EP MAE0699 - Tópicos de probabilidades

Fabricio Kassardjian nusp:2234961
Robert Mota dos Santos nusp:9039927

15 de maio de 2019

Introdução

O trabalho se refere a estudar as curvas de distribuição para $T(v)$ e $C(v, w)$, onde **T** representa o caminho mais curto de retorno ao vértice v e **C** o caminho mais curto entre os vértices v e w. O modelo de Erdo-Rényi é utilizado para gerar os grafos aleátorios com n vértices e probabilidade p de ligação entre cada par de vértices.

Modelo 2

O modelo escolhido para o teste foi usando conexões não direcionadas e preguiçoso. O fato de ser preguiçoso implica que existem conexões para continuar no mesmo vértice. Além disso fica determinado que cada connexão só pode ser usada uma unica vez para cada caminho testado, assim evita-se que a distribuição **T** tenha apenas valores 1 e 2. Como as conexões são aleatórias podem existir vértices isolados e também como não pode ser utilizado a mesma conexão para voltar podem existir valores de **T** e **C** que podemos considerar inf.

Metodologia

Estimação das distribuições

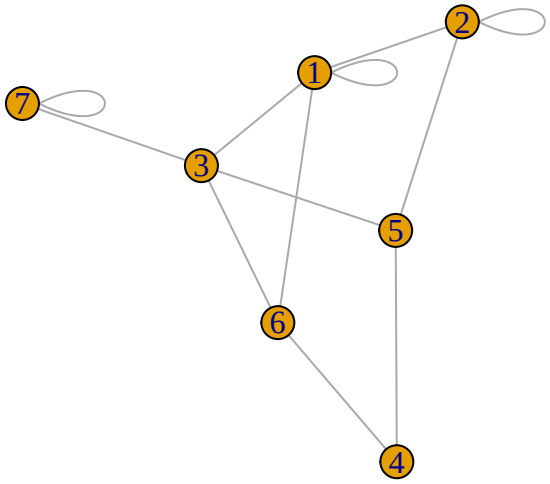
Para o teste primeiro inicia-se uma matriz A representando com **TRUE** quando a ligação entre os vértices está presente e **FALSE** quando não há ligação. A linha i da matriz representa o vértice de saída e a coluna j representa o vértice de chegada. Como o modelo é preguiçoso pode existir **TRUE** na diagonal principal da matriz, e pelo fato das conexões não serem direcionadas a matriz é simétrica.

Exemplo de matriz de conexões para 7 vértices com probabilidade de conexão 0.4:

```
n = 7
A = generateMatrix(n, 0.4)
print(1*A) #1* para deixar em formato numerico
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,]    1    1    1    0    0    1    0
## [2,]    1    1    0    0    1    0    0
## [3,]    1    0    0    0    1    1    1
## [4,]    0    0    0    0    1    1    0
## [5,]    0    1    1    1    0    0    0
## [6,]    1    0    1    1    0    0    0
## [7,]    0    0    1    0    0    0    1
```

```
plot(graph_from_adjacency_matrix(A, mode = 'undirected', weighted = TRUE))
```



Para cada mariz gerada é testado para cada vértice o menor caminho de volta, usando uma busca em profundidade dos caminhos possíveis da matriz e armazenado o vetor com a contagem de cada valor para **T** encontrado. O mesmo é feito para cada combinação de vértices possíveis para encontrar os valores de **C**. Os caminhos podem ter tamanhos até n e iremos considerar o valor $n+1$ como sendo infinito.

Exemplo de valores de $T(v)$ para cada vértice de **A**

```
for(i in 1:n) {
  Ti = findPath(i,i,A,0,n+1)
  cat(sprintf("T(%d) = %d\n",i, Ti))
}
```

```
## T(1) = 1
## T(2) = 1
## T(3) = 3
## T(4) = 4
## T(5) = 4
## T(6) = 3
## T(7) = 1
```

Será gerado para cada tamanho de $n \in \{6, 7, 8, 9, 10, 11, 12\}$ uma amostra de 300 matrizes e feita uma contagem para cada valor de **T** encontrado. A distribuição é estimada tirando a média da contagem por $n * 300$. Assim:

$$\hat{P}(T = k) = \frac{1}{n * 300} \sum_{i=1}^{n*300} \mathbb{1}_{(T=k)}$$

Para a distribuição de C é usado processo similar mas como temos as combinações entre os pares serão estimados $n * n$ valores para cada matriz, assim:

$$\hat{P}(C = k) = \frac{1}{n^2 * 300} \sum_{i=1}^{n^2*300} \mathbb{1}_{(C=k)}$$

Tamanho da amostra

Para determinar um tamanho bom de amostra para a aproximação da estimação, fixamos $n = 8$ e geramos a distribuição e o gráfico para alguns tamanhos de amostra ($N \in \{25, 50, 75, 100, 150, 200, 250, ..., 750, 800\}$). Depois calculamos a soma das diferenças ao quadrado entre os valores de cada distribuição e colocamos em um gráfico. No gráfico pode ser verificado se houve convergência e com que tamanho de amostra podemos considerar a convergência.

$$erro = \sum_{i=1}^{n+1} \big(P(T = k) - P'(T = k)\big)^2$$

onde $P(T = k)$ é a probabilidade para o tamando de amostra atual e $P'(T = k)$ a probabilidade da amostra anterior.

Teste de aderência

Após determinar um função de densidade de probabilidade que define a distribuição para os valore de $P(T = k)$ e $P(C = k)$, usaremos um teste de aderência para validar a hipótese. Testamos a H_0: população segue distribuição proposta contra a H_1: população tem outra distribuição. Para ta usaremos a seguinte estatística de teste:

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

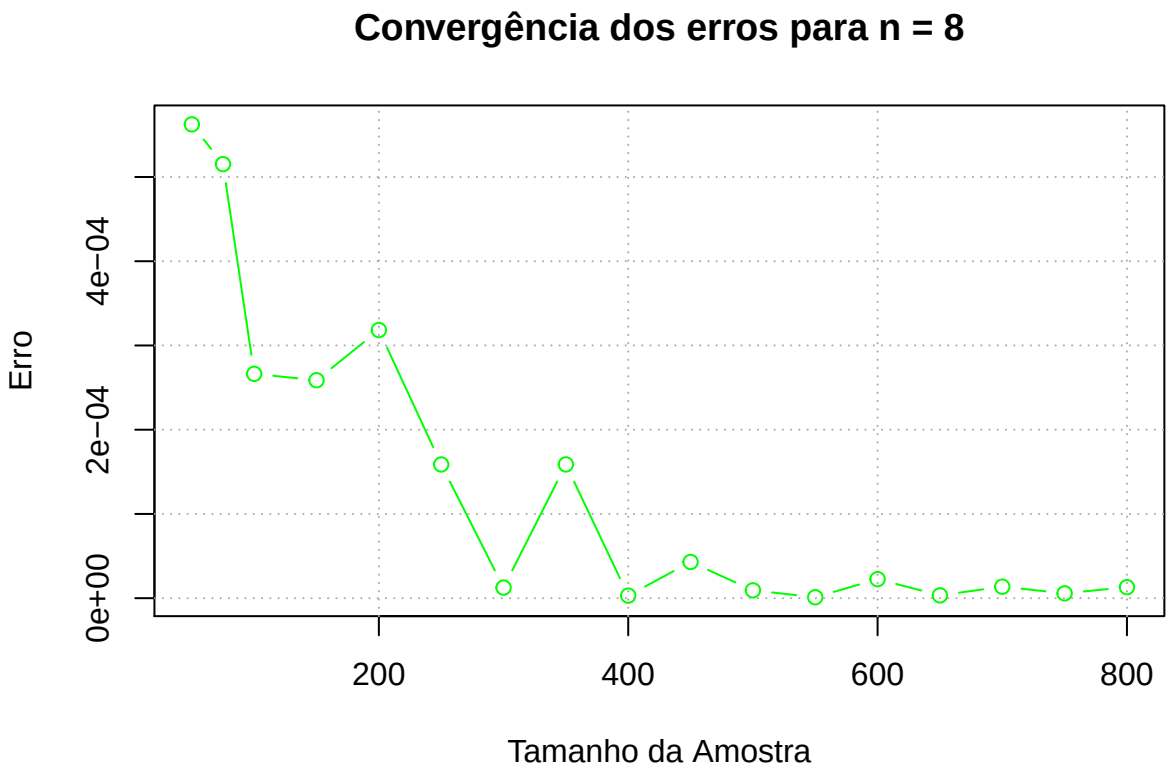
Onde O_i são as frequências observadas na simulação, $E_i = P(T = i) \times N$ a frequência esperada e K representa a quantidade de pontos da distribuição. Assim χ^2 tem uma distribuição chi-quadrada com $K - 1$ graus de liberade. Com o teste calculado definimos a região critica como $RC = (c, \infty)$ onde $P(\chi^2_{s-1} > c) = \alpha$, sendo α o nível de significância para o nosso teste.

Dessa forma após as simulações e escolhida uma distribuição provável que explique a frequencia observada, é feita a soma dos calculos das diferenças, e o resultado comparado com c. Se o valor calculado for maior que c rejeitamos nossa hipótese H_0 com o nível de significância escolhido, e a distribuição selecionada não é aderente as observações. Mas se o valor calculado for menor que c então não rejeitamos H_0 e podemos considerar que a variável aleatória T tem a distribuição do modelo sugerido.

Simulação

Tamanho da amostra

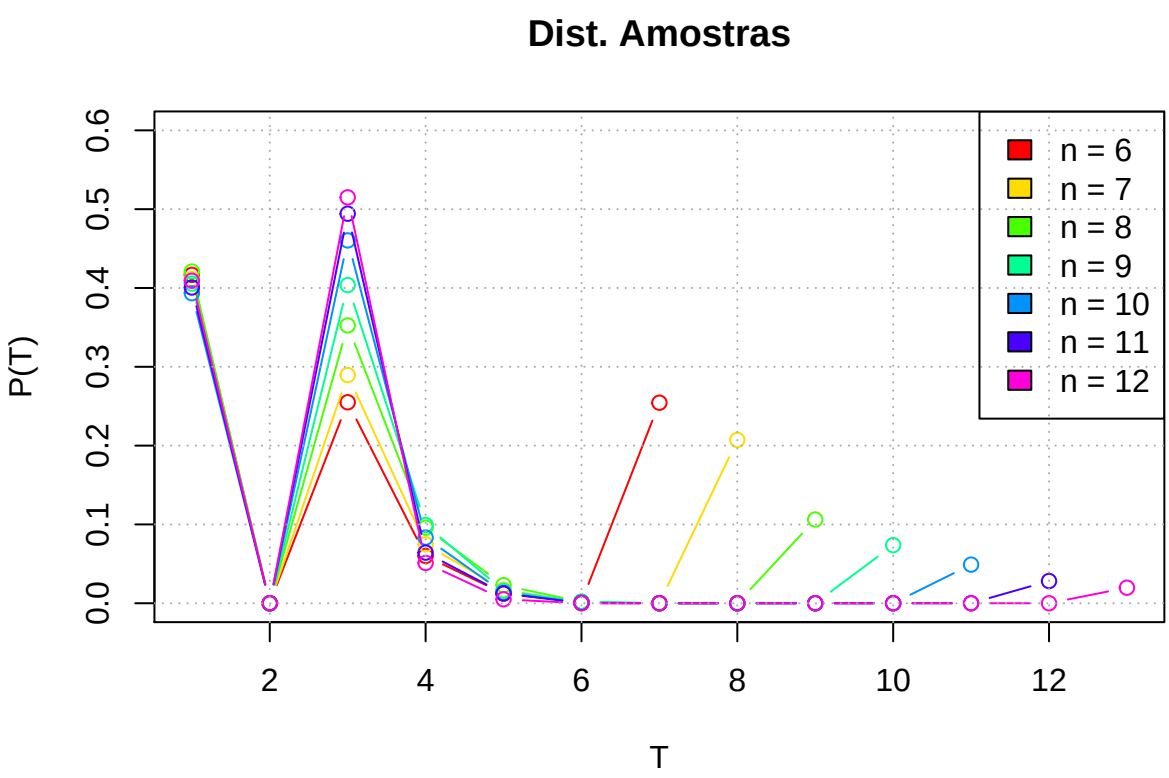
Gráfico para vários tamanhos de amostra com $p = 0.4$:



Pelo gráfico podemos considerar uma amostra com tamanho 300 razoável para as estimações de distribuição.

Distribuições para T

Valores estimados e gráficos da distribuição, usando 300 amostras, e $p = 0.4$



Proposta de distribuição proposta para T

Ao realizarmos os testes, percebemos que nossa distribuição aparenta ter caracteristicas e formas de algumas distribuições conhecidas, sendo elas: Poisson, Exponencial e Geométrica. Neste caso vamos tentar identificar qual distribuição é mais 'próxima' da nossa distribuição e quais são os melhores paramêtros da distribuição escolhida.

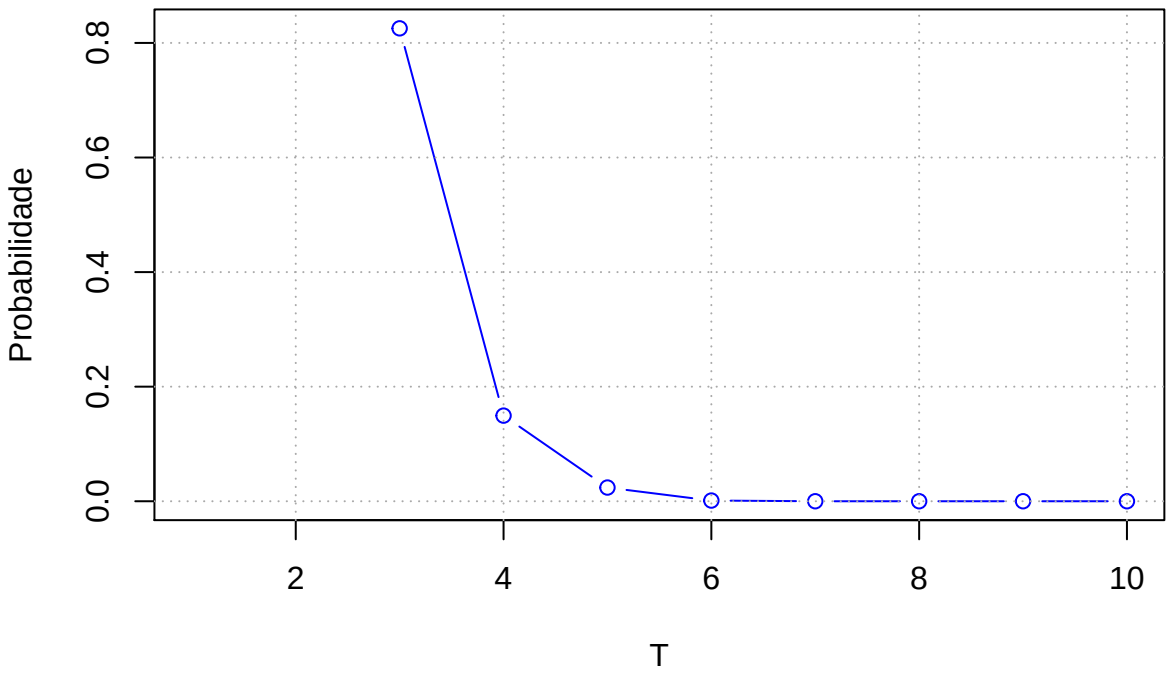
Premissas iniciais:

Nossa simulação é um processo discreto, além disso realizamos algumas mudanças em nossas distribuições de T. Primeiramente, retiramos a observação 2 de nosso processo, porque o problema definido não permite que o tempo minimo de volta seja 2, assim: $O_2 = 0$.Retiramos, também, a observação 1 porque, de fato, $P(O_1) = p$, pois o problema é definido para podermos voltar ao mesmo nó com probabilidade p , fica bem definido assim $P(O_1) = p$ e por fim e não menos importante, retiramos T_{inf} por ser um ponto degenerado na nossa distribuição esperada.

Caracteristicas da simulação:

- Número de vértices: 10
- Número de repetições: 300
- Probabilidade de conexão entre vértices: 0.4

A distribuição sem as observações citadas tem uma distribuição reescalada na forma. A escala foi modificada na forma $P(T = k|k \in \{3, 4, 5, 6, 7, 8, 9, 10\})$, ou seja condicionamos as probabilidades aos valores de interesse.



Parâmetros da distribuição:

Para encontrar os parametros das distribuição vamos usar as leis dos grandes números que diz que a média da amostra converge para a média da população quando o tamanho da amostra cresce, em outras palavras:

$$\lim_{n->\infty} \hat{\mu_n} \to \mu$$

Usaremos estas premissas para encontrar os parâmetros da distribuição . Sendo assim devemos conhecer a esperança da distribuição geométrica,poisson e exponêncial:

$$E(X \sim Geo(p)) = \frac{1}{p}$$
$$E(X \sim Exp(\lambda)) = \lambda$$
$$E(X \sim Pois(\lambda)) = \lambda$$

A partir da simulação encontramos os parâmetros das distribuições como se segue:

Geométrica	Exponêncial	Poisson
E(x) = 0.833	E(X) = 1.201	E(x) = 1.201

Realizando simulações com as premissas citadas e com os parâmetros encontrados obtemos os seguintes distribuições:

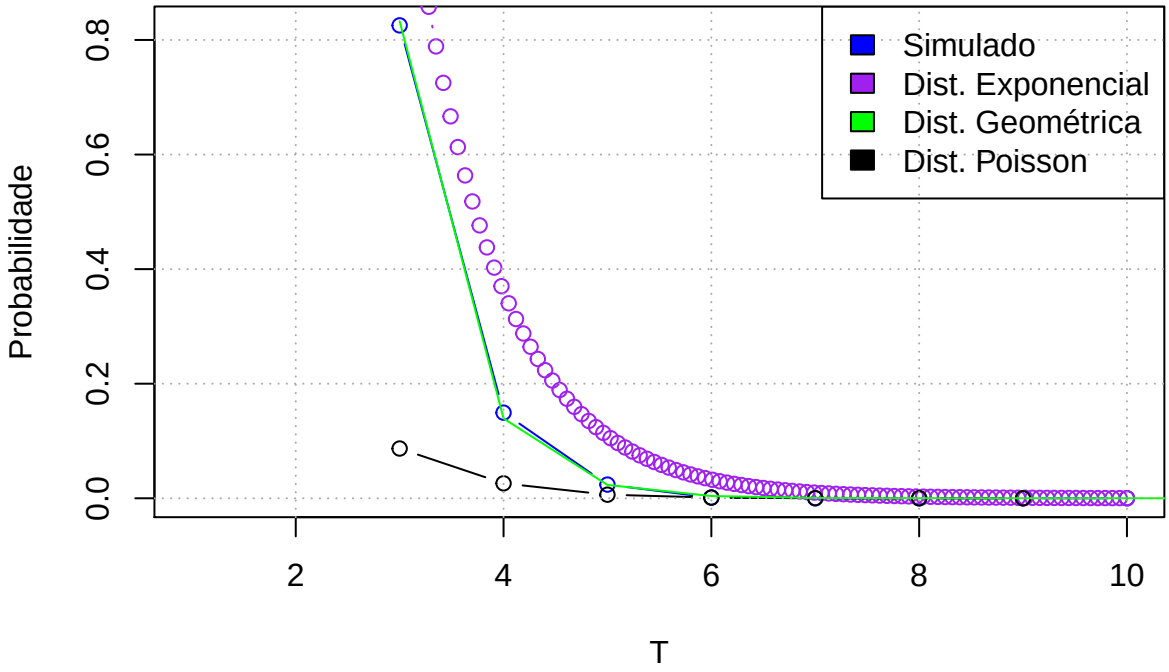
x	3	4	5	6	7	8	9	10
Exponencial	1.2008	0.3614	0.1088	0.0327	0.0098	0.0030	8.92e-04	2.68e-04
Geométrica	0.8328	0.1393	0.0233	0.0039	0.0007	0.0001	1.82e-05	3.0481e-06
Poisson	0.3009	0.3614	0.2170	0.0869	0.0261	0.0063	1.25e-03	2.1500e-04
Dist. Simulada	0.8255	0.1494	0.0239	0.0012	0.0000	0.0000	0.0000	0.0000

Para chegar nos resultados da tabela acima foram utilizados o seguinte calcúlos:

$$E(T) = \sum_{i=1}^{n} T_i * P(T = i) \rightarrow p = \frac{1}{E(T)}$$

para cada uma das distribuições.

A seguir o resultado de cada uma das distribuições com os parâmetros encontrados:



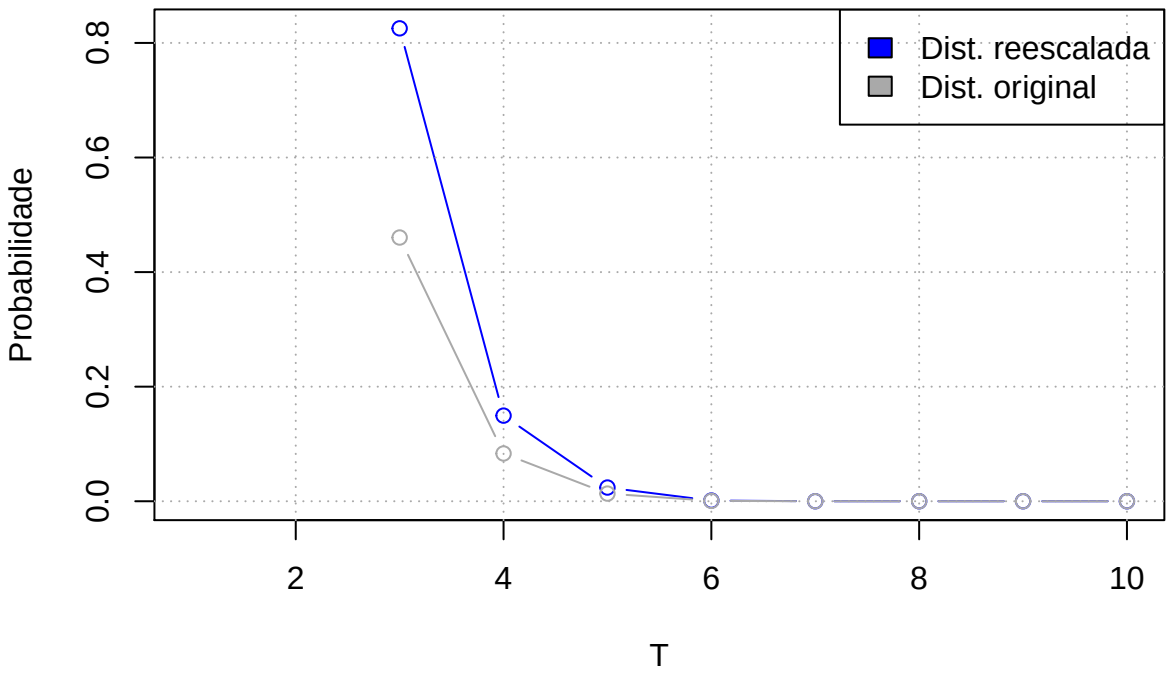
A partir destas distribuições tomamos aquela que possui menor distância entre a distribuição simulada e as distribuições propostas.

Como pode ser visto na imagem acima a distribuição que será assumida será a Geométrica, pois é a que se aproxima mais da nossa distribuição simulada.

Em seguida iremos realizar o teste de χ^2 para avaliarmos se estamos cometendo o erro na escolha da distruibuição.

Distribuição reescalada e distribuição proposta

Queremos agora, voltar para distribuição original a partir da distribuição reescalada. Veja abaixo ambas as distribuição:



Foi necessario reescalar a distribuição e retirar dados pois a mesma iria interferir na comparação com uma distribuição já conhecida, que neste caso foi a geométrica. Portanto, temos que voltar a distribuição reescalada para distribuição original.

```
#n = 10
#dT = distT[3:n] ##Retiramos os nós 1, 2 e inf (Premissa)
#dTp = dT / sum(dT) ##Distribuição reescalada (Azul)
#dTP = dT / sum(distT) ##Distribuição Original considerando todos os valores.
```

Como queremos sair da distribuição Azul(reescalada) para a distribuição cinza(orginal), devemos multiplicar por um fator $(1 - p - T_\infty)$:

Realizando esses passos podemos encontrar a distribuição reescalada para a forma original:

Agora é possivel reeconstruir toda a distribuição da simulação,porque sabe-se que $P(T = 2) = 0$ e $P(T = 1) = p = 0.4$ e que $P(T =' \infty') = 1 - P(T <= n)$, todas as informações foram citadas na seção de premissas.

E finalmente temos nossa distribuição reconstruida completamente.

A distribuição proposta é:

$$T \sim \begin{cases} P(T = 1) = p \\ P(T = 2) = 0 \\ P(T = k) \sim Geo(g) * (1 - p - T_\infty), \text{usando valor k-3 para } k \in \{3, 4, .., n\} \\ P(T = \infty) = 1 - p - \sum_{i=3}^{n} P(T = i) \end{cases}$$

Sendo $g = 1/E[T - 2]$ para $T \in \{3, 4, 5, ..., n\}$, ou seja, é a média da amostra observada do ponto 3 em diante, porém deslocados em duas unidades.

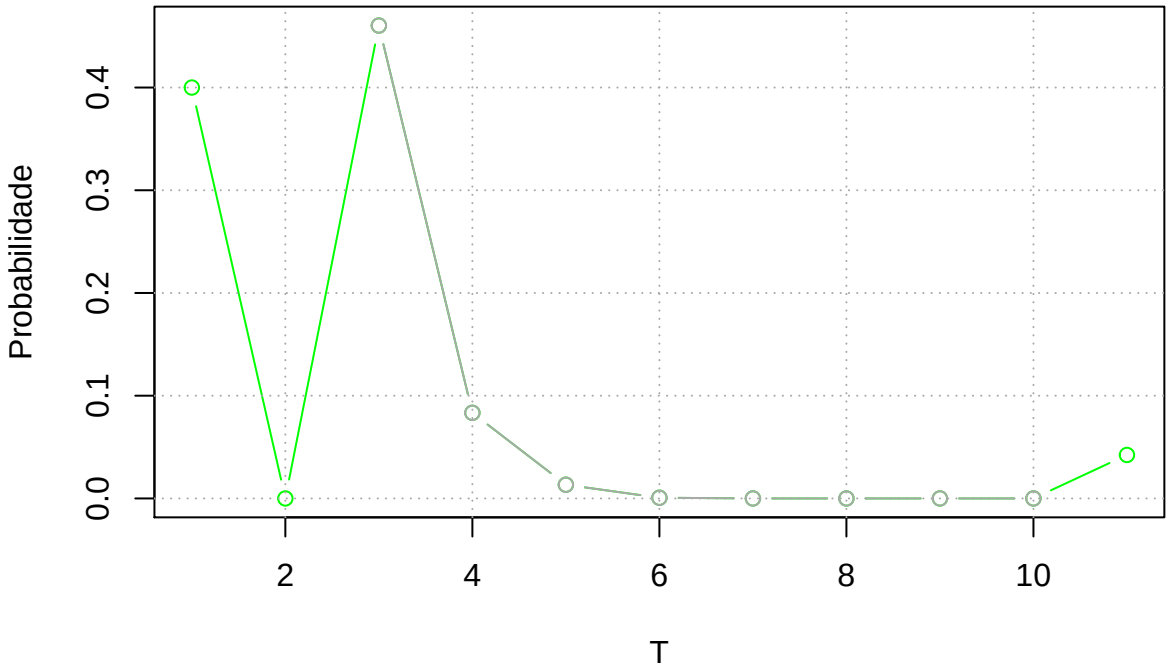


Figure 1: Distribuição completa com base na distribuição geométrica

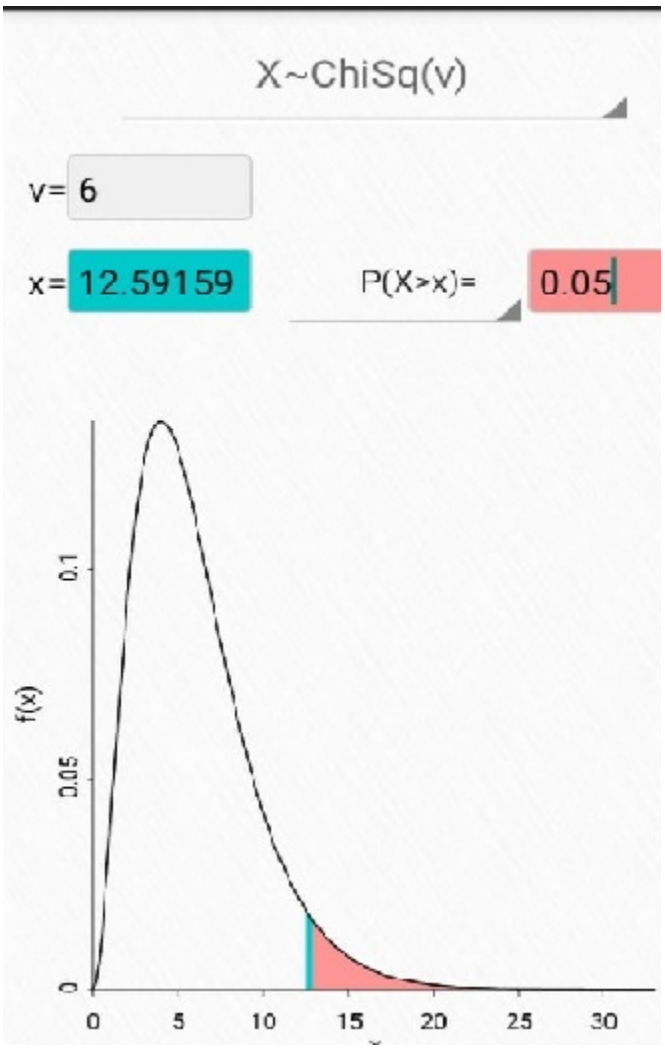
Teste de aderência:

Com nossa função proposta em mãos, podemos realizar o teste χ^2 com $\alpha = 5\%$ de tolerância. Para avaliarmos se estamos ou não comentendo o erro de escolher a distribuição Geométrica com parâmetro ($p = 0.8327526$).

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

Onde O_i são as frequências observadas na simulação e $E_i = P(T = i) \times N$ com $T \sim (Geometrica)$. K representa o número de repetições.

Teste de aderência para T



Hipótese nula: Distribuição proposta seja igual a distribuição T simulada.

Hipótese alternativa: Distribuição proposta seja diferente da distribuição T simulada.

Nossa região critica para $\alpha = 5\%$ é $RC = \{x : \mathbb{R}, (16.9189776, +\infty)\}$

O valor de $\chi^2(s = (9)) = 0.0021527$. Sendo assim assumimos que nossa distribuição proposta tem boa aderência com significância de $\alpha = 5\%$. Como nosso $\chi^2 \notin RC$ sendo assim não rejeitamos a nossa hipótese nula.

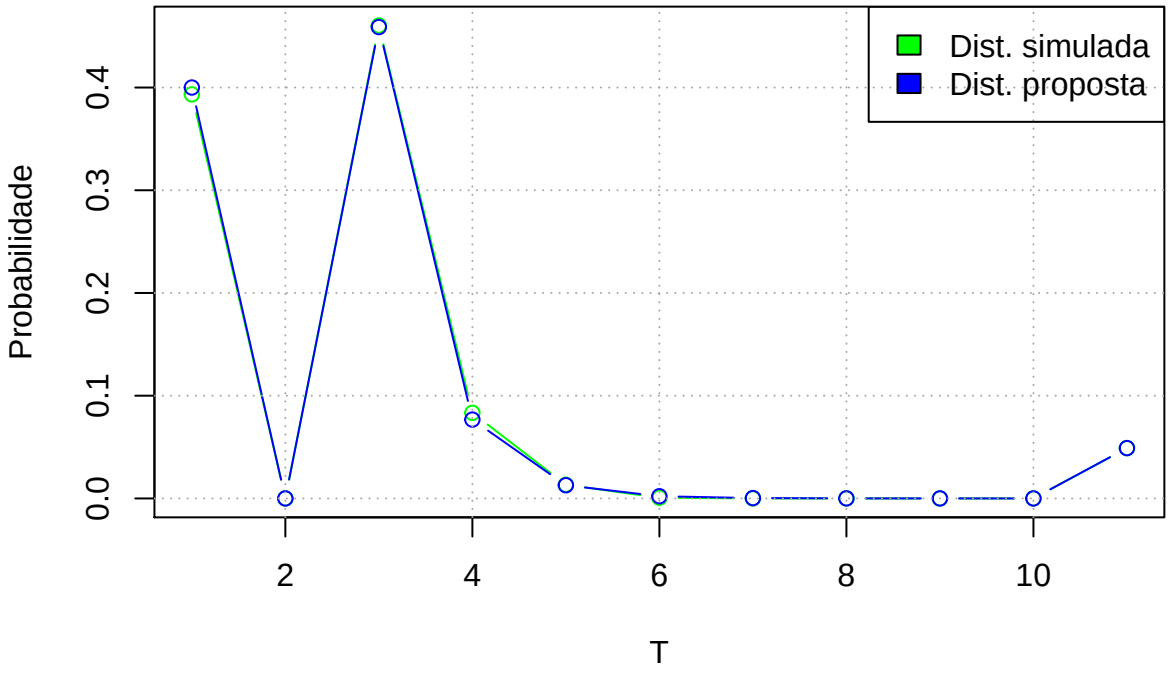
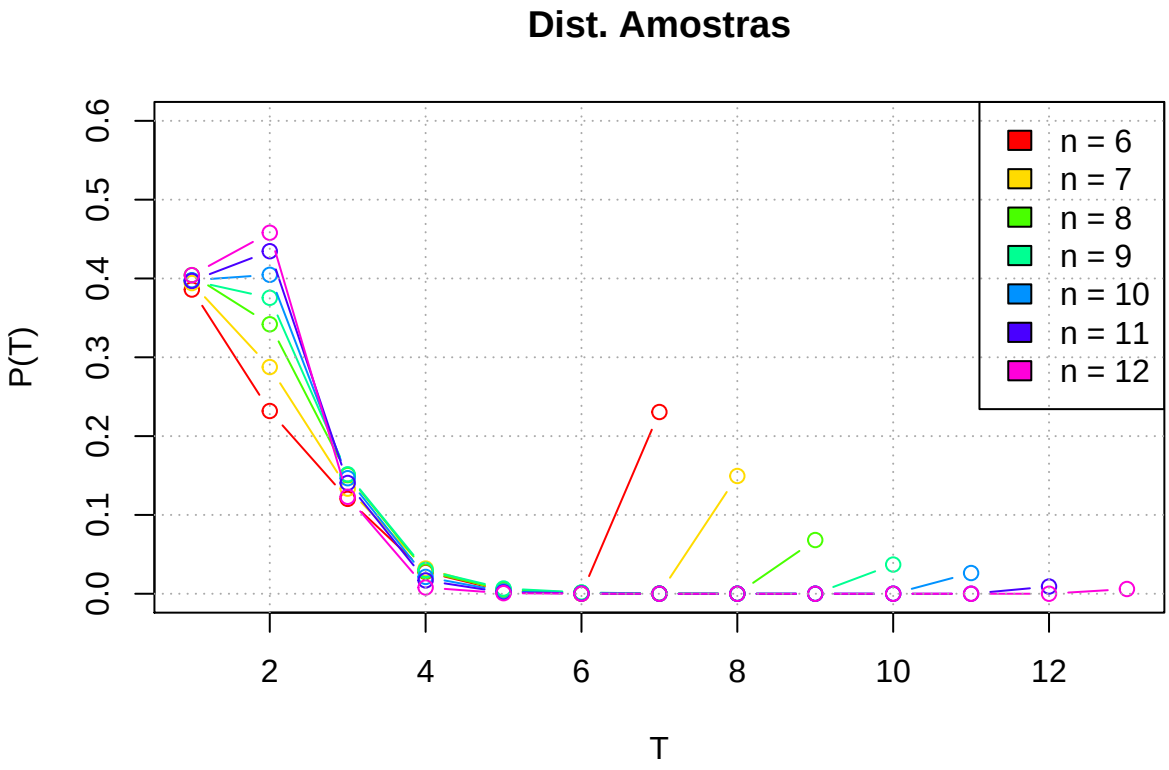


Figure 2: Proposta x Simulada

Notas: Consideramos 9 graus de liberdade pois retiramos $T = 2$

Distribuições para C

Valores estimados e gráficos da distribuição. Como a quantidade de testes aumenta em n vezes para a distribuiçao C, reduzimos a quantidade de matrizes geradas pela metade, e portanto usamos uma amostra de tamanho 150 com $p = 0.4$.



Premissas

De forma análoga a distribuição de T, retiramos desta vez apenas a observação 1, ou seja $C(v, v') = 1$, porque $P(C = 1) = p = 0.4$ e a observação 11 (C_∞).

Caracteristicas da simulação:

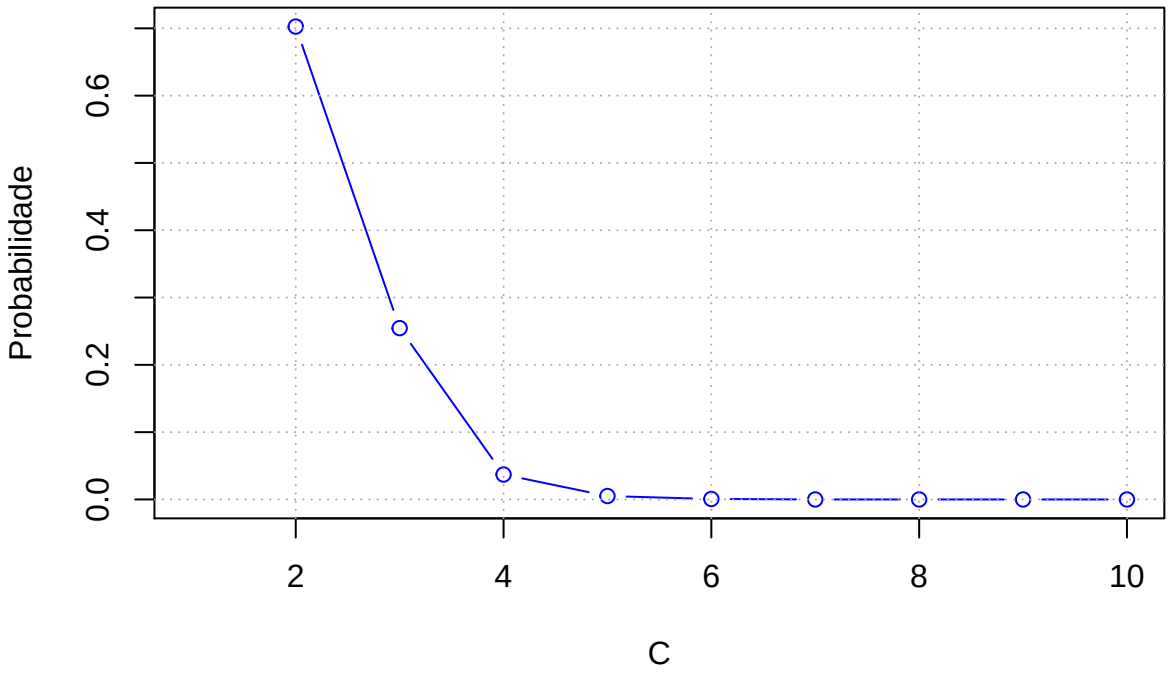
- Número de vértices: 10
- Número de repetições: 150
- Probabilidade de conexão entre vértices: 0.4

Distribuição para C

Podemos seguir o mesmo raciocinio da distribuição T para inferirmos sobre a distribuição C. Portanto podemos calcular diretamente:

$$E(X \sim Geo(p)) = \frac{1}{p}$$

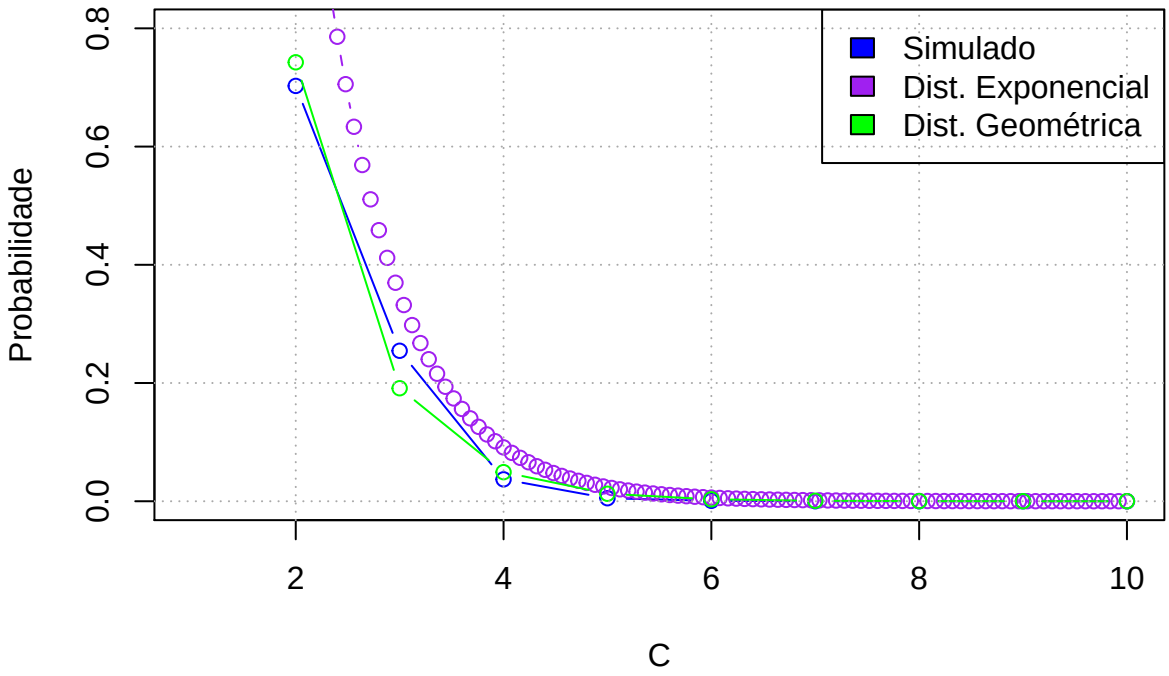
E verificar se essa distribuição se aproxima da nossa distribuição simulada. Veja inicialmente, que a distribuição para C com as premissas citadas possui o seguinte gráfico:



Realizando as simulações conseguimos encontrar o parâmetro p por meio da esperança

$$E(X \sim Geo(p)) = 1.3466836 = \frac{1}{p} \rightarrow p = 0.7425649$$

Assim podemos comparar a distribuição simulada com a distribuição geométrica com paramêtro $p = 0.7425649$ e a exponêncial com $\lambda = 1.3466836$. Obtemos assim os seguintes gráficos:



Tomamos como hipótese de distribuição aquela que obteve a menor distância com a distribuição simulada, que neste caso foi a distribuição Geométrica.

Distribuição reescalada e distribuição proposta:

Queremos agora voltar a distribuição original da simulação, ou seja, sem retirar as observações 1 e 11 como haviamos feito anteriormente.Seguindo o mesmo raciocinio da distribuição T.

Usando a mesma metodologia que na distribuição C, podemos reescalar a distribuição geométrica para ficar compatível com a simulada. É possivel reeconstruir toda a distribuição da simulação,porque sabe-se que $P(C = 1) = p = 0.4$ e que $P(C = \infty) = 1 - P(C <= n)$

E finalmente temos nossa distribuição reconstruida completamente.

A distribuição proposta é:

$$C \sim \begin{cases} P(C = 1) = p \\ P(C = k) \sim Geo(g) * (1 - p - C_{\infty}), \text{usando valor k-2 para } k \in \{2, 4, .., n\} \\ P(C = \infty) = 1 - p - \sum_{i=2}^{n} P(C = i) \end{cases}$$

Sendo $g = 1/E[C - 1]$ para $C \in \{2, 4, 5, ..., n\}$, ou seja, é a média da amostra observada do ponto 2 em diante, porém deslocados em uma unidade.

Teste de aderência para C

Tendo como hipótese nossa dsitribuição geométrica podemos realizar o teste de aderência para verificar a adequabilidade da distribuição simulada com a distribuição proposta.

De modo análogo podemos calcular: χ^2 com $\alpha = 5\%$ de significância, para avaliarmos se estamos ou não comentendo o erro de escolher a distribuição proposta.

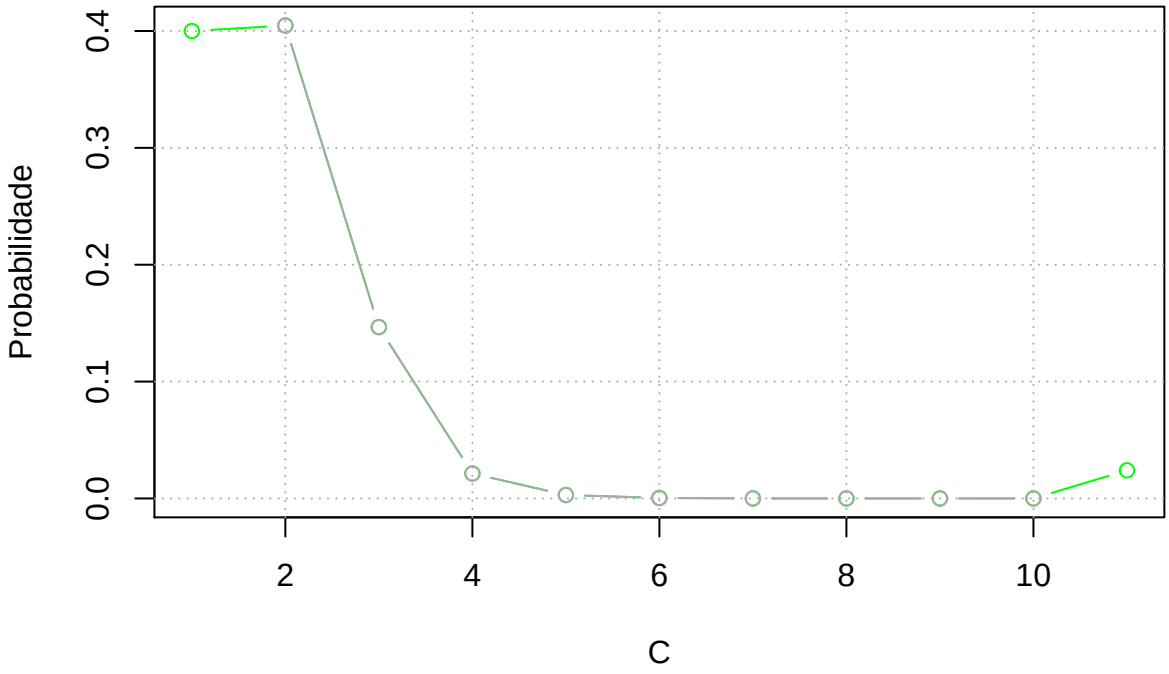


Figure 3: Distribuição completa com base na distribuição geométrica

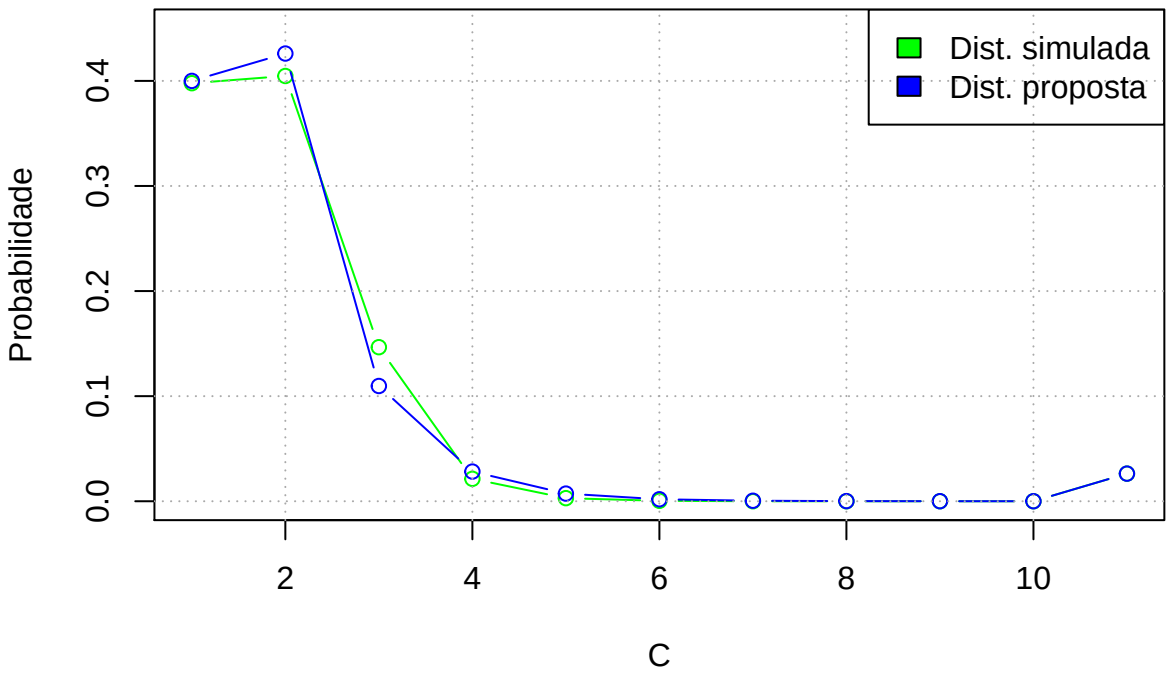


Figure 4: Proposta x Simulada

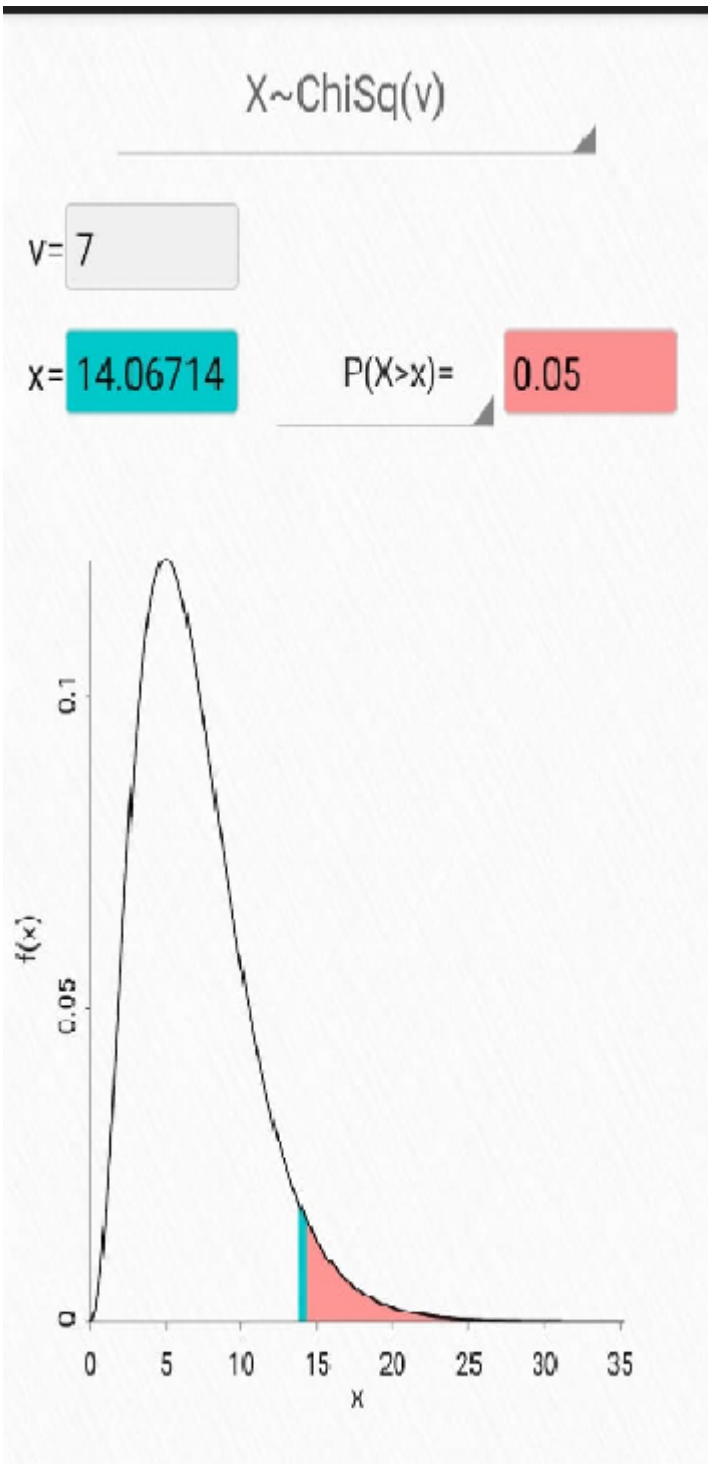
$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

Podemos definir nossas hipóteses para este modelo da seguinte forma:

Hipótese nula: Distribuição proposta seja igual a distribuição C simulada.

Hipótese alternativa: Distribuição proposta seja diferente da distribuição C simulada. Nossa região critica para $\alpha = 5\%$ é $RC = \{x : \mathbb{R}, (18.3070381, +\infty)\}$

O valor de $\chi^2(s = (10)) = 0.0195897$. Sendo assim assumimos que nossa distribuição proposta tem boa aderência com significância de $\alpha = 5\%$. Como nosso $\chi^2 \notin RC$ sendo assim não rejeitamos a nossa hipótese nula.



Notas: Consideramos 10 graus de liberdade pois temos 9 elementos comparados.