Projeto 2

Estatística Computacional e Simulação – 41165

2022/2023

Docente: Professora Doutora Isabel Pereira

Gonçalo Freitas - 98012

Tiago Alvim - 95584

Vasco Costa – 97746

Exercício 1

a) Pretende-se calcular o seguinte integral: $I_x = \int_0^x e^{-\frac{t^2}{2}}\, dt$, considerando uma mudança de variável $u = \frac{t}{x}$ pelo que $t = ux$ em que $dt = xdu$. Ficando então com

$$I_x = \int_0^x xe^{-\frac{(xu)^2}{2}}\, du$$

O Método de Integração de Monte Carlo, baseia-se na aproximação de que, dado um integral, que se pretende calcular, $\theta = \int_{-\infty}^{+\infty} g(x)f(x)dx$ em que $\theta = E[g(X)]$ onde X é uma variável aleatória continua com uma função densidade de probabilidade $f(x)$ e $g(x)$ uma função real. Este método baseia-se na Lei dos Grandes Números em que $E[g(X)] = lim\frac{1}{N}\sum_{i=1}^{N} g(X_i)$ sendo válido para quando $N \to \infty$, na prática efetua-se a aproximação de que $E[g(X)] \approx \frac{1}{N}\sum_{i=1}^{N} g(X_i)$ para valores de N grandes, com os X_i a serem gerados pela função densidade de probabilidade de f.

Posto isto, sabendo que $f(x) = 1$ traduz a função densidade de probabilidade de uma distribuição $U(0,1)$ pois, para uma distribuição $U(a,b)$, $f_U(x) = \frac{1}{b-a}1_{(a,b)}(x)$. Assim, podemos modelar o integral para obter o seguinte.

$$I_x = \int_0^x xe^{-\frac{(xu)^2}{2}}\, du = \int_0^x xe^{-\frac{(xu)^2}{2}} \times 1\, du = \int_0^x g(u)f_U(u)\, du$$

Em que $g(u) = x\exp\left\{-\frac{(xu)^2}{2}\right\}$ e $f(u) = 1$. Assim, a partir do Método de Integração de Monte Carlo, obtemos, tal como queríamos demonstrar:

$$I_x = E\left[x\exp\left\{-\frac{(xu)^2}{2}\right\}\right], u\sim U(0,1)$$

b) Simulação do valor do integral e construção do intervalo de confiança para o mesmo com comparação ao valor exato

i. ***Valor estimado:***

Tomando $x = 10$ e $N = 10^7$ é possível efetuar-se o método de Monte Carlo como demonstrado anteriormente em a). Primeiramente gerando N números aleatórios que seguem a distribuição Uniforme (0,1) e depois efetuando o valor médio dos valores de $g(u)$ com $x = 10$ e com cada um dos valores de u a ser então um dos valores da distribuição.

Assim, o código proposto para obter uma estimativa para o valor de $\hat{I}_{10}$ pode ser descrito nos seguintes passos, resumidamente, como de seguida:

1. Gerar N valores de u, que seguem uma distribuição $U(0,1)$.
2. Calcular os N valores para g, usando os N valores de u gerados, guardando todos os valores num vetor, I_{all}.
3. Calcular a estimativa para I_{10} como $\hat{I}_{10} = mean(I_{all})$.

De notar que o valor desta estimativa irá diferir de tentativa em tentativa. Para combater este problema da variação dos resultados de cada vez que era corrido o código, fixou-se o valor da SEED como 20, tendo a estimativa obtida sido $\hat{I}_{10} \approx 1.253898$, para este valor de seed.

Analisando a variância do estimador de Monte Carlo, que será $V\left[\frac{1}{N}\sum_{i=1}^{N} g(X_i)\right] = \frac{V[g(X)]}{N}$ chega-se a que o erro estimado seja $\sqrt{\frac{s^2_{g(x)}}{N}}$ que será a raiz quadrada do quociente da variância dos valores simulados pelo número de simulações. Resultando num erro aproximadamente igual a 8.5417×10^{-4}.

Na Figura 1 apresenta-se o histograma dos valores de $g(u)$ em que se pode verificar que apesar de existirem valores dispersos no xx apresentam uma frequência relativa muito baixa. Visualizando-se que existe pouca variância dos valores como seria de esperar pelo que se obteve na estimação do erro do método de Monte Carlo.

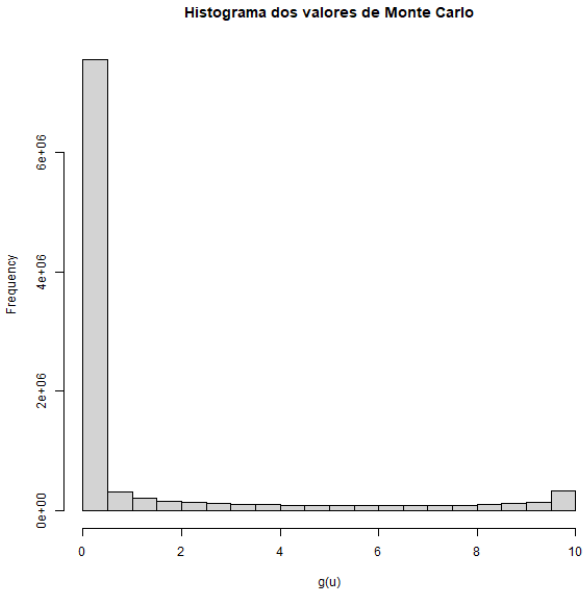

Figura 1 - Histograma com os valores de g(u)

ii. ***Valor em R:***

Sabendo-se que a função de densidade de probabilidade de uma distribuição Normal (μ, σ) de média μ e variância σ^2 é $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, caso seja de uma Normal (0, 1) chega-se a $f_{N(0,1)}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, algo bastante parecido ao que se tem no interior de I_x. Além do mais sabe-se que $\int_{-\infty}^{+\infty} f(x)dx = 1$ e dado que a função é simétrica em $x = 0$ pode-se afirmar que $\frac{1}{2} + \int_0^{+\infty} f(x)dx = 1$.

Contudo, caso se procure uma probabilidade, $P(X \leq x)$ em que X segue uma dada distribuição, neste caso $X \sim N(0,1)$, esta probabilidade será então $\int_{-\infty}^{x} f(x)dx$ que pode ser decomposto em $P(X \leq x) = \frac{1}{2} + \int_0^x f(x)dx = \frac{1}{2} + \int_0^x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{x^2}{2}} dx$. Como $I_x = \int_0^x e^{-\frac{t^2}{2}} dt$ este pode ser transportado para que $P(X \leq x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} * I_x$ que será equivalente a que $I_x = \left(P(X \leq x) - \frac{1}{2}\right) * \sqrt{2\pi}$. Em R é possível usar a inicial *p* que devolve o valor do integral de $-\infty$ até ao valor desejado para uma dada distribuição pelo que a instrução $pnorm(x)$ irá devolver o valor do integral de $-\infty$ até x de uma distribuição Normal (0,1). Resultando em $I_x = \left(pnorm(x) - \frac{1}{2}\right) * \sqrt{2\pi}$. Com $x = 10$, chega-se a que $I_{10} = \left(pnorm(10) - \frac{1}{2}\right) * \sqrt{2\pi}$ como se pretendia demonstrar.

Calculando esta expressão no R obtém-se que $I_{10} = 1.253314$ sendo este então o valor exato.

iii. ***Intervalo de confiança:***

Primeiramente pode-se verificar o $Bias$ do valor calculado que ao obter-se $\hat{I}_{10} \approx 1.253767$ leva a que $Bias = \hat{I}_{10} - I_{10} \approx 5.839 \times 10^{-4}$. Verificando-se assim que o valor estimado é uma sobrestimação, com um erro relativo igual a $Er(\%) = \frac{|I_{10} - \hat{I}_{10}|}{I_{10}} \times 100 \approx 0.046\ \%$.

O intervalo de confiança pode ser construído a partir do teorema de Slutsky em que $\sqrt{N}\frac{\hat{I}-\mu}{S_T} \sim N(0,1)$ onde $S_T^2 = \frac{1}{N}\sum_{i=1}^{N}(g(u_i) - \hat{I})^2$. Considerando $z_{1-\frac{\alpha}{2}}$ o quantil de ordem $1 - \frac{\alpha}{2}$ da distribuição Normal (0, 1) tem-se que $P\left(-z_{1-\frac{\alpha}{2}} \leq \sqrt{N}\frac{\hat{I}-\mu}{S_T} \leq z_{1-\frac{\alpha}{2}}\right) \approx 1 - \alpha$. Desenvolvendo estes termos para isolar μ chega-se a que este se encontra no intervalo $\left[\hat{I} - z_{1-\frac{\alpha}{2}}\frac{S_T}{\sqrt{N}}, \hat{I} + z_{1-\frac{\alpha}{2}}\frac{S_T}{\sqrt{N}}\right]$. Tomando um $\alpha = 0.05$ chega-se a um intervalo de confiança de

$$[1.252224, 1.255572]$$

Uma vez que o valor exato, I_{10}, se encontra dentro do intervalo de confiança pode-se considerar que o método efetuou uma boa estimação.

Exercício 2

De referir que neste exercício também se fixou o valor da SEED, novamente para um valor de 20, de forma a obter os mesmos resultados sempre que se corra o código.

Considere-se a série temporal $x_t = ax_{t-1} + bx_{t-1}y_{t-1} + y_t$ onde $Y \sim N(0,1)$. Uma série temporal é um conjunto de observações que são feitas sequencialmente em que a ordem apresenta uma importância e existe dependência entre os termos.

a) Sabendo que $S = \frac{\sum_{t=2}^{n} x_t x_{t-1}}{\sum_{t=2}^{n} x_{t-1}^2}$ é um estimador para a. E utilizando os parâmetros $a = 0.4$ e $b = 0.1$.

i. De forma a estimar o parâmetro a foi aplicado o método das réplicas 1000 vezes, $N = 1000$, para amostras de dimensão 100, $obs = 100$. O método das réplicas consiste em utilizar N amostras independentes calculando o valor da estatística de teste em cada uma e analisando assim o conjunto das variáveis de teste.

De forma prática é então necessário criar dois vetores, um para armazenar os valores de x e outro os de y, este último pode ser logo criado com os obs, 100, valores da distribuição Normal (0, 1) que se pretende. Dando ao primeiro valor de x, x_1, o mesmo de valor y_1, para a primeira observação. De seguida, faz-se a iteração ao longo de t, do segundo até ao último, neste caso 100 (valor de obs), em que se calcula o valor de x_t e se pode já ir efetuando a soma dos valores do numerador e do denominador para se conseguir calcular S ao final das obs desejadas. Armazenando-se a estimativa para a nesta repetição e efetuando este processo N vezes, aplicando assim o método das réplicas.

Assim, os passos que descrevem o código proposto para obter uma estimativa para o valor de â podem ser descritos, resumidamente, como de seguida:

1. Criação de um vetor y com obs (100) valores que seguem uma distribuição $N(0,1)$.
2. Inicializar um vetor x de comprimento obs.
3. Definir a semente, $x_1 = y_1$.
4. Ciclo for de 2 até obs, responsável por
    a. Calcular o valor de $x_t, x_t = ax_{t-1} + b(x_{t-1} \times y_{t-1}) + y_t$
5. Calcular o valor de S, $S = \frac{\sum_{t=2}^{n} x_t x_{t-1}}{\sum_{t=2}^{n} x_{t-1}^2}$, guardando o valor num vetor R
6. Repetir os pontos de 1 a 5, N vezes
7. Calcular a média dos valores presentes no vetor R, $\hat{a} = mean(R)$

Considerando então N amostras independentes com elementos $\left(X_1^{(1)}, X_2^{(1)}, \dots, X_{obs}^{(1)}\right), \dots, \left(X_1^{(N)}, X_2^{(N)}, \dots, X_{obs}^{(N)}\right)$ é possível calcular para cada $X^{(i)}$ uma estimativa para a estatística de teste T, em que $T^{(1)} = \left(X_1^{(1)}, X_2^{(1)}, \dots, X_{obs}^{(1)}\right), \dots, T^{(N)} = \left(X_1^{(N)}, X_2^{(N)}, \dots, X_{obs}^{(N)}\right)$.

Assim, a estimativa para a será estimado pelo valor médio dos â calculados nas réplicas que resultou num valor de 0.413343 que se pode comparar face ao valor teórico de 0.4.

O desvio padrão das réplicas também pode ser calculado com o comando *sd* e chegou-se a um valor de $sd(\hat{a}) = 0.101615$. Efetuando o cálculo do erro quadrático médio, ou seja, $\frac{1}{N}\sum_{i=1}^{N}(\hat{a}^i - a)^2$ resulta num valor de 0.010493.

Analisando o viés da estimativa, ou seja $E[\hat{a} - a]$, que será igual a $\frac{1}{N}\sum_{i=1}^{N}(\hat{a}^i - a)$, efetuando esta conta chega-se a um valor de 0.013343, ou seja um valor superior ao valor medido. Levando a que se conclua que este método levou a uma sobrestimação do valor real.

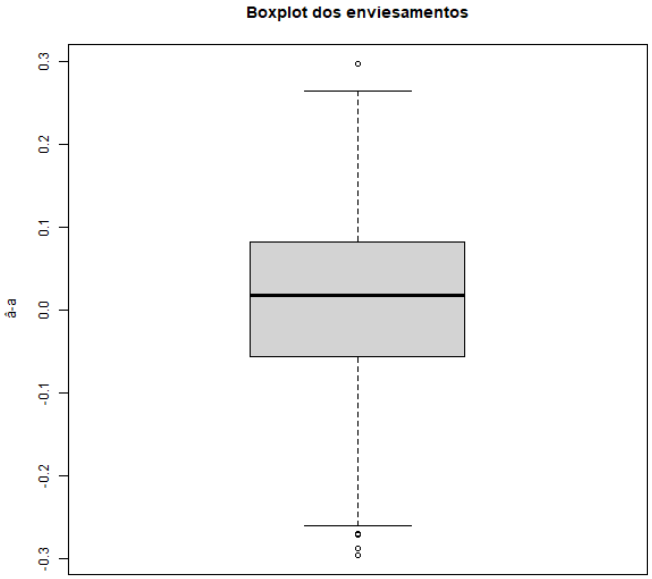

Figura 2 - Boxplot dos enviesamentos da amostra, $\hat{a} - a$

De uma forma visual, na Figura 2, pode-se efetuar o *boxplot* dos enviesamentos, ou seja, os valores de $\hat{a}^i - a$, que permite assim comparar relativamente a 0, sendo que ao ser maior é caso de uma sobrestimação e caso seja menor a uma subestimação. Analisando assim os resultados obtidos verifica-se que este método sobrestimou o valor real.

ii. Esperando que os valores de â seguem uma distribuição normal de média μ e de variância conhecida σ^2 sabe-se que um intervalo de confiança para o valor de μ será $[\bar{\hat{a}} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{obs}},\ \bar{\hat{a}} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{obs}}]$, como não se sabe σ utiliza-se o desvio padrão de â, da amostra. Resultando assim num intervalo de confiança a 95% em $[0.407045, 0.419641]$. Dado que a, o valor exato, não se encontra dentro do intervalo de confiança pode-se considerar que a estimativa encontrada, não é uma boa estimativa. Sendo o valor exato inferior aos do intervalo de confiança, indo assim de encontra à conclusão anterior de se obter uma sobrestimação.

Métodos de reamostragem, *bootstrap*, são técnicas estatísticas que fazem uso de um só conjunto de dados e criam diversas amostras simuladas partindo deste. O propósito passa por se utilizarem os próprios dados para se conseguir efetuar o estudo de variações estatísticas efetuadas dos próprios dados, permitindo assim estimar intervalos de confiança em vez de estimativas pontuais.

Amostragem é o processo pelo qual se escolhe um certo número de elementos ao acaso de um conjunto, podendo este ser feito com ou sem substituição. A diferença encontra-se no ponto em que ao fazer-se sem substituição não existe reposição pelo que não se terão elementos repetidos. Com substituição, já será possível ter elementos repetidos pois há reposição.

O *bootstrap* empírico parte da ideia de que o conjunto de dados em análise, a amostra, ser grande o suficiente para, através da lei dos grandes números, se conseguir obter uma representação da população. Através da reamostragem dos dados originais $x_1, x_2, \dots, x_n$ no conjunto $x_1^*, x_2^*, \dots, x_n^*$ de tamanho n escolhendo valores do conjunto original, podendo-se definir uma estatística nova, v^*, para qualquer estatística v original através dos valores da

reamostragem. O tamanho dos dados reamostrados poderiam ser menores a n, contudo, querendo-se controlar a variação da estatística, como esta depende do número de elementos em teste necessita de ser igual.

Assim para calcular o *bootstrap* empírico parte-se do conjunto de valores de $\hat{a}$ e aceita-se o seu valor médio, $\bar{\hat{a}}$ ou $\bar{x}$, como sendo a verdadeira média da população, $\bar{a}$. De seguida, efetua-se um certo número de vezes suficientemente grande a reamostragem de tamanho N com reposição, e calcula-se o desvio da média desta amostra, $\bar{x}^*$, a $\bar{\hat{a}}$, $\bar{x}^* - \bar{\hat{a}}$. Guardando-se esta diferença num vetor.

O histograma destes valores pode ser verificado na Figura 3. Onde se verifica a semelhança a uma distribuição normal de média perto de 0. Analisando de facto, o valor médio este foi de $-1.283748e-05$ com uma variância de $1.035682e-05$.

Para se obter um intervalo de confiança a 95% escolhe-se $\alpha = 0.05$ pelo que ao efetuarem-se os quantis $q_1 = \frac{\alpha}{2}$ e $q_2 = 1 - \frac{\alpha}{2}$ dos chega-se ao intervalo de confiança $[\bar{x}^* - q_2,\ \bar{x}^* - q_1]$, que resulta no intervalo de $[0.406728, 0.419503]$.

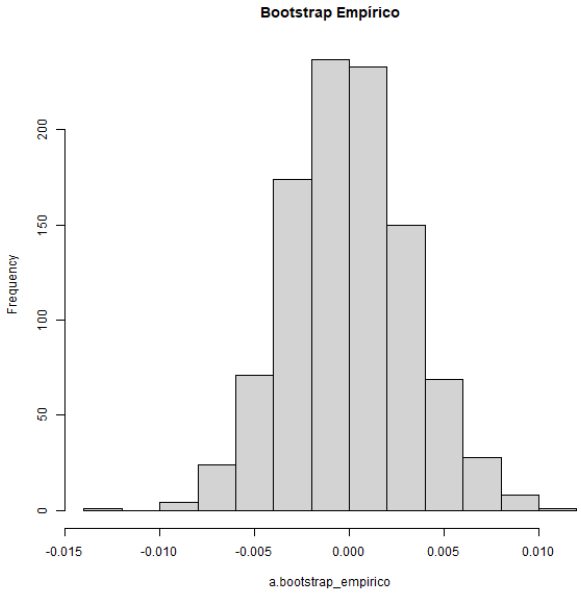

Figura 3 – Histograma dos valores do bootstrap empírico

O método de *bootstrap* percentil em vez de se efetuar o cálculo das diferenças utiliza a distribuição das amostras de *bootstrap* como aproximação direta. Efetuando esta alteração em relação ao anterior obtém-se um valor médio de 0.413330 com variância $1.035682e-05$ e um intervalo de confiança de $[0.407182, 0.419958]$. O seu histograma encontra-se apresentado na Figura 4.

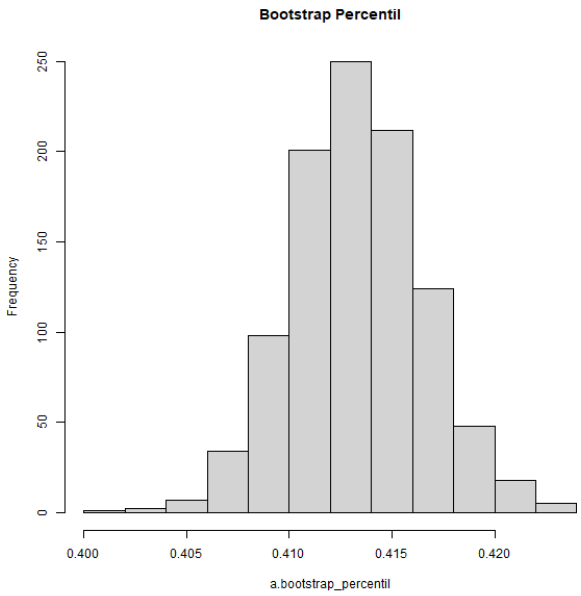

Figura 4 – Histograma dos valores do bootstrap percentil

Pela literatura disponibilizada seria de esperar que estes intervalos tivessem amplitudes diferentes, levando a que o método percentil fosse menos exato. Contudo a amplitude de ambos, calculada com os valores exatos no R, tem o valor, aproximado, de 0.012776. Mostrando-se assim que nesta situação não existiu diferença significativa entre um método e o outro.

Apesar deste último método ser mais apelativo devido a ser mais simples este irá depender mais fortemente de $\bar{x}^*$ e de se a sua amostra é uma boa aproximação a $\bar{x}$. Contudo, existem situações em que a sua utilização poderá ser útil, quando é possível visualizar o viés da distribuição das diferenças, como é apresentado na Figura 3. Verificando-se a existência de outros métodos de *bootstrap* que apresentam uma melhor precisão.[1]

b) Pretendendo-se agora testar o teste de hipótese $H_0: \sigma^2 = 0.4(4)$ vs $H_1: \sigma^2 < 0.4(4)$ em que $X_t \sim N(0, \frac{a}{1-b})$ sob a condição de estacionaridade. Considerando-se a estatística de teste da variância de uma população normal que $T = \frac{(n-1)S_c^2}{\sigma^2} \sim \chi^2_{n-1}$, estimando-se ao nível de significação de $\alpha = 0.05$ usando 1000 réplicas no modelo X_t. A dizima infinita periódica $0.4(4)$ pode ser representada na forma fracionária por $\frac{4}{9}$.

O princípio desta técnica passa por se assumir a estatística de teste, H_0, e comparar com a regra de teste, aceitando ou rejeitando a decisão. Na prática, calcula-se o valor do quantil α da distribuição χ^2_{n-1}, em que n será o número de observações. É se testado face ao quantil α devido a ser um teste para se é inferior.[2] Para cada uma das réplicas assume-se H_0 gerando um número de obs, neste caso segundo uma distribuição Normal $(0, \frac{a}{1-b})$, e contabilizam-se os casos em que não se aceita H_0 sendo H_1 verdade, ou seja, se ao aplicar $\frac{(n-1)S_c^2}{\sigma^2}$ aos valores da distribuição, o valor que se obtém é menor ou igual ao quantil da distribuição de χ^2_{n-1}.

Contabilizando as vezes que o valor de teste não é aceite com o valor de 1 e as vezes que é aceite com 0, efetuando assim a média consegue-se obter um valor que a percentagem de rejeições. Quanto mais perto de 0 este for maior confiança se poderá ter em não rejeitar que H_0 é verdade.

Desta forma, o código proposto para a resolução deste problema pode ser descrito, resumidamente, pelos seguintes passos:

1. Inicialização das variáveis, $\alpha = 0.05$ e $\sigma = \frac{4}{9}$
2. Criação do valor critico, *vcrit*, quantil $\chi^2_{\alpha,n-1}$
3. Criação de um vetor, *testrule* com N valores, sendo N o número de réplicas
4. Ciclo de 1 a N:
    a. Retirar um valor, x, ao acaso da distribuição pretendida
    b. Decidir se se rejeita ou não, 1 ou 0, caso o valor de $\frac{(n-1)var(x)}{\sigma}$ seja inferior ao valor critico, *vcrit*, sendo inferior o valor é rejeitado.
    c. *Testrule* fica com o valor 0 ou 1 se é rejeitado ou não, respetivamente.
5. A percentagem de rejeição é dada pela média do vetor *testrule*.

Deste modo procedeu-se à aplicação deste algoritmo para os valores sugeridos no enunciado:

[1] Mais fontes foram consultadas como https://stats.stackexchange.com/q/357498 e https://journals.sagepub.com/doi/full/10.1177/2515245920911881

[2] Consultou-se o seguinte website para adaptar o que se desenvolveu na aula em que se efetuou o teste para se era superior https://www.itl.nist.gov/div898/handbook/eda/section3/eda358.htm

i. Com $a = 0.4$ e $b = 0.1$

Nesta situação obteve-se que a média das contabilizações das vezes em que o valor de teste não foi aceite foi de 0.053 ou seja um valor muito perto de α, o nível de significância. Ao testar-se este método mais vezes e com outros valores de SEED este alterou-se e por vezes era inferior a α, levando assim a situações em que não se rejeitava H_0, ao contrário desta, bem como houve outras vezes que o valor era superior a α, rejeitando então a hipótese H_0 da igualdade $\sigma^2 = 0.4(4)$. Na Figura 5 é possível verificar o histograma da contabilização das vezes em que este se aceita $H_0(0)$ e das vezes em que é rejeitado segundo $H_1(1)$.

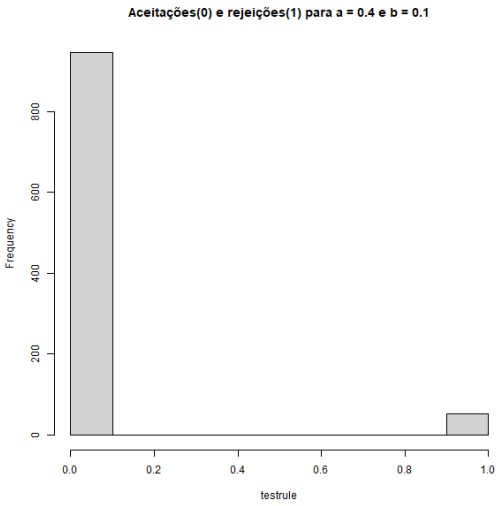

Figura 5 - Contabilização das aceitações (0) e das rejeições (1) do teste de hipótese para a = 0.4 e b=0.1

ii. Com $a = 0.1$ e $b = 0.1$

Para este caso a média das contabilizações das vezes em que o valor de teste não foi aceite foi de 1, ou seja, rejeitou-se H_0 como sendo verdade em todos os casos. Podendo-se verificar na Figura 6 o histograma aqui obtido para estas contabilizações.

Dado que $X_t \sim N(0, \frac{a}{1-b})$, ao utilizar-se $a = 0.1$ e $b = 0.1$ obtém-se uma $N(0, \frac{1}{9})$, ou seja, $\sigma^2 = 0.1(1)$ o que é, por sua vez, muito menor que $0.4(4)$ pelo que faz sentido que se rejeite mais vezes a hipótese H_0. Seguindo a mesma linha de raciocínio, considerando-se $a = 0.4$ e $b = 0.1$ obtém-se uma $N(0, \frac{0.4}{1-0.1})$, ou seja, $\sigma^2 = 0.4(4)$, um valor igual ao qual se pretende testar contra, levando há existência de situações em que se não se rejeite e outras em que se rejeite H_0. Isto deve-se aos valores das amostras obtidas através da distribuição de base que não conseguem generalizar a distribuição geral.

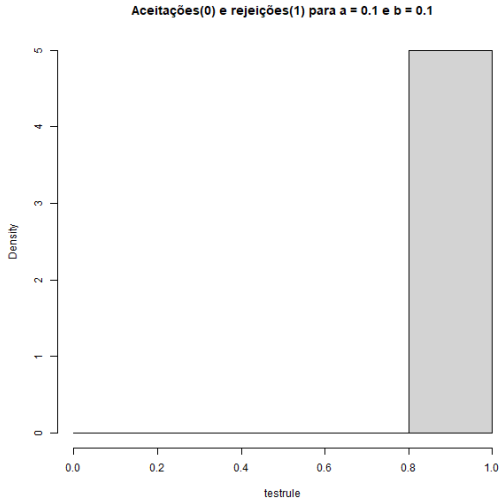

Figura 6 – Contabilização das aceitações (0) e das rejeições (1) do teste de hipótese para a = 0.1 e b=0.1