Comparando medidas de correlação

Anaih Pastana e Arthur Rocha

Dezembro de 2017

1 Introdução

Uma das maiores preocupações ao se medir variávies é saber o quanto elas estão relacionadas. No caso de variáveis quantitativas, uma das formas mais comum é a correlação, calculada por meio de coeficientes. Esses coeficientes são quantidades entre -1 e 1, sendo que quanto mais próximo de -1 maior a correlação negativa, em contrapartida, ao se aproximar de 1 maior a correlação positiva.

Nesse trabalho considerou-se o uso das 3 principais medidas de correlação, a de Pearson, Spearman e Kendall. A correlação foi interpretada na sua forma mais simples, a quantidade de variabilidade que as variáveis tem em conjunto, desta forma partiu-se do pressuposto que os 3 coeficientes estavam medindo a mesma coisa.

O intuito foi verificar através de simulações qual coeficiente cometia menos erros considerando as situações do tamanho de amostra, correlação negativa, positiva e quase nula.

2 Metodologia

Para a geração de valores com determinada correlação utilizou-se a biblioteca *mvtnorm* e para se calcular as correlações, a função *cor* do pacote básico (*stats*) do R foi utilizada. Os códigos para geração dos valores e cálculo das correlações são dados abaixo:

```
library("mvtnorm")

s=p=k=NULL  ##Vetores vazios
v<-c(5,10,11:1000) ##Tamanho das amostras

set.seed(12345) ##Pra conseguir gerar os mesmos valores

for(t in 1:length(v)){

  ##Gerando valores normais (media 0) com uma correlacao
     especifica (0.72)
valores <- rmvnorm(n = v[t],
                   mean = c(0,0),
                   matrix(c(1,0.72,0.72,1),2,2)) ##Matrix=
                      matriz de correlacao


s[t]<-cor(valores[,1],valores[,2],method = "spearman")  ##
   Correlacao spearman
p[t]<-cor(valores[,1],valores[,2],method = "pearson")   ##
   Correlacao Pearson
k[t]<-cor(valores[,1],valores[,2],method = "kendall")   ##
   Correlacao kendall

}
```

De forma análoga gerou-se valores para amostras de tamanho 2 a 20, utilizados na segunda parte desse trabalho, além de valores com correlações diferentes a fim de comparação. Em seguida foram medidos os erros ($|\rho - cor|$) e assim, com auxílio do pacote *ggplot2* gerou-se gráficos para melhor representação dos resultados.

3 Resultados

Nessa seção foram abordadas algumas situações de simulações com o objetivo de verificar o comportamento dos coeficientes em cada uma delas.

3.1 Correlação positiva

A Figura 1 representa a distribuição do erro em modulo, isto é, a diferença em módulo da verdadeira correlação (0.72) com as quantidades estimadas pelos 3 métodos conforme o tamanho de amostra. A partir da Figura 2, em conjunto com a Figura 1, é possível observar que Kendall tem um erro superior aos demais, enquanto que Pearson e Spearman parecem não ter muita diferença.

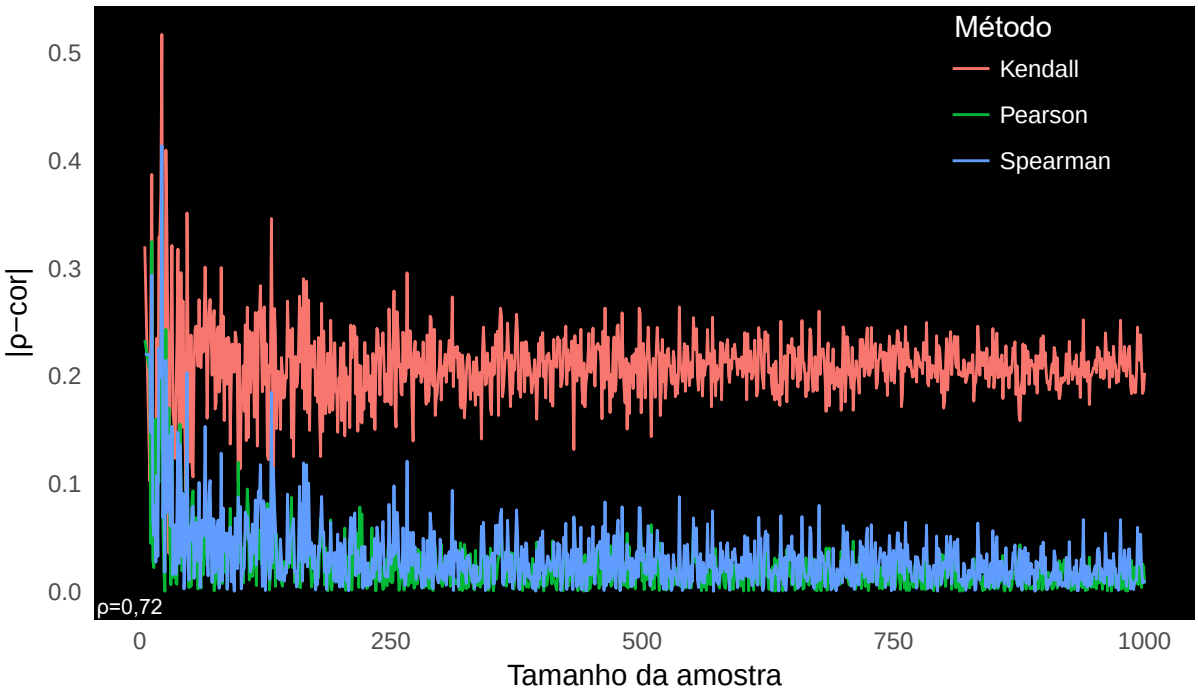

Figura 1: Erro (em módulo) das correlações estimadas conforme o tamanho da amostra.

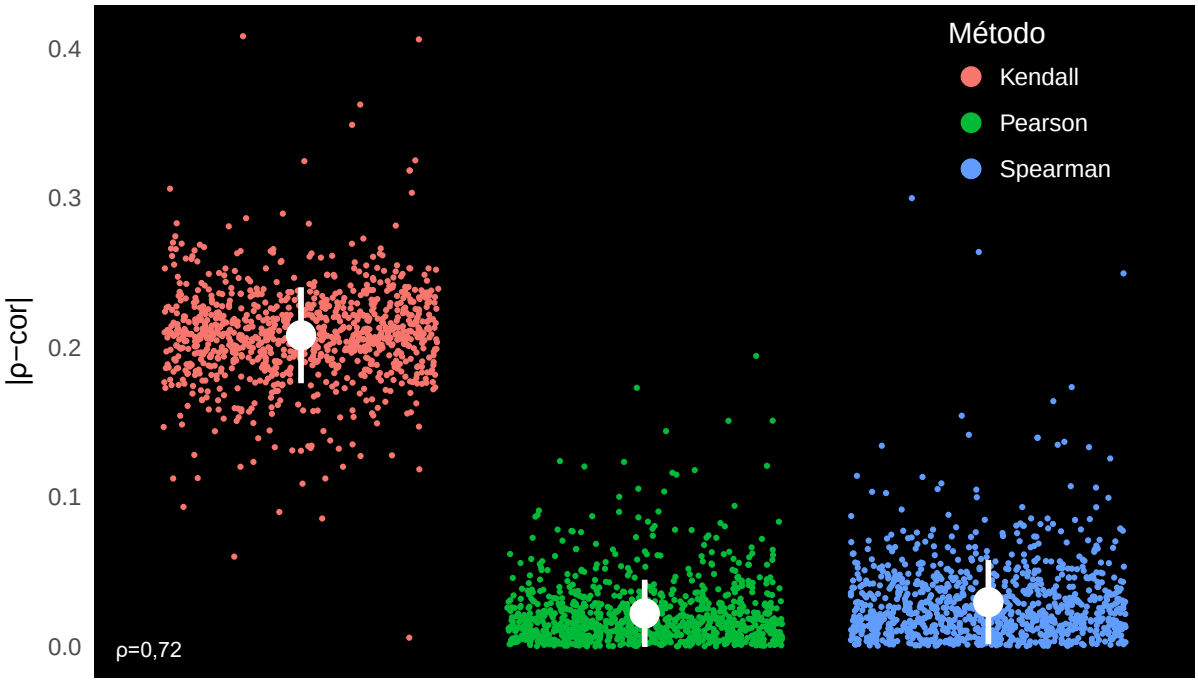

Figura 2: Distribuição do erro (em módulo) das correlações estimadas.

A Figura 3 representa a distribiução dos coeficientes de correlação calculados conforme o tamanho da amostra. É perceptível que a medida de Kendall tende a subestimar o valor da correlação.

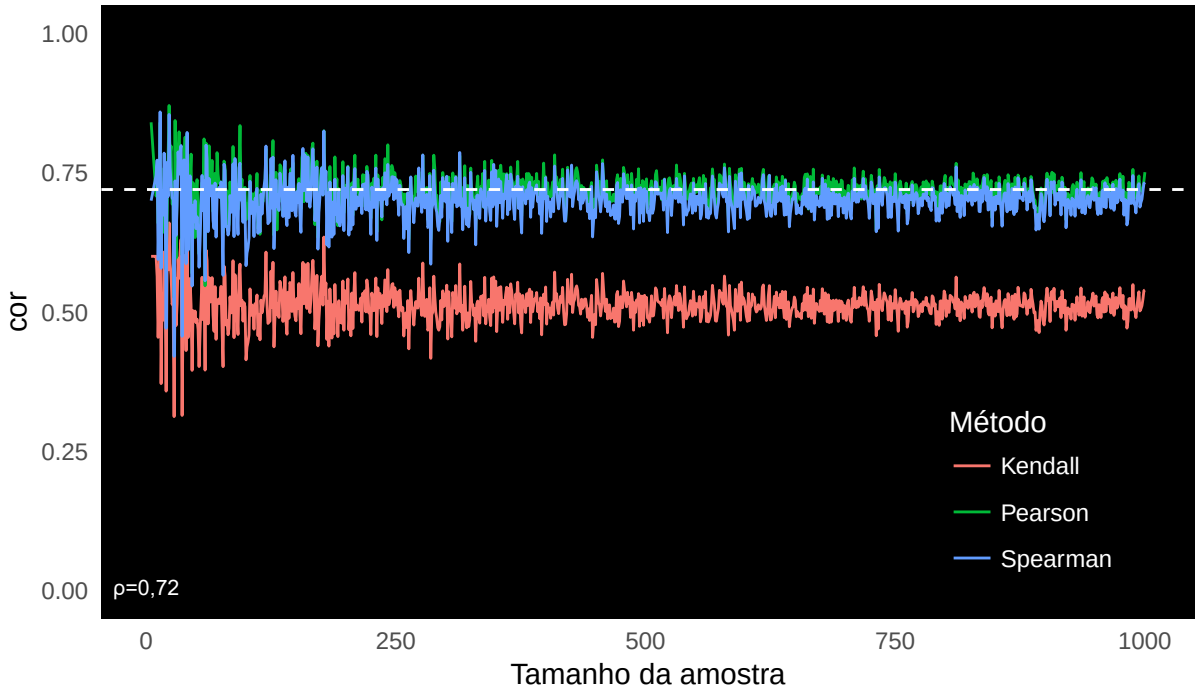

Figura 3: Distribuição das correlações estimadas conforme o tamanho da amostra.

3.2 Correlação negativa

Avaliando o caso da verdadeira correlação ser negativa (-0,85), a partir da Figura 4 e da Figura 5 é possível enxergar que acontece o mesmo que a situação de correlação positiva visto anteriormente. O coeficiente de Kendall apresenta erros maiores, enquanto que os outros dois ficam próximos nesse quesito.

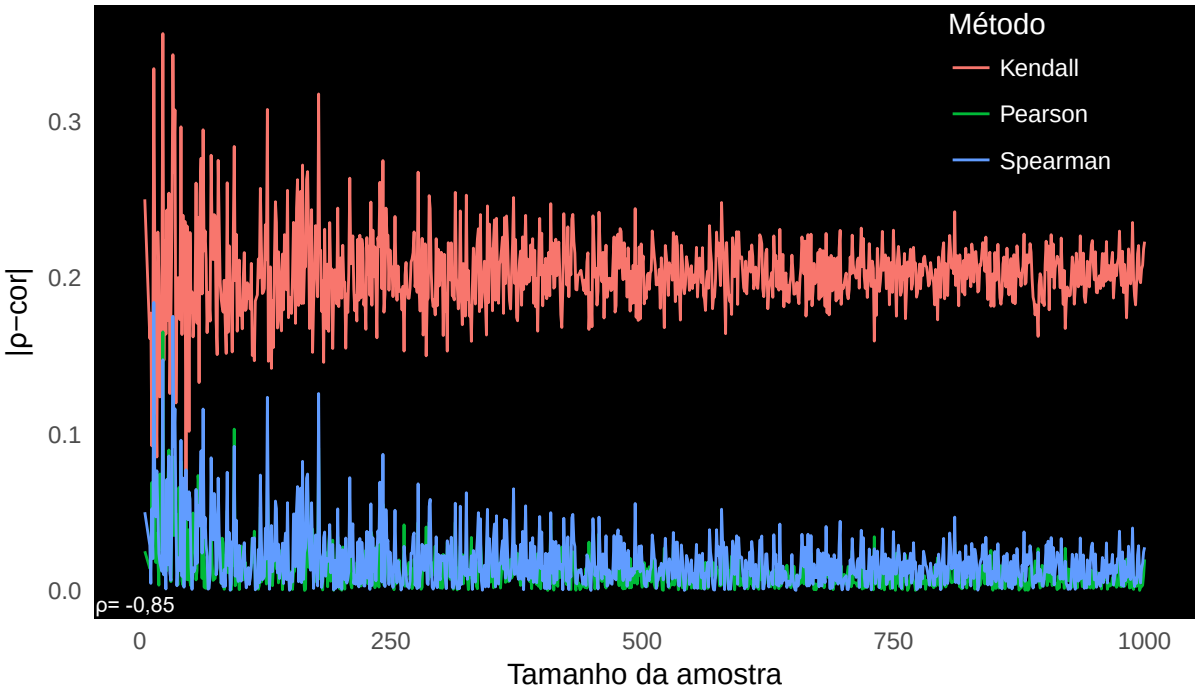

Figura 4: Erro (em módulo) das correlações estimadas conforme o tamanho da amostra.

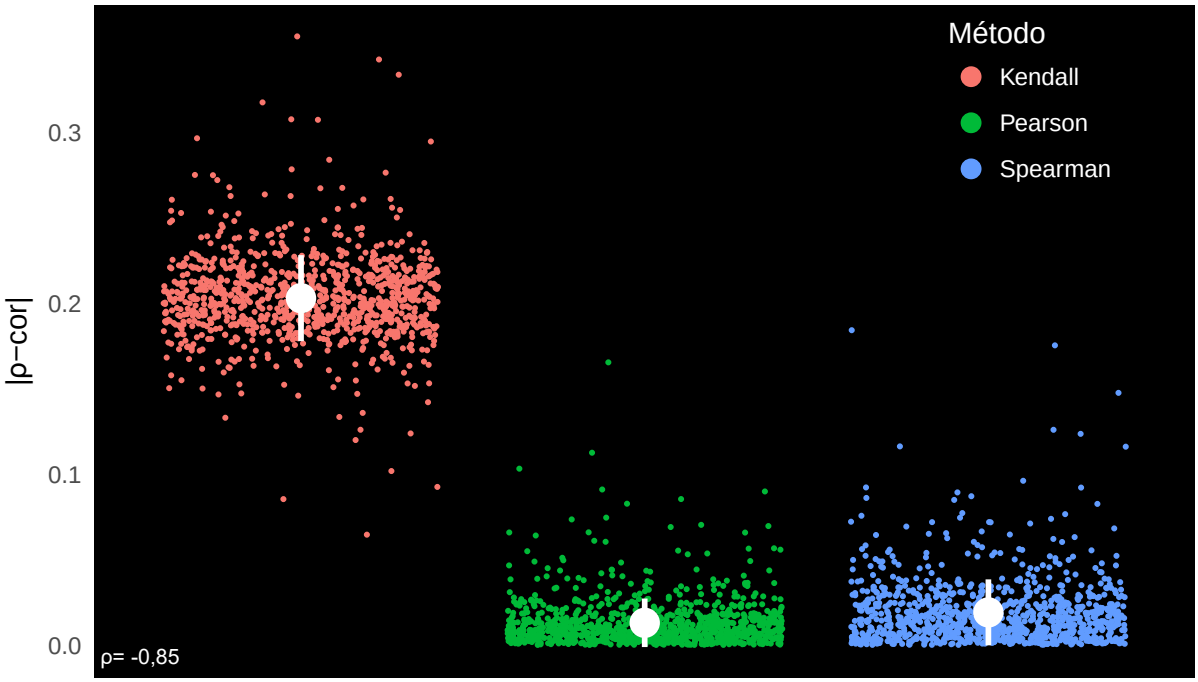

Figura 5: Distribuição do erro (em módulo) das correlações estimadas.

Da mesma forma, pela Figura 6 é perceptível que a medida de Kendall subestima o valor (manitude) da correlação.

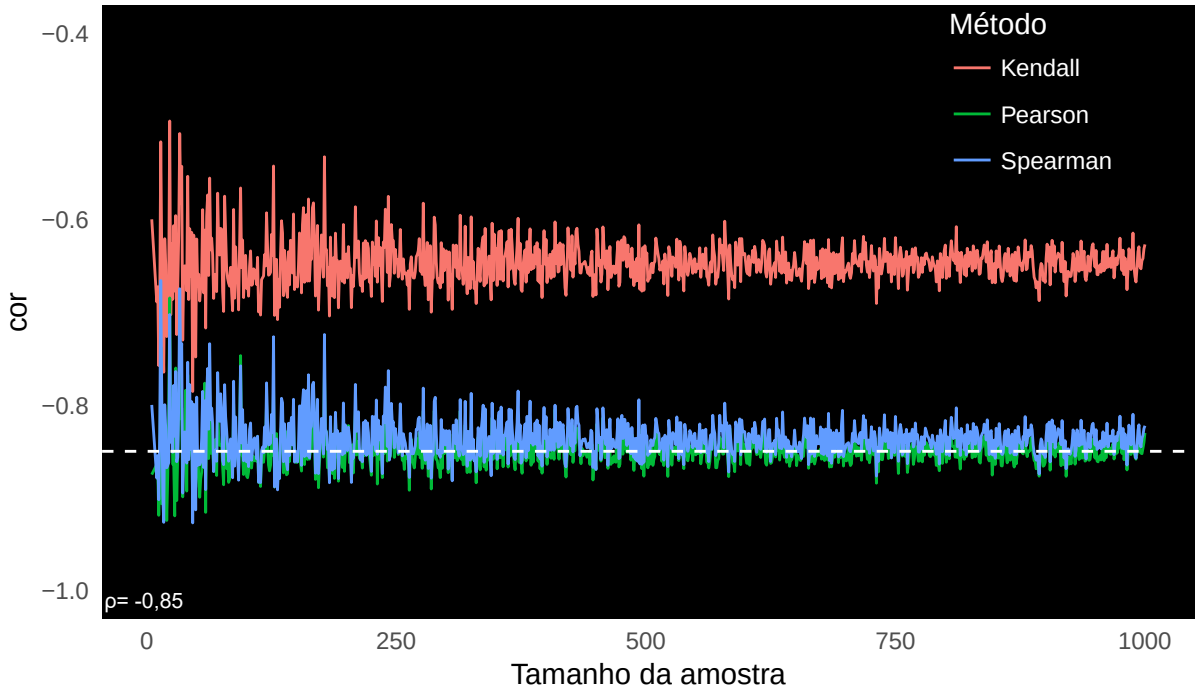

Figura 6: Distribuição das correlações estimadas conforme o tamanho da amostra.

3.3 Correlação quase nula

Considerando agora a simulação de valores com correlação quase nula (0,12). Nota-se que os 3 métodos se equivalem nessa situação, como visto na Figura 7, Figura 8 e Figura 9.

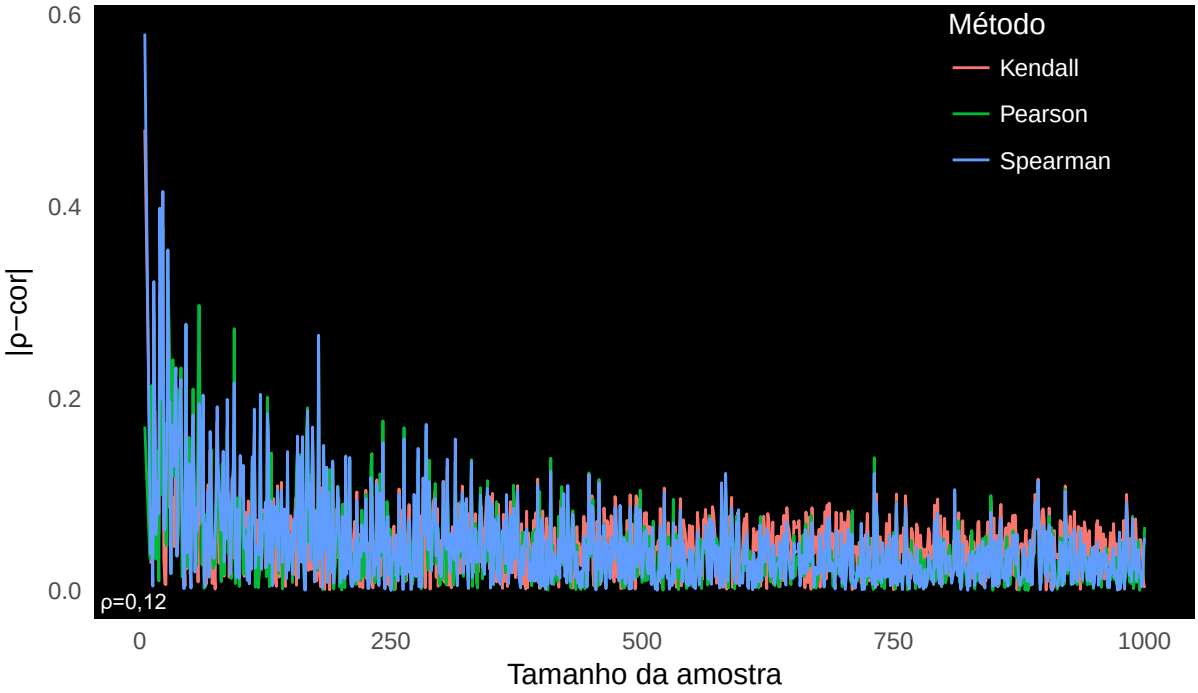

Figura 7: Erro (em módulo) das correlações estimadas conforme o tamanho da amostra.

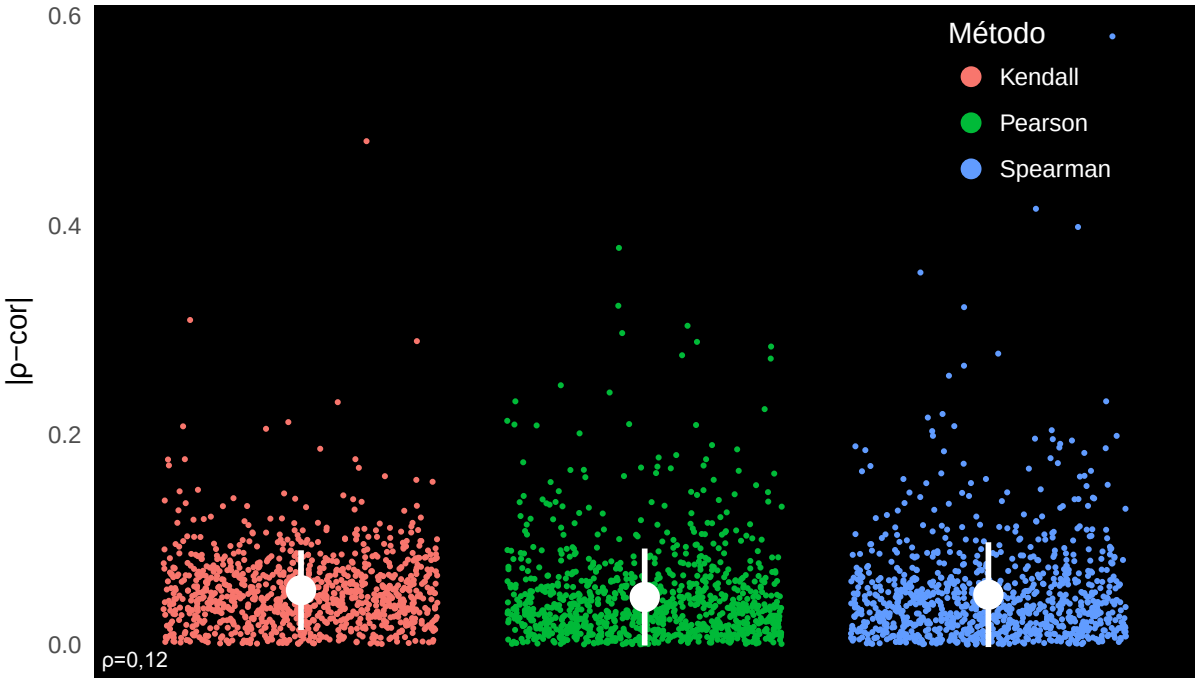

Figura 8: Distribuição do erro (em módulo) das correlações estimadas.

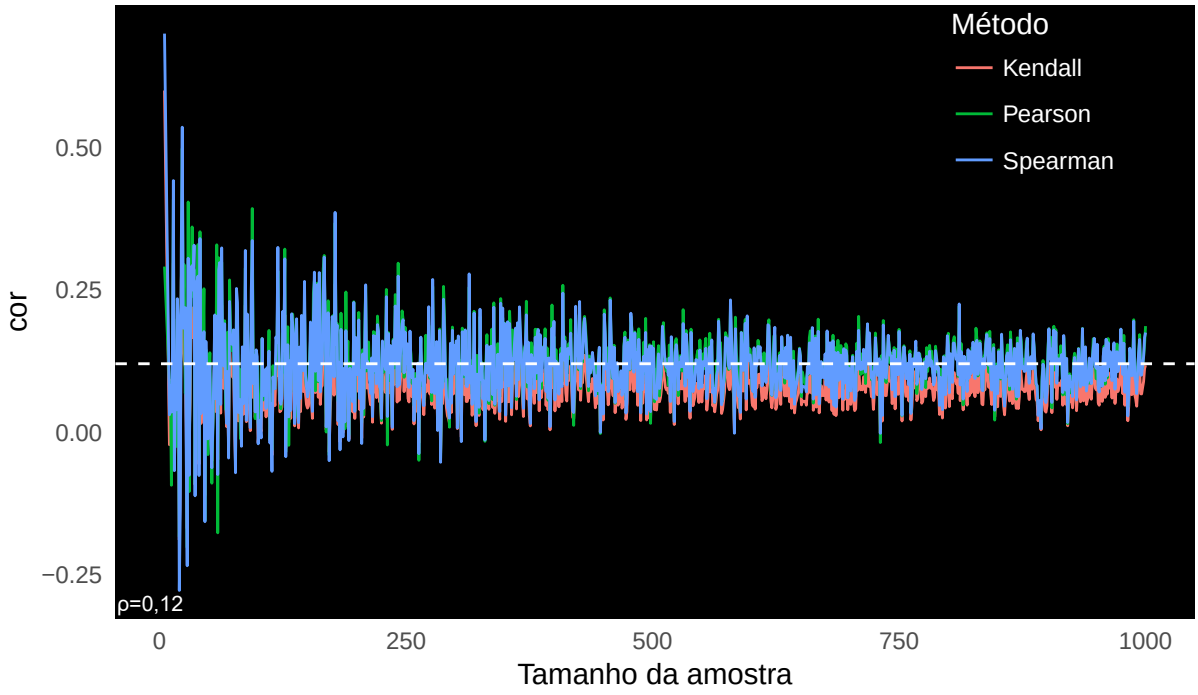

Figura 9: Distribuição das correlações estimadas conforme o tamanho da amostra.

3.4 Amostra pequena

Com intuito de observar o comportamento dos erros em amostras pequenas, é possível averiguar a partir da Figura 10 e Figura 11 que há apenas uma ligeira diferença nos erros, sendo o coeficiente de Kendall novamente o de erro maior.

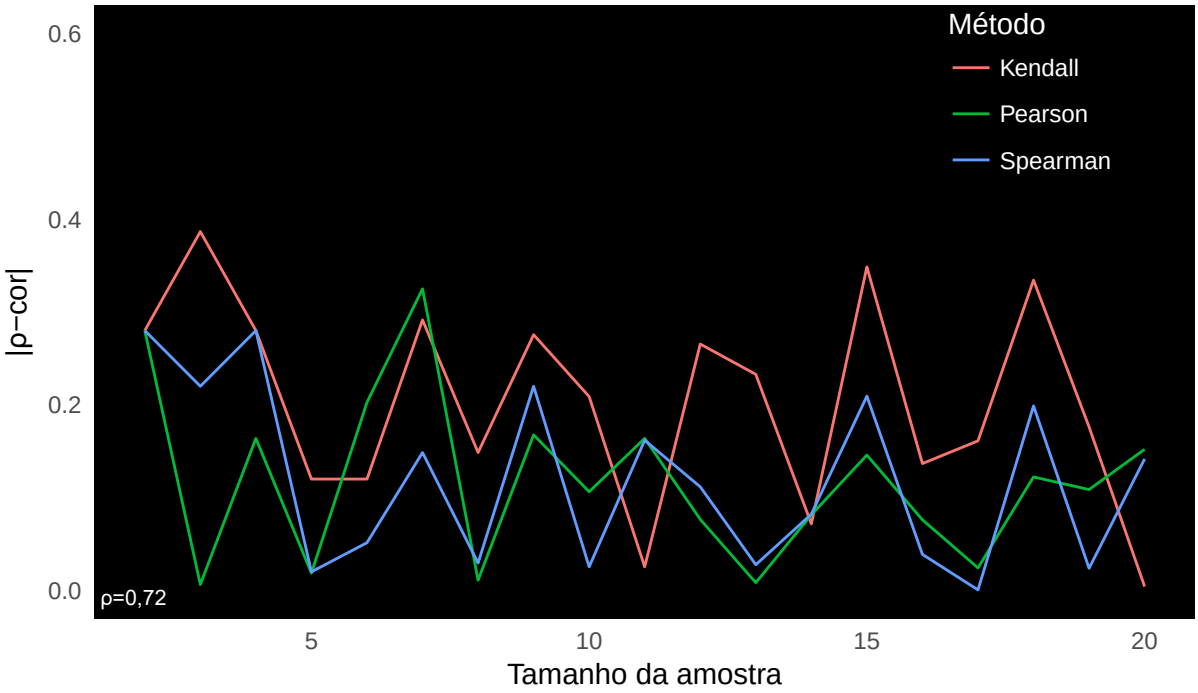

Figura 10: Erro (em módulo) das correlações estimadas conforme o tamanho da amostra.

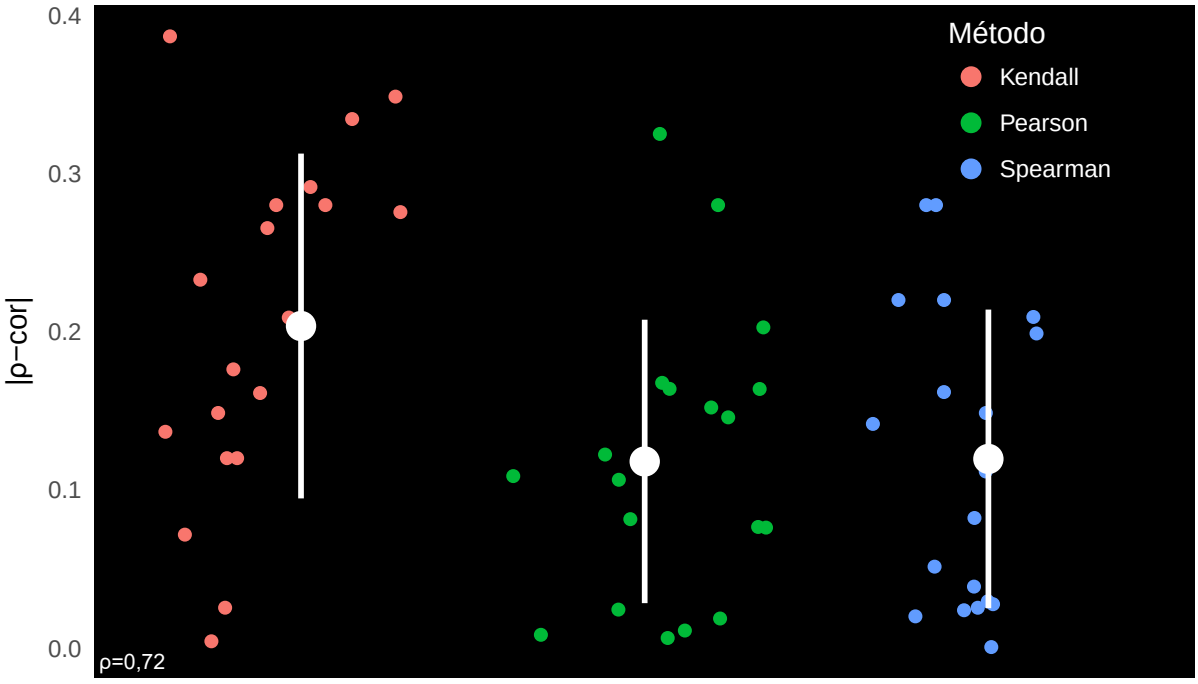

Figura 11: Distribuição do erro (em módulo) das correlações estimadas.

Observando a Figura 12 nota-se que, com amostras muito pequenas (menores ou iguais a 5), os três métodos possuem erros grandes, porém o coeficiente de Pearson pareceu se sair melhor nesse caso. É possível notar ainda que, conforme o tamanho da amostra aumenta, maior a tendência da medida de Kendall subestimar o valor da correlação.

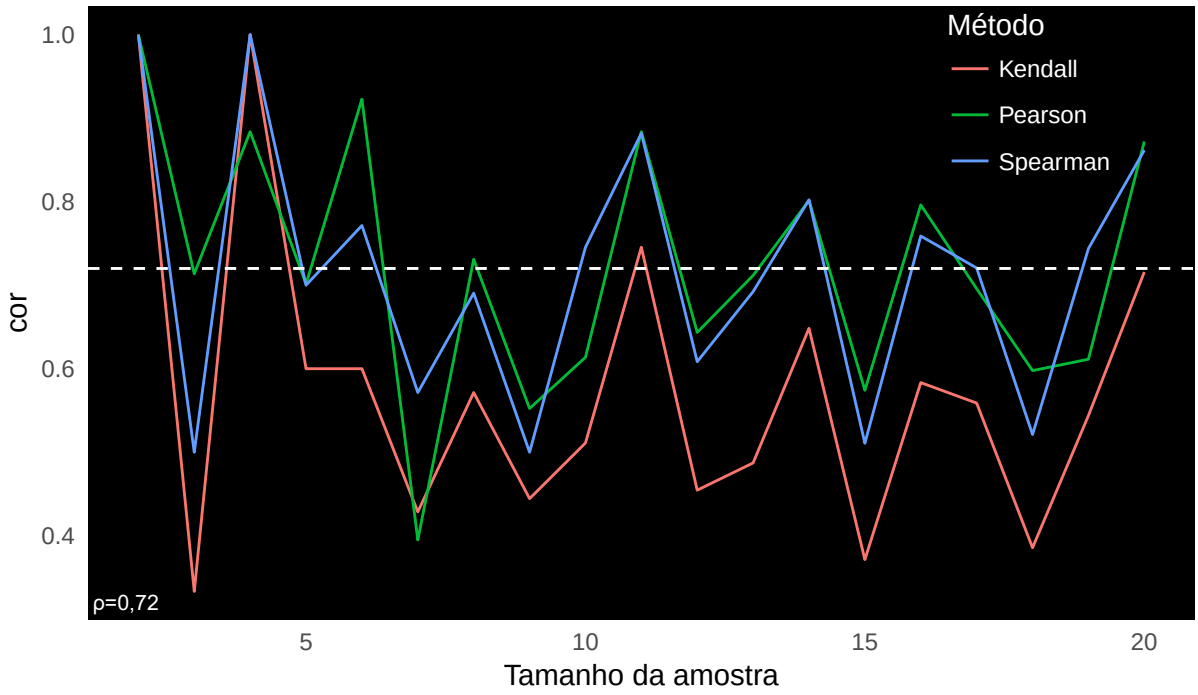

Figura 12: Distribuição das correlações estimadas conforme o tamanho da amostra.

4 Conclusão

Verificou-se a partir de simulações o comportamento dos coeficientes de correlações propostos (Pearson, Spearman e Kendall). Foi possível observar que o coeficiente de Kendall tende a errar mais, subestimando o valor da verdadeira correlação, considerando as circunstâncias de correlação positiva, negativa e amostra pequena. Por outro lado, as medidas de Pearson e Spearman pareceram concordar em todas as situações testadas, não havendo uma grande divergência entre elas, apenas uma ligeira vantagem (erro menor) do coeficiente de Pearson em algumas situações. Quando submetidos ao quadro de correlação quase nula, os 3 métodos praticamente se equivaleram.