

Desvantagem salarial dos empregados na hotelaria por região

Glauber Santos

2023-02-06

Este exercício busca analisar a diferença entre o salário médio na hotelaria e nas demais atividades produtivas. Em geral, a remuneração na hotelaria é mais baixa do que no restante da economia. Contudo, a diferença salarial não é homogênea no espaço, sendo mais acentuada em alguns lugares ou regiões.

Para analisar a variação regional da desvantagem salarial dos hoteleiros por região, iremos utilizar dados sobre a remuneração média dos trabalhadores da hotelaria e das demais atividades produtivas por município do Brasil. Compararemos a desvantagem salarial dos hoteleiros nos municípios das regiões Sul e Sudeste do Brasil (mais ricas) com aquela observada nos municípios das demais regiões do país. A ideia é entender se a desvantagem salarial dos hoteleiros é ainda maior nas regiões mais pobres do país.

Preparação de dados

O salário médio na hotelaria e na economia como um todo será computado a partir de dados do relatório de [Estatísticas do Cadastro Central de Empresas (CEMPRE)] (<https://www.ibge.gov.br/estatisticas/economicas/comercio/9016-estatisticas-do-cadastro-central-de-empresas.html>) do IBGE. Os dados do número de empregados e a massa salarial de cada município brasileiro por atividade econômica podem ser baixados através da Tabela 6450 do banco de dados SIDRA do IBGE.

Para baixar os dados de forma fácil e direta, vamos utilizar o pacote `sidrar`. Para instalar esse pacote, utilize o comando `install.packages("sidrar")`. Depois, execute o comando `get_sidra`. O parâmetro para utilizado nessa função pode ser obtido diretamente no site do SIDRA. Na Tabela 6450, selecione as variáveis **Pessoal ocupado assalariado** e **Salários e outras remunerações**, a atividade econômica **55.10-8 Hotéis e similares**, bem como o **Total** das atividades econômicas, o ano de **2019** e **todos os municípios** do país em unidade territorial. O ano de 2019 foi selecionado por ser o mais recente representativo de um momento de não pandemia. Ao invés de clicar em visualizar, vamos selecionar o botão **links de compartilhar** na parte inferior da página. Em seguida, copie o conteúdo de **Parâmetros para a API** a partir do trecho `\t\`. O download pode demorar alguns segundos. Uma vez baixados os dados, vamos dar uma olhada no resultado da importação utilizando o comando `head`.

```
library(sidrar)

empregos <- get_sidra(api = "/t/6450/n6/all/v/662,708/p/2019/c12762/117546,117897/d/v662%200")

head(empregos, 10)
```

##	Nível Territorial (Código)	Nível Territorial	Unidade de Medida (Código)
## 2	6	Município	40
## 3	6	Município	40
## 4	6	Município	45
## 5	6	Município	45
## 6	6	Município	40
## 7	6	Município	40

## 8		6	Município	45
## 9		6	Município	45
## 10		6	Município	40
## 11		6	Município	40
##	Unidade de Medida	Valor	Município (Código)	Município
## 2	Mil Reais	101	1100015 Alta Floresta D'Oeste - RO	
## 3	Mil Reais	58805	1100015 Alta Floresta D'Oeste - RO	
## 4	Pessoas	8	1100015 Alta Floresta D'Oeste - RO	
## 5	Pessoas	2443	1100015 Alta Floresta D'Oeste - RO	
## 6	Mil Reais	NA	1100023 Ariquemes - RO	
## 7	Mil Reais	440332	1100023 Ariquemes - RO	
## 8	Pessoas	NA	1100023 Ariquemes - RO	
## 9	Pessoas	16718	1100023 Ariquemes - RO	
## 10	Mil Reais	NA	1100031 Cabixi - RO	
## 11	Mil Reais	13352	1100031 Cabixi - RO	
##	Variável (Código)		Variável	Ano (Código) Ano
## 2	662 Salários e outras remunerações			2019 2019
## 3	662 Salários e outras remunerações			2019 2019
## 4	708 Pessoal ocupado assalariado			2019 2019
## 5	708 Pessoal ocupado assalariado			2019 2019
## 6	662 Salários e outras remunerações			2019 2019
## 7	662 Salários e outras remunerações			2019 2019
## 8	708 Pessoal ocupado assalariado			2019 2019
## 9	708 Pessoal ocupado assalariado			2019 2019
## 10	662 Salários e outras remunerações			2019 2019
## 11	662 Salários e outras remunerações			2019 2019
##	Classificação Nacional de Atividades Econômicas (CNAE 2.0)		(Código)	
## 2				117546
## 3				117897
## 4				117546
## 5				117897
## 6				117546
## 7				117897
## 8				117546
## 9				117897
## 10				117546
## 11				117897
##	Classificação Nacional de Atividades Econômicas (CNAE 2.0)			
## 2		55.10-8 Hotéis e similares		
## 3		Total		
## 4		55.10-8 Hotéis e similares		
## 5		Total		
## 6		55.10-8 Hotéis e similares		
## 7		Total		
## 8		55.10-8 Hotéis e similares		
## 9		Total		
## 10		55.10-8 Hotéis e similares		
## 11		Total		

Note que cada município consta em quatro linhas diferentes e todos os valores estão na mesma variável intitulada “Valor”. Veja que o objeto “empregos” tem 22280 linhas, mas o Brasil tem apenas 5570 municípios. Cada linha se refere a uma combinação de número de empregados ou massa salarial (Variável) e hotelaria ou economia como um todo (Classificação CNAE). Para calcular os salários médios e a diferença salarial dos hoteleiros, é conveniente que os quatro valores estejam em variáveis separadas e que cada município conste

em apenas uma linha. Vamos fazer isso utilizando o comando `pivot_wider`. Antes disso, temos que carregar o pacote `tidyverse`. Se você ainda não tem esse pacote instalado, utilize o comando `install.packages`.

```
library(tidyverse)

empregos <- pivot_wider(empregos,
                        names_from = c("Classificação Nacional de Atividades Econômicas (CNAE 2.0)", "Município (Código)"),
                        values_from = Valor,
                        id_cols = c("Município (Código)", "Município"))

head(empregos)
```

```
## # A tibble: 6 x 6
##   Município (Código) Município 55.10~1 Total~2 55.10~3 Total~4
##   <chr> <chr> <dbl> <dbl> <dbl> <dbl>
## 1 1100015 Alta Floresta D'Oeste - ~ 101 58805 8 2443
## 2 1100023 Ariquemes - RO NA 440332 NA 16718
## 3 1100031 Cabixi - RO NA 13352 NA 536
## 4 1100049 Cacoal - RO 3303 436894 222 17058
## 5 1100056 Cerejeiras - RO 333 50928 25 2022
## 6 1100064 Colorado do Oeste - RO NA 47580 NA 2110
## # ... with abbreviated variable names
## # 1: '55.10-8 Hotéis e similares_Salários e outras remunerações',
## # 2: 'Total_Salários e outras remunerações',
## # 3: '55.10-8 Hotéis e similares_Pessoal ocupado assalariado',
## # 4: 'Total_Pessoal ocupado assalariado'
```

Agora temos uma tabela no formato desejado. Note que apenas as colunas relevantes da tabela foram mantidas, sendo as demais descartadas em meio ao uso do comando `pivot_wider`. Note também que agora a tabela `empregos` passou a ter 5570 linhas, o que equivale ao número de municípios do país.

Em seguida, vamos calcular o salário médio na hotelaria e na economia como um todo para cada município do país. Nos municípios em que não há registro de empregados no setor hoteleiro, o número de empregados e a massa salarial constam como NA (não disponível). O resultado do cálculo do salário médio nesses casos também será NA. Após o cálculo dos salários médios, vamos calcular a diferença salarial percentual.

```
empregos$salmed_hotel <- empregos$`55.10-8 Hotéis e similares_Salários e outras remunerações` / empregos$`Total_Pessoal ocupado assalariado`

empregos$salmed_geral <- empregos$`Total_Salários e outras remunerações` / empregos$`Total_Pessoal ocupado assalariado`

empregos$diferenca_salarial <- empregos$salmed_hotel/empregos$salmed_geral-1

head(empregos)
```

```
## # A tibble: 6 x 9
##   Município (C~1 Munic~2 55.10~3 Total~4 55.10~5 Total~6 salme~7 salme~8 difer~9
##   <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 1100015 Alta F~ 101 58805 8 2443 12.6 24.1 -0.476
## 2 1100023 Arique~ NA 440332 NA 16718 NA 26.3 NA
## 3 1100031 Cabixi~ NA 13352 NA 536 NA 24.9 NA
## 4 1100049 Cacoal~ 3303 436894 222 17058 14.9 25.6 -0.419
## 5 1100056 Cereje~ 333 50928 25 2022 13.3 25.2 -0.471
## 6 1100064 Colora~ NA 47580 NA 2110 NA 22.5 NA
```

```
## # ... with abbreviated variable names 1: 'Município (Código)', 2: Município,
## # 3: '55.10-8 Hotéis e similares_Salários e outras remunerações',
## # 4: 'Total_Salários e outras remunerações',
## # 5: '55.10-8 Hotéis e similares_Pessoal ocupado assalariado',
## # 6: 'Total_Pessoal ocupado assalariado', 7: salmed_hotel, 8: salmed_geral,
## # 9: diferenca_salarial
```

Podemos ver o resultado na última coluna da tabela. Note que os valores estão registrados em decimais. No município 1100015, por exemplo, os empregados na hotelaria ganham, em média, 47,55% a menos do que os empregados da economia como um todo.

Por fim, vamos criar uma variável para identificar os municípios das regiões Sul e Sudeste do país. A região do país pode ser identificada pelo primeiro dígito do código do município. Municípios da região Norte começam com o dígito 1, Nordeste 2, Sudeste 3, Sul 4 e Centro Oeste 5. Sendo assim, vamos primeiro criar uma variável identificando a região e depois diferenciar Sul e Sudeste das demais regiões. Para tanto, será preciso considerar o código do município como variável numérica, embora originalmente ela tenha sido importada como texto.

```
empregos$regiao <- floor(as.numeric(empregos$`Município (Código)`)/1000000)

empregos$sulsudeste <- ifelse((empregos$regiao==3 | empregos$regiao==4), 1, 0)

head(empregos)
```

```
## # A tibble: 6 x 11
##   Munic~1 Munic~2 55.10~3 Total~4 55.10~5 Total~6 salme~7 salme~8 difer~9 regiao
##   <chr>   <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 1100015 Alta F~    101   58805      8    2443    12.6    24.1  -0.476      1
## 2 1100023 Arique~    NA  440332     NA   16718     NA    26.3    NA      1
## 3 1100031 Cabixi~    NA  13352     NA    536     NA    24.9    NA      1
## 4 1100049 Cacoal~   3303  436894    222  17058    14.9    25.6  -0.419      1
## 5 1100056 Cereje~   333   50928     25   2022    13.3    25.2  -0.471      1
## 6 1100064 Colora~    NA  47580     NA   2110     NA    22.5    NA      1
## # ... with 1 more variable: sulsudeste <dbl>, and abbreviated variable names
## # 1: 'Município (Código)', 2: Município,
## # 3: '55.10-8 Hotéis e similares_Salários e outras remunerações',
## # 4: 'Total_Salários e outras remunerações',
## # 5: '55.10-8 Hotéis e similares_Pessoal ocupado assalariado',
## # 6: 'Total_Pessoal ocupado assalariado', 7: salmed_hotel, 8: salmed_geral,
## # 9: diferenca_salarial
```

Análise de dados

Vamos começar a exploração dos dados com medidas descritivas da diferença salarial da hotelaria por região. Para calcular estatísticas, vamos utilizar o pacote `rstatix`.

```
library(rstatix)

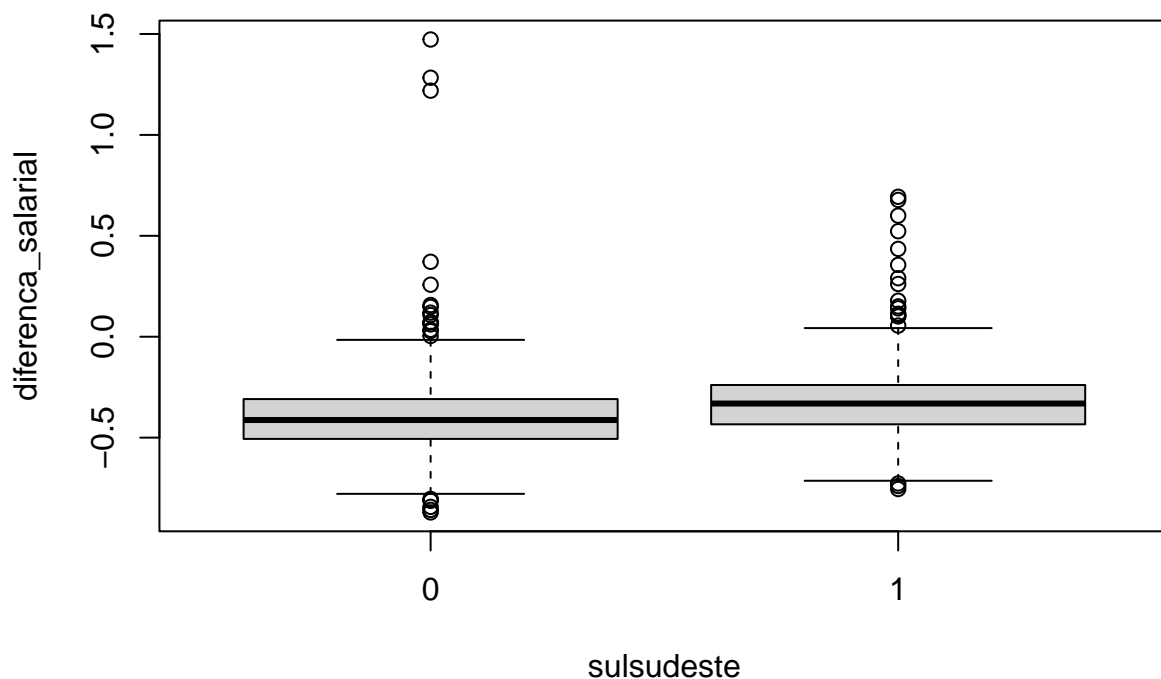
empregos %>%
  group_by(sulsudeste) %>%
  get_summary_stats(diferenca_salarial, type = "mean_sd")
```

```
## # A tibble: 2 x 5
##   sulsudeste variable      n  mean   sd
##   <dbl> <fct>      <dbl> <dbl> <dbl>
## 1         0 diferenca_salarial  548 -0.389 0.212
## 2         1 diferenca_salarial  746 -0.324 0.174
```

Os cálculos mostram que os empregados na hotelaria ganham cerca de 32,4% a menos do que a média dos trabalhadores na região Sul/Sudeste. No restante do país, a desvantagem salarial na hotelaria chega a 38,9%. A heterogeneidade da diferença salarial entre municípios é menor na região Sul/Sudeste, pois o desvio padrão (sd) é menor nessa parte do país.

Os dados da diferença salarial da hotelaria por região podem ser visualizados no gráfico boxplot apresentado a seguir.

```
boxplot(diferenca_salarial ~ sulsudeste, data = empregos)
```



Os dados observados revelam uma realidade concreta, permitindo-nos elaborar um enunciado particular. Contudo, na ciência, geralmente estamos interessados em enunciados gerais. A partir da análise concreta, não podemos saber se a desvantagem salarial é menor nas regiões Sul e Sudeste em razão de uma simples distribuição aleatória de valores entre municípios, ou se efetivamente podemos atribuir essa diferença à região. Para afirmar que a diferença é devida à região, temos que comparar a diferença de médias com aquela que seria esperada se as realidades municipais fossem aleatoriamente distribuídas no país. Em outras palavras, vamos testar a hipótese:

H1: A desvantagem salarial dos hoteleiros é menor nas regiões mais ricas do que nas regiões mais pobres.

A hipótese nula referente a H1 pode ser redigida como:

H0: A desvantagem salarial dos hoteleiros é a mesma nas regiões mais ricas e nas regiões mais pobres.

Para tanto, vamos começar testando a normalidade da distribuição de municípios segundo o valor da desvantagem salarial de hoteleiros.

```
empregos %>%
  group_by(sulsudeste) %>%
  shapiro_test(diferenca_salarial)
```

```
## # A tibble: 2 x 4
##   sulsudeste variable      statistic      p
##   <dbl> <chr>          <dbl>    <dbl>
## 1       0 diferenca_salarial  0.802 2.45e-25
## 2       1 diferenca_salarial  0.936 2.29e-17
```

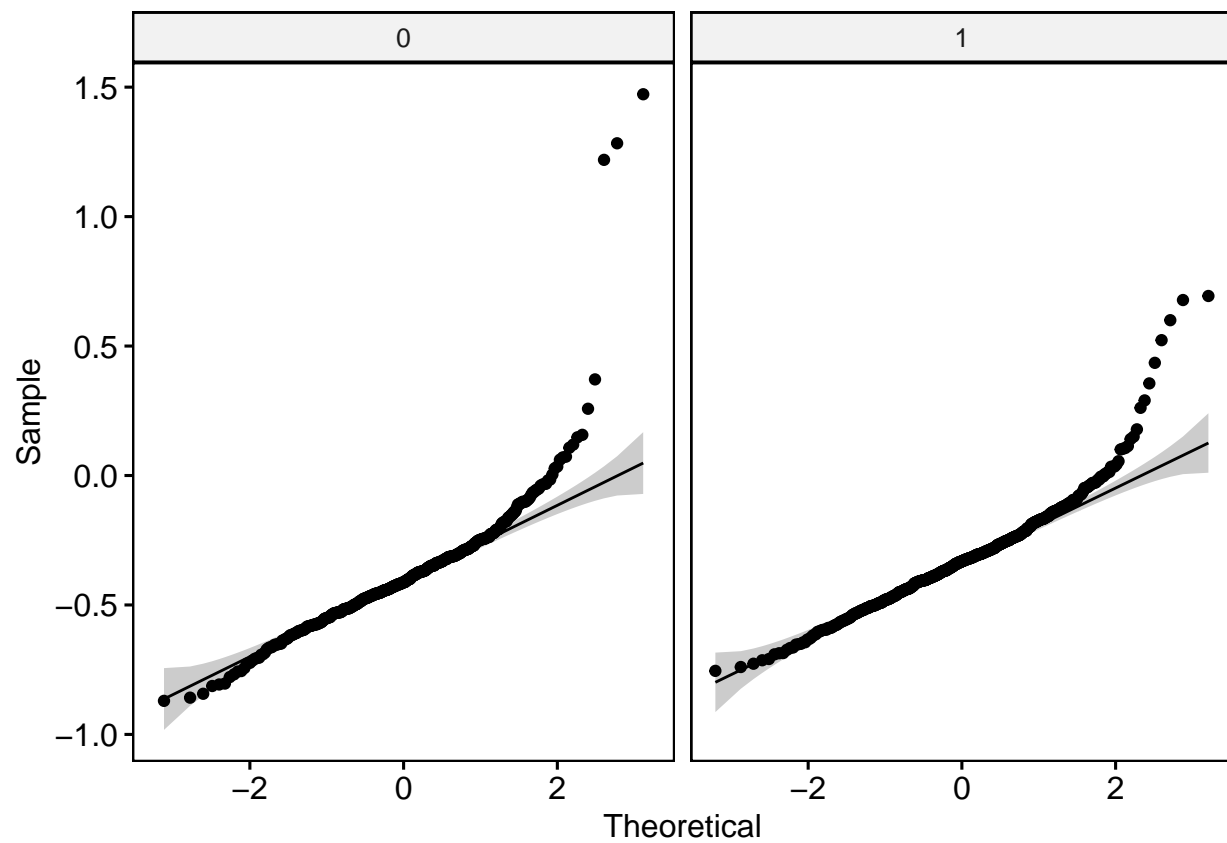
Os resultados indicam que as distribuições das duas partes do país se afastam da distribuição Normal, confirmando a impressão de assimetria positiva sugerida pelo boxplot. Uma alternativa a ser examinada neste caso é a consideração do logaritmo da diferença salarial.

```
empregos$log_diferenca_salarial <- log(empregos$diferenca_salarial)
empregos %>%
  group_by(sulsudeste) %>%
  shapiro_test(log_diferenca_salarial)
```

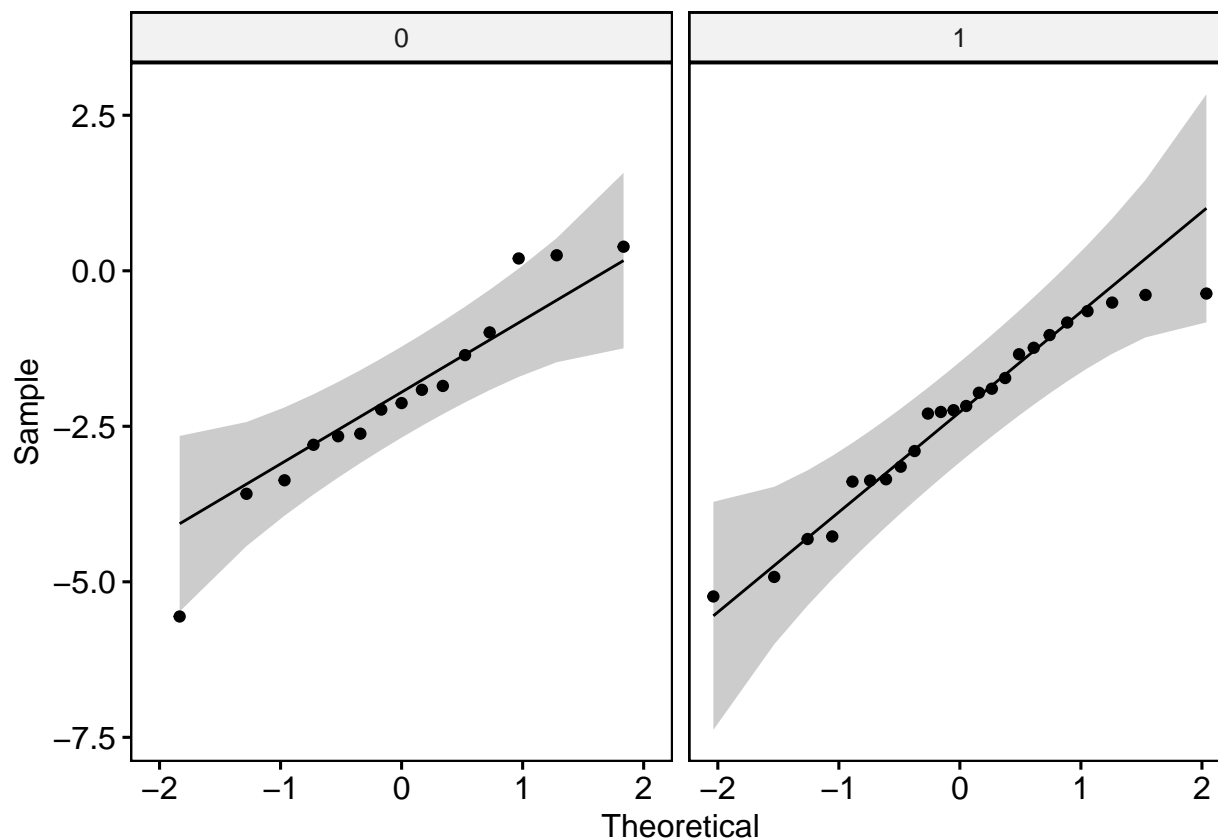
```
## # A tibble: 2 x 4
##   sulsudeste variable      statistic      p
##   <dbl> <chr>          <dbl>    <dbl>
## 1       0 log_diferenca_salarial  0.950 0.520
## 2       1 log_diferenca_salarial  0.948 0.248
```

Refazendo o teste de normalidade Shapiro-Wilk para o logaritmo da diferença salarial, chega-se à conclusão de que pode-se assumir que os municípios seguem uma distribuição Normal. As distribuições com e sem transformação logarítmica são apresentadas a seguir. Para elaborar os gráficos, vamos utilizar o pacote ggpubr.

```
library(ggpubr)
ggqqplot(empregos, x = "diferenca_salarial", facet.by = "sulsudeste")
```



```
ggqqplot(empregos, x = "log_diferenca_salarial", facet.by = "sulsudeste")
```



Além da normalidade, a escolha do teste de médias requer a análise da igualdade de variâncias da desvantagem salarial para os dois grupos em questão. A igualdade das variâncias para as duas partes do país é analisada com auxílio do teste de Levene.

```
empregos$sulsudeste <- as.character(empregos$sulsudeste)
empregos %>% levene_test(log_diferenca_salarial ~ sulsudeste)
```

```
## # A tibble: 1 x 4
##   df1 df2 statistic    p
##   <int> <int>     <dbl> <dbl>
## 1     1    37  0.00426 0.948
```

O resultado do teste de igualdade de variâncias indica que a hipótese de igualdade não deve ser descartada. Sendo assim, para a comparação de médias, utiliza-se o teste t de Student, que assume variâncias iguais entre os dois grupos.

```
empregos %>%
  t_test(log_diferenca_salarial ~ sulsudeste, var.equal = TRUE) %>%
  add_significance()
```

```
## # A tibble: 1 x 9
##   .y. group1 group2 n1 n2 statistic df p p.sig-1
##   <chr> <chr> <chr> <int> <int> <dbl> <dbl> <dbl> <chr>
## 1 log_diferenca_salarial 0 1 15 24 0.630 37 0.533 ns
## # ... with abbreviated variable name 1: p.signif
```


Os resultados apontam que a diferença entre regiões no logaritmo da diferença salarial não é significativa, sugerindo a rejeição de H1 e a aceitação de H0. Logo, conclui-se que não há evidências para descartar a hipótese de que a desvantagem salarial dos empregados do turismo é a mesma nas regiões Sul e Sudeste e no restante do país.

Por fim, apenas a título de exemplo, já que o teste t revelou a igualdade de médias, estima-se a dimensão da diferença encontrada entre as duas partes do país. Utilizando-se a estatística d de Cohen, conclui-se que o tamanho do efeito é 0,207, valor que é tradicionalmente classificado como pequeno.

```
empregos %>% cohens_d(log_diferenca_salarial ~ sulsudeste, var.equal = TRUE)
```

```
## # A tibble: 1 x 7
##   .y.          group1 group2 effsize    n1    n2 magnitude
## * <chr>      <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 log_diferenca_salarial 0      1      0.207   15    24 small
```

Referência <https://www.datanovia.com/en/lessons/t-test-in-r/>