,2532	1,9915	2,3552	0,7157	0,9608	2,3712
População 2	0,8461	2,3680	0,2711	1,3543	2,9736	1,6409
	0,4499	0,2574	2,4050	1,5506	1,5869	1,3390
	0,8492	2,4832	1,9878			

Considerando a priori que α_1 e $\alpha_2 \sim Gama(\frac{1}{1000};\ \frac{1}{1000})$, β_1 e $\beta_2 \sim Gama(\frac{1}{1000};\ \frac{1}{1000})$ tem-se que, a posteriori, $(\alpha_1,\ \alpha_2,\ \beta_1,\ \beta_2)|x$ segue a distribuição definida em (5.5), com $a_1 = a_2 = b_1 = b_2 = c_1 = c_2 = d_1 = d_2 = \frac{1}{1000}$, $n_1 = 18$, $n_2 = 15$ e:

$$\sum_{i=1}^{n_1} x_{1i} = 21,9619,$$
$$\sum_{i=1}^{n_2} x_{2i} = 22,36322,$$
$$\sum_{i=1}^{n_1} log(x_{1i}) = 1,8866 \text{ e}$$
$$\sum_{i=1}^{n_2} log(x_{2i}) = 2,6254.$$

A Figura 18 apresenta as densidades marginais a posteriori de ϕ para ambas as populações e a Tabela 9 apresenta as inferências dos parâmetros $\alpha_1,\ \alpha_2,\ \beta_1,\ \beta_2,\ \phi_1$ e ϕ_2 (coeficiente de variação).

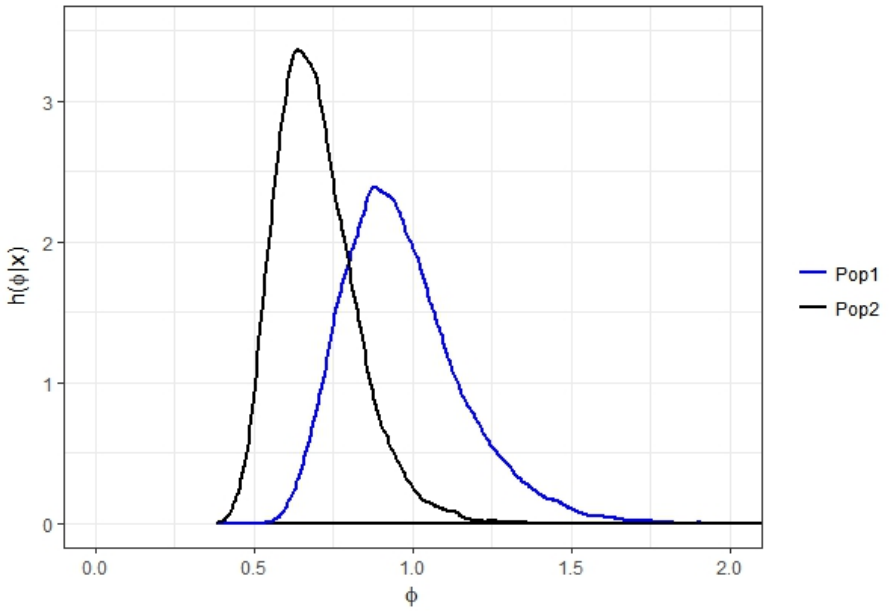

Figura 18 – Densidades marginais a posteriori de ϕ_1 e ϕ_2, populações independentes - Gama

Tabela 9 – Inferência Bayesiana de α, β e ϕ das duas populações de distribuição Gama

Parâmetros	População 1		População 2	
	Média a posteriori	IC *HPD* 95%	Média a posteriori	IC *HPD* 95%
α	1,1876	[0,4399; 2,0837]	2,2643	[0,8738; 3,8541]
β	0,6499	[0,1366; 1,2275]	1,5203	[0,4546; 2,6841]
ϕ	0,9674	[0,6423; 1,3559]	0,6969	[0,4747; 0,9588]

*Inferências geradas com base em uma cadeia de 100000 e Burn-in=10000

De acordo com os resultados apresentados na Figura 18, nota-se que, em geral, o coeficiente de variação da População 1 assume valores maiores que o da População 2, visto que a curva de densidade da População 2 está deslocada à direita em relação a da População 1. Esse fato pode ser verificado pelos dados da Tabela 9, visto que, em média, $\phi_1 = 0,9674$ e $\phi_2 = 0,6969$.

Para comparar os parâmetros α_1 e α_2 e, consequentemente o coeficiente de variação, utilizou-se o Teste de Significância Genuinamente Bayesiano, ou *FBST*:

$$\begin{cases} H_0 : \phi_1 = \phi_2 \\ H_a : \phi_1 \neq \phi_2 \end{cases} \Rightarrow \begin{array}{l} H_0 : \alpha_1 = \alpha_2 \\ H_a : \alpha_1 \neq \alpha_2 \end{array}$$

O logaritmo da posteriori conjunta de $\alpha_1, \beta_1, \alpha_2, \beta_2 | x_1, x_2$, dado $H_0 : \alpha_1 = \alpha_2$ é dada por:

$$
\begin{aligned}
log(h_0(\alpha_1, \beta_1, \beta_2 | x_1, x_2)) = & \left\{ \left[n_1 \alpha_1 \log(\beta_1) - n_1 \log(\Gamma(\alpha_1)) + (\alpha_1 - 1) \sum_{i=1}^{n_1} \log(x_{1i}) - \beta_1 \sum_{i=1}^{n_1} x_{1i} \right] + \right. \\
& \left. log\left(\frac{a_1^{b_1} \, \alpha_1^{a_1 - 1} \, e^{-b_1 \alpha_1}}{\Gamma(a_1)} \right) + log\left(\frac{c_1^{d_1} \, \beta_1^{c_1 - 1} \, e^{-d_1 \beta_1}}{\Gamma(c_1)} \right) \right\} + \\
& \left\{ \left[n_2 \alpha_1 \log(\beta_2) - n_2 \log(\Gamma(\alpha_1)) + (\alpha_1 - 1) \sum_{i=1}^{n_2} \log(x_{2i}) - \beta_2 \sum_{i=1}^{n_2} x_{2i} \right] + \right. \\
& \left. log\left(\frac{a_2^{b_2} \, \alpha_1^{a_2 - 1} \, e^{-b_2 \alpha_1}}{\Gamma(a_2)} \right) + log\left(\frac{c_2^{d_2} \, \beta_2^{c_2 - 1} \, e^{-d_2 \beta_2}}{\Gamma(c_2)} \right) \right\}
\end{aligned}
\tag{5.7}
$$

O valor $(\alpha_1^*, \beta_1^*, \beta_2^*)$ que maximiza (5.7) não pode ser obtido analiticamente, mas o mesmo pode ser encontrado via métodos numéricos, em particular o método Newton-Raphson. Para tal amostra, tem-se que $(\alpha_1^*, \beta_1^*, \beta_2^*) = (1,2733;\ 0,6501;\ 0,8094)$. Assim, o máximo do logaritmo da posteriori sob H_0 é dado por:

$$
log(h_0^*(\alpha_1, \beta_1, \beta_2 | x_1, x_2)) = log(h_0(1,2733;\ 0,6501;\ 0,8094 | x_1, x_2)) = -66.4777
$$

Dessa maneira, a região tangente à hipótese H_0 é dada pelos valores de $\theta = (\alpha_1, \beta_1, \beta_2)$ cujo logaritmo da posteriori é maior do que $log(h_0^*(\alpha_1, \beta_1, \beta_2 | x_1, x_2)) = -66.4777$, isto é:

$$
T^*(x_1, x_2) = \{ \theta = (\alpha_1, \alpha_2, \beta_1, \beta_2) \in \mathbb{R}_+^4 : log(\pi(\alpha_1, \alpha_2, \beta_1, \beta_2 | x_1, x_2)) > -66.4777 \}
$$

Assim, o valor-e do *FBST* é dado por:

$$
\text{valor-}e = 1 - P((\alpha_1, \alpha_2, \beta_1, \beta_2) \in T(x_1, x_2) X |) = 0,8513
$$

Pode-se concluir que não há evidências suficientes para rejeitar a hipótese de $\theta_1 = \theta_2$ e, consequentemente, não há evidências suficientes para afirmar que $\phi_1 \neq \phi_2$, visto que valor-$e > 0,05$.

5.5 Log-Normal

Seja $x_1, x_2, x_3, ..., x_n$ uma amostra que, dado σ e μ, segue uma distribuição Log-Normal(μ; σ). Assim, a função de verossimilhança é dada por:

$$L(\theta;X) = \prod_{i=1}^{n} \frac{1}{x_i \sigma\sqrt{2\pi}} exp\Big\{ \frac{-(log(x_i)-\mu)^2}{2\sigma^2} \Big\},\ x > 0$$

Considerando a priori que $\mu \sim Normal(a;\ b)$ e $\sigma \sim Gama(\alpha;\ \beta)$, e que a, b, α e β são hiper-parâmetros positivos conhecidos e, além disso, μ e σ são independentes, temos que o *log* da distribuição a posteriori é dado por:

$$\begin{gathered} log(\pi(\mu,\sigma|x)) = \Bigg[\alpha log(\beta)\ + (\alpha-1) log(\sigma)\ - \beta\sigma\ - log(\Gamma(\alpha)) \Bigg] + \\ \Bigg[-\sum_{i=1}^{n} log(x_i)\ - n log(\sigma)\ - \frac{n}{2} log(2\pi)\ - \frac{1}{2\sigma^2} \Bigg[\sum_{i=1}^{n} [log(x_i)]^2\ - 2\mu \sum_{i=1}^{n} log(x_i)\ + n\mu^2 \Bigg] \Bigg] + \\ \Bigg[-\frac{1}{2} log(2\pi)\ - log(b)\ + \frac{1}{2b^2} \Big[\mu^2\ - 2\mu a\ + a^2\ \Big] \Bigg] \end{gathered} \tag{5.8}$$

Seja, agora, $x_{11}, x_{12}, x_{13}, ..., x_{1n_1}$ uma amostra que, dado μ_1 e σ_1, segue uma distribuição *Log- Normal*$(\mu_1;\ \sigma_1)$ e $x_{21}, x_{22}, x_{23}, ..., x_{2n_2}$ uma segunda amostra que, dado μ_2 e σ_2, segue uma distribuição *Log*(-)*Normal*$(\mu_2;\ \sigma_2)$. Considere a priori que $\mu_1 \sim Normal(a_1, b_1)$, $\sigma_1 \sim Gama(\alpha_1, \beta_1)$, $\mu_2 \sim Normal(a_2, b_2)$ e $\sigma_2 \sim Gama(\alpha_2, \beta_2)$, em que a_1, b_1, α_1, β_1, a_2, b_2, α_2, β_2 são hiper-parâmetros positivos conhecidos e, além disso, μ_1, σ_1, μ_2, σ_2 são independentes, a priori. Assim, temos que o *log* da distribuição a posteriori é dado por:

$$\begin{gathered} log(\pi(\mu_1,\sigma_1,\mu_2,\sigma_2|x)) = \Bigg\{ \Bigg[\alpha_1 log(\beta_1)\ + (\alpha_1-1) log(\sigma_1)\ - \beta_1\sigma_1\ - log(\Gamma(\alpha_1)) \Bigg] + \\ \Bigg[-\sum_{i=1}^{n_1} log(x_{1i})\ - n_1 log(\sigma_1)\ - \frac{n_1}{2} log(2\pi)\ - \frac{1}{2\sigma_1^2} \Bigg[\sum_{i=1}^{n_1} [log(x_{1i})]^2\ - 2\mu_1 \sum_{i=1}^{n_1} log(x_{1i})\ + n_1\mu_1^2 \Bigg] \Bigg] + \\ \Bigg[-\frac{1}{2} log(2\pi)\ - log(b_1)\ + \frac{1}{2b_1^2} \Big[\mu_1^2\ - 2\mu_1 a_1\ + a_1^2\ \Big] \Bigg] \Bigg\} + \\ \Bigg\{ \Bigg[\alpha_2 log(\beta_2)\ + (\alpha_2-1) log(\sigma_2)\ - \beta_2\sigma_2\ - log(\Gamma(\alpha_2)) \Bigg] + \\ \Bigg[-\sum_{i=1}^{n_2} log(x_{2i})\ - n_2 log(\sigma_2)\ - \frac{n_2}{2} log(2\pi)\ - \frac{1}{2\sigma_2^2} \Bigg[\sum_{i=1}^{n_2} [log(x_{2i})]^2\ - 2\mu_2 \sum_{i=1}^{n_2} log(x_{2i})\ + n_2\mu_2^2 \Bigg] \Bigg] + \\ \Bigg[-\frac{1}{2} log(2\pi)\ - log(b_2)\ + \frac{1}{2b_2^2} \Big[\mu_2^2\ - 2\mu_2 a_2\ + a_2^2\ \Big] \Bigg] \Bigg\} \end{gathered} \tag{5.9}$$

Nota-se que, em ambos os casos, a posteriori não assemelha-se a nenhuma distribuição conhecida. Nesse caso, faz-se necessária a utilização do método MCMC (Markov Chain Monte Carlo) [8].

O Coeficiente de Variação da distribuição Log-Normal foi definido no Capítulo 4 como:

$$\phi = \sqrt{exp(\sigma^2) - 1} \tag{5.10}$$

A obtenção da distribuição de ϕ de maneira analítica é muito complicada. Sendo assim, a mesma pode ser obtida numericamente por meio da geração de μ; σ da posteriori conjunta de x. Abaixo segue o algoritmo para obtenção da distribuição à posteriori de $\phi|x$, bem como para geração dos dados necessários para as inferências:

1. Gerar M valores de $(\mu, \sigma|x)$ segundo a distribuição a posteriori obtida;
2. Obter o gráfico da densidade a posteriori de $\phi = CV$ (fórmula 5.10);
3. Obter a estimativa intervalar e pontual de μ, σ e ϕ;
4. Obter o valor-*e* resultante do teste a ser realizado;
5. Analisar os resultados gerados.

5.5.1 Exemplo 1

Considere uma amostra de tamanho $n = 10$ de uma variável aleatória X que, dado μ e σ, segue uma *Log- Normal*$(\mu; \sigma)$. Além disso, considere que, hipoteticamente, $\phi \geq 0,9$ é considerado um coeficiente de variação alto.

5,2789	6,4359	18,8257	7,7084	7,9851
20,6774	9,7431	3,4589	4,8934	5,6554

Se, a priori, $\mu \sim Normal(2;\ 1)$ e $\sigma \sim Gama(\frac{1}{1000};\ \frac{1}{1000})$, tem-se que, a posteriori, $(\mu, \sigma)|x$ segue a distribuição definida em 5.8, com $a = 2,\ b = 1,\ \alpha = \beta = \frac{1}{1000}$ e, além disso:

$$\sum_{i=1}^{n} x_i = 90,6620 \text{ e}$$
$$\sum_{i=1}^{n} log(x_i) = 20,4478$$

A Figura 19 apresenta a densidade a posteriori de ϕ.

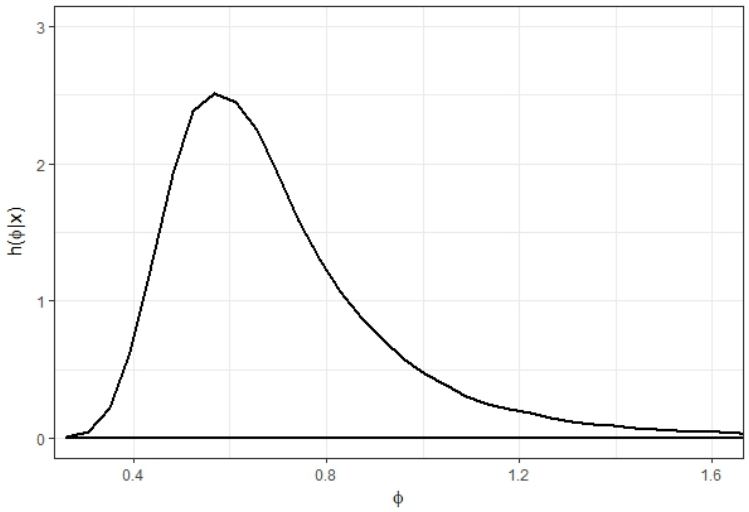

Figura 19 – Densidade a posteriori de ϕ - Log- Normal

A Tabela 10 apresenta as inferências dos parâmetros μ, σ e ϕ, na qual ϕ representa o coeficiente de variação.

Tabela 10 – Inferência Bayesiana dos parâmetros μ, σ e ϕ da distribuição Log- Normal

Parâmetro	Média a posteriori	IC *HPD* 95%
μ	2,0423	[1,6438; 2,4349]
σ	0,6279	[0,3616; 0,9804]
ϕ	0,7244	[0,3539; 1,2418]

Ao analisar a Tabela 10 nota-se que, em média, $\phi = 0,7244$, variando entre $0,3539$ e $1,2418$ com 95% de credibilidade, Além disso, a probabilidade de obter um alto coeficiente, nesse caso, é de $P(\phi \geq 0,9|x) = 0,1675$.

5.5.2 Exemplo 2: Comparação do CV de duas populações Log- Normais independentes

Considere uma amostra de uma população que, dado μ_1 e σ_1, segue uma distribuição *Log- Normal*$(\mu_1; \sigma_1)$. Além disso, retira-se uma segunda amostra, independente da primeira, a priori, que dado μ_2 e σ_2, segue uma distribuição *Log- Normal*$(\mu_2; \sigma_2)$.

População 1	2,2302	8,2796	0,3609	4,1563	10,8397	5,2940
	0,8768	0,3260	8,4344	4,9326	5,0775	4,0966
População 2	2,4176	6,7658	0,7746	3,8694	8,4959	4,6884
	1,2855	0,7354	6,8715	4,4304	4,5340	3,8258
	2,4262	7,0947	5,6795			

Considere a priori que $\mu_1 \sim Normal(2;\ 1)$, $\sigma_1 \sim Gama(\frac{1}{1000};\ \frac{1}{1000})$, $\mu_2 \sim Normal(4;\ 0,6)$ e $\sigma_2 \sim Gama(\frac{1}{1000};\ \frac{1}{1000})$, tem-se que, a posteriori, $(\mu_1,\ \sigma_1,\ \mu_2,\ \sigma_2)|x$ segue a distribuição definida em (5.9), com $a_1 = 2$; $b_2 = 1$; $a_2 = 4$; $b_2 = 0,6$; $\alpha_1 = \alpha_2 = \beta_1 = \beta_2$ e:

$$\sum_{i=1}^{n_1} x_{1i} = 54,9048,$$
$$\sum_{i=1}^{n_2} x_{2i} = 63,8949,$$
$$\sum_{i=1}^{n_1} log(x_{1i}) = 12,8821 \text{ e}$$
$$\sum_{i=1}^{n_2} log(x_{2i}) = 18,3728.$$

A Figura 20 apresenta as densidades marginais a posteriori de ϕ para ambas as populações e a Tabela 11 apresenta as inferências dos parâmetros μ_1, μ_2, σ_1, σ_2 e ϕ (coeficiente de variação) para ambas as amostras.

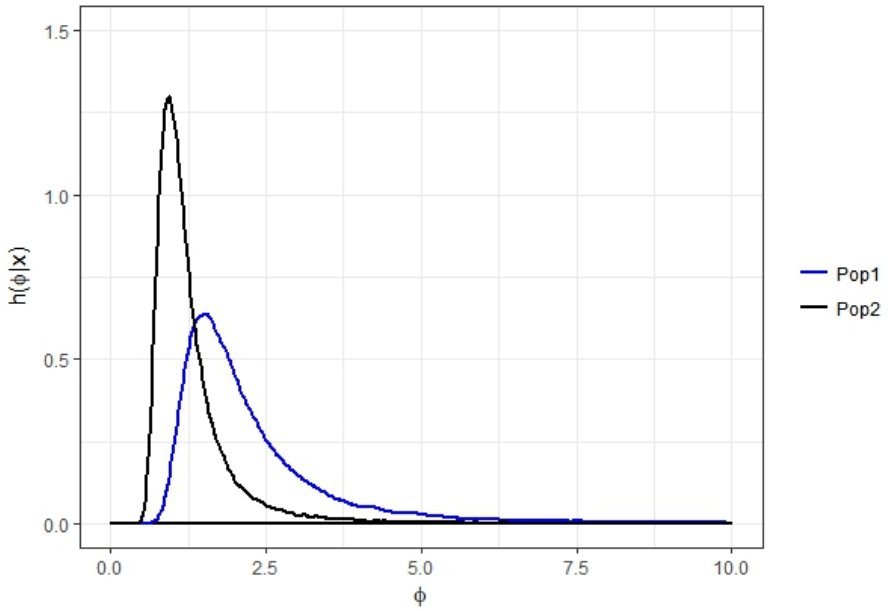

Figura 20 – Densidades marginais a posteriori de ϕ, populações independentes - Log- Normal

Tabela 11 – Inferência Bayesiana de μ, σ e ϕ das duas populações de distribuição Log- Normal

Parâmetros	População 1		População 2	
	Média a posteriori	IC *HPD* 95%	Média a posteriori	IC *HPD* 95%
μ	1,1897	[0,4836; 1,8868]	1,6213	[1,1010; 2,2099]
σ	1,2850	[0,7693; 1,8621]	0,9408	[0,5525; 1,4410]
ϕ	2,5067	[0,7755; 5,3243]	1,3195	[0,5639; 2,5727]

*Inferências geradas com base em uma cadeia de 100000 e Burn-in=10000

De acordo com os resultados apresentados na Figura 20, nota-se que o coeficiente de variação da População 1 tende a assumir valores maiores que o da População 2, o que é

confirmado na Tabela 9, visto que, em média, $\phi_1 = 2,5067$ e $\phi_2 = 1,3195$. Além disso, ϕ_1 varia entre 0,7755 e 5,3243, enquanto ϕ_2 varia entre 0,5639 e 2,5727, ambos com 95% de credibilidade.

Para comparar os parâmetros σ_1 e σ_2 e, consequentemente o coeficiente de variação, utilizou-se o Teste de Significância Genuinamente Bayesiano, ou *FBST*:

$$\begin{cases} H_0 : \phi_1 = \phi_2 & \Rightarrow \quad H_0 : \sigma_1 = \sigma_2 \\ H_a : \phi_1 \neq \phi_2 & \qquad\quad H_a : \sigma_1 \neq \sigma_2 \end{cases}$$

O logaritmo da posteriori conjunta de $(\mu_1,\ \sigma_1,\ \mu_2,\ \sigma_2)|x_1,\ x_2$, dado $H_0 : \sigma_1 = \sigma_2$, é dado por:

$$\begin{aligned} log(\pi(\mu_1,\sigma_1,\mu_2,\sigma_1|x)) = \Bigg\{ \Bigg[\alpha_1 log(\beta_1) + (\alpha_1 - 1)log(\sigma_1) - \beta_1\sigma_1 - log(\Gamma(\alpha_1)) \Bigg] + \\ \Bigg[-\sum_{i=1}^{n_1} log(x_{1i}) - n_1 log(\sigma_1) - \frac{n_1}{2} log(2\pi) - \frac{1}{2\sigma_1^2} \Bigg[\sum_{i=1}^{n_1} [log(x_{1i})]^2 - 2\mu_1 \sum_{i=1}^{n_1} log(x_{1i}) + n_1\mu_1^2 \Bigg] \Bigg] + \\ \Bigg[-\frac{1}{2} log(2\pi) - log(b_1) + \frac{1}{2b_1^2} \Big[\mu_1^2 - 2\mu_1 a_1 + a_1^2 \Big] \Bigg] \Bigg\} + \\ \Bigg\{ \Bigg[\alpha_2 log(\beta_2) + (\alpha_2 - 1)log(\sigma_1) - \beta_2\sigma_1 - log(\Gamma(\alpha_2)) \Bigg] + \\ \Bigg[-\sum_{i=1}^{n_2} log(x_{2i}) - n_2 log(\sigma_1) - \frac{n_2}{2} log(2\pi) - \frac{1}{2\sigma_1^2} \Bigg[\sum_{i=1}^{n_2} [log(x_{2i})]^2 - 2\mu_2 \sum_{i=1}^{n_2} log(x_{2i}) + n_2\mu_2^2 \Bigg] \Bigg] + \\ \Bigg[-\frac{1}{2} log(2\pi) - log(b_2) + \frac{1}{2b_2^2} \Big[\mu_2^2 - 2\mu_2 a_2 + a_2^2 \Big] \Bigg] \Bigg\} \end{aligned} \tag{5.11}$$

O valor $(\mu_1^*, \sigma_1^*, \mu_2^*)$ que maximiza (5.11) não pode ser obtido analiticamente, mas o mesmo pode ser encontrado via métodos numéricos, em particular o método Newton-Raphson. Para tal amostra, tem-se que $(\mu_1^*, \sigma_1^*, \mu_2^*) = (1,1404;\ 0,9641;\ 1,6323)$. Assim, o máximo do logaritmo da posteriori sob H_0 é dado por:

$$log(h_0^*(\mu_1,\sigma_1,\mu_2|x_1,x_2)) = log(h_0(1,1404;\ 0,9641;\ 1,6323|x_1,x_2)) = -92,8181$$

Dessa maneira, a região tangente à hipótese H_0 é dada pelos valores de $\theta = (\mu_1,\sigma_1,\mu_2)$ cujo logaritmo da posteriori é maior do que $log(h_0^*(\mu_1,\sigma_1,\mu_2|x_1,x_2)) = -92,8181$, isto é:

$$T^*(x_1,x_2) = \{\theta = (\mu_1,\sigma_1,\mu_2,\sigma_2) \in \mathbb{R}_+^4 : log(\pi(\mu_1,\sigma_1,\mu_2,\sigma_2|x_1,x_2)) > -92,8181\}$$

Assim, o valor-*e* do *FBST* é dado por:

$$\text{valor-}e = 1 - P((\mu_1, \mu_2, \sigma_1, \sigma_2) \in T(x_1, x_2|X)) = 0,8805$$

Pode-se concluir que não há evidências suficientes para rejeitar a hipótese de $\theta_1 = \theta_2$ e, consequentemente, não há evidências suficientes para afirmar que $\phi_1 \neq \phi_2$, visto que valor-$e > 0,05$.

6 Aplicação em dados reais

A Pesquisa Nacional por Amostra de Domicílios (PNAD) obtém informações anuais sobre características demográficas e socioeconômicas da população, como sexo, idade, educação, trabalho e rendimento, e características dos domicílios, e, com periodicidade variável, informações sobre migração, fecundidade, nupcialidade, entre outras, tendo como unidade de coleta os domicílios. Temas específicos abrangendo aspectos demográficos, sociais e econômicos também são investigados [IBGE].

A fim de comparar a desigualdade econômica no Brasil em 2015, utilizou-se os dados da PNAD do ano em questão. Além disso, fez-se subamostragens dos domicílios amostrados no ano em questão com intuito de analisar o comportamento do coeficiente de variação como uma medida de desigualdade econômica. Nesse caso, não é possível detectar qual a magnitude da desigualdade, mas sim, se há ou não desigualdade entre os estados. Em adição, aqui não há interesse em saber se há desigualdade interna, ou seja, se a renda dentro do estado é distribuída de forma igualitária ou não.

Foram utilizadas as variáveis:

- UF: unidade da federação;
- V0102: número de controle;
- V0103: número de série;
- V0301: número de ordem;
- V4742: rendimento mensal domiciliar per capita

A primeira variável foi utilizada como forma de seleção da amostra de acordo com os estados; as variáveis 2 a 4 serviram para garantir que não havia repetição de um indivíduo dentro da amostra; a variável renda é aquela na qual se tem interesse.

Vale ressaltar que foram excluídos os indivíduos que não declararam a renda e aqueles aos quais não se aplicava a variável em questão. Além disso, pessoas que declararam ter renda igual a R\$0,00 foram consideradas como tendo renda igual a R\$1,00 para que pudesse ser ajustado um modelo Log-Normal aos dados.

6.1 Comparação entre São Paulo e Amazonas

Nessa seção, será analisada a desigualdade econômica entre o Amazonas (AM) e São Paulo (SP). Foram selecionadas amostras de tamanho 30, 50 e 200, de forma aleatória, sem

considerar os pesos da PNAD.

Considere as amostras disponíveis em B, C, D, E, F e G, as quais são independentes em relação ao estado, mas uma é subamostra da outra, ou seja, a de tamanho 30 é subamostra da de 50, que por sua vez é uma subamostra da de 200. Para que possa ser ajustado um modelo *Log- Normal* aos dados de cada amostra faz-se necessário testar se os mesmos seguem tal distribuição. Contudo, se $X \sim Log\text{-}Normal(\mu; \sigma)$, então $log(X) \sim Normal(\mu; \sigma)$. As figuras abaixo permitem a visualização da densidade do logaritmo dos dados obtidos e da distribuição Normal.

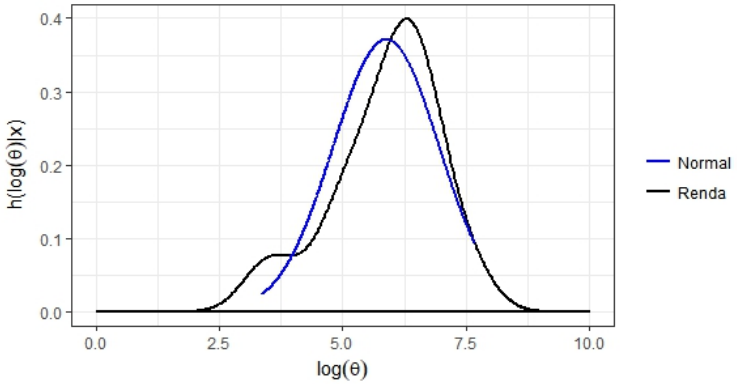

Figura 21 – Densidade empírica do logaritmo da renda- Amazonas- Amostra de tamanho 30
Fonte: PNAD 2015

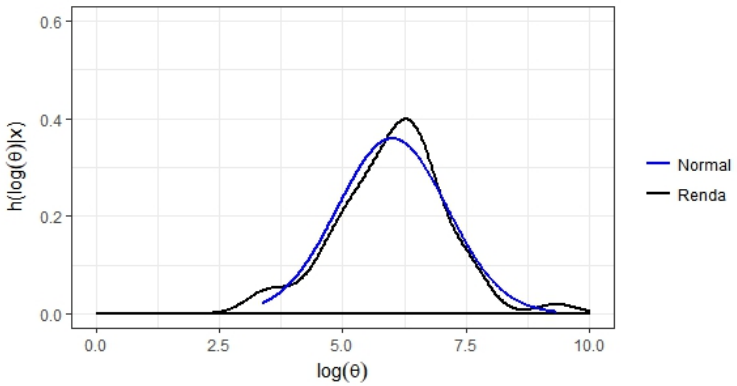

Figura 22 – Densidade empírica do logaritmo da renda- Amazonas- Amostra de tamanho 50
Fonte: PNAD 2015

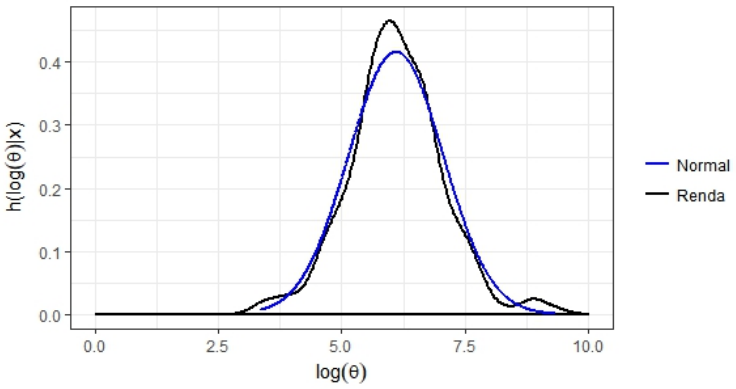

Figura 23 – Densidade empírica do logaritmo da renda- Amazonas- Amostra de tamanho 200
Fonte: PNAD 2015

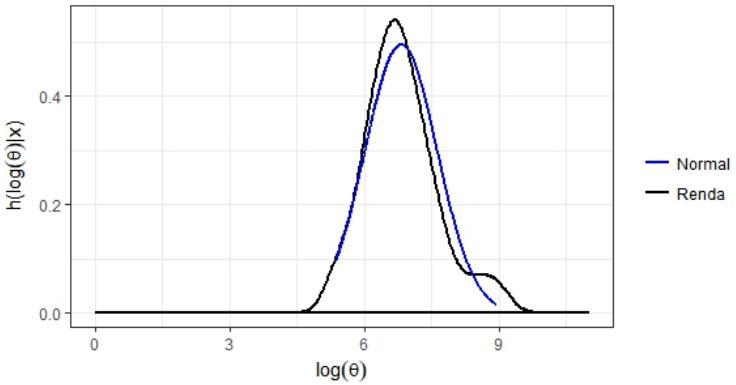

Figura 24 – Densidade empírica do logaritmo da renda- São Paulo- Amostra de tamanho 30
Fonte: PNAD 2015

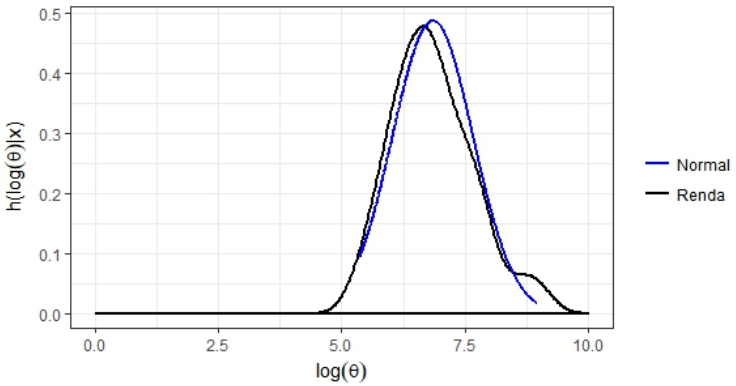

Figura 25 – Densidade empírica do logaritmo da renda- São Paulo- Amostra de tamanho 50
Fonte: PNAD 2015

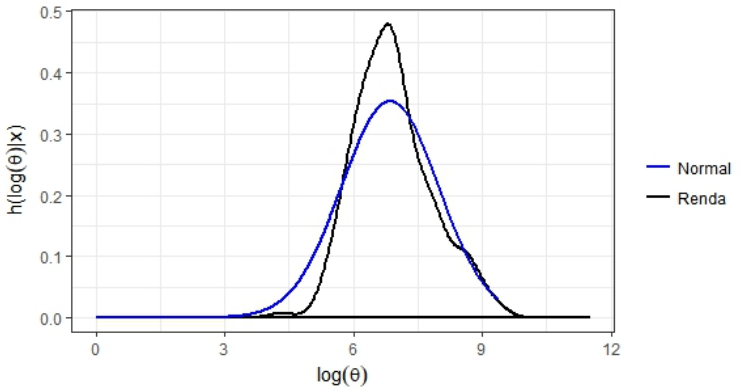

Figura 26 – Densidade empírica do logaritmo da renda- São Paulo- Amostra de tamanho 200
Fonte: PNAD 2015

A Tabela 12 apresenta os resultados do teste de Kolmogorov-Sminorv, utilizado para verificar se, em cada caso, o logaritmo dos dados segue a distribuição Normal e, consequentemente, se os dados seguem a distribuição Log-Normal. As hipóteses são:

$$\begin{cases} H_0 : X \sim \text{ Log-Normal} & \rightarrow \quad H_0 : log(X) \sim \text{Normal} \\ H_a : X \text{ segue outro modelo} & \quad\quad\;\; H_a : log(X) \text{ segue outro modelo} \end{cases}$$

Tabela 12 – Resultados do teste de Kolmogorov- Sminorv para logaritmo dos dados - Comparação entre São Paulo e Amazonas

Amostra	Tamanho	Média	Variância	Teste de Kolmogorov-Sminorv	
				Estatística do teste	Valor-p
Amazonas	30	5,8837	1,0756	0,1343	0,6511
	50	6,0074	1,1109	0,0895	0,8176
	200	6,1019	0,9597	0,0629	0,4069
São Paulo	30	6,8279	0,8058	0,0991	0,9021
	50	6,8487	0,8179	0,0892	0,8207
	200	6,8553	1,1289	0,0913	0,0714

*Fonte: PNAD 2015

Ao analisar os resultados obtidos nota-se que, tanto pelos gráficos quanto pelo teste, o modelo *Log-Normal* é adequado para os dados em questão, visto que o teste de Kolmogorov-Sminorv não rejeitou a hipótese nula a um nível de significância $\alpha = 0,05$ para o logaritmos dos dados.

Considere a priori que $\mu_{AM} \sim Normal(0;20)$, $\sigma_{AM} \sim Gama(\frac{1}{1000};\ \frac{1}{1000})$, $\mu_{SP} \sim Normal(0;20)$, $\sigma_{SP} \sim Gama(\frac{1}{1000};\ \frac{1}{1000})$. Tem-se que, a posteriori, $(\mu_{AM},\ \sigma_{AM},\ \mu_{SP},\ \sigma_{SP})|(x_{AM},\ x_{SP})$ segue a distribuição definida em 5.9, com:

Amostra de tamanho 30:
$\sum_{i=1}^{n_1} x_{AM.i} = 16618,00$
$\sum_{i=1}^{n_2} x_{SP.i} = 40150,00$
$\sum_{i=1}^{n_1} log(x_{AM.i}) = 176,5096$
$\sum_{i=1}^{n_2} log(x_{SP.i}) = 204,8364$

Amostra de tamanho 50:
$\sum_{i=1}^{n_1} x_{AM.i} = 39619,00$
$\sum_{i=1}^{n_2} x_{SP.i} = 68522,00$
$\sum_{i=1}^{n_1} log(x_{AM.i}) = 300,3678$
$\sum_{i=1}^{n_2} log(x_{SP.i}) = 342,4374$

Amostra de tamanho 200:
$\sum_{i=1}^{n_1} x_{AM.i} = 150015,00$
$\sum_{i=1}^{n_2} x_{SP.i} = 313931,20$
$\sum_{i=1}^{n_1} log(x_{AM.i}) = 1220,384$
$\sum_{i=1}^{n_2} log(x_{SP.i}) = 1366,458$

As Figuras 27, 28 e 29 apresentam a densidade a posteriori de ϕ para ambos estados e para as amostras de tamanho 30, 50 e 200, respectivamente.

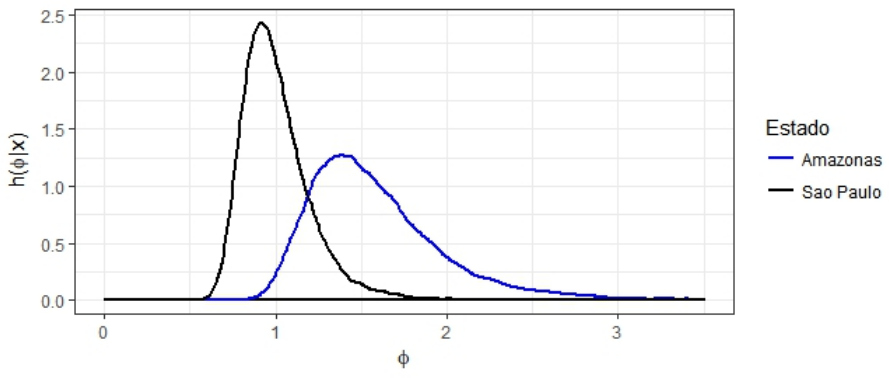

Figura 27 – Densidade a posteriori de ϕ, para o Amazonas e São Paulo - Amostra de tamanho 30

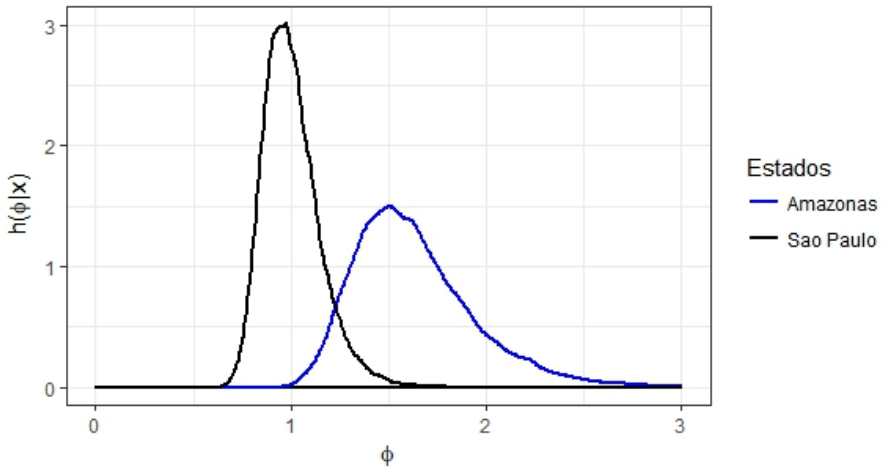

Figura 28 – Densidade a posteriori de ϕ, para o Amazonas e São Paulo - Amostra de tamanho 50

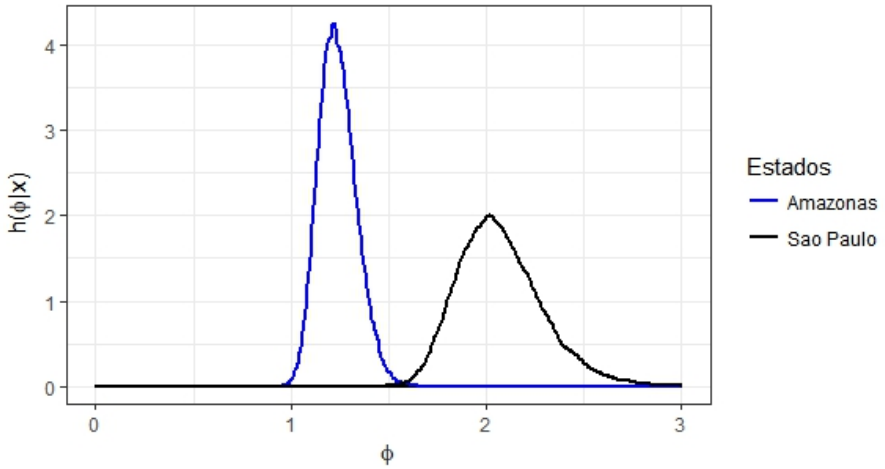

Figura 29 – Densidade a posteriori de ϕ, para o Amazonas e São Paulo - Amostra de tamanho 200

As Tabelas 13, 14 e 14 apresentam as inferências dos parâmetros μ_{AM}, μ_{SP}, σ_{AM}, σ_{SP} e ϕ (coeficiente de variação) para ambos estados, nos diferentes tamanhos de amostra. Vale ressaltar que ξ e δ representam a média e o desvio padrão da distribuição Log-Normal, respectivamente, dadas por:

$$\xi = exp\Big\{\mu + \frac{\sigma^2}{2}\Big\},$$

$$\delta = \sqrt{exp\{2\mu+\sigma^2\}(exp\{\sigma^2\}-1)} \text{ e}$$

Tabela 13 – Inferência Bayesiana de μ, σ e ϕ para ambos estados- Amostra de tamanho 30

Parâmetros	Amazonas		São Paulo	
	Média a posteriori	IC *HPD* 95%	Média a posteriori	IC *HPD* 95%
μ	5,8844	[5,4757; 6,2831]	6,8274	[6,5301; 7,1308]
σ	1,1050	[0,8220; 1,3918]	0,8277	[0,6272; 1,0568]
ϕ	1,5932	[0,9566; 2,4060]	1,0057	[0,6817; 1,4149]
ξ	695,8800	[378,1030; 1097,8090]	1330,2410	[905,9511; 1838,4955]
δ	1171,8930	[373,9276; 2369,1469]	1367,0520	[669,9736; 2352,3139]

*inferências geradas com base em uma cadeia de 100000

Tabela 14 – Inferência Bayesiana de μ, σ e ϕ para ambos estados- Amostra de tamanho 50

Parâmetros	Amazonas		São Paulo	
	Média a posteriori	IC *HPD* 95%	Média a posteriori	IC *HPD*95%
μ	6,0100	[5,6969; 6,3221]	6,8495	[6,6199; 7,0782]
σ^2	1,1291	[0,9109; 1,3595]	0,8310	[0,6686; 0,9967]
ϕ	1,6346	[1,1045; 2,2651]	1,0049	[0,7475; 1,3001]
ξ	794,3812	[503,9539; 1154,5619]	1350,2430	[1007,2550; 1726,5360]
δ	1339,1340	[569,3714; 2406,4002]	1372,3110	[809,9764; 2096,7331]

*inferências geradas com base em uma cadeia de 100000

Tabela 15 – Inferência Bayesiana de μ, σ e ϕ para ambos estados- Amostra de tamanho 200

Parâmetros	Amazonas		São Paulo	
	Média a posteriori	IC *HPD* 95%	Média a posteriori	IC *HPD*95%
μ	6,1028	[5,9733; 6,2401]	6,8319	[6,6588; 7,0152]
σ^2	0,9631	[0,8728; 1,0593]	1,2875	[1,1643; 1,4147]
ϕ	1,2394	[1,0686; 1,4393]	2,0729	[1,6887; 2,5181]
ξ	714,2052	[600,4775; 830,3343]	2144,0170	[1663,1010; 2692,4930]
δ	888,4638	[657,4717; 1144,1212]	4484,2690	[2810,5620; 6393,1460]

*inferências geradas com base em uma cadeia de 100000

Ao analisar os resultados para a mostra de tamanho 30 percebe-se que no Amazonas as pessoas recebem, em média, R$ 695,88, variando entre R$ 378,10 e R$ 1097,81 com 95% de credibilidade. Já em São Paulo, o salário médio, nesse caso, é de R$ 1330,24, com um intervalo de 95% de credibilidade que vai de R$ 905,95 a R$ 1838,50. O coeficiente de variação obtido para o Amazonas foi igual a 1,59 e para São Paulo foi de 1,01, o que mostra que o segundo é mais homogêneo que o primeiro, visto que o valor dessa medida para o Amazonas é 1,58 vezes o valor para São Paulo.

Para a amostra de tamanho 50 o cenário permanece o mesmo: maior média salarial ocorre em São Paulo (R$ 1350,24) quando comparado ao Amazonas (R$ 794,38) e os dados do Amazonas são mais heterogêneos que os de São Paulo, pois $CV_{AM} = 1,63 > 1,01 = CV_{SP}$.

Por fim, para a amostra de tamanho 200 pode-se notar que o cenário manteve-se igual ao das amostras anteriores em relação á media salarial: no Amazonas as pessoas recebem, em

média, menos que em São Paulo, visto que a média obtida para o Amazonas foi de R$ 714,21 e para São Paulo foi de R$ 2144,02. Em relação ao coeficiente de variação, o cenário inverteu quando comparado aos valores obtidos nas demais amostras: nesse caso, os dados do Amazonas são mais homogêneos que os de São Paulo, visto que o CV obtido para o primeiro foi igual a 1,64 e, para o segundo, 2,07.

Com intuito de comparar se os coeficientes obtidos para os estados em análise são diferentes, em cada caso, aplicou-se o *FBST*. As hipóteses do teste são iguais para ambos tamanhos de amostra:

$$\begin{cases} H_0 : \phi_{AM} = \phi_{SP} & \Rightarrow \quad H_0 : \sigma_{AM} = \sigma_{SP} \\ H_a : \phi_{AM} \neq \phi_{SP} & \qquad\quad H_a : \sigma_{AM} \neq \sigma_{SP} \end{cases}$$

O logaritmo da posteriori conjunta de $(\mu_{AM},\ \sigma_{AM},\ \mu_{SP},\ \sigma_{SP})|x_{AM},\ x_{SP}$, dado $H_0 : \sigma_{AM} = \sigma_{SP}$, é dado por:

$$\begin{aligned} log(\pi(\mu_{AM},\sigma_{AM},\mu_{SP},\sigma_{AM}|x)) = \Bigg\{ \Bigg[\alpha_{AM} log(\beta_{AM}) + \\ (\alpha_{AM}-1)log(\sigma_{AM}) - \beta_{AM}\sigma_{AM} - log(\Gamma(\alpha_{AM})) \Bigg] + \Bigg[-\sum_{i=1}^{n_1} log(x_{AM.i}) - \\ n_1 log(\sigma_{AM}) - \frac{n_1}{2} log(2\pi) - \frac{1}{2\sigma_{AM}^2} \Bigg[\sum_{i=1}^{n_1} [log(x_{AM.i})]^2 - \\ 2\mu_{AM} \sum_{i=1}^{n_1} log(x_{AM.i}) + n_1 \mu_{AM}^2 \Bigg] \Bigg] + \Bigg[-\frac{1}{2} log(2\pi) - log(b_{AM}) + \\ \frac{1}{2b_{AM}^2} \Big[\mu_{AM}^2 - 2\mu_{AM} a_{AM} + a_{AM}^2 \Big] \Bigg] \Bigg\} + \\ \Bigg\{ \Bigg[\alpha_{SP} log(\beta_{SP}) + (\alpha_{SP}-1) log(\sigma_{AM}) - \beta_{SP}\sigma_{AM} - log(\Gamma(\alpha_{SP})) \Bigg] + \\ \Bigg[-\sum_{i=1}^{n_2} log(x_{SP.i}) - n_2 log(\sigma_{AM}) - \frac{n_2}{2} log(2\pi) - \\ \frac{1}{2\sigma_{AM}^2} \Bigg[\sum_{i=1}^{n_2} [log(x_{SP.i})]^2 - 2\mu_{SP} \sum_{i=1}^{n_2} log(x_{SP.i}) + n_2 \mu_{SP}^2 \Bigg] \Bigg] + \\ \Bigg[-\frac{1}{2} log(2\pi) - log(b_{SP}) + \frac{1}{2b_{SP}^2} \Big[\mu_{SP}^2 - 2\mu_{SP} a_{SP} + a_{SP}^2 \Big] \Bigg] \Bigg\} \end{aligned} \tag{6.1}$$

Na qual $n_1 = n_2 = 30$ no caso da amostra de tamanho 30, $n_1 = n_2 = 50$ no caso da amostra de tamanho 50 e $n_1 = n_2 = 200$ para a de tamanho 200.

O valor $(\mu^*_{AM},\, \mu^*_{SP},\, \sigma^*)$ que maximiza 6.1 não pode ser obtido analiticamente, mas o mesmo pode ser encontrado via métodos numéricos, em particular o método Newton-Raphson. Para a amostra de tamanho 30 tem-se:

$$(\mu^*_{AM},\, \mu^*_{SP},\, \sigma^*) = (5,8837;\, 0,9194;\, 6,8267)$$

E o máximo do logaritmo da posteriori sob H_0 é dado por:

$$\begin{gathered} log(h^*_0(\mu^*_{AM},\, \mu^*_{SP},\, \sigma^*|x_{AM},x_{SP})) = \\ log(h_0(5,8837;\, 0,9194;\, 6,8267|x_{AM},x_{SP})) = -484,0175 \end{gathered} \tag{6.2}$$

Dessa maneira, a região tangente à hipótese H_0 é dada pelos valores de $\theta = (\mu_{AM},\, \sigma_{AM},\, \mu_{SP},\, \sigma_{SP})$ cujo logaritmo da posteriori é maior do que 6.2. Isto é:

$$T(x_{AM},x_{SP}) = \{\theta = (\mu_{AM},\sigma_{AM},\mu_{SP},\sigma_{SP}) \in \mathbb{R}^4_+ :$$
$$log(\pi(\mu_{AM},\sigma_{AM},\mu_{SP},\sigma_{SP}|x_{AM},x_{SP})) > -484,0175\},\ \text{no caso da amostra de tamanho 30}$$

Analogamente, para as amostras de tamanho 50 e 200, tem-se, respectivamente:

$$T(x_{AM},x_{SP}) = \{\theta = (\mu_{AM},\sigma_{AM},\mu_{SP},\sigma_{SP}) \in \mathbb{R}^4_+ :$$
$$log(\pi(\mu_{AM},\sigma_{AM},\mu_{SP},\sigma_{SP}|x_{AM},x_{SP})) > -802,8880\},\ \text{no caso da amostra de tamanho 50}$$

$$T(x_{AM},x_{SP}) = \{\theta = (\mu_{AM},\sigma_{AM},\mu_{SP},\sigma_{SP}) \in \mathbb{R}^4_+ :$$
$$log(\pi(\mu_{AM},\sigma_{AM},\mu_{SP},\sigma_{SP}|x_{AM},x_{SP})) > -3225,2340\},\ \text{no caso da amostra de tamanho 200}$$

Assim, o valor-e do *FBST* é dado por:

Amostra de tamanho 30
$$\text{valor-}e = 1 - P((\mu_{AM},\mu_{SP},\sigma_{AM},\sigma_{SP}) \in T(x_{AM},x_{SP})) = 0,6621$$

Amostra de tamanho 50:
$$\text{valor-}e = 1 - P((\mu_{AM},\mu_{SP},\sigma_{AM},\sigma_{SP}) \in T(x_{AM},x_{SP})) = 0,3318$$

Amostra de tamanho 200:
$$\text{valor-}e = 1 - P((\mu_{AM},\mu_{SP},\sigma_{AM},\sigma_{SP}) \in T(x_{AM},x_{SP})) = 0,0021$$

Pode-se concluir que, para as amostras de tamanho 30 e 50 não há evidências suficientes para rejeitar a hipótese de $\theta_{AM} = \theta_{SP}$ e, consequentemente, não se pode afirmar que $\phi_{AM} \neq \phi_{SP}$,

visto que *e-valor*$> 0,05$ em ambos os casos. Consequentemente, não é possível afirmar que há desigualdade entre os estados do Amazonas e de São Paulo.

No caso da amostra de tamanho 200, pode-se concluir que há evidências suficientes para rejeitar a hipótese de $\theta_{AM} = \theta_{SP}$ e, consequentemente, $\phi_{AM} \neq \phi_{SP}$, visto que valor-$e < 0,05$. Sendo assim, pode-se afirmar que há desigualdade econômica entre o estado do Amazonas e São Paulo. Contudo, não é possível quantificar essa desigualdade e nem se pode afirmar que a distribuição de renda é diferente em cada estado, pois o coeficiente de variação depende apenas do parâmetro σ, e a distribuição a posteriori depende do parâmetro μ, além de σ.

Vale ressaltar que, como é possível observar nos dados apresentados, ao aumentar o tamanho da amostra, o teste ganha credibilidade e consegue detectar, no caso da amostra de tamanho 200, a diferença existente entre o coeficiente do Amazonas e o de São Paulo, além da estimação também ser mais precisa.

6.2 Panorama geral da distribuição de renda em São Paulo e no Amazonas

Nessa seção, será analisada a desigualdade econômica entre o Amazonas (AM) e São Paulo (SP) utilizando todos os dados disponíveis na PNAD. Nesse caso, a mostra para o Amazonas foi de 11961 domicílios e, para São Paulo, 37917 domicílios. Considere toda a amostra da PNAD de 2015 e, além disso, considere também que essas amostras são independentes umas das outras. As figuras abaixo permitem a visualização da densidade empírica dos dados obtidos e da distribuição Normal.

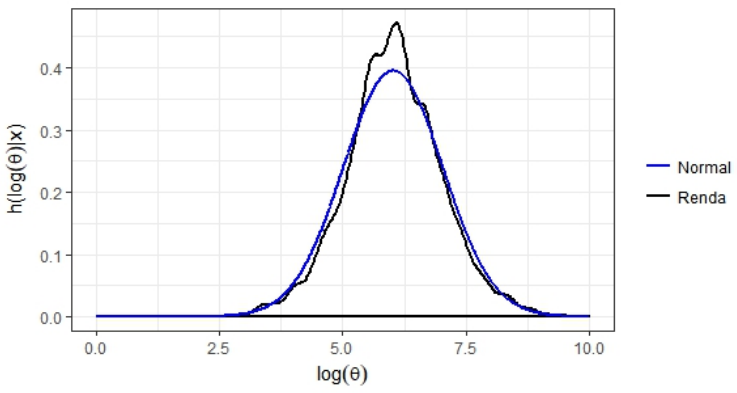

Figura 30 – Densidade empírica do logaritmo da renda - Amazonas, PNAD 2015

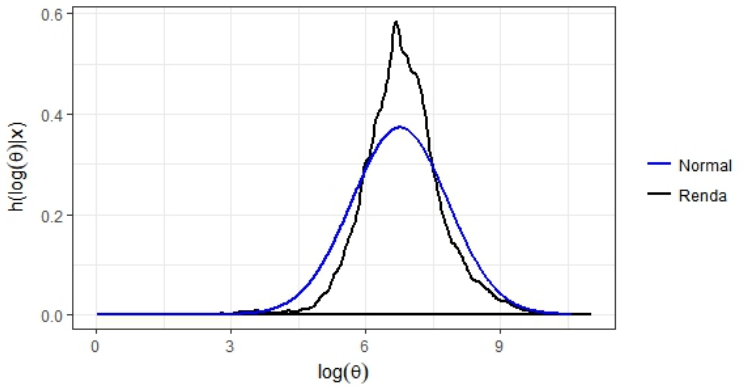

Figura 31 – Densidade empírica do logaritmo da renda - São Paulo, PNAD 2015

Considere a priori que $\mu_{AM} \sim Normal(0,20)$, $\sigma_{AM} \sim Gama(\frac{1}{1000}; \frac{1}{1000})$, $\mu_{SP} \sim Normal(0,20)$, $\sigma_{SP} \sim Gama(\frac{1}{1000}; \frac{1}{1000})$. Tem-se que, a posteriori, $(\mu_{AM}, \sigma_{AM}, \mu_{SP}, \sigma_{SP})|(x_{AM}, x_{SP})$ segue a distribuição definida em 5.9, com:

$$\sum_{i=1}^{n_1} x_{am.i} = 8120717,00$$
$$\sum_{i=1}^{n_2} x_{sp.i} = 51383117,00$$
$$\sum_{i=1}^{n_1} log(x_{am.i}) = 71975,61$$
$$\sum_{i=1}^{n_2} log(x_{sp.i}) = 255527,20$$

A Figura 32 apresenta a densidade a posteriori de ϕ para ambos estados. A Tabela 16 apresenta as inferências dos parâmetros μ_{AM}, μ_{SP}, σ_{AM}, σ_{SP} e ϕ (coeficiente de variação) para o logaritmo dos dados de ambos estados.

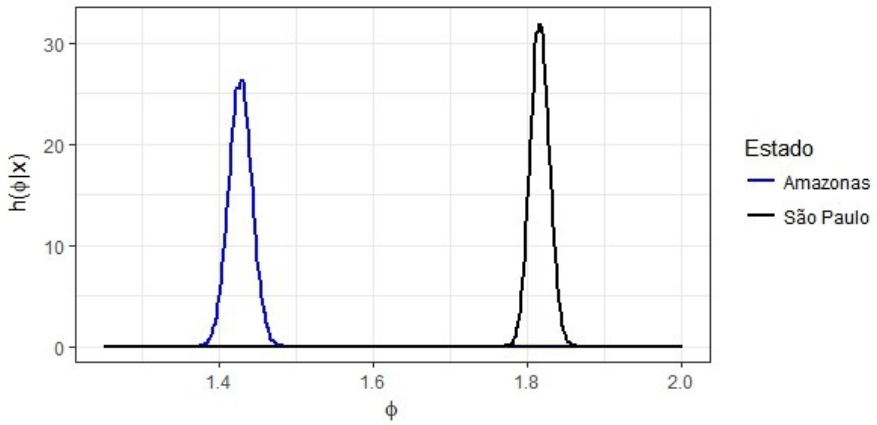

Figura 32 – Densidade a posteriori de ϕ, para o Amazonas e São Paulo - Amostra da PNAD 2015

Tabela 16 – Inferência Bayesiana de μ, σ e ϕ para ambos estados- Amostra da PNAD 2015

Parâmetros	Amazonas		São Paulo	
	Média a posteriori	IC *HPD* 95%	Média a posteriori	IC *HPD* 95%
μ	6,0176	[5,9988; 6,0368]	6,7374	[6,7253; 6,7492]
σ	1,0539	[1,0409; 1,0672]	1,2077	[1,1992; 1,2160]
ϕ	1,4273	[1,3982; 1,4572]	1,8165	[1,7922; 1,8405]
ξ	715,5601	[698,4917; 732,1817]	1748,7300	[1721,8160; 1776,1510]
δ	1021,3580	[981,4443; 1062,2503]	3176,6770	[3092,3950; 3258,9990]

*inferências geradas com base em uma cadeia de 100000

Ao analisar os resultados da Tabela 16 percebe-se que no Amazonas as pessoas recebem, em média, R$ 715,56, variando entre R$ 698,49 e R$ 732,18 com 95% de credibilidade. Já em São Paulo, o salário médio, nesse caso, é de R$ 1748,66, com um intervalo de 95% de credibilidade que vai de R$ 1721,82 a R$ 1776,15. O coeficiente de variação obtido para o Amazonas foi igual a 1,4273 e para São Paulo foi de 1,8165, o que mostra que o primeiro é mais homogêneo que o segundo, visto que o valor dessa medida para São Paulo é 1,28 vezes o valor para o Amazonas.

Por fim, a Figura 32 mostra que a densidade do coeficiente de variação de São Paulo está deslocada à direita em relação à do Amazonas, estando ambas nitidamente separadas, o que induz a pensar que de fato os coeficientes são diferentes. Contudo, para analisar se realmente isso ocorre, aplicou-se o *FBST*. As hipóteses do teste são:

$$\begin{cases} H_0 : \phi_{AM} = \phi_{SP} & \Rightarrow \quad H_0 : \sigma_{AM} = \sigma_{SP} \\ H_a : \phi_{AM} \neq \phi_{SP} & \qquad\quad H_a : \sigma_{AM} \neq \sigma_{SP} \end{cases}$$

O logaritmo da posteriori conjunta de $(\mu_{AM},\ \sigma_{AM},\ \mu_{SP},\ \sigma_{SP})|x_{AM},\ x_{SP}$, dado $H_0 : \sigma_{AM} = \sigma_{SP}$, é dado pela fórmula definida em 6.1, na qual $n_1 = n_{AM} = 11961$ e $n_2 = n_{SP} = 37927$.

O valor $(\mu^*_{AM},\ \mu^*_{SP},\ \sigma^*)$ que maximiza 6.1 não pode ser obtido analiticamente, mas o mesmo pode ser encontrado via métodos numéricos, em particular o método Newton-Raphson. Para a amostra da PNAD 2015, tem-se:

$$(\mu^*_{AM},\ \mu^*_{SP},\ \sigma^*) = (6,0166;\ 1,1724;\ 6,7375)$$

E o máximo do logaritmo da posteriori sob H_0 é dado por:

$$\begin{gathered} log(h^*_0(\mu^*_{AM},\ \mu^*_{SP},\ \sigma^*|x_{AM},x_{SP})) = \\ log(h_0(6,0166;\ 1,1724;\ 6,7375|x_{AM},x_{SP})) = -406256,1000 \end{gathered} \tag{6.3}$$

Dessa maneira, a região tangente à hipótese H_0 é dada pelos valores de $\theta = (\mu_{AM},\ \sigma_{AM},\ \mu_{SP},\ \sigma_{SP})$ cujo logaritmo da posteriori é maior do que 6.3. Isto é:

$$T(x_{AM}, x_{SP}) = \{\theta = (\mu_{AM}, \sigma_{AM}, \mu_{SP}, \sigma_{SP}) \in \mathbb{R}^4_+ :$$
$$log(\pi(\mu_{AM}, \sigma_{AM}, \mu_{SP}, \sigma_{SP} | x_{AM}, x_{SP})) > -406256,1\}$$

Assim, o valor-e do *FBST* é dado por:

$$\text{valor-}e = 1 - P((\mu_{AM}, \mu_{SP}, \sigma_{AM}, \sigma_{SP}) \in T(x_{AM}, x_{SP})) = 0,0000,$$

Pode-se concluir que há evidências suficientes para rejeitar a hipótese de $\theta_{AM} = \theta_{SP}$ e, consequentemente, conclui-se que $\phi_{AM} \neq \phi_{SP}$, visto que *e-valor*$< 0,05$ em ambos os casos. Sendo assim, pode-se afirmar que há desigualdade econômica entre o estado do Amazonas e São Paulo. Contudo, não é possível quantificar essa desigualdade pelo que já foi explicitado anteriormente.

7 Visão Clássica e Bayesiana

Nesse capítulo será feita uma comparação entre a visão Clássica e a Bayesiana para geração do intervalo de confiança/credibilidade do coeficiente de variação. Aqui, mesmo não sendo correto assumir que os dados seguem uma distribuição Normal pelo fato dessa medida não poder ser utilizada em casos nos quais os dados podem assumir valores negativos, será feita essa consideração pelo fato de não ter sido encontrada nenhuma referência que assumisse outra distribuição que não a Normal ou *T-Student*.

Segundo Hervé Abdi [10], se assumir que os dados seguem uma distribuição Normal é possível encontrar o intervalo de confiança para o coeficiente de variação através da seguinte fórmula:

$$t_{cv} = C_v \pm t_{\alpha,v}\, S_{c_v} \tag{7.1}$$

Na qual $t_{\alpha,v}$ é o valor crítico da distribuição T-Student com $N-1$ graus de liberdade e α o nível de significância; e $S_{c_v} = \frac{C_V}{\sqrt{2N}}$, com N = tamanho da amostra.

Assim, aplicando aos dados da PNAD 2015, obteve-se o seguinte intervalo com 95% de confiança:

Tabela 17 – Inferência Clássica para o coeficiente de variação em ambos estados - Amostra da PNAD 2015

Estado	Média	Desvio padrão	Coeficiente de Variação	Desvio padrão do CV	Intervalo de Confiança para o CV
Amazonas	678,93	986,63	1,4532	0,0094	[1,4348; 1,4716]
São Paulo	1354,79	1774,05	1,3094	0,0048	[1,3188; 1,3002]

Comparando os resultados obtidos na visão Clássica e na Bayesiana é possível notar que a média salarial para ambos estados é maior na Bayesiana do que na clássica: para o Amazonas a renda obtida na visão Bayesiana é 1,05 vezes a da Clássica; para São Paulo essa proporção foi igual a 1,29. Além disso, no caso Bayesiano a estimativa do coeficiente de variação obtida para São Paulo foi maior que na visão Frequentista e para o Amazonas foi menor. Por fim, o cenário inverteu-se no caso Clássico: a distribuição de renda no Amazonas é mais heterogênea que em São Paulo; enquanto na visão Bayesiana a de São Paulo é mais heterogênea que a do Amazonas.

8 Conclusão

O coeficiente de variação é uma medida que exclui a influência da ordem de grandeza dos dados, permitindo a análise da dispersão dos mesmos e comparação da precisão de dois instrumentos de medição ou bancos de dados com unidades diferentes. Essa medida representa a dispersão dos dados em relação à média. Vale ressaltar que só deve ser usado em casos nos quais a variável de interesse é positiva e do tipo razão.

Atualmente há pouco conteúdo disponível em relação a essa medida, principalmente quando se trata da utilização do paradigma Bayesiano para realizar as inferências e os trabalhos que existem e que usam a inferência clássica consideram que os dados seguem uma distribuição Normal ou *T-Student*, o que é incorreto, visto que, devido à definição do coeficiente de variação, não se pode calcular essa medida nos casos em que os dados assumem valores negativos e as duas distribuições citadas possuem suporte real (estão definidas em toda a reta real). Tendo isso em vista, este trabalho teve como objetivo realizar inferência Bayesiana do coeficiente de variação, sem considerar que os dados seguem tais distribuições. Além disso, como não existe nenhum pacote no *software R* que realize inferência do coeficiente de variação e gere o gráfico do *Full Bayesian Significance Test (FBST)* ou Teste de Significância Genuinamente Bayesiano, este trabalho deixa disponível o código em linguagem *R* para tal.

Foram utilizados dados simulados para exemplificar a metodologia e aplicou-a em dados reais obtidos através da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2015 com intuito de verificar qual estados, Amazonas ou São Paulo, apresenta maior desigualdade de renda. Para esse fim foi utilizado a variável "rendimento mensal domiciliar per capita". A metodologia também foi ilustrada por meio de dados artificiais gerados por distribuições com 1 (um) ou 2 (dois) parâmetros: Binomial, Binomial Negativa, Poisson, Gama e Log- Normal; mas o procedimento não se restringe apenas a essas distribuições: o mesmo pode ser aplicado a toda e qualquer distribuição definida na reta real positiva.

É importante ressaltar que:

- Nos casos em que a distribuição possui apenas um parâmetro, o CV depende do parâmetro e o teste para duas amostras dê significativo é correto concluir que as distribuições também são diferentes;
- Distribuições diferentes não implicam em CV's diferentes: pode-se ter distribuições diferentes que resultem em um mesmo CV;
- Além disso, distribuições iguais implicam em CV's iguais;
- CV's iguais não implicam em distribuições iguais;

- Nos casos em que a distribuição tem mais de um parâmetro, e o CV não envolve todos eles, pode-se concluir que as distribuições são diferentes se os CV's derem diferentes, no entanto não se pode concluir que as distribuições são iguais caso os CV's derem iguais.

As tabelas a seguir resumem os possíveis resultados acima citados. Vale ressaltar que na tabela 18 o resultado é em relação ao que uma distribuição implica no CV e a tabela 19 mostra os possíveis resultados de uma distribuição quando se analisa o CV, para diferentes quantidades de parâmetros e resultados do FBST:

Tabela 18 – Distribuições implicam em CV's:

Distribuições	Conclusão Coeficientes de variação
Diferentes	Não necessariamente diferentes
Iguais	Iguais

Tabela 19 – CV's implicam em distribuições

Número de parâmetros	Número de parâmetros da distribuição no CV	FBST significativo?	Conclusão	
			CV's	Distribuições
1	1	Sim	Diferentes	Diferentes
1	1	Não	Iguais	Iguais
2 ou mais	Menos que o total	Sim	Diferentes	Diferentes
2 ou mais	Menos que o total	Não	Iguais	Não necessariamente

Aplicando a metodologia proposta aos dados da PNAD 2015, pode-se notar que a mesma é capaz de detectar a desigualdade econômica existente entre ambos estados, mas não se há desigualdade interna ou qual desiguais eles são. Aqui, pode-se notar que de fato São Paulo e Amazonas apresentam níveis diferentes de desigualdade de renda, visto que os dois estados apresentam CV's diferentes. Assim, apesar de São Paulo ter uma maior variação de renda, ele se mostrou menos desigual que o Amazonas, pois $CV_{SP} < CV_{AM}$. Vale ressaltar que, nesse caso, a amostra de São Paulo não ficou tão próxima da distribuição Log- Normal, mas o afastamento não foi extremo.

Referências

[1] Amaral, A.M.; Muniz, J.A.; Souza, M. *"Avaliação do coeficiente de variação como medida da precisão na experimentação com citros"*. Pesq. Agropec. Bras., Brasília, vol.32, n.12, p.1221-1225, 1997

[2] Blanxart, M.F.; Cosialls, L.S.; Olmos, J.G. et al. *"Análisis exploratorio de datos: nuevas técnicas estadísticas"*. Barcelona: Promociones y Publicaciones Universitarias, p. 296, 1992.

[3] Berger,J. O.; Selke, T. *"Testing a point null hypothesis: The irreconciability of p-values and evidence"*. Journal of the American Statistical Association, (82):112–130, 1987.

[4] Conover, W.J. *"Practical Nonparametric Statistics"*. United Sates of America, 1980.

[5] Costa, N.H.A.D., Seraphin, J.C., Zimmermann, F.J.P. *"Novo método de classificação de coeficientes de variação para a cultura do arroz de terras altas"*. Pesq. Agropec. Bras., vol.37, n.3, p.243-249, 2002.

[6] Dobson, A. J.; Barnett, A. G. *"An Introduction to Generalized Linear Models"*. CRC Press, 2008.

[7] Ehlers, R. S. *"Inferência Bayesiana"*. (Apostila de curso) 2007. URL: www2.icmc.usp.br/ ehlers/bayes/.

[8] Gamerman, D.; Lopes, H.F. *"Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference"*. Ed. 2 Chapman and Hall, U.K.: p.342, 2006.

[9] Garcia, C. H. *"Tabelas para classificação do coeficiente de variação"*. Piracicaba: IPEF 1989. 12p. (Circular técnica n. 171).

[10] Hérve, A. *"Coefficient of Variation"*. Encyclopedia of Research Desig, Thousand Oaks, CA, 2010.

[11] Hoaglin, D.C.; Mosteller, F.; Tuckey, J.W. *"Understanding robust and exploratory data analysis"*. New York: J. Wiley, p. 477, 1983.

[12] Jeffreys, H. *"Theory of Probability"*. Ed. 3. Oxford, U.K.: Oxford University Press, 1961.

[13] Judice, M. G.; Miniz, J. A.; Carvalheiro, R. *"Avaliação do coeficiente de variação na experimentação com suínos"*. Ciênc. e Agrotec., vol. 23, n.1, p.170-173, 1999.

[14] Lindley, D. V. *"Some Comments on Bayes Factors"*. Journal of Statistical Planning and Inference, (61):181–189, 1997.

[15] Mead, R.; Curnow, R.N. *"Statistical methods in agriculture and experimental biology"*. New York: Chapman and Hall, p. 335, 1983.

[16] Paulino, C. D.; Turkman, M. A. A.; Murteira, B. *"Estatística Bayesiana"*. Lisboa: Fundação Calouste Gulbenkian, 2003.

[17] Pereira, C. A. de B.; Stern, J. M. *"Can a Significance Test Be Genuinely Bayesian?"*. Bayesian Analysis, v.3, n.1, p. 79-100, 2008.

[18] Pereira, C. A. de B.; Stern, J. M. *"Evidence and credibility: full bayesian significance test of precise hypothesis"*. Entropy, 1:99–110, 1999.

[19] Gomes, P. *"Curso de Estatística Experimental"*. Piracicaba: Nobel, ed.12, 1990. 467p.

[20] Gomes, P. *"Curso de Estatística Experimental"*. Piracicaba: Degaspari, ed.14, 2000. 477p.

[21] R Development Core Team (2009) *"R: A language and environment for statistical computing"*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[22] Shafer, G. *"Lindley's paradox by Glenn Shafer"*. Nota técnica 125, Stanford University, Califórnia, 1976.

[23] Triola, M. F. *"Introdução a estatística"*. LTC, ed. 7, 1999.

ANEXO A – Tabela dos quantis do teste de Kolmogorov-Sminorv

	α						α				
n	0,20	0,10	0,05	0,025	0,01	n	0,20	0,10	0,05	0,025	0,01
1	0,900	0,950	0,975	0,990	0,995	21	0,226	0,259	0,287	0,321	0,344
2	0,684	0,776	0,842	0,900	0,929	22	0,221	0,253	0,281	0,314	0,337
3	0,565	0,636	0,708	0,785	0,829	23	0,216	0,247	0,275	0,307	0,330
4	0,493	0,565	0,624	0,689	0,734	24	0,212	0,242	0,269	0,301	0,323
5	0,447	0,509	0,563	0,627	0,669	25	0,208	0,238	0,264	0,295	0,317
6	0,410	0,468	0,519	0,577	0,617	26	0,204	0,233	0,259	0,290	0,311
7	0,381	0,436	0,483	0,538	0,576	27	0,200	0,229	0,254	0,284	0,305
8	0,358	0,410	0,454	0,507	0,542	28	0,197	0,225	0,250	0,279	0,300
9	0,339	0,387	0,430	0,480	0,513	29	0,193	0,221	0,246	0,275	0,295
10	0,323	0,369	0,409	0,457	0,489	30	0,190	0,218	0,242	0,270	0,290
11	0,308	0,352	0,391	0,437	0,468	31	0,187	0,214	0,238	0,266	0,285
12	0,296	0,338	0,375	0,419	0,449	32	0,184	0,211	0,234	0,262	0,281
13	0,285	0,325	0,361	0,404	0,432	33	0,182	0,208	0,231	0,258	0,277
14	0,275	0,314	0,349	0,390	0,418	34	0,179	0,205	0,227	0,254	0,273
15	0,266	0,304	0,338	0,377	0,404	35	0,177	0,202	0,224	0,251	0,269
16	0,258	0,295	0,327	0,366	0,392	36	0,174	0,199	0,221	0,247	0,265
17	0,250	0,286	0,318	0,355	0,381	37	0,172	0,196	0,218	0,244	0,262
18	0,244	0,279	0,309	0,346	0,371	38	0,170	0,194	0,215	0,241	0,258
19	0,237	0,271	0,301	0,337	0,361	39	0,168	0,191	0,213	0,238	0,255
20	0,232	0,265	0,294	0,329	0,352	40	0,165	0,189	0,210	0,235	0,252
						Aproximação para n>40	$\frac{1,07}{\sqrt{n}}$	$\frac{1,22}{\sqrt{n}}$	$\frac{1,36}{\sqrt{n}}$	$\frac{1,52}{\sqrt{n}}$	$\frac{1,63}{\sqrt{n}}$

ANEXO B – Amostra de tamanho 30 - Amazonas

Amazonas						
ID	UF	Número de controle	Número de série	Número de ordem	Sexo	Rendimento mensal domiciliar per capta
14673	13	13000772	12	1	4	162
20767	13	13001930	2	3	2	566
16150	13	13001043	1	3	4	1769
21899	13	13002155	22	2	4	1225
22586	13	13002309	23	4	2	252
11710	13	13000128	12	3	4	866
17614	13	13001370	6	1	2	280
22012	13	13002198	3	2	4	788
17903	13	13001426	8	5	2	535
16737	13	13001167	12	3	4	400
22782	13	13002333	8	3	4	500
16697	13	13001159	12	2	2	400
19424	13	13001701	9	5	4	100
18162	13	13001477	4	3	2	2166
12400	13	13000284	4	1	2	800
22100	13	13002228	1	1	2	684
14174	13	13000683	16	2	4	850
11669	13	13000110	16	1	4	394
15160	13	13000870	1	2	4	925
22754	13	13002333	1	5	4	522
21977	13	13002180	6	2	2	140
19614	13	13001736	17	2	4	50
18980	13	13001620	14	7	2	31
23240	13	13002449	13	4	2	278
19163	13	13001655	14	1	4	166
19803	13	13001779	9	1	2	750
17815	13	13001400	14	1	2	477
18421	13	13001523	11	1	4	156
14692	13	13000772	16	3	2	29
12947	13	13000381	14	2	4	357

ANEXO C – Amostra de tamanho 30 - São Paulo

São Paulo						
ID	UF	Número de controle	Número de série	Número de ordem	Sexo	Rendimento mensal domiciliar per capta
227930	35	35000716	3	1	4	817
257877	35	35008270	11	3	4	7900
262879	35	35009675	17	1	2	1666
235984	35	35002743	4	2	2	1192
231915	35	35001739	2	2	4	460
226560	35	35000406	20	3	4	675
232296	35	35001836	3	4	2	912
250659	35	35006420	4	1	2	900
226119	35	35000287	6	1	2	1066
225535	35	35000112	9	2	4	5000
240909	35	35003928	11	2	4	350
257718	35	35008210	14	2	4	1400
240281	35	35003774	10	1	4	788
240454	35	35003812	12	5	4	1260
235805	35	35002697	17	1	2	750
242752	35	35004371	11	1	2	430
243449	35	35004541	1	3	4	575
246696	35	35005432	10	4	2	612
251582	35	35006692	6	4	4	215
229624	35	35001100	14	1	2	925
233836	35	35002190	7	2	2	700
256612	35	35007907	11	1	2	587
229053	35	35000970	6	1	4	1576
253430	35	35007125	5	7	4	469
233816	35	35002182	18	1	4	2000
235930	35	35002727	11	3	2	260
245292	35	35005084	19	3	4	425
232668	35	35001925	16	1	2	3000
242755	35	35004371	12	1	2	2129
251774	35	35006757	3	5	4	1111

ANEXO D – Amostra de tamanho 50 - Amazonas

Amazonas						
ID	UF	Número de controle	Número de série	Número de ordem	Sexo	Rendimento mensal domiciliar per capta
14673	13	13000772	12	1	4	162
20767	13	13001930	2	3	2	566
16150	13	13001043	1	3	4	1769
21899	13	13002155	22	2	4	1225
22586	13	13002309	23	4	2	252
11710	13	13000128	12	3	4	866
17614	13	13001370	6	1	2	280
22012	13	13002198	3	2	4	788
17903	13	13001426	8	5	2	535
16737	13	13001167	12	3	4	400
22782	13	13002333	8	3	4	500
16697	13	13001159	12	2	2	400
19424	13	13001701	9	5	4	100
18162	13	13001477	4	3	2	2166
12400	13	13000284	4	1	2	800
22100	13	13002228	1	1	2	684
14174	13	13000683	16	2	4	850
11669	13	13000110	16	1	4	394
15160	13	13000870	1	2	4	925
22754	13	13002333	1	5	4	522
21977	13	13002180	6	2	2	140
19614	13	13001736	17	2	4	50
18980	13	13001620	14	7	2	31
23240	13	13002449	13	4	2	278
19163	13	13001655	14	1	4	166
19803	13	13001779	9	1	2	750
17815	13	13001400	14	1	2	477
18421	13	13001523	11	1	4	156
14692	13	13000772	16	3	2	29
12947	13	13000381	14	2	4	357
22866	13	13002350	2	4	2	105
22130	13	13002228	5	7	4	269
19589	13	13001736	10	6	2	362
20852	13	13001949	13	7	2	85
11460	13	13000080	10	1	2	424

Amazonas						
ID	UF	Número de controle	Número de série	Número de ordem	Sexo	Rendimento mensal domiciliar per capta
16991	13	13001221	7	5	2	557
20400	13	13001876	15	3	2	200
13815	13	13000586	1	1	2	657
15043	13	13000845	9	2	4	11300
14000	13	13000632	4	3	4	300
12894	13	13000381	1	1	2	806
16221	13	13001051	7	2	4	525
16212	13	13001051	4	1	2	2040
15665	13	13000950	12	2	2	197
13011	13	13000390	11	1	2	1433
12836	13	13000373	14	2	2	597
14017	13	13000632	16	5	4	131
16849	13	13001191	9	4	4	225
14415	13	13000730	6	1	4	788
21598	13	13002090	8	1	2	2000

ANEXO E – Amostra de tamanho 50 - São Paulo

São Paulo						
ID	UF	Número de controle	Número de série	Número de ordem	Sexo	Rendimento mensal domiciliar per capta
227930	35	35000716	3	1	4	817
257877	35	35008270	11	3	4	7900
262879	35	35009675	17	1	2	1666
235984	35	35002743	4	2	2	1192
231915	35	35001739	2	2	4	460
226560	35	35000406	20	3	4	675
232296	35	35001836	3	4	2	912
250659	35	35006420	4	1	2	900
226119	35	35000287	6	1	2	1066
225535	35	35000112	9	2	4	5000
240909	35	35003928	11	2	4	350
257718	35	35008210	14	2	4	1400
240281	35	35003774	10	1	4	788
240454	35	35003812	12	5	4	1260
235805	35	35002697	17	1	2	750
242752	35	35004371	11	1	2	430
243449	35	35004541	1	3	4	575
246696	35	35005432	10	4	2	612
251582	35	35006692	6	4	4	215
229624	35	35001100	14	1	2	925
233836	35	35002190	7	2	2	700
256612	35	35007907	11	1	2	587
229053	35	35000970	6	1	4	1576
253430	35	35007125	5	7	4	469
233816	35	35002182	18	1	4	2000
235930	35	35002727	11	3	2	260
245292	35	35005084	19	3	4	425
232668	35	35001925	16	1	2	3000
242755	35	35004371	12	1	2	2129
251774	35	35006757	3	5	4	1111
234720	35	35002433	15	2	4	1144
260960	35	35009144	2	1	2	394
260532	35	35009020	13	2	2	7000
257725	35	35008229	1	3	4	2553
250336	35	35006358	10	2	4	683

Amazonas						
ID	UF	Número de controle	Número de série	Número de ordem	Sexo	Rendimento mensal domiciliar per capta
262821	35	35009667	14	1	4	306
252762	35	35006960	11	1	4	502
258965	35	35008580	15	5	2	1840
240604	35	35003855	5	1	2	917
240849	35	35003910	6	1	4	394
248605	35	35005920	9	2	2	2500
246292	35	35005335	14	5	2	800
264237	35	35010029	8	1	4	263
233012	35	35001992	23	1	2	560
258988	35	35008598	9	2	4	2333
228946	35	35000937	13	2	2	499
240467	35	35003820	1	2	2	2340
227023	35	35000511	9	2	4	1466
231256	35	35001569	12	2	4	1050
256080	35	35007770	19	1	2	828

ANEXO F – Amostra de tamanho 200 - Amazonas

Amazonas						
ID	UF	Número de controle	Número de série	Número de ordem	Sexo	Rendimento mensal domiciliar per capta
14673	13	13000772	12	1	4	162
20767	13	13001930	2	3	2	566
16150	13	13001043	1	3	4	1769
21899	13	13002155	22	2	4	1225
22586	13	13002309	23	4	2	252
11710	13	13000128	12	3	4	866
17614	13	13001370	6	1	2	280
22012	13	13002198	3	2	4	788
17903	13	13001426	8	5	2	535
16737	13	13001167	12	3	4	400
22782	13	13002333	8	3	4	500
16697	13	13001159	12	2	2	400
19424	13	13001701	9	5	4	100
18162	13	13001477	4	3	2	2166
12400	13	13000284	4	1	2	800
22100	13	13002228	1	1	2	684
14174	13	13000683	16	2	4	850
11669	13	13000110	16	1	4	394
15160	13	13000870	1	2	4	925
22754	13	13002333	1	5	4	522
21977	13	13002180	6	2	2	140
19614	13	13001736	17	2	4	50
18980	13	13001620	14	7	2	31
23240	13	13002449	13	4	2	278
19163	13	13001655	14	1	4	166
19803	13	13001779	9	1	2	750
17815	13	13001400	14	1	2	477
18421	13	13001523	11	1	4	156
14692	13	13000772	16	3	2	29
12947	13	13000381	14	2	4	357
22866	13	13002350	2	4	2	105
22130	13	13002228	5	7	4	269
19589	13	13001736	10	6	2	362
20852	13	13001949	13	7	2	85
11460	13	13000080	10	1	2	424
16991	13	13001221	7	5	2	557
20400	13	13001876	15	3	2	200
13815	13	13000586	1	1	2	657
15043	13	13000845	9	2	4	11300
14000	13	13000632	4	3	4	300

Amazonas						
ID	UF	Número de controle	Número de série	Número de ordem	Sexo	Rendimento mensal domiciliar per capta
12894	13	13000381	1	1	2	806
16221	13	13001051	7	2	4	525
16212	13	13001051	4	1	2	2040
15665	13	13000950	12	2	2	197
13011	13	13000390	11	1	2	1433
12836	13	13000373	14	2	2	597
14017	13	13000632	16	5	4	131
16849	13	13001191	9	4	4	225
14415	13	13000730	6	1	4	788
21598	13	13002090	8	1	2	2000
11714	13	13000128	13	1	4	121
16564	13	13001132	11	1	2	418
20894	13	13001965	1	5	2	148
12627	13	13000357	2	2	2	1912
18022	13	13001442	5	1	2	302
13696	13	13000551	9	3	4	450
12701	13	13000365	2	4	2	127
20338	13	13001876	3	3	2	333
22043	13	13002201	1	1	2	788
15732	13	13000969	14	3	4	441
19275	13	13001671	12	1	4	266
12304	13	13000268	8	2	4	86
15851	13	13000985	14	1	4	326
14515	13	13000748	17	4	2	300
21082	13	13001990	18	3	2	223
16640	13	13001140	15	2	2	157
21027	13	13001990	4	1	2	107
21055	13	13001990	12	2	4	170
20839	13	13001949	12	5	2	259
16536	13	13001132	2	7	2	330
20352	13	13001876	5	9	4	500
18845	13	13001590	11	6	4	500
19822	13	13001779	16	1	4	300
11173	13	13000012	1	8	4	318
16961	13	13001213	14	1	2	752
13859	13	13000594	1	5	4	613
15801	13	13000985	3	1	4	725
18644	13	13001566	16	3	2	500
15461	13	13000918	22	1	4	1429
12499	13	13000314	1	2	4	242
14143	13	13000683	2	2	4	2000
19310	13	13001680	9	3	2	257
16262	13	13001060	10	1	2	7500
20765	13	13001930	2	1	2	566
12400.1	13	13000284	4	1	2	800

Amazonas						
ID	UF	Número de controle	Número de série	Número de ordem	Sexo	Rendimento mensal domiciliar per capta
16477	13	13001124	1	3	4	872
23129	13	13002430	6	3	4	433
22019	13	13002198	6	3	4	40
21941	13	13002171	3	1	2	128
13281	13	13000454	13	1	2	644
12739	13	13000365	13	4	4	1140
19131	13	13001655	6	1	4	994
15349	13	13000900	1	4	2	271
19175	13	13001663	1	1	4	100
15069	13	13000853	8	5	4	800
13439	13	13000497	9	5	4	133
20695	13	13001922	6	1	2	394
12289	13	13000268	3	4	4	61
16859	13	13001191	12	1	2	750
17407	13	13001329	14	3	4	2000
18491	13	13001531	13	2	2	1429
15219	13	13000888	3	3	4	279
17130	13	13001256	4	6	4	1516
22754.1	13	13002333	1	5	4	522
17062	13	13001248	3	5	4	217
21987	13	13002180	8	6	2	224
22275	13	13002252	13	2	4	788
18596	13	13001558	16	1	2	1800
16172	13	13001043	8	3	4	925
12947.1	13	13000381	14	2	4	357
22525	13	13002309	9	4	2	475
14836	13	13000802	1	3	2	760
11896	13	13000152	1	4	2	394
22673	13	13002317	10	4	2	400
19947	13	13001795	19	3	2	866
12888	13	13000373	25	4	4	400
17877	13	13001426	1	5	2	166
22749	13	13002325	14	1	4	788
18316	13	13001507	13	3	2	483
16096	13	13001027	19	3	2	733
19068	13	13001647	6	3	4	306
15063	13	13000853	7	1	4	400
14914	13	13000810	2	2	4	309
13855	13	13000594	1	1	2	613
15673	13	13000950	14	3	2	1200
23120	13	13002430	4	2	4	275
13032	13	13000390	16	3	2	236
12258	13	13000250	1	3	2	102
12884	13	13000373	24	2	4	1394
19581	13	13001736	5	1	4	666
18726	13	13001582	11	1	4	402

Amazonas						
ID	UF	Número de controle	Número de série	Número de ordem	Sexo	Rendimento mensal domiciliar per capta
22000	13	13002198	1	1	2	430
19370	13	13001698	7	2	2	2600
20144	13	13001833	9	3	2	676
17530	13	13001353	9	3	2	411
19212	13	13001663	12	1	4	463
21167	13	13002015	4	2	2	306
20742	13	13001922	14	5	4	247
23067	13	13002406	10	6	2	53
16531	13	13001132	2	2	2	330
14965	13	13000829	8	2	4	690
16158	13	13001043	4	1	4	533
11291	13	13000063	7	3	4	452
13393	13	13000489	11	2	4	391
21418	13	13002066	13	2	4	788
13994	13	13000624	9	2	4	390
14089	13	13000667	7	2	4	1394
12087	13	13000195	3	1	4	103
14169	13	13000683	13	1	2	700
20085	13	13001825	14	2	4	394
21474	13	13002074	9	4	2	269
17237	13	13001272	10	5	2	500
15898	13	13000993	7	1	2	1150
14178	13	13000691	2	4	2	1057
12498	13	13000314	1	1	2	242
15923	13	13001000	1	2	4	457
18154	13	13001469	16	3	2	360
13821	13	13000586	2	2	2	925
16595	13	13001132	20	3	4	163
13834	13	13000586	6	11	2	261
17294	13	13001299	5	1	2	1617
15486	13	13000926	6	1	2	2192
19093	13	13001647	14	2	2	466
15735	13	13000969	14	6	2	441
15505	13	13000926	14	3	4	350
17684	13	13001388	4	9	2	427
20183	13	13001841	4	4	4	197
13871	13	13000594	5	2	4	3777
16200	13	13001043	16	2	4	825
14411	13	13000730	3	2	4	197
18854	13	13001604	1	2	2	1182
13393.1	13	13000489	11	2	4	391
21668	13	13002112	3	2	4	1250
20257	13	13001850	15	1	2	850
19313	13	13001680	9	6	2	257

Amazonas						
ID	UF	Número de controle	Número de série	Número de ordem	Sexo	Rendimento mensal domiciliar per capta
18711	13	13001582	6	1	2	2500
15705	13	13000969	8	2	2	1166
17634	13	13001370	10	3	2	744
21800	13	13002139	15	1	4	833
18271	13	13001493	10	2	4	262
21382	13	13002066	1	1	4	500
14974	13	13000829	10	2	4	6500
19800	13	13001779	8	1	2	323
14403	13	13000730	2	1	4	156
18424	13	13001523	12	2	2	220
17036	13	13001230	9	3	4	7233
14403.1	13	13000730	2	1	4	156
18066	13	13001450	7	4	2	274
22260	13	13002252	6	8	2	292
22125	13	13002228	5	2	2	269
14513	13	13000748	17	2	4	300
15082	13	13000853	14	2	4	1150
23137	13	13002430	8	3	2	201
18735	13	13001582	12	3	4	100
22549	13	13002309	14	1	4	315
16856	13	13001191	10	4	4	496
16124	13	13001035	9	4	2	829
19205	13	13001663	9	3	2	798
13010	13	13000390	10	5	4	560
18165	13	13001477	7	2	4	61

ANEXO G – Amostra de tamanho 200 - São Paulo

São Paulo						
ID	UF	Número de controle	Número de série	Número de ordem	Sexo	Rendimento mensal domiciliar per capta
227930	35	35000716	3	1	4	817
257877	35	35008270	11	3	4	7900
262879	35	35009675	17	1	2	1666
235984	35	35002743	4	2	2	1192
231915	35	35001739	2	2	4	460
226560	35	35000406	20	3	4	675
232296	35	35001836	3	4	2	912
250659	35	35006420	4	1	2	900
226119	35	35000287	6	1	2	1066
225535	35	35000112	9	2	4	5000
240909	35	35003928	11	2	4	350
257718	35	35008210	14	2	4	1400
240281	35	35003774	10	1	4	788
240454	35	35003812	12	5	4	1260
235805	35	35002697	17	1	2	750
242752	35	35004371	11	1	2	430
243449	35	35004541	1	3	4	575
246696	35	35005432	10	4	2	612
251582	35	35006692	6	4	4	215
229624	35	35001100	14	1	2	925
233836	35	35002190	7	2	2	700
256612	35	35007907	11	1	2	587
229053	35	35000970	6	1	4	1576
253430	35	35007125	5	7	4	469
233816	35	35002182	18	1	4	2000
235930	35	35002727	11	3	2	260
245292	35	35005084	19	3	4	425
232668	35	35001925	16	1	2	3000
242755	35	35004371	12	1	2	2129
251774	35	35006757	3	5	4	1111
234720	35	35002433	15	2	4	1144
260960	35	35009144	2	1	2	394
260532	35	35009020	13	2	2	7000
257725	35	35008229	1	3	4	2553
250336	35	35006358	10	2	4	683
262821	35	35009667	14	1	4	306
252762	35	35006960	11	1	4	502
258965	35	35008580	15	5	2	1840
240604	35	35003855	5	1	2	917
240849	35	35003910	6	1	4	394

São Paulo						
ID	UF	Número de controle	Número de série	Número de ordem	Sexo	Rendimento mensal domiciliar per capta
248605	35	35005920	9	2	2	2500
246292	35	35005335	14	5	2	800
264237	35	35010029	8	1	4	263
233012	35	35001992	23	1	2	560
258988	35	35008598	9	2	4	2333
228946	35	35000937	13	2	2	499
240467	35	35003820	1	2	2	2340
227023	35	35000511	9	2	4	1466
231256	35	35001569	12	2	4	1050
256080	35	35007770	19	1	2	828
240993	35	35003952	2	3	2	1000
239710	35	35003626	16	1	4	1600
242044	35	35004177	18	3	2	75
238032	35	35003200	13	1	2	894
247161	35	35005548	2	1	4	6500
263995	35	35009950	18	1	4	1234
247399	35	35005610	1	1	2	1666
248044	35	35005777	14	3	2	325
250647	35	35006412	16	1	2	466
258170	35	35008334	11	4	4	575
255078	35	35007532	11	1	4	1680
242575	35	35004312	11	2	2	672
264360	35	35010061	18	2	4	2500
247311	35	35005572	12	1	2	716
236958	35	35003006	5	2	4	380
231520	35	35001623	11	2	2	1000
230786	35	35001445	12	1	4	1500
255166	35	35007559	15	1	2	1500
239368	35	35003545	10	4	4	326
255654	35	35007680	18	2	2	1016
237709	35	35003138	2	1	4	693
253623	35	35007176	17	1	2	480
249361	35	35006102	4	3	4	3333
240567	35	35003847	5	1	2	1475
237785	35	35003146	15	3	2	1166
263294	35	35009799	17	1	4	1050
256603	35	35007907	5	5	4	478
253115	35	35007052	12	3	4	575
231444	35	35001615	4	1	4	1300
257785	35	35008245	11	4	2	4250

São Paulo						
ID	UF	Número de controle	Número de série	Número de ordem	Sexo	Rendimento mensal domiciliar per capta
246740	35	35005440	12	3	4	500
260330	35	35008962	12	3	4	984
230063	35	35001232	5	1	2	875
234798	35	35002468	8	2	4	425
253762	35	35007214	11	2	4	5294
235680	35	35002670	8	2	4	0
227389	35	35000600	12	2	4	694
260332	35	35008962	13	1	4	910
240380	35	35003790	17	4	4	612
229202	35	35001011	6	3	4	325
251920	35	35006781	7	1	2	840
231323	35	35001585	5	3	2	1950
231711	35	35001674	10	1	2	465
239905	35	35003677	14	1	2	2666
248089	35	35005793	6	1	2	333
244923	35	35004908	12	2	2	780
255799	35	35007729	3	4	4	645
240110	35	35003731	7	2	2	1116
237609	35	35003111	1	2	4	500
238890	35	35003413	20	3	4	400
244434	35	35004789	12	9	4	179
258855	35	35008555	4	2	4	3000
243355	35	35004517	14	3	2	2000
259624	35	35008750	12	3	2	4975
252020	35	35006790	15	3	2	1000
254178	35	35007303	14	3	2	562
264263	35	35010029	15	3	2	380
232921	35	35001984	2	2	2	1063
231021	35	35001496	9	1	2	293
245037	35	35004959	2	2	2	2350
255841	35	35007729	13	3	2	333
236787	35	35002956	12	3	4	4666
253252	35	35007087	5	1	2	833
228479	35	35000830	5	3	4	2196
228921	35	35000937	4	2	4	1000
248514	35	35005904	10	1	2	866
241279	35	35004002	11	1	4	4400
251012	35	35006528	5	3	2	991
249941	35	35006269	12	2	4	1166
262183	35	35009519	13	1	4	953

São Paulo						
ID	UF	Número de controle	Número de série	Número de ordem	Sexo	Rendimento mensal domiciliar per capta
245176	35	35005050	6	5	2	566
246267	35	35005335	7	7	2	613
229184	35	35001011	1	2	4	390
239931	35	35003693	2	2	4	1313
241999	35	35004169	14	3	2	525
245673	35	35005173	16	3	2	4833
260333	35	35008962	14	1	2	517
263283	35	35009799	13	3	2	1197
244743	35	35004851	3	1	2	1356
250877	35	35006480	7	2	4	733
257770	35	35008237	14	1	4	5750
235580	35	35002654	4	3	4	600
242378	35	35004240	20	1	4	966
255408	35	35007613	5	2	4	6666
259903	35	35008849	11	1	2	1727
230606	35	35001399	6	3	2	750
236947	35	35002999	15	2	4	1533
241699	35	35004088	16	2	4	1400
231845	35	35001712	6	2	4	1646
253711	35	35007206	9	2	4	300
250851	35	35006471	10	2	4	11500
237058	35	35003014	9	2	2	253
231633	35	35001658	9	1	2	800
236456	35	35002867	12	4	4	197
257774	35	35008237	16	3	4	833
245502	35	35005149	10	6	4	744
247133	35	35005530	12	3	2	712
235511	35	35002638	13	1	2	1147
259190	35	35008652	3	3	4	1201
239746	35	35003634	1	4	4	833
239891	35	35003677	9	2	4	4833
257922	35	35008288	12	1	4	3000
251958	35	35006781	21	1	2	3300
227127	35	35000546	12	1	2	7100
234830	35	35002476	1	3	4	1060
233867	35	35002190	14	2	4	445
241441	35	35004037	15	1	4	970
265163	35	35010266	15	1	4	2655
231418	35	35001607	10	3	2	882
234750	35	35002441	13	3	4	562

ANEXO H – Programação em R

H.1 Distribuição Binomial

```
library(MCMCpack)
library("TeachingDemos")
library(ggplot2)
library(tidyr)

#######################################
### Inferencia para uma populacao ###
#######################################

#### gerando uma amostra
m<-10
p<-0.3
n<-30
set.seed(123)
x<-rbinom(n,m,p)

#### hiperparametros da priori de theta
a<-1
b<-1

#### parametros da posteriori de theta|x
A<-a+sum(x)
B<-b+sum(m-x)

#### obtendo a densidade a posteriori de theta
theta<-rbeta(1000000,A,B)
theta_data=data.frame(theta)

ggplot(theta_data, aes(theta)) +
  geom_density( size=1)+
  coord_cartesian(xlim = c(0.25, 0.45),ylim = c(0,15)) +
  ylab(expression(paste("h(",theta,"|x)")))+
  xlab(expression(theta))+
  theme_bw()

### obtendo a densidade a posteriori de fi=CV
```

```
37 fi<-((1-theta)/(m*theta))
38 fi_data=data.frame(fi)
39
40 ggplot(fi_data, aes(fi)) +
41   geom_density( size=1)+
42   coord_cartesian(xlim = c(0.1, 0.3),ylim = c(0,18)) +
43   ylab(expression(paste("h(",phi,"|x)")))+
44   xlab(expression(phi))+
45   theme_bw()
46
47 #### estimativa pontual e intervalar de theta
48 mean(theta)
49 emp.hpd(theta,conf=0.95)
50
51 #### estimativa pontual e intervalar de fi
52 mean(fi)
53 emp.hpd(fi,conf=0.95)
54
55 ###Fator de bayes
56 set.seed(123)
57 priori=rbeta(1000000,a,b)
58 fi_priori<-((1-priori)/(m*priori))
59 fb=(mean(fi>0.3)/mean(fi<0.3))/mean(fi_priori>0.3)/mean(fi_priori<0.3)
60 ev=fb^(-1)
61
62 ########################################
63 ### Inferencia para duas populacoes ###
64 ########################################
65
66 #### gerando amostras
67 m<-10
68 p1<-0.3
69 n1<-10
70 p2<-0.6
71 n2<-15
72 set.seed(123)
73 x1<-rbinom(n1,m,p1)
74 set.seed(123)
75 x2<-rbinom(n2,m,p2)
76
77 #### hiperparametros da priori de theta1 e theta2
78 a1<-1
```

```
79  b1<-1
80  a2<-1
81  b2<-1
82
83  #### parametros da posteriori de theta1|x1 theta2|x2
84  A1<-a1+sum(x1)
85  B1<-b1+sum(m-x1)
86  A2<-a2+sum(x2)
87  B2<-b2+sum(m-x2)
88
89  #### obtendo a densidade a posteriori de theta=(theta1,theta2)|x
90  theta1<-rbeta(1000000,A1,B1)
91  theta2<-rbeta(1000000,A2,B2)
92  theta1=sort(theta1)
93  theta2=sort(theta2)
94  theta_data=data.frame(theta1, theta2)
95  theta_data1=gather(theta_data,"tipo","valores",1:2)
96
97  legend_title=""
98  ggplot(data=theta_data1, aes(x=valores))+
99    geom_density(aes(color=tipo), show.legend=F, size=1) +
100   stat_density(aes(x=valores, color=tipo), geom="line",position="identity",
      size=1)+
101   xlim(0.1, 0.75)+ylim(0,10) +
102   ylab(expression(paste("h(",theta,"|x)")))+
103   xlab(expression(theta))+
104   theme_bw()+
105   scale_colour_manual(legend_title,values=c("blue", "black"),
106                       labels=c("Pop1", "Pop2"))
107
108 ### obtendo a densidade a posteriori e estimativas de fi
109 fi1<-((1-theta1)/(m*theta1))
110 fi2<-((1-theta2)/(m*theta2))
111 fi_data=data.frame(fi1,fi2)
112 fi_data1=gather(fi_data,"tipo","valores",1:2)
113
114 legend_title=''
115 ggplot(data=fi_data1, aes(x=valores))+
116   geom_density(aes(color=tipo), show.legend=F, size=1) +
117   stat_density(aes(x=valores, color=tipo), geom="line",position="identity",
      size=1)+
118   coord_cartesian(xlim = c(0,0.4),ylim = c(0,30)) +
```

```
119   ylab(expression(paste("h(",phi,"|x)")))+
120   xlab(expression(phi))+
121   theme_bw()+
122   scale_colour_manual(legend_title ,values=c("blue", "black"),
123                       labels=c("Pop1", "Pop2"))
124
125 ### estimativa pontual e intervalar de theta1 e theta2
126 mean(theta1)
127 emp.hpd(theta1 ,conf=0.95)
128 mean(theta2)
129 emp.hpd(theta2 ,conf=0.95)
130
131 mean(fi1)
132 emp.hpd(fi1 ,conf=0.95)
133 mean(fi2)
134 emp.hpd(fi2 ,conf=0.95)
135
136 #### FBST - H:theta1=theta2
137 ### max da posteriori sob H0
138 theta.max<-(A1+A2-2)/(A1+A2+B1+B2-4)
139 post.h0<-dbeta(theta.max,A1,B1)*dbeta(theta.max,A2,B2)
140 post<-dbeta(theta1 ,A1,B1)*dbeta(theta2 ,A2,B2)
141 mean(post<post.h0)     ### Valor e
142
143 ### figura da regiao tangente do FBST
144 theta1<-rep(seq(.001,1,0.001),1000)
145 theta2 = rep(1:1000/1000,each= 1000)
146
147 post.h0<-dbeta(theta.max,A1,B1)*dbeta(theta.max,A2,B2)
148 post<-dbeta(theta1 ,A1,B1)*dbeta(theta2 ,A2,B2)
149
150 tangente<-which(post>post.h0&post<(post.h0+0.1))
151 ds = data.frame(t1=theta1[tangente],t2=theta2[tangente])
152 ggplot(ds,aes(t1 ,t2))+geom_point()+
153   coord_cartesian(xlim = c(0, 1),ylim = c(0,1)) +
154   geom_abline(intercept =  0, slope =  1, size = 1,col = 'darkgray')+
155   annotate('point',x=theta.max,y=theta.max,color='blue', size=3)+
156   annotate("text", x = .45, y = 0.3, label = "Max. posteriori \n sob H0",
      color="blue")+
157   annotate("text", x = 0.77, y = 0.6, label = "H0", color="darkgray")+
158   ylab(expression(paste(theta[2])))+
159   xlab(expression(paste(theta[1])))+
```

```
160   theme_bw()
```

H.2 Distribuição Binomial Negativa

```
1  library(MCMCpack)
2  library("TeachingDemos")
3  library(ggplot2)
4  library(tidyr)
5
6  ######################################
7  ### Inferencia para uma populacao ###
8  ######################################
9
10 #### gerando amostra
11 set.seed(123)
12 m<-10
13 p<-0.4
14 n<-30
15 x<-rnbinom(n,m,p)
16
17 #### hiperparametros da priori de theta
18 a<-2
19 b<-1
20
21 #### parametros da posteriori de theta|x
22 A<-a+n*m
23 B<-b+sum(x)
24
25 #### obtendo a densidade a posteriori de theta
26 theta<-rbeta(1000000,A,B)
27 theta_data=data.frame(theta)
28
29 ggplot(theta_data, aes(theta)) +
30   geom_density( size=1)+
31   coord_cartesian(xlim = c(0.35, 0.5),ylim = c(0,25)) +
32   ylab(expression(paste("h(",theta,"|x)")))+
33   xlab(expression(theta))+
34   theme_bw()
35
36 ### obtendo a densidade a posteriori de fi=CV
37 fi<-1/(sqrt(m*(1-theta)))
38 fi_data=data.frame(fi)
```

```
39
40 ggplot(fi_data, aes(fi)) +
41   geom_density( size=1)+
42   coord_cartesian(xlim = c(0.4, 0.45),ylim = c(0,60)) +
43   ylab(expression(paste("h(",phi,"|x)")))+
44   xlab(expression(phi))+
45   theme_bw()
46
47 #### estimativa pontual e intervalar de theta
48 mean(theta)
49 emp.hpd(theta,conf=0.95)
50
51 #### estimativa pontual e intervalar de fi
52 mean(fi)
53 emp.hpd(fi,conf=0.95)
54
55 ###Fator de bayes
56 set.seed(135)
57 priori=rbeta(1000000,a,b)
58 fi_priori<-((1-priori)/(m*priori))
59 fb=(mean(fi>0.45)/mean(fi<0.45))/mean(fi_priori>0.45)/mean(fi_priori<0.45)
60 ev=fb^(-1)
61
62 ########################################
63 ### Inferencia para duas populacoes ###
64 ########################################
65
66 #### gerando amostras
67 m<-10
68 p1<-0.9
69 n1<-10
70 p2<-0.7
71 n2<-15
72 set.seed(123)
73 x1<-rnbinom(n1,m,p1)
74 set.seed(123)
75 x2<-rnbinom(n2,m,p2)
76
77 #### hiperparametros da priori de theta1 e theta2
78 a1<-3
79 b1<-2
80 a2<-3
```

```
81  b2<-2
82
83  #### parametros da posteriori de theta1|x1 theta2|x2
84  A1<-a1+n1*m
85  B1<-b1+sum(x1)
86  A2<-a2+n2*m
87  B2<-b2+sum(x2)
88
89  #### obtendo a densidade a posteriori de theta=(theta1,theta2)|x
90  theta1<-rbeta(1000000,A1,B1)
91  theta2<-rbeta(1000000,A2,B2)
92  theta1=sort(theta1)
93  theta2=sort(theta2)
94  theta_data=data.frame(theta1, theta2)
95  theta_data1=gather(theta_data,"tipo","valores",1:2)
96
97
98  legend_title=""
99  ggplot(data=theta_data1, aes(x=valores))+
100   geom_density(aes(color=tipo), show.legend=F, size=1) +
101   stat_density(aes(x=valores, color=tipo), geom="line",position="identity",
        size=1)+
102   coord_cartesian(xlim = c(0.5,1),ylim = c(0,15)) +
103   ylab(expression(paste("h(",theta,"|x)")))+
104   xlab(expression(theta))+
105   theme_bw()+
106   scale_colour_manual(legend_title,values=c("blue", "black"),
107                       labels=c("Pop1", "Pop2"))
108
109 ### obtendo a densidade a posteriori e estimativas de fi
110 fi1<-1/(sqrt(m*(1-theta1)))
111 fi2<-1/(sqrt(m*(1-theta2)))
112 fi_data=data.frame(fi1,fi2)
113 fi_data1=gather(fi_data,"tipo","valores",1:2)
114
115 legend_title=""
116 ggplot(data=fi_data1, aes(x=valores))+
117   geom_density(aes(color=tipo), show.legend=F, size=1) +
118   stat_density(aes(x=valores, color=tipo), geom="line",position="identity",
        size=1)+
119   coord_cartesian(xlim = c(0.4,1.1),ylim = c(0,16)) +
120   ylab(expression(paste("h(",phi,"|x)")))+
```

```
121     xlab(expression(phi))+
122     theme_bw()+
123     scale_colour_manual(legend_title ,values=c("blue", "black"),
124                         labels=c("Pop1", "Pop2"))
125
126 ### estimativa pontual e intervalar de theta1 e theta2
127 mean(theta1)
128 emp.hpd(theta1 ,conf=0.95)
129 mean(theta2)
130 emp.hpd(theta2 ,conf=0.95)
131
132 mean(fi1)
133 emp.hpd(fi1 ,conf=0.95)
134 mean(fi2)
135 emp.hpd(fi2 ,conf=0.95)
136
137 #### FBST - H:theta1=theta2
138 ### max da posteriori sob H0
139 theta.max<-(A1+A2-2)/(A1+A2+B1+B2-4)
140 post.h0<-dbeta(theta.max,A1,B1)*dbeta(theta.max,A2,B2)
141 post<-dbeta(theta1 ,A1,B1)*dbeta(theta2 ,A2,B2)
142
143 mean(post<post.h0)      ### Valor e
144
145 ### figura da regiao tangente do FBST
146 theta1<-rep(seq(.001,1,0.001),1000)
147 theta2 = rep(1:1000/1000,each= 1000)
148
149 post.h0<-dbeta(theta.max,A1,B1)*dbeta(theta.max,A2,B2)
150 post<-dbeta(theta1 ,A1,B1)*dbeta(theta2 ,A2,B2)
151
152 tangente<-which(post>post.h0&post<(post.h0+0.1))
153 ds = data.frame(t1=theta1[tangente],t2=theta2[tangente])
154 ggplot(ds,aes(t1 ,t2))+geom_point()+
155     coord_cartesian(xlim = c(0, 1),ylim = c(0,1)) +
156     geom_abline(intercept =  0, slope =  1, size = 1,col = 'darkgray')+
157     annotate('point',x=theta.max,y=theta.max,color='blue', size=3)+
158     annotate("text", x = .6, y = 0.77, label = "Max. posteriori \n sob H0",
        color="blue")+
159     annotate("text", x = 0.5, y = 0.4, label = "H0", color="darkgray")+
160     ylab(expression(paste(theta[2])))+
161     xlab(expression(paste(theta[1])))+
```

```
162   theme_bw()
```

H.3 Distribuição Poisson

```
1  library(MCMCpack)
2  library("TeachingDemos")
3  library(ggplot2)
4  library(tidyr)
5
6  ######################################
7  ### Inferencia para uma populacao ###
8  ######################################
9
10 set.seed(123)
11 lambda<-0.7
12 n<-20
13 x<-rpois(n,lambda)
14
15
16 x=c(0, 1, 0, 2, 2, 0, 1, 2, 1, 0, 2, 0, 1, 1, 0, 2, 0, 0, 0, 2)
17
18 #### hiperparametros da priori de theta
19 a<-4
20 b<-2
21
22 #### parametros da posteriori de theta|x
23 A<-a+sum(x)
24 B<-b+n
25
26 #### obtendo a densidade a posteriori de theta
27 theta<-rgamma(1000000,A,B)
28
29 theta_data=data.frame(theta)
30 ggplot(theta_data, aes(theta)) +
31   geom_density( size=1)+
32   coord_cartesian(xlim = c(0.3, 2),ylim = c(0,2)) +
33   ylab(expression(paste("h(",theta,"|x)")))+
34   xlab(expression(theta))+
35   theme_bw()
36
37 ## obtendo a densidade a posteriori de fi=CV
38 fi<-1/(sqrt(theta))
```

```
39 fi_data=data.frame(fi)
40
41 ggplot(fi_data, aes(fi)) +
42   geom_density( size=1)+
43   coord_cartesian(xlim = c(0.6, 1.6),ylim = c(0,3.5)) +
44   ylab(expression(paste("h(",phi,"|x)")))+
45   xlab(expression(phi))+
46   theme_bw()
47
48 #### estimativa pontual e intervalar de theta
49 mean(theta)
50 emp.hpd(theta,conf=0.95)
51
52 #### estimativa pontual e intervalar de fi
53 mean(fi)
54 emp.hpd(fi,conf=0.95)
55
56 ###Fator de bayes
57 set.seed(135)
58 priori=rgamma(1000000,a,b)
59 fi_priori<-1/(sqrt(priori))
60 fb=(mean(fi<0.8)/mean(fi>0.8))/(mean(fi_priori<0.8)/mean(fi_priori>0.8))
61 ev=fb^(-1)
62
63 ########################################
64 ### Inferencia para duas populacoes ###
65 ########################################
66
67 #### gerando amostras
68 lambda1<-0.6
69 n1<-13
70 lambda2<-0.8
71 n2<-22
72 set.seed(123)
73 x1<-rpois(n1,lambda1)
74 set.seed(123)
75 x2<-rpois(n2,lambda2)
76
77 #### hiperparametros da priori de theta1 e theta2
78 a1<-3
79 b1<-2
80 a2<-3
```

```
81  b2<-2
82
83  #### parametros da posteriori de theta1|x1 theta2|x2
84  A1<-a1+sum(x1)
85  B1<-b1+n1
86  A2<-a2+sum(x2)
87  B2<-b2+n2
88
89  #### obtendo a densidade a posteriori de theta=(theta1,theta2)|x
90  theta1<-as.numeric(rgamma(1000000,A1,B1))
91  theta2<-as.numeric(rgamma(1000000,A2,B2))
92  theta1=sort(theta1)
93  theta2=sort(theta2)
94  theta_data=data.frame(theta1, theta2)
95  theta_data1=gather(theta_data,"tipo","valores",1:2)
96
97  legend_title=""
98  ggplot(data=theta_data1, aes(x=valores))+
99    geom_density(aes(color=tipo), show.legend=F, size=1) +
100   stat_density(aes(x=valores, color=tipo), geom="line",position="identity",
      size=1)+
101   xlim(0, 2.5)+ylim(0,2) +
102   ylab(expression(paste("h(",theta,"|x)")))+
103   xlab(expression(theta))+
104   theme_bw()+
105   scale_colour_manual(legend_title,values=c("blue", "black"),
106                       labels=c("Pop1", "Pop2"))
107
108 ### obtendo a densidade a posteriori e estimativas de fi
109 fi1<-1/(sqrt(theta1))
110 fi2<-1/(sqrt(theta2))
111 fi_data=data.frame(fi1,fi2)
112 fi_data1=gather(fi_data,"tipo","valores",1:2)
113
114 legend_title=""
115 ggplot(data=fi_data1, aes(x=valores))+
116   geom_density(aes(color=tipo), show.legend=F, size=1) +
117   stat_density(aes(x=valores, color=tipo), geom="line",position="identity",
      size=1)+
118   coord_cartesian(xlim = c(0.5, 2),ylim = c(0,5)) +
119   ylab(expression(paste("h(",phi,"|x)")))+
120   xlab(expression(phi))+
```

```
121   theme_bw()+
122   scale_colour_manual(legend_title ,values=c("blue", "black"),
123                       labels=c("Pop1", "Pop2"))
124
125 ### estimativa pontual e intervalar de theta1 e theta2
126 mean(theta1)
127 emp.hpd(theta1 ,conf=0.95)
128 mean(theta2)
129 emp.hpd(theta2 ,conf=0.95)
130
131 mean(fi1)
132 emp.hpd(fi1 ,conf=0.95)
133 mean(fi2)
134 emp.hpd(fi2 ,conf=0.95)
135
136 #### FBST - H:theta1=theta2
137 ### max da posteriori sob H0
138 theta.max<-(A1+A2-2)/(B1+B2)
139 post.h0<-dgamma(theta.max,A1,B1)*dgamma(theta.max,A2,B2)
140 post<-dgamma(theta1 ,A1,B1)*dgamma(theta2 ,A2,B2)
141
142 mean(post<post.h0)     ### Valor e
143
144 ### figura da regiao tangente do FBST
145 theta1<-rep(seq(.001,3,0.001),1000)
146 theta2 = rep(1:3000/1000,each= 1000)
147
148 post.h0<-dgamma(theta.max,A1,B1)*dgamma(theta.max,A2,B2)
149 post<-dgamma(theta1 ,A1,B1)*dgamma(theta2 ,A2,B2)
150
151 tangente<-which(post>post.h0&post<(post.h0+0.1))
152 ds = data.frame(t1=theta1[tangente],t2=theta2[tangente])
153 ggplot(ds,aes(t1 ,t2))+geom_point()+
154   coord_cartesian(xlim = c(0, 1.5),ylim = c(0,1.5)) +
155   geom_abline(intercept =  0, slope =  1, size = 1,col = 'darkgray')+
156   annotate('point',x=theta.max,y=theta.max,color='blue', size=3)+
157   annotate("text", x = .6, y = 0.77, label = "Max. posteriori \n sob H0",
        color="blue")+
158   annotate("text", x = 0.5, y = 0.4, label = "H0", color="darkgray")+
159   ylab(expression(paste(theta[2])))+
160   xlab(expression(paste(theta[1])))+
161   theme_bw()
```

H.4 Distribuição Gama

```
library(MCMCpack)
library("TeachingDemos")
library(ggplot2)
library(tidyr)

###################
### Uma gama  ###
###################

### gerando amostra
set.seed(123)
x1<-rgamma(15,2,1)

###definindo o log da posteriori
log.post<-function(par,x1){
  alpha1<-par[1]
  beta1<-par[2]
  n1<-length(x1)
  X1<-sum(x1)
  LX1<-sum(log(x1))
  if ((alpha1>0)&&(beta1>0)) return ( alpha1*n1*log(beta1) - n1*log(gamma(
    alpha1)) + (alpha1-1)*LX1 -
      beta1*X1 + dgamma(alpha1,1/1000,1/1000,log=T) + dgamma(beta1,1/1000,1/
    1000,log=T))
  else return(-Inf)
}

###gerando valores da posterior
chute<-c(1,1)
theta<-MCMCmetrop1R(log.post,chute,x1=x1,mcmc=100000,logfun=T)

###obtendo as estimativas dos parametros
mean(theta[,1])
mean(theta[,2])

fi<-(1/theta[,1])^.5
fi_data=data.frame(fi)

ggplot(fi_data, aes(var1)) +
  geom_density( size=1)+
```

```
39    coord_cartesian(xlim = c(0.5, 1.6),ylim = c(0,3)) +
40    ylab(expression(paste("h(",phi,"|x)")))+
41    xlab(expression(phi))+
42    theme_bw()
43
44  #### estimativa pontual e intervalar de alfa, beta e fi
45  #alfa
46  mean(theta[,1])
47  emp.hpd(theta[,1],conf=0.95)
48
49  #beta
50  mean(theta[,2])
51  emp.hpd(theta[,2],conf=0.95)
52
53  #fi=cv
54  mean(fi)
55  emp.hpd(fi,conf=0.95)
56
57  ###################
58  ### Duas gamas ###
59  ###################
60
61  ### gerando amostras
62  set.seed(123)
63  x1<-rgamma(12,2,1)
64  set.seed(123)
65  x2<-rgamma(15,3,2)
66
67  ###definindo o log da posteriori
68  log.post<-function(par,x1,x2){
69    alpha1<-par[1]
70    beta1<-par[2]
71    alpha2<-par[3]
72    beta2<-par[4]
73    n1<-length(x1)
74    X1<-sum(x1)
75    LX1<-sum(log(x1))
76    n2<-length(x2)
77    X2<-sum(x2)
78    LX2<-sum(log(x2))
79    if ((alpha1>0)&&(beta1>0)&&(alpha2>0)&&(beta2>0))
80      return (
```

```
81              alpha1*n1*log(beta1) - n1*log(gamma(alpha1))
82            + (alpha1-1)*LX1 - beta1*X1 + dgamma(alpha1,1/1000,1/1000,log=T)
83            + dgamma(beta1,1/1000,1/1000,log=T) +
84
85            alpha2*n2*log(beta2) - n2*log(gamma(alpha2))
86            + (alpha2-1)*LX2 - beta2*X2  + dgamma(alpha2,1/1000,1/1000,log=T)
87            + dgamma(beta2,1/1000,1/1000,log=T))
88    else return(-Inf)
89  }
90
91  ###gerando valores da posteriori
92  chute<-c(1,1,1,1)
93  theta<-MCMCmetrop1R(log.post,chute,x1=x1,x2=x2,mcmc=100000,logfun=T)
94
95  ###obtendo as estimativas dos parametros
96  mean(theta[,1])
97  emp.hpd(theta[,1],conf=0.95)
98  mean(theta[,2])
99  emp.hpd(theta[,2],conf=0.95)
100 mean(theta[,3])
101 emp.hpd(theta[,3],conf=0.95)
102 mean(theta[,4])
103 emp.hpd(theta[,4],conf=0.95)
104
105 fi1<-(1/theta[,1])^.5    ### CV=raiz(alpha)
106 fi2<-(1/theta[,3])^.5    ### CV=raiz(alpha)
107 fi1=sort(fi1)
108 fi2=sort(fi2)
109 fi_data=data.frame(fi1,fi2)
110 fi_data1=gather(fi_data,"tipo","valores",1:2)
111
112 legend_title=""
113 ggplot(data=fi_data1, aes(x=valores))+
114   geom_density(aes(color=tipo), show.legend=F, size=1) +
115   stat_density(aes(x=valores, color=tipo), geom="line",position="identity",
      size=1)+
116   coord_cartesian(xlim = c(0,2),ylim = c(0,3.5)) +
117   ylab(expression(paste("h(",phi,"|x)")))+
118   xlab(expression(phi))+
119   theme_bw()+
120   scale_colour_manual(legend_title,values=c("blue", "black"),
121                       labels=c("Pop1", "Pop2"))
```

```
122
123 mean(fi1)
124 emp.hpd(fi1,conf=0.95)
125 mean(fi2)
126 emp.hpd(fi2,conf=0.95)
127
128 log.post.h0<-function(par,x1,x2){
129   alpha1<-par[1]
130   beta1<-par[2]
131   beta2<-par[3]
132   n1<-length(x1)
133   X1<-sum(x1)
134   LX1<-sum(log(x1))
135   n2<-length(x2)
136   X2<-sum(x2)
137   LX2<-sum(log(x2))
138   if ((alpha1>0)&&(beta1>0)&&(beta2>0))
139     return (-1*( alpha1*n1*log(beta1) - n1*log(gamma(alpha1)) +
140           (alpha1-1)*LX1 - beta1*X1 + dgamma(alpha1,1/1000,1/1000,log=T) +
141            dgamma(beta1,1/1000,1/1000,log=T)+alpha1*n2*log(beta2) -
142            n2*log(gamma(alpha1)) + (alpha1-1)*LX2 - beta2*X2 +
143            dgamma(alpha1,1/1000,1/1000,log=T) + dgamma(beta2,1/1000,1/1000,
    log=T)))
144   else return(-Inf)
145 }
146
147 chute=c(1,1,1)
148 a=optim(chute, log.post.h0, x1=x1, x2=x2)
149
150 max.post.h0=-a$value
151
152 valor.post= numeric()
153 for (i in 1:100000){
154   valor.post[i] = log.post(theta[i,], x1,x2)
155 }
156
157 valor.e=mean(valor.post<max.post.h0)
```

H.5 Distribuição Log- Normal

```
1 install.packages('tidyr')
2 install.packages("MCMCpack")
```

```
3  library(MCMCpack)
4  library(tidyr)
5  library(ggplot2)
6  library("TeachingDemos")
7
8  ###################
9  ## Uma log normal##
10 ###################
11
12 ###gerando a amostra para ilustrar o procedimento
13 set.seed(123)
14 x1<-rlnorm(10,2,.6)
15
16 ### definindo o log da posteriori
17 log.post<-function(par,x1){
18   mu1<-par[1]
19   sigma1<-par[2]
20   n1<-length(x1)
21   X1<-sum(x1)
22   LX1<-sum(log(x1))
23   if ((sigma1>0)) return ((-LX1 - n1*log(sigma1) -(n1/2)*log(2*pi) - 1/(2*
       sigma1^2)
24                           *(sum(log(x1)^2) -2*mu1*LX1 + n1*mu1^2)) +
25                            dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu1
       ,2,1,log=T))
26   else return(-Inf)
27 }
28
29 ### gerando valores da posteriori
30 chute<-c(1,1)
31 theta<-MCMCmetrop1R(log.post,chute,x1=x1,mcmc=100000,logfun=T)
32
33 fi<-sqrt(exp(theta[,2]^2)-1)
34 fi_data=data.frame(fi)
35 ggplot(fi_data, aes(var1)) +
36   geom_density( size=1)+
37   coord_cartesian(xlim = c(0.3, 1.6),ylim = c(0,3)) +
38   ylab(expression(paste("h(",phi,"|x)")))+
39   xlab(expression(phi))+
40   theme_bw()
41
42 #### estimativa pontual e intervalar de alfa, beta e fi
```

```
43 #mu
44 mean(theta[,1])
45 emp.hpd(theta[,1],conf=0.95)
46
47 #sigma
48 mean(theta[,2])
49 emp.hpd(theta[,2],conf=0.95)
50
51 #fi=cv
52 mean(fi)
53 emp.hpd(fi,conf=0.95)
54
55 ###################
56 ### Duas log normais ###
57 ###################
58
59 ### gerando a amostra para ilustrar o procedimento
60 set.seed(123)
61 x1<-rgamma(12,2,0.4)
62 set.seed(123)
63 x2<-rgamma(15,3,.7)
64
65 ### definindo o log da posteriori
66 log.post<-function(par,x1, x2){
67   mu1<-par[1]
68   sigma1<-par[2]
69   mu2<-par[3]
70   sigma2<-par[4]
71   n1<-length(x1)
72   n2<- length(x2)
73   X1<-sum(x1)
74   LX1<-sum(log(x1))
75   X2<-sum(x2)
76   LX2<-sum(log(x2))
77   if ((sigma1>0)) return (
78
79       ((-LX1 - n1*log(sigma1) -(n1/2)*log(2*pi) - 1/(2*sigma1^2)
80     * (sum(log(x1)^2) -2*mu1*LX1 + n1*mu1^2))
81     + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu1,2,1,log=T))
82
83     + ((-LX2 - n2*log(sigma2) -(n2/2)*log(2*pi) - 1/(2*sigma2^2)
84     * (sum(log(x2)^2) -2*mu2*LX2 + n2*mu2^2))
```

```
85      + dgamma(sigma2,1/1000,1/1000,log=T) + dnorm(mu2,4,0.6,log=T))
86      )
87    else return(-Inf)
88  }
89
90  ### gerando valores da posteriori
91  chute<-c(1,1,1,1)
92  theta<-MCMCmetrop1R(log.post,chute,x1=x1,x2=x2,mcmc=100000,logfun=T)
93
94  fi1<-sqrt(exp(theta[,2]^2)-1)
95  fi2<-sqrt(exp(theta[,4]^2)-1)
96  fi_data=data.frame(fi1,fi2)
97  fi_data1=gather(fi_data,"tipo","valores",1:2)
98  legend_title=''
99  ggplot(data=fi_data1, aes(x=valores))+
100   geom_density(aes(color=tipo), show.legend=F, size=1) +
101   stat_density(aes(x=valores, color=tipo), geom="line",position="identity",
        size=1)+
102   xlim(0, 10)+ylim(0,1.5) +
103   ylab(expression(paste("h(",phi,"|x)")))+
104   xlab(expression(phi))+
105   theme_bw()+
106   scale_colour_manual(legend_title,values=c("blue", "black"),
107                       labels=c("Pop1", "Pop2"))
108
109
110 ### obtendo as estimativas dos parametros ###
111 mean(theta[,1])
112 emp.hpd(theta[,1],conf=0.95)
113
114 mean(theta[,2])
115 emp.hpd(theta[,2],conf=0.95)
116
117 mean(theta[,3])
118 emp.hpd(theta[,3],conf=0.95)
119
120 mean(theta[,4])
121 emp.hpd(theta[,4],conf=0.95)
122
123 mean(fi1)
124 emp.hpd(fi1,conf=0.95)
125 mean(fi2)
```

```
126 emp.hpd(fi2 ,conf=0.95)
127
128
129 log.post.h0<-function(par,x1, x2){
130   mu1<-par[1]
131   sigma1<-par[2]
132   mu2<-par[3]
133   n1<-length(x1)
134   n2<- length(x2)
135   X1<-sum(x1)
136   LX1<-sum(log(x1))
137   X2<-sum(x2)
138   LX2<-sum(log(x2))
139   if ((sigma1>0)) return (-1*(
140
141       ((-LX1 - n1*log(sigma1) -(n1/2)*log(2*pi) - 1/(2*sigma1^2)
142       * (sum(log(x1)^2) -2*mu1*LX1 + n1*mu1^2))
143       + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu1,2,1,log=T))
144
145     + ((-LX2 - n2*log(sigma1) -(n2/2)*log(2*pi) - 1/(2*sigma1^2)
146       * (sum(log(x2)^2) -2*mu2*LX2 + n2*mu2^2))
147       + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu2,4,0.6,log=T))
148   ))
149   else return(-Inf)
150 }
151
152 chute=c(1,1,1)
153 a=optim(chute, log.post.h0, x1=x1, x2=x2)
154
155 max.post.h0=-a$value
156
157 valor.post= numeric()
158 for (i in 1:100000){
159   valor.post[i] = log.post(theta[i,], x1,x2)
160 }
161
162 valor.e=mean(valor.post<max.post.h0)
```

H.6 PNAD- Comparação por SP e AM com amostras

```
1 library(MCMCpack)
2 library("TeachingDemos")
```

```
3  library(ggplot2)
4  library(tidyr)
5
6
7  data = read.table("../dados.txt")
8  dados= data[which(is.na(data$V4742)==F),]
9  dados$renda=as.numeric(dados$V4742)
10 dados=dados[,-6]
11
12 ###SEPARAR ENTRE OS ESTADOS:
13 amdados=dados[dados$UF==13,]
14 spdados=dados[dados$UF==35,]
15
16 ##############################################################################
17 ##############################TAMANHO 30######################################
18 ##############################################################################
19
20 set.seed(123)
21 amostraam30=amdados[runif(30, 1, nrow(amdados)),]
22 set.seed(12)
23 amostrasp30=spdados[runif(30, 1, nrow(spdados)),]
24
25 logam30<-log(amostraam30$renda)
26 logsp30<-log(amostrasp30$renda)
27 logam30[which(logam30==-Inf)]<-log(1)
28 logsp30[which(logsp30==-Inf)]<-log(1)
29
30 muam30<-mean(logam30)
31 sigmaam30<-sd(logam30)
32 xxam30<-seq(min(logam30),max(logam30),.01)
33 yyam30<-dnorm(xxam30,muam30,sigmaam30)
34
35 musp30<-mean(logsp30)
36 sigmasp30<-sd(logsp30)
37 xxsp30<-seq(min(logsp30),max(logsp30),.01)
38 yysp30<-dnorm(xxsp30,musp30,sigmasp30)
39
40
41 ##Grafico distribuicao log(tehta) e normal
42 rep=rep("renda", length(logam30))
43 dataam30=data.frame(logam30, rep)
44 normal=rep("normal", rep(length(xxam30)))
```

```
45 dataam30_normal=data.frame(xxam30, yyam30, normal)
46
47 legend_title=""
48 ggplot() +
49   geom_density(data=dataam30,aes(logam30, color=rep), show.legend=F, size
     =1)+
50   geom_line(data = dataam30_normal,aes(x=xxam30,y=yyam30,color=normal), size
     =1) +
51   xlim(0.00001, 10)+
52   ylab(expression(paste("h(",log(theta),"|x)"))) +
53   xlab(expression(log(theta))) +
54   theme_bw()+
55   scale_colour_manual(legend_title ,values=c("blue", "black"),
56                       labels=c("Normal", "Renda"))
57
58 rep=rep("renda", length(logsp30))
59 datasp30=data.frame(logsp30, rep)
60 normal=rep("normal", rep(length(xxsp30)))
61 datasp30_normal=data.frame(xxsp30, yysp30, normal)
62
63 legend_title=""
64 ggplot() +
65   geom_density(data=datasp30,aes(logsp30, color=rep), show.legend=F, size
     =1)+
66   geom_line(data = datasp30_normal,aes(x=xxsp30,y=yysp30,color=normal), size
     =1) +
67   xlim(0,11)+
68   ylab(expression(paste("h(",log(theta),"|x)"))) +
69   xlab(expression(log(theta))) +
70   theme_bw()+
71   scale_colour_manual(legend_title ,values=c("blue", "black"),
72                       labels=c("Normal", "Renda"))
73
74
75 ###Teste KS
76 ks.test(logam30,pnorm,muam30,sigmaam30)
77 ks.test(logsp30,pnorm,musp30,sigmasp30)
78
79 ###############BAYESIANA################
80 ###########PRIORI E POSTERIORI##########
81 ########################################
82
```

```
x1<-amostraam30$renda
x2<-amostrasp30$renda
x1[which(x1==0)]<-0.1
x2[which(x2==0)]<-0.1

### definindo o log da posteriori ###
log.post<-function(par,x1, x2){
  mu1<-par[1]
  sigma1<-par[2]
  mu2<-par[3]
  sigma2<-par[4]
  n1<-length(x1)
  n2<- length(x2)
  X1<-sum(x1)
  LX1<-sum(log(x1))
  X2<-sum(x2)
  LX2<-sum(log(x2))
  if ((sigma1>0) && (sigma2>0)) return (

    ((-LX1 - n1*log(sigma1) -(n1/2)*log(2*pi) - 1/(2*sigma1^2)
      * (sum(log(x1)^2) -2*mu1*LX1 + n1*mu1^2))
     + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu1,0,20,log=T))

    + ((-LX2 - n2*log(sigma2) -(n2/2)*log(2*pi) - 1/(2*sigma2^2)
        * (sum(log(x2)^2) -2*mu2*LX2 + n2*mu2^2))
       + dgamma(sigma2,1/1000,1/1000,log=T) + dnorm(mu2,0,20,log=T))
  )
  else return(-Inf)
}

### gerando valores da posteriori ###
chute<-c(mean(log(x1)),sd(log(x1)),mean(log(x2)),sd(log(x2)))
theta<-MCMCmetrop1R(log.post,chute,x1=x1,x2=x2,mcmc=100000,logfun=T)

##Definindo o coeficiente de variacao e a distribuicao a posteriori
fi1<-sqrt(exp(theta[,2]^2)-1)
fi2<-sqrt(exp(theta[,4]^2)-1)
fi_data=data.frame(fi1,fi2)
fi_data1=gather(fi_data,"tipo","valores",1:2)

legend_title='Estado'
ggplot(data=fi_data1, aes(x=valores))+
```

```
125   geom_density(aes(color=tipo), show.legend=F, size=1) +
126   stat_density(aes(x=valores, color=tipo), geom="line",position="identity",
        size=1)+
127   xlim(0, 3.5) +
128   ylab(expression(paste("h(",phi,"|x)")))+
129   xlab(expression(phi))+
130   theme_bw()+
131   scale_colour_manual(legend_title ,values=c("blue", "black"),
132                       labels=c("Amazonas", "Sao Paulo"))
133
134
135 ### obtendo as estimativas dos parametros ###
136 mean(theta[,1])
137 emp.hpd(theta[,1],conf=0.95)
138
139 mean(theta[,2])
140 emp.hpd(theta[,2],conf=0.95)
141
142 mean(theta[,3])
143 emp.hpd(theta[,3],conf=0.95)
144
145 mean(theta[,4])
146 emp.hpd(theta[,4],conf=0.95)
147
148 mean(fi1)
149 emp.hpd(fi1,conf=0.95)
150 mean(fi2)
151 emp.hpd(fi2,conf=0.95)
152
153 ##media e desvio padrao da log-normal
154 media1=exp(theta[,1]+theta[,2]^2/2)
155 sd1=sqrt(exp(2*theta[,1] + theta[,2]^2)*(exp(theta[,2]^2)-1))
156 mean(media1)
157 emp.hpd(media1,conf=0.95)
158 mean(sd1)
159 emp.hpd(sd1,conf=0.95)
160
161
162 media2=exp(theta[,3]+theta[,4]^2/2)
163 sd2=sqrt(exp(2*theta[,3] + theta[,4]^2)*(exp(theta[,4]^2)-1))
164 mean(media2)
165 emp.hpd(media2,conf=0.95)
```

```
166 mean(sd2)
167 emp.hpd(sd2,conf=0.95)
168
169
170 log.post.h0<-function(par,x1, x2){
171   mu1<-par[1]
172   sigma1<-par[2]
173   mu2<-par[3]
174   n1<-length(x1)
175   n2<- length(x2)
176   X1<-sum(x1)
177   LX1<-sum(log(x1))
178   X2<-sum(x2)
179   LX2<-sum(log(x2))
180   if ((sigma1>0)) return (-1*(
181
182     ((-LX1 - n1*log(sigma1) -(n1/2)*log(2*pi) - 1/(2*sigma1^2)
183       * (sum(log(x1)^2) -2*mu1*LX1 + n1*mu1^2))
184      + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu1,0,20,log=T))
185
186     + ((-LX2 - n2*log(sigma1) -(n2/2)*log(2*pi) - 1/(2*sigma1^2)
187         * (sum(log(x2)^2) -2*mu2*LX2 + n2*mu2^2))
188        + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu2,0,20,log=T))
189   ))
190   else return(-Inf)
191 }
192
193 chute<-c(mean(log(x1)),sd(log(x1)),mean(log(x2)))
194 a=optim(chute, log.post.h0, x1=x1, x2=x2)
195
196 max.post.h0=-a$value
197
198 valor.post= numeric()
199 for (i in 1:100000){
200   valor.post[i] = log.post(theta[i,], x1,x2)
201 }
202 valor.e=mean(valor.post<max.post.h0)
203 valor.e
204
205
206 ##########################################################################
207 ############################TAMANHO 50####################################
```

```
208 ##############################################################################
209
210 set.seed(123)
211 amostraam50=amdados[runif(50, 1, nrow(amdados)),]
212 set.seed(12)
213 amostrasp50=spdados[runif(50, 1, nrow(spdados)),]
214
215 logam50<-log(amostraam50$renda)
216 logsp50<-log(amostrasp50$renda)
217 logam50[which(logam50==-Inf)]<-log(1)
218 logsp50[which(logsp50==-Inf)]<-log(1)
219
220 muam50<-mean(logam50)
221 sigmaam50<-sd(logam50)
222 xxam50<-seq(min(logam50),max(logam50),.01)
223 yyam50<-dnorm(xxam50,muam50,sigmaam50)
224
225 musp50<-mean(logsp50)
226 sigmasp50<-sd(logsp50)
227 xxsp50<-seq(min(logsp50),max(logsp50),.01)
228 yysp50<-dnorm(xxsp50,musp50,sigmasp50)
229
230
231 ##Grafico distribuicao log(tehta) e normal
232 rep=rep("renda", length(logam50))
233 dataam50=data.frame(logam50, rep)
234 normal=rep("normal", rep(length(xxam50)))
235 dataam50_normal=data.frame(xxam50, yyam50, normal)
236
237 legend_title=""
238 ggplot() +
239   geom_density(data=dataam50,aes(logam50, color=rep), show.legend=F, size
      =1)+
240   geom_line(data = dataam50_normal,aes(x=xxam50,y=yyam50,color=normal), size
      =1) +
241   xlim(0.0001,10) + ylim(0,.6) +
242   ylab(expression(paste("h(",log(theta),"|x)"))) +
243   xlab(expression(log(theta))) +
244   theme_bw()+
245   scale_colour_manual(legend_title ,values=c("blue", "black"),
246                       labels=c("Normal", "Renda"))
247
```

```
248 rep=rep("renda", length(logsp50))
249 datasp50=data.frame(logsp50, rep)
250 normal=rep("normal", rep(length(xxsp50)))
251 datasp50_normal=data.frame(xxsp50, yysp50, normal)
252
253 legend_title=""
254 ggplot() +
255   geom_density(data=datasp50,aes(logsp50, color=rep), show.legend=F, size
      =1)+
256   geom_line(data = datasp50_normal,aes(x=xxsp50,y=yysp50,color=normal), size
      =1) +
257   xlim(0.001,10)+
258   ylab(expression(paste("h(",log(theta),"|x)"))) +
259   xlab(expression(log(theta))) +
260   theme_bw()+
261   scale_colour_manual(legend_title,values=c("blue", "black"),
262                       labels=c("Normal", "Renda"))
263
264
265 ###Teste KS
266 ks.test(logam50,pnorm,muam50,sigmaam50)
267 ks.test(logsp50,pnorm,musp50,sigmasp50)
268
269 ############BAYESIANA######################
270 ##########Priori e posteriori#############
271 ##########################################
272
273 x1<-amostraam50$renda
274 x2<-amostrasp50$renda
275 x1[which(x1==0)]<-0.1
276 x2[which(x2==0)]<-0.1
277
278 ### definindo o log da posteriori ###
279 log.post<-function(par,x1, x2){
280   mu1<-par[1]
281   sigma1<-par[2]
282   mu2<-par[3]
283   sigma2<-par[4]
284   n1<-length(x1)
285   n2<- length(x2)
286   X1<-sum(x1)
287   LX1<-sum(log(x1))
```

```
288   X2<-sum(x2)
289   LX2<-sum(log(x2))
290   if ((sigma1>0) && (sigma2>0)) return (
291
292     ((-LX1 - n1*log(sigma1) -(n1/2)*log(2*pi) - 1/(2*sigma1^2)
293       * (sum(log(x1)^2) -2*mu1*LX1 + n1*mu1^2))
294      + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu1,0,20,log=T))
295
296     + ((-LX2 - n2*log(sigma2) -(n2/2)*log(2*pi) - 1/(2*sigma2^2)
297         * (sum(log(x2)^2) -2*mu2*LX2 + n2*mu2^2))
298        + dgamma(sigma2,1/1000,1/1000,log=T) + dnorm(mu2,0,20,log=T))
299   )
300   else return(-Inf)
301 }
302
303 ### gerando valores da posteriori ###
304 chute<-c(mean(log(x1)),sd(log(x1)),mean(log(x2)),sd(log(x2)))
305 theta<-MCMCmetrop1R(log.post,chute,x1=x1,x2=x2,mcmc=100000,logfun=T)
306
307 ##Coeficiente de variacao
308 fi1<-sqrt(exp(theta[,2]^2)-1)
309 fi2<-sqrt(exp(theta[,4]^2)-1)
310 fi_data=data.frame(fi1,fi2)
311 fi_data1=gather(fi_data,"tipo","valores",1:2)
312
313 legend_title='Estados'
314 ggplot(data=fi_data1, aes(x=valores))+
315   geom_density(aes(color=tipo), show.legend=F, size=1) +
316   stat_density(aes(x=valores, color=tipo), geom="line",position="identity",
      size=1)+
317   xlim(0, 3) +
318   ylab(expression(paste("h(",phi,"|x)")))+
319   xlab(expression(phi))+
320   theme_bw()+
321   scale_colour_manual(legend_title,values=c("blue", "black"),
322                       labels=c("Amazonas", "Sao Paulo"))
323
324 ##Estimativa e IC
325 mean(theta[,1])
326 emp.hpd(theta[,1],conf=0.95)
327
328 mean(theta[,2])
```

```
329 emp.hpd(theta[,2],conf=0.95)
330
331 mean(theta[,3])
332 emp.hpd(theta[,3],conf=0.95)
333
334 mean(theta[,4])
335 emp.hpd(theta[,4],conf=0.95)
336
337 mean(fi1)
338 emp.hpd(fi1,conf=0.95)
339 mean(fi2)
340 emp.hpd(fi2,conf=0.95)
341
342 ##media e desvio padrao da log-normal
343 media1=exp(theta[,1]+theta[,2]^2/2)
344 sd1=sqrt(exp(2*theta[,1] + theta[,2]^2)*(exp(theta[,2]^2)-1))
345 mean(media1)
346 emp.hpd(media1,conf=0.95)
347 mean(sd1)
348 emp.hpd(sd1,conf=0.95)
349
350
351 media2=exp(theta[,3]+theta[,4]^2/2)
352 sd2=sqrt(exp(2*theta[,3] + theta[,4]^2)*(exp(theta[,4]^2)-1))
353 mean(media2)
354 emp.hpd(media2,conf=0.95)
355 mean(sd2)
356 emp.hpd(sd2,conf=0.95)
357
358 log.post.h0<-function(par,x1, x2){
359   mu1<-par[1]
360   sigma1<-par[2]
361   mu2<-par[3]
362   n1<-length(x1)
363   n2<- length(x2)
364   X1<-sum(x1)
365   LX1<-sum(log(x1))
366   X2<-sum(x2)
367   LX2<-sum(log(x2))
368   if ((sigma1>0)) return (-1*(
369
370     ((-LX1 - n1*log(sigma1) -(n1/2)*log(2*pi) - 1/(2*sigma1^2)
```

```
371         * (sum(log(x1)^2) -2*mu1*LX1 + n1*mu1^2))
372       + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu1,0,20,log=T))
373
374     + ((-LX2 - n2*log(sigma1) -(n2/2)*log(2*pi) - 1/(2*sigma1^2)
375           * (sum(log(x2)^2) -2*mu2*LX2 + n2*mu2^2))
376         + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu2,0,20,log=T))
377   ))
378   else return(-Inf)
379 }
380
381 chute<-c(mean(log(x1)),sd(log(x1)),mean(log(x2)))
382 a=optim(chute, log.post.h0, x1=x1, x2=x2)
383
384 max.post.h0=-a$value
385
386 valor.post= numeric()
387 for (i in 1:100000){
388   valor.post[i] = log.post(theta[i,], x1,x2)
389 }
390
391 valor.e=mean(valor.post<max.post.h0)
392 valor.e
393
394 ##############################################################################
395 ##############################TAMANHO 200#####################################
396 ##############################################################################
397
398 set.seed(123)
399 amostraam200=amdados[runif(200, 1, nrow(amdados)),]
400 set.seed(12)
401 amostrasp200=spdados[runif(200, 1, nrow(spdados)),]
402
403 logam200<-log(amostraam200$renda)
404 logsp200<-log(amostrasp200$renda)
405 logam200[which(logam200==-Inf)]<-log(1)
406 logsp200[which(logsp200==-Inf)]<-log(1)
407
408 muam200<-mean(logam200)
409 sigmaam200<-sd(logam200)
410 xxam200<-seq(min(logam200),max(logam200),.01)
411 yyam200<-dnorm(xxam200,muam200,sigmaam200)
412
```

```
musp200<-mean(logsp200)
sigmasp200<-sd(logsp200)
xxsp200<-seq(min(logsp200),max(logsp200),.01)
yysp200<-dnorm(xxsp200,musp200,sigmasp200)



##Grafico distribuicao log(tehta) e normal
rep=rep("renda", length(logam200))
dataam200=data.frame(logam200, rep)
normal=rep("normal", rep(length(xxam200)))
dataam200_normal=data.frame(xxam200, yyam200, normal)

legend_title=""
ggplot() +
  geom_density(data=dataam200,aes(logam200, color=rep), show.legend=F, size
    =1)+
  geom_line(data = dataam200_normal,aes(x=xxam200,y=yyam200,color=normal),
    size=1) +
  xlim(0.0001,10) +
  ylab(expression(paste("h(",log(theta),"|x)"))) +
  xlab(expression(log(theta))) +
  theme_bw()+
  scale_colour_manual(legend_title ,values=c("blue", "black"),
                      labels=c("Normal", "Renda"))

rep=rep("renda", length(logsp200))
datasp200=data.frame(logsp200, rep)
normal=rep("normal", rep(length(xxsp200)))
datasp200_normal=data.frame(xxsp200, yysp200, normal)

legend_title=""
ggplot() +
  geom_density(data=datasp200,aes(logsp200, color=rep), show.legend=F, size
    =1)+
  geom_line(data = datasp200_normal,aes(x=xxsp200,y=yysp200,color=normal),
    size=1) +
  xlim(0.00001,11.5)+
  ylab(expression(paste("h(",log(theta),"|x)"))) +
  xlab(expression(log(theta))) +
  theme_bw()+
  scale_colour_manual(legend_title ,values=c("blue", "black"),
```

```
451                          labels=c("Normal", "Renda"))
452
453 ###Teste KS
454 ks.test(logam200,pnorm,muam200,sigmaam200)
455 ks.test(logsp200,pnorm,musp200,sigmasp200)
456
457 ############BAYESIANA#####################
458 ########Priori e posteriori###############
459 ##########################################
460
461 x1<-amostraam200$renda
462 x2<-amostrasp200$renda
463 x1[which(x1==0)]<-0.1
464 x2[which(x2==0)]<-0.1
465
466 ### definindo o log da posteriori ###
467 log.post<-function(par,x1, x2){
468   mu1<-par[1]
469   sigma1<-par[2]
470   mu2<-par[3]
471   sigma2<-par[4]
472   n1<-length(x1)
473   n2<- length(x2)
474   X1<-sum(x1)
475   LX1<-sum(log(x1))
476   X2<-sum(x2)
477   LX2<-sum(log(x2))
478   if ((sigma1>0) && (sigma2>0)) return (
479
480     ((-LX1 - n1*log(sigma1) -(n1/2)*log(2*pi) - 1/(2*sigma1^2)
481       * (sum(log(x1)^2) -2*mu1*LX1 + n1*mu1^2))
482      + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu1,0,20,log=T))
483
484     + ((-LX2 - n2*log(sigma2) -(n2/2)*log(2*pi) - 1/(2*sigma2^2)
485         * (sum(log(x2)^2) -2*mu2*LX2 + n2*mu2^2))
486        + dgamma(sigma2,1/1000,1/1000,log=T) + dnorm(mu2,0,20,log=T))
487   )
488   else return(-Inf)
489 }
490
491 ### gerando valores da posteriori ###
492 chute<-c(mean(log(x1)),sd(log(x1)),mean(log(x2)),sd(log(x2)))
```

```
493 theta<-MCMCmetrop1R(log.post,chute,x1=x1,x2=x2,mcmc=100000,logfun=T)
494
495 ##Coeficiente de variacao
496 fi1<-sqrt(exp(theta[,2]^2)-1)
497 fi2<-sqrt(exp(theta[,4]^2)-1)
498 fi_data=data.frame(fi1,fi2)
499 fi_data1=gather(fi_data,"tipo","valores",1:2)
500
501 legend_title='Estados'
502 ggplot(data=fi_data1, aes(x=valores))+
503   geom_density(aes(color=tipo), show.legend=F, size=1) +
504   stat_density(aes(x=valores, color=tipo), geom="line",position="identity",
      size=1)+
505   xlim(0, 3) +
506   ylab(expression(paste("h(",phi,"|x)")))+
507   xlab(expression(phi))+
508   theme_bw()+
509   scale_colour_manual(legend_title,values=c("blue", "black"),
510                       labels=c("Amazonas", "Sao Paulo"))
511
512 ##Estimativa e IC
513 mean(theta[,1])
514 emp.hpd(theta[,1],conf=0.95)
515
516 mean(theta[,2])
517 emp.hpd(theta[,2],conf=0.95)
518
519 mean(theta[,3])
520 emp.hpd(theta[,3],conf=0.95)
521
522 mean(theta[,4])
523 emp.hpd(theta[,4],conf=0.95)
524
525 mean(fi1)
526 emp.hpd(fi1,conf=0.95)
527 mean(fi2)
528 emp.hpd(fi2,conf=0.95)
529
530 ##media e desvio padrao da log-normal
531 media1=exp(theta[,1]+theta[,2]^2/2)
532 sd1=sqrt(exp(2*theta[,1] + theta[,2]^2)*(exp(theta[,2]^2)-1))
533 mean(media1)
```

```
534 emp.hpd(media1,conf=0.95)
535 mean(sd1)
536 emp.hpd(sd1,conf=0.95)
537
538
539 media2=exp(theta[,3]+theta[,4]^2/2)
540 sd2=sqrt(exp(2*theta[,3] + theta[,4]^2)*(exp(theta[,4]^2)-1))
541 mean(media2)
542 emp.hpd(media2,conf=0.95)
543 mean(sd2)
544 emp.hpd(sd2,conf=0.95)
545
546 log.post.h0<-function(par,x1, x2){
547   mu1<-par[1]
548   sigma1<-par[2]
549   mu2<-par[3]
550   n1<-length(x1)
551   n2<- length(x2)
552   X1<-sum(x1)
553   LX1<-sum(log(x1))
554   X2<-sum(x2)
555   LX2<-sum(log(x2))
556   if ((sigma1>0)) return (-1*(
557
558     ((-LX1 - n1*log(sigma1) -(n1/2)*log(2*pi) - 1/(2*sigma1^2)
559       * (sum(log(x1)^2) -2*mu1*LX1 + n1*mu1^2))
560      + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu1,0,20,log=T))
561
562     + ((-LX2 - n2*log(sigma1) -(n2/2)*log(2*pi) - 1/(2*sigma1^2)
563         * (sum(log(x2)^2) -2*mu2*LX2 + n2*mu2^2))
564        + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu2,0,20,log=T))
565   ))
566   else return(-Inf)
567 }
568
569 chute<-c(mean(log(x1)),sd(log(x1)),mean(log(x2)))
570 a=optim(chute, log.post.h0, x1=x1, x2=x2)
571
572 max.post.h0=-a$value
573
574 valor.post= numeric()
575 for (i in 1:100000){
```

```
576     valor.post[i] = log.post(theta[i,], x1,x2)
577   }
578
579   valor.e=mean(valor.post<max.post.h0)
580   valor.e
581
582   ##################################PNAD 2015#################################
583   ############################Priori e posteriori#############################
584   ##########################################################################
585
586   amostraamgeral=amdados
587   amostraspgeral=spdados
588
589   logamgeral<-log(amostraamgeral$renda)
590   logspgeral<-log(amostraspgeral$renda)
591   logamgeral[which(logamgeral==-Inf)]<-log(1)
592   logspgeral[which(logspgeral==-Inf)]<-log(1)
593
594
595   muamgeral<-mean(logamgeral)
596   sigmaamgeral<-sd(logamgeral)
597   xxamgeral<-seq(min(logamgeral),max(logamgeral),.01)
598   yyamgeral<-dnorm(xxamgeral,muamgeral,sigmaamgeral)
599
600   muspgeral<-mean(logspgeral)
601   sigmaspgeral<-sd(logspgeral)
602   xxspgeral<-seq(min(logspgeral),max(logspgeral),.01)
603   yyspgeral<-dnorm(xxspgeral,muspgeral,sigmaspgeral)
604
605
606   ks.test(logamgeral,pnorm,muamgeral,sigmaamgeral)
607   ks.test(logspgeral,pnorm,muspgeral,sigmaspgeral)
608
609
610
611   ggplot() +
612     geom_density(data=data.frame(logamgeral),aes(logamgeral), size=1)+
613     xlim(0.00001, 10)+
614     geom_line(data = data.frame(xxamgeral,yyamgeral),aes(x=xxamgeral,y=
        yyamgeral),color='blue', size=1) +
615     ylab(expression(paste("h(",theta,"|x)"))) +
616     xlab(expression(theta)) +
```

```
617     theme_bw()
618
619 ggplot() +
620     geom_density(data=data.frame(logspgeral),aes(logspgeral), size=1)+
621     geom_line(data = data.frame(xxspgeral,yyspgeral),aes(x=xxspgeral,y=
          yyspgeral),color='blue', size=1) +
622     xlim(2,11)+
623     ylab(expression(paste("h(",theta,"|x)"))) +
624     xlab(expression(theta)) +
625     theme_bw()
626
627
628 x1<-amostraamgeral$renda
629 x2<-amostraspgeral$renda
630 x1[which(x1==0)]<-0.1
631 x2[which(x2==0)]<-0.1
632
633 x1
634 x2
635 sum(x1)
636 sum(x2)
637 sum(log(x1))
638 sum(log(x2))
639
640 ### definindo o log da posteriori ###
641 log.post<-function(par,x1, x2){
642     mu1<-par[1]
643     sigma1<-par[2]
644     mu2<-par[3]
645     sigma2<-par[4]
646     n1<-length(x1)
647     n2<- length(x2)
648     X1<-sum(x1)
649     LX1<-sum(log(x1))
650     X2<-sum(x2)
651     LX2<-sum(log(x2))
652     if ((sigma1>0) && (sigma2>0)) return (
653
654       ((-LX1 - n1*log(sigma1) -(n1/2)*log(2*pi) - 1/(2*sigma1^2)
655         * (sum(log(x1)^2) -2*mu1*LX1 + n1*mu1^2))
656        + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu1,0,20,log=T))
657
```

```
658       + ((-LX2 - n2*log(sigma2) -(n2/2)*log(2*pi) - 1/(2*sigma2^2)
659             * (sum(log(x2)^2) -2*mu2*LX2 + n2*mu2^2))
660           + dgamma(sigma2,1/1000,1/1000,log=T) + dnorm(mu2,0,20,log=T))
661     )
662     else return(-Inf)
663   }
664
665   ### gerando valores da posteriori ###
666   chute<-c(mean(log(x1)),sd(log(x1)),mean(log(x2)),sd(log(x2)))
667   theta<-MCMCmetrop1R(log.post,chute,x1=x1,x2=x2,mcmc=100000,logfun=T)
668
669   ##Definindo o coeficiente de variacao e a distribuicao a posteriori
670   fi1<-sqrt(exp(theta[,2]^2)-1)
671   fi2<-sqrt(exp(theta[,4]^2)-1)
672
673   fi_data=data.frame(fi1,fi2)
674   fi_data1=gather(fi_data,"tipo","valores",1:2)
675
676
677   legend_title='Estado'
678   ggplot(data=fi_data1, aes(x=valores))+
679     geom_density(aes(color=tipo), show.legend=F, size=1) +
680     stat_density(aes(x=valores, color=tipo), geom="line",position="identity",
        size=1)+
681     xlim(1.25, 2) +
682     ylab(expression(paste("h(",phi,"|x)")))+
683     xlab(expression(phi))+
684     theme_bw()+
685     scale_colour_manual(legend_title,values=c("blue", "black"),
686                         labels=c("Amazonas", "Sao Paulo"))
687
688   ### obtendo as estimativas dos parametros ###
689   mam=mean(theta[,1])
690   icmam=emp.hpd(theta[,1],conf=0.95)
691
692   sdam=mean(theta[,2])
693   icsdam=emp.hpd(theta[,2],conf=0.95)
694
695   msp=mean(theta[,3])
696   icmsp=emp.hpd(theta[,3],conf=0.95)
697
698   sdsp=mean(theta[,4])
```

```
699 icsdsp=emp.hpd(theta[,4],conf=0.95)
700
701 mfi1=mean(fi1)
702 icfi1=emp.hpd(fi1,conf=0.95)
703
704 mfi2=mean(fi2)
705 icmfi2=emp.hpd(fi2,conf=0.95)
706
707
708 log.post.h0<-function(par,x1, x2){
709   mu1<-par[1]
710   sigma1<-par[2]
711   mu2<-par[3]
712   n1<-length(x1)
713   n2<- length(x2)
714   X1<-sum(x1)
715   LX1<-sum(log(x1))
716   X2<-sum(x2)
717   LX2<-sum(log(x2))
718   if ((sigma1>0)) return (-1*(
719
720     ((-LX1 - n1*log(sigma1) -(n1/2)*log(2*pi) - 1/(2*sigma1^2)
721       * (sum(log(x1)^2) -2*mu1*LX1 + n1*mu1^2))
722      + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu1,0,20,log=T))
723
724     + ((-LX2 - n2*log(sigma1) -(n2/2)*log(2*pi) - 1/(2*sigma1^2)
725         * (sum(log(x2)^2) -2*mu2*LX2 + n2*mu2^2))
726        + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu2,0,20,log=T))
727   ))
728   else return(-Inf)
729 }
730
731 chute<-c(mean(log(x1)),sd(log(x1)),mean(log(x2)))
732 a=optim(chute, log.post.h0, x1=x1, x2=x2)
733
734 max.post.h0=-a$value
735
736 valor.post= numeric()
737 for (i in 1:100000){
738   valor.post[i] = log.post(theta[i,], x1,x2)
739 }
740 valor.e=mean(valor.post<max.post.h0)
```

```
741 valor.e
```

H.7 PNAD- Comparação por SP e AM com dados completos

```
1  library(MCMCpack)
2  library("TeachingDemos")
3  library(ggplot2)
4  library(tidyr)
5
6  data = read.table(".../dados.txt")
7  dados= data[which(is.na(data$V4742)==F),]
8  dados$renda=as.numeric(dados$V4742)
9  dados=dados[,-6]
10
11 ###SEPARAR ENTRE OS ESTADOS:
12 amdados=dados[dados$UF==13,]
13 spdados=dados[dados$UF==35,]
14
15 amostraamgeral=amdados
16 amostraspgeral=spdados
17
18 logamgeral<-log(amostraamgeral$renda)
19 logspgeral<-log(amostraspgeral$renda)
20 logamgeral[which(logamgeral==-Inf)]<-log(1)
21 logspgeral[which(logspgeral==-Inf)]<-log(1)
22
23 muamgeral<-mean(logamgeral)
24 sigmaamgeral<-sd(logamgeral)
25 xxamgeral<-seq(min(logamgeral),max(logamgeral),.01)
26 yyamgeral<-dnorm(xxamgeral,muamgeral,sigmaamgeral)
27
28 muspgeral<-mean(logspgeral)
29 sigmaspgeral<-sd(logspgeral)
30 xxspgeral<-seq(min(logspgeral),max(logspgeral),.01)
31 yyspgeral<-dnorm(xxspgeral,muspgeral,sigmaspgeral)
32
33
34 ##Grafico distribuicao log(tehta) e normal
35 rep=rep("renda", length(logamgeral))
36 dataamgeral=data.frame(logamgeral, rep)
37 normal=rep("normal", rep(length(xxamgeral)))
38 dataamgeral_normal=data.frame(xxamgeral, yyamgeral, normal)
```

```
39
40 legend_title=""
41 ggplot() +
42   geom_density(data=dataamgeral,aes(logamgeral, color=rep), show.legend=F,
       size=1)+
43   geom_line(data = dataamgeral_normal,aes(x=xxamgeral,y=yyamgeral,color=
       normal), size=1) +
44   xlim(0.00001, 10)+
45   ylab(expression(paste("h(",log(theta),"|x)"))) +
46   xlab(expression(log(theta))) +
47   theme_bw()+
48   scale_colour_manual(legend_title ,values=c("blue", "black"),
49                       labels=c("Normal", "Renda"))
50
51
52
53 rep=rep("renda", length(logspgeral))
54 dataspgeral=data.frame(logspgeral, rep)
55 normal=rep("normal", rep(length(xxspgeral)))
56 dataspgeral_normal=data.frame(xxspgeral, yyspgeral, normal)
57
58 legend_title=""
59 ggplot() +
60   geom_density(data=dataspgeral,aes(logspgeral, color=rep), show.legend=F,
       size=1)+
61   geom_line(data = dataspgeral_normal,aes(x=xxspgeral,y=yyspgeral,color=
       normal), size=1) +
62   xlim(0.01, 11)+
63   ylab(expression(paste("h(",log(theta),"|x)"))) +
64   xlab(expression(log(theta))) +
65   theme_bw()+
66   scale_colour_manual(legend_title ,values=c("blue", "black"),
67                       labels=c("Normal", "Renda"))
68
69
70 x1<-amostraamgeral$renda
71 x2<-amostraspgeral$renda
72 x1[which(x1==0)]<-0.1
73 x2[which(x2==0)]<-0.1
74
75 ### definindo o log da posteriori ###
76 log.post<-function(par,x1, x2){
```

```
77    mu1<-par[1]
78    sigma1<-par[2]
79    mu2<-par[3]
80    sigma2<-par[4]
81    n1<-length(x1)
82    n2<- length(x2)
83    X1<-sum(x1)
84    LX1<-sum(log(x1))
85    X2<-sum(x2)
86    LX2<-sum(log(x2))
87    if ((sigma1>0) && (sigma2>0)) return (
88
89      ((-LX1 - n1*log(sigma1) -(n1/2)*log(2*pi) - 1/(2*sigma1^2)
90        * (sum(log(x1)^2) -2*mu1*LX1 + n1*mu1^2))
91       + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu1,0,20,log=T))
92
93      + ((-LX2 - n2*log(sigma2) -(n2/2)*log(2*pi) - 1/(2*sigma2^2)
94          * (sum(log(x2)^2) -2*mu2*LX2 + n2*mu2^2))
95         + dgamma(sigma2,1/1000,1/1000,log=T) + dnorm(mu2,0,20,log=T))
96    )
97    else return(-Inf)
98  }
99
100 ### gerando valores da posteriori ###
101 chute<-c(mean(log(x1)),sd(log(x1)),mean(log(x2)),sd(log(x2)))
102 theta<-MCMCmetrop1R(log.post,chute,x1=x1,x2=x2,mcmc=100000,logfun=T)
103
104 ##Definindo o coeficiente de variacao e a distribuicao a posteriori
105 fi1<-sqrt(exp(theta[,2]^2)-1)
106 fi2<-sqrt(exp(theta[,4]^2)-1)
107 fi_data=data.frame(fi1,fi2)
108 fi_data1=gather(fi_data,"tipo","valores",1:2)
109
110 legend_title='Estado'
111 ggplot(data=fi_data1, aes(x=valores))+
112   geom_density(aes(color=tipo), show.legend=F, size=1) +
113   stat_density(aes(x=valores, color=tipo), geom="line",position="identity",
        size=1)+
114   xlim(1.25, 2) +
115   ylab(expression(paste("h(",phi,"|x)")))+
116   xlab(expression(phi))+
117   theme_bw()+
```

```
118   scale_colour_manual(legend_title ,values=c("blue", "black"),
119                       labels=c("Amazonas", "Sao Paulo"))
120
121
122 ### obtendo as estimativas dos parametros ###
123 mean(theta[,1])
124 emp.hpd(theta[,1],conf=0.95)
125 mean(theta[,2])
126 emp.hpd(theta[,2],conf=0.95)
127
128 mean(theta[,3])
129 emp.hpd(theta[,3],conf=0.95)
130 mean(theta[,4])
131 emp.hpd(theta[,4],conf=0.95)
132
133 mean(fi1)
134 emp.hpd(fi1,conf=0.95)
135 mean(fi2)
136 emp.hpd(fi2,conf=0.95)
137
138 ##media e desvio padrao da log-normal
139 media1=exp(theta[,1]+theta[,2]^2/2)
140 sd1=sqrt(exp(2*theta[,1] + theta[,2]^2)*(exp(theta[,2]^2)-1))
141 mean(media1)
142 emp.hpd(media1,conf=0.95)
143 mean(sd1)
144 emp.hpd(sd1,conf=0.95)
145
146
147 media2=exp(theta[,3]+theta[,4]^2/2)
148 sd2=sqrt(exp(2*theta[,3] + theta[,4]^2)*(exp(theta[,4]^2)-1))
149 mean(media2)
150 emp.hpd(media2,conf=0.95)
151 mean(sd2)
152 emp.hpd(sd2,conf=0.95)
153 ###Definindo log da posteriori
154 log.post.h0<-function(par,x1, x2){
155   mu1<-par[1]
156   sigma1<-par[2]
157   mu2<-par[3]
158   n1<-length(x1)
159   n2<- length(x2)
```

```
160   X1<-sum(x1)
161   LX1<-sum(log(x1))
162   X2<-sum(x2)
163   LX2<-sum(log(x2))
164   if ((sigma1>0)) return (-1*(
165
166     ((-LX1 - n1*log(sigma1) -(n1/2)*log(2*pi) - 1/(2*sigma1^2)
167       * (sum(log(x1)^2) -2*mu1*LX1 + n1*mu1^2))
168      + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu1,0,20,log=T))
169
170     + ((-LX2 - n2*log(sigma1) -(n2/2)*log(2*pi) - 1/(2*sigma1^2)
171         * (sum(log(x2)^2) -2*mu2*LX2 + n2*mu2^2))
172        + dgamma(sigma1,1/1000,1/1000,log=T) + dnorm(mu2,0,20,log=T))
173   ))
174   else return(-Inf)
175 }
176
177 chute<-c(mean(log(x1)),sd(log(x1)),mean(log(x2)))
178 a=optim(chute, log.post.h0, x1=x1, x2=x2)
179
180 max.post.h0=-a$value
181
182 valor.post= numeric()
183 for (i in 1:100000){
184   valor.post[i] = log.post(theta[i,], x1,x2)
185 }
186 valor.e=mean(valor.post<max.post.h0)
187
188
189 ####Visao classica
190
191 am_n=length(amdados$renda)
192 sp_n=length(spdados$renda)
193
194 am_media=mean(amdados$renda)
195 sp_media=mean(spdados$renda)
196
197 am_sd=sd(amdados$renda)
198 sp_sd=sd(spdados$renda)
199
200 am_cv= am_sd/am_media
201 sp_cv= sp_sd/sp_media
```

```
202
203 am_cv_sd=am_cv/sqrt(2*am_n)
204 sp_cv_sd=sp_cv/sqrt(2*sp_n)
205
206 #Intervalo de confianca
207 am_cv+qt(.975, df=am_n-1)*am_cv_sd
208 am_cv+qt(.025, df=am_n-1)*am_cv_sd
209
210 sp_cv+qt(.975, df=sp_n-1)*sp_cv_sd
211 sp_cv+qt(.025, df=sp_n-1)*sp_cv_sd
```