Testes de hipóteses para duas populações independentes

Adriana Eva Fernandes da Silva

Tópicos

- Descrição do conjunto de dados;
- Tabela dos dados analisados;
- Breve descrição e justificativa da análise a ser realizada;
- Análise descritiva;
- Suposições e metodologia do método paramétrico;
- Verificação das suposições do método paramétrico;
- Testes de hipóteses via método paramétrico;
- Suposições e metodologia do método de permutação;
- Verificação das suposições do método de permutação;
- Testes de hipóteses via método de permutação;
- Suposições e metodologia do método de bootstrap;
- Verificação das suposições do método de bootstrap;
- Testes de hipóteses via método de bootstrap;
- Conclusão;
- Referências bibliográficas.

Descrição do conjunto de dados

O conjunto de dados utilizado no trabalho foi retirado do site dasl.datadescription.com. Trata-se de um experimento conduzido com 52 nuvens para determinar se a semeadura com iodeto de prata aumenta a precipitação.

As nuvens foram aleatoriamente selecionadas de forma que 26 nuvens fossem semeadas e 26 não fossem semeadas, e a quantidade de chuva foi medida em acre-pés.

Variáveis aleatórias:

- X_i: precipitação das nuvens não semeadas;
- Y_i: precipitação das nuvens semeadas;

Como não temos o local e a forma de coleta dos dados, consideramos que o processo foi feito adequadamente.

Tabela dos dados analisados

Tabela: Dados do problema

Nuvens	x_i	y_i
1	1202.6	2745.6
2	830.1	1697.8
3	372.4	1656.0
4	345.5	978.0
5	321.2	703.4
6	244.3	489.1
7	163.0	430.0
8	147.8	334.1
9	95.0	302.8
10	87.0	274.7
11	81.2	274.7
12	68.5	255.0
13	47.3	242.5
14	41.1	200.7
15	36.6	198.6
16	29.0	129.6
17	28.6	119.0
18	26.3	118.3
19	26.1	115.3
20	24.4	92.4
21	21.7	40.6
22	17.3	32.7
23	11.5	31.4
24	4.9	17.5
25	4.9	7.7
26	1.0	4.1

Breve descrição e justificativa da análise a ser realizada

Um experimento foi conduzido para comparar a precipitação de nuvens. O interesse estava em comparar se, em média, as nuvens semeadas com iodeto de prata geraram mais chuva.

Como as nuvens foram escolhidas de forma aleatória para serem ou não semeadas, foram conduzidos testes para duas populações independentes com o objetivo de verificar a significância das diferenças entre as médias de precipitação.

Por se tratar de uma amostra pequena e por não termos conhecimento sobre a sua distribuição, temos um indicativo de que devemos realizar apenas testes não paramétricos, mas por fins didáticos e de comparação, realizaremos também testes paramétricos.

Análise Descritiva

Com o objetivo de verificar o interesse do experimento, inicialmente foi feita uma análise descritiva dos dados.

Análise Descritiva: Boxplot

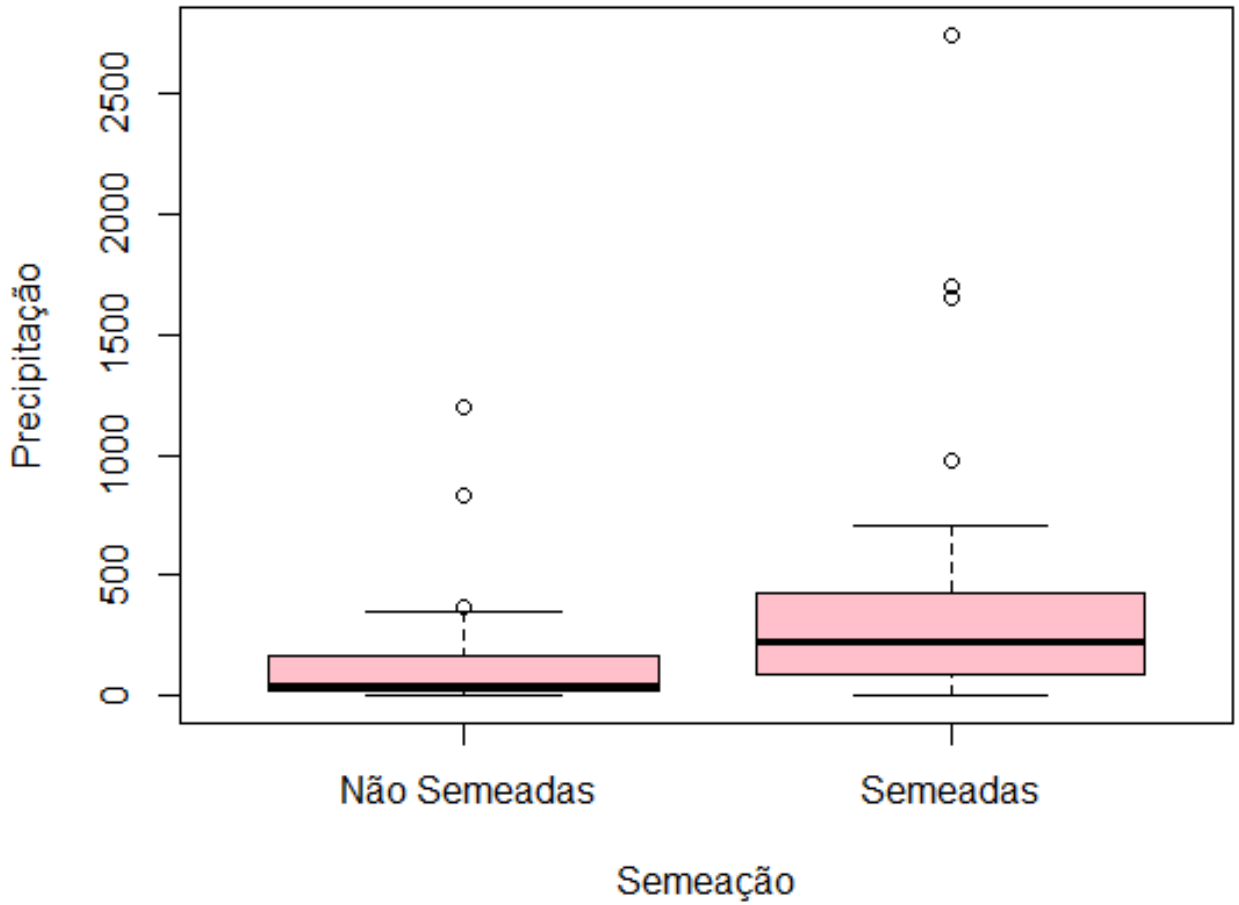


Figura: Boxplot dos dados

Análise Descritiva: Medidas de resumo

Tabela: Medidas de resumo

Semeação	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Não Semeadas	1.00	24.82	44.20	164.59	159.20	1202.60
Semeadas	4.10	98.12	221.60	441.98	406.02	2745.60

Análise Descritiva: Resultados

Através do boxplot e da tabela de medidas de resumo, é possível ver que em termos de variabilidade as distribuições das variáveis não são parecidas. Também verificamos a presença de outliers em ambos os casos (nuvens semeadas e nuvens não semeadas), o que distorce a simetria.

As medidas de resumo também nos dão um indicativo de que a quantidade média (em acre-pés) de chuva das nuvens semeadas, é maior do que a das nuvens não semeadas.

Teste paramétrico para média com duas populações independentes

Para seguir com o teste para duas médias de populações independentes é necessário verificar se as variâncias são iguais ou não. Sendo assim, será realizado o teste F para comparar essas variâncias.

Teste paramétrico para variância com duas populações independentes: Suposições

- Indivíduos sorteados aleatoriamente da população
- As observações de X e Y devem ser independentes
- As amostras devem ser retiradas de populações com distribuição normal

Teste paramétrico para variância com duas populações independentes: Verificação das suposições com o Q-Q Norm para a população X

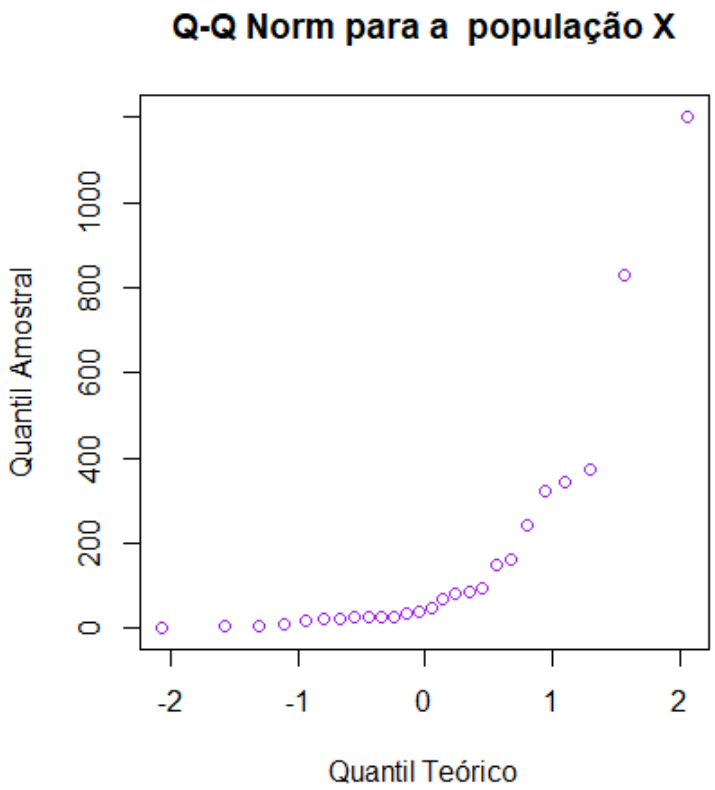


Figura: Gráfico Q-Q Norm para a população X

Teste paramétrico para variância com duas populações independentes: Verificação das suposições com testes de normalidade para a população X

Tabela: Teste de normalidade para X com $\alpha = 5\%$

Teste	p-valor
Shapiro-Wilk	0.0000003
Anderson-Darling	9.049e-10
Kolmogorov-Smirnov	0.02448
Lilliefors	0.0000046

Teste paramétrico para variância com duas populações independentes: Resultados das suposições

Podemos notar, pelo gráfico de normalidade (QQ-Norm), que os pontos não seguem uma reta, nos resultando em evidências de que a população X não possui distribuição normal. Também podemos constatar que todos os testes apresentaram p-valores menores que 0.05, nos levando, portanto, a um nível de significância de 5% , rejeitar a hipótese de normalidade. Segue-se, então, que não há normalidade na população X.

Teste paramétrico para variância com duas populações independentes: Verificação das suposições com o Q-Q Norm para a população Y

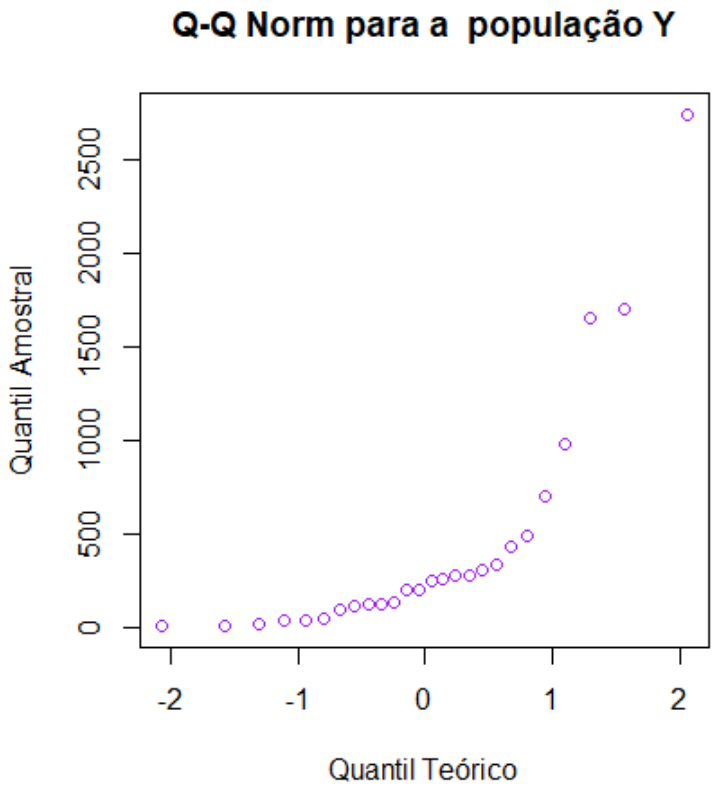


Figura: Gráfico Q-Q Norm para a população Y

Teste paramétrico para variância com duas populações independentes: Verificação das suposições com testes de normalidade para a população Y

Tabela: Teste de normalidade para Y com $\alpha = 5\%$

Teste	**p-valor**
Shapiro-Wilk	0.0000014
Anderson-Darling	0.0000000
Kolmogorov-Smirnov	0.02062
Lilliefors	0.0000026

Teste paramétrico para variância com duas populações independentes: Resultados das suposições

Podemos notar, pelo gráfico de normalidade (QQ-Norm), que os pontos não seguem uma reta, nos resultando em evidências de que a população Y não possui distribuição normal. Também podemos constatar que todos os testes apresentaram p-valores menores que 0.05, nos levando, portanto, a um nível de significância de 5% , rejeitar a hipótese de normalidade. Segue-se, então, que não há normalidade na população Y.

Como não temos o tempo de coleta dos dados e sendo que a execução da coleta dos dados foi feita por meio de uma amostragem aleatorizada, concluíremos, então, que a independência de X e Y é atendida.

Logo, o teste paramétrico não é válido. Porém, por fins didáticos seguiremos conduzindo o teste para a variância com duas populações independentes.

Teste paramétrico para variância com duas populações independentes: Teste de hipóteses

- Hipóteses

$H_0 : \sigma_x = \sigma_y$, (A variabilidade da precipitação das nuvens não semeadas é igual a variabilidade da precipitação das nuvens semeadas)

$H_1 : \sigma_x \neq \sigma_y$, (A variabilidade da precipitação das nuvens não semeadas é diferente da variabilidade da precipitação das nuvens semeadas)

- Nível de significância

$\alpha = 5\%$

Teste paramétrico para variância com duas populações independentes: Teste de hipóteses

- Estatística de teste sob H_0

$F = \frac{S_X^2}{S_Y^2}$

$F_{obs} = 0.183$

- Regra de decisão:

Rejeitamos H_0 ao nível de significância $\alpha = 5\%$, se valor-p < 0.05 e não rejeitamos H_0 ao nível de significância $\alpha = 5\%$, caso contrário.

- Conclusão

Com um valor-p $= 0.00006695$ rejeitamos H_0 . Assim, a um nível de significância $\alpha = 5\%$ temos evidências de que as variâncias são diferentes.

Desta forma seguiremos com o teste de hipóteses para duas médias com variâncias diferentes.

Teste paramétrico para médias com duas populações independentes e variâncias diferentes: Suposições

- Indivíduos sorteados aleatoriamente da população
- As observações de X e Y devem ser independentes
- As amostras devem ser retiradas de populações com distribuição normal

As mesmas suposições já foram verificadas anteriormente para conduzir o teste para variância. Logo, o teste não é válido. Porém por meios didáticos seguiremos conduzindo o teste para as duas médias.

Teste paramétrico para médias com duas populações independentes e variâncias diferentes: Teste de hipóteses

- Hipóteses

$H_0 : \mu_x = \mu_y$, (A precipitação média das nuvens não semeadas é igual a precipitação média das nuvens semeadas)

$H_1 : \mu_x < \mu_y$, (A precipitação média das nuvens não semeadas é menor do que a precipitação média das nuvens semeadas)

- Nível de significância

$\alpha = 5\%$

Teste paramétrico para médias com duas populações independentes e variâncias diferentes: Teste de hipóteses

- Estatística de teste sob H_0

$T = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}}$

T_{obs} = -1.998

- Regra de decisão:

Rejeitamos H_0 ao nível de significância $\alpha = 5\%$, se $valor - p < 5\%$ e não rejeitamos H_0 ao nível de significância $\alpha = 5\%$, caso contrário.

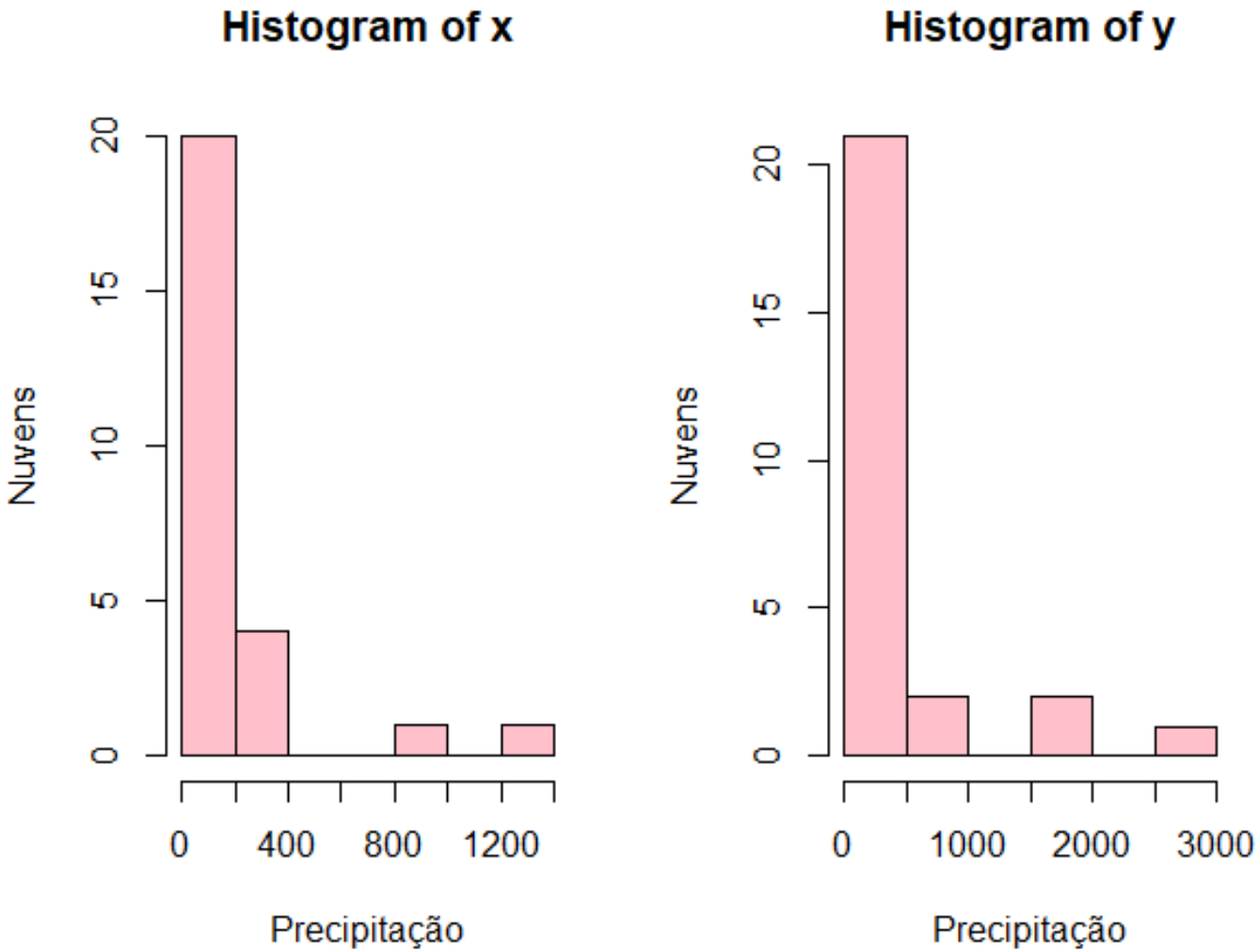
- Conclusão

Com um valor-p = 0.027 rejeitamos H_0 . Assim, a um nível de significância α = 5% temos evidências de que a precipitação média das nuvens não semeadas é menor a precipitação média das nuvens semeadas.

Teste de permutação para duas populações independentes: Suposições

- 1- $\mathbf{X} = (X_1,..., X_n)$ e $\mathbf{Y} = (Y_1..., Y_m)$ são amostras aleatórias de suas respectivas populações;
- 2 - X_s e Y_s são mutuamente independentes;
- 3 - as variáveis aleatórias X e Y são contínuas;
- 4 - a função distribuição de X e Y são idênticas ou uma tem média (no caso de querermos testar sobre as médias populacionais) maior que da outra (variâncias iguais).

Teste de permutação para duas populações independentes: Verificação das suposições



Teste de permutação para duas populações independentes: Verificação das suposições

Como não temos o local e a forma de coleta dos dados, consideramos que o processo de aleatorização para o tipo de experimento foi feito adequadamente. As nuvens foram aleatoriamente selecionadas de forma que 26 nuvens fossem semeadas e 26 não fossem semeadas. Sendo assim, as suposições 1,2,3 estão sendo atendidas.

Porém, podemos notar pelo histograma das variáveis X e Y, que as distribuições dos dados para cada variável não são iguais, um indício disso é a variabilidade da precipitação das nuvens não semeadas ser menor que a variabilidade da precipitação das nuvens semeadas.

Logo, o teste de permutação também não é válido. Porém, por fins didáticos seguiremos conduzindo o teste para as duas médias.

Teste de permutação para duas populações independentes: Teste de hipóteses

- Hipóteses

$H_0 : E(X) = E(Y)$, (A média de precipitação das nuvens não semeadas é igual a média de precipitação das nuvens semeadas)

$H_1 : E(X) < E(Y)$, (A média de precipitação das nuvens não semeadas é menor do que a média de precipitação das nuvens semeadas)

Teste de permutação para duas populações independentes: Teste de hipóteses

- Nível de significância

$\alpha = 5\%$

- Estatística de teste sob H_0

$T = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}}$

T_{obs} = -1.998

- Regra de decisão:

Rejeitamos H_0 ao nível de significância $\alpha = 5\%$, se valor-p $< 5\%$ e não rejeitamos H_0 ao nível de significância $\alpha = 5\%$, caso contrário.

- Conclusão

Com um valor-p = 0.020 rejeitamos H_0 . Assim, a um nível de significância $\alpha = 5\%$ temos evidências de que a precipitação média das nuvens não semeadas é menor do que a precipitação média das nuvens semeadas.

Teste de permutação para duas populações independentes: Tamanho da amostra

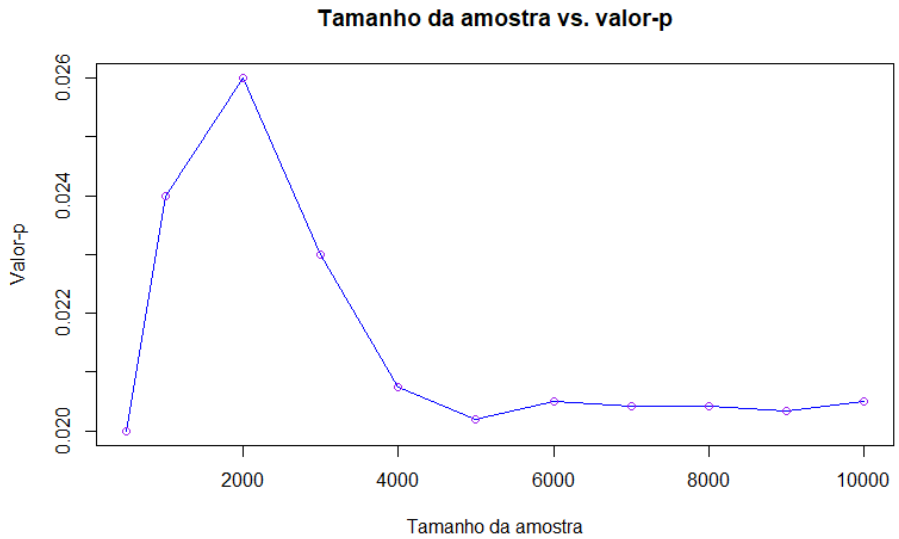


Figura: Gráfico de linha para tamanho da amostra vs. valor-p

Teste de permutação para duas populações independentes: Tamanho da amostra

Tabela: Tamanho da amostra e valor-p

Amostra	Valor-p
500	0.020
1000	0.024
2000	0.026
3000	0.023
4000	0.021
5000	0.020
6000	0.020
7000	0.020
8000	0.020
9000	0.020
10000	0.020

Teste de permutação para duas populações independentes: Tamanho da amostra

Com base no gráfico de linha para o tamanho da amostra vs. valor-p e na tabela dos p-valores com relação ao tamanho da amostra podemos notar que para um B (tamanho da amostra) de 5000 os p-valores convergem. Portanto, o melhor B é de 5000.

Método de reamostragem via bootstrap: Suposições

- Se temos duas amostras aleatórias independentes, essas amostras precisam ser aleatórias e independentes entre si.

Método de reamostragem via bootstrap: Verificação das suposições

Como não temos o tempo de coleta dos dados e a execução da coleta dos dados foi feita por meio de uma amostragem aleatorizada, as variáveis são independentes entre si. Além disso, as nuvens foram aleatoriamente selecionadas de forma que 26 nuvens fossem semeadas e 26 não fossem semeadas, concluímos, então, que a independência de X e Y.

Método de reamostragem via bootstrap: Teste de hipóteses

- Hipóteses

$H_0 : \mu_X = \mu_Y$, (A média de precipitação das nuvens não semeadas é igual a média de precipitação das nuvens semeadas)

$H_1 : \mu_X < \mu_Y$, (A média de precipitação das nuvens não semeadas é menor do que a média de precipitação das nuvens semeadas)

Método de reamostragem via bootstrap: Teste de hipóteses

- Nível de significância

$\alpha = 5\%$

- Estatística de teste sob H_0

$T(z_b^*) = \frac{\bar{x}_b^* - \bar{y}*_b}{\sqrt{\frac{s^{2*}_x}{n} + \frac{s^{2*}_y}{m}}}$

$T(z) = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$

T_{obs} = -1.998

- Regra de decisão:

Rejeitamos H_0 ao nível de significância $\alpha = 5\%$, se valor-p $< 5\%$ e não rejeitamos H_0 ao nível de significância $\alpha = 5\%$, caso contrário.

- Conclusão Com um valor-p $= 0.009$ rejeitamos H_0 . Assim, a um nível de significância $\alpha = 5\%$ temos evidências de que a precipitação média das nuvens não semeadas é menor do que a precipitação média das nuvens semeadas.

Método de reamostragem via bootstrap: Tamanho da amostra

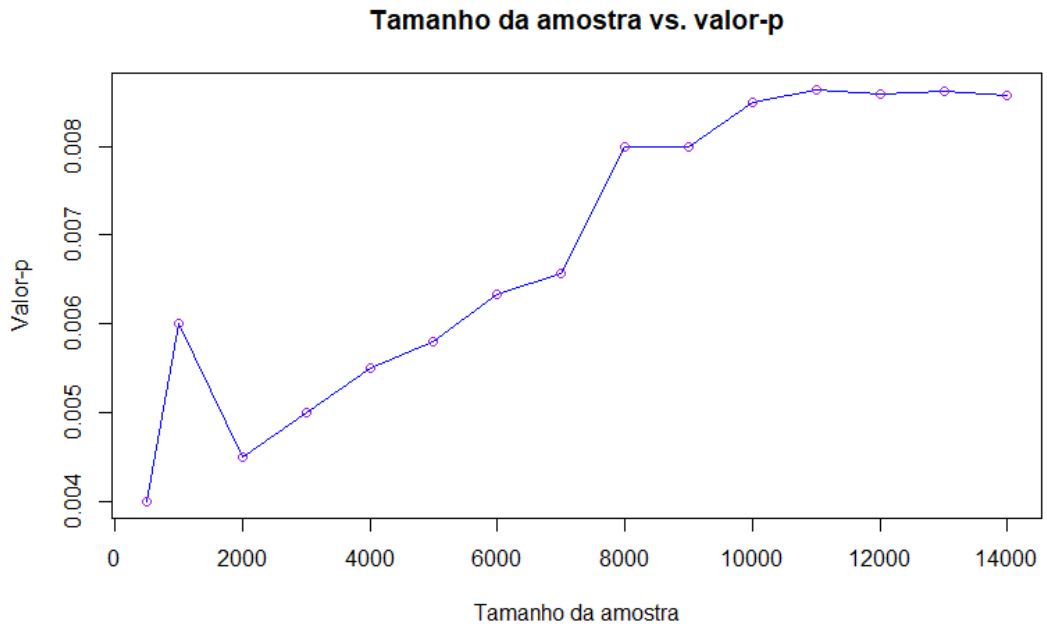


Figura: Gráfico do tamanho da amostra vs. valor-p

Método de reamostragem via bootstrap: Tamanho da amostra

Tabela: Tamanho da amostra e valor-p

Amostra	Valor-p
500	0.004
1000	0.006
2000	0.004
3000	0.005
4000	0.005
5000	0.006
6000	0.006
7000	0.007
8000	0.008
9000	0.008
10000	0.009
11000	0.009
12000	0.009
13000	0.009

Método de reamostragem via bootstrap: Tamanho da amostra

Com base no gráfico de linha para o tamanho da amostra vs. valor-p e na tabela dos p-valores com relação ao tamanho da amostra podemos notar que para um B (tamanho da amostra) de 10000 os p-valores convergem. Portanto, o melhor B é de 10000.

Conclusão: Diferenças entre os testes

Teste paramétrico

- O **teste-t paramétrico** é utilizado quando os dados são provenientes de uma distribuição normal ou se a variância é conhecida e o tamanho da amostra é suficientemente grande para aproximarmos pelo Teorema do Limite Central, da distribuição normal, quando o interesse está em comparar médias.

Testes não paramétricos

- O **teste de permutação** é utilizado quando não conhecemos de qual distribuição os dados são provenientes e quando temos as suas suposições satisfeitas e só podem ser usados para hipóteses que envolvam comparações.
- O **teste de bootstrap** é utilizado quando não conhecemos de qual distribuição os dados são provenientes, quando temos as suas suposições satisfeitas e conseguimos verificar a hipótese H_0 baseada na estatística teste de interesse.

Conclusão: Comparando os testes

A princípio, foi feito o **teste-t paramétrico** para testar a igualdade da precipitação entre as nuvens não semeadas e as nuvens semeadas com o iodeto de prata, mas como as suposições do teste não foram verificadas para os dados, invalidamos o teste paramétrico.

Em seguida, foi feito o teste não paramétrico para comparação de igualdade de médias das variáveis. As suposições necessárias para a realização do **teste de permutação** também não foram verificadas, logo, fizemos a realização do teste de bootstrap.

Este último, por se tratar de um método de reamostragem não paramétrico mais flexível, as suposições necessárias para o nosso conjunto de dados foram verificadas e foi decidido que o **método de reamostragem via booststrap** é o que melhor se adequada a nossa amostra.

Referências bibliográficas

- Slides da disciplina Métodos Computacionalmente Intensivos - Daiane Aparecida Zuanetti.
- https://dasl.datadescription.com/datafile/cloud-seeding/?_sfm_methods=Comparing+Two+Groups&_sfm_cases=4+59943