SME 0810 - Métodos Não Paramétricos

Trabalho Prático 1/3

Testes para Duas Amostras Independentes

Aluno(a) 1: Ada Maris Pereira Mário_____ Nº USP: 12725432
Aluno(a) 2: Emanuel Victor da Silva Favorato_____ Nº USP: 12558151

RELATÓRIO

1 Base de Dados

A base de dados utilizada neste estudo foi extraída do conjunto disponibilizado por Arnab Chaki no Kaggle, intitulado Data Science Salaries 2023 [Kaggle, 2023]. Essa base contém informações detalhadas sobre salários de profissionais de tecnologia, suas funções, localizações e níveis de experiência, coletadas ao longo de 2023.

Para a realização dos testes apresentados neste relatório, consideramos as seguintes variáveis principais:

- `salary_in_usd`: salário anual em dólares americanos, utilizado para análises de distribuição salarial e cálculo de medidas de tendência central.
- `employee_residence`: região geográfica de onde o profissional atua, permitindo comparações entre os Estados Unidos e a União Europeia.
- `experience_level`: nível de experiência do trabalhador, aqui tomadas as categorias *Senior* ou *Executive*.

As estatísticas da variáveis utilizadas estão descritas na Tabela 1, em que as colunas 4 a 8 representam as estatísticas de `salary_in_usd` em função das varíaveis categóricas.

Variável		Proporção	Min	Max	Média	Mediana	Variância
`experience_level`	SE	0.67	15000	416000	194930.93	196000	4.993108e+09
	EX	0.03	8000	423834	153051.07	14600	3.237185e+09
`employee_residence`	US	0.80	24000	450000	152822.01	145000	3.064662e+09
	EU	0.06	6304	214618	62281.73	59020.0	8.684043e+08

Tabela 1: Proporções e estatísticas salariais por nível de experiência e região.

2 Testes

Para a execução do projeto, consideramos um nível de significância $\alpha = 5\%$, já na formulação das hipóteses, foi considerada a visualização dos histogramas das variáveis de interesse, bem como o debate atual acerca da discrepância salarial entre os Estados Unidos e a União Europeia no setor de tecnologia. Para testar nossas suspeitas acerca dos problemas em questão, utilizamos os testes não-paramétricos para Duas Amostras Independentes, cujos resultados serão discutidos a seguir:

2.1 Teste Qui-Quadrado de Independência

O Teste Qui-Quadrado de Independência foi utilizado com o intuito de verificar se a proporção de trabalhadores no nível senior(SE) e executivo(EX) com salários acima de US$100.000 é semelhante, para tal, foram formuladas as hipóteses a seguir:

- H_0: $p_1 = p_2$ (A proporção de empregados com salários acima de US$100.000 é igual para os níveis de experiência SE e EX)
- H_1: $p_1 \neq p_2$ (A proporção de empregados com salários acima de US$100.000 é diferente para os níveis de experiência SE e EX)

Onde:

- p_1: proporção de empregados no nível SE com salários superiores a US$100.000
- p_2: proporção de empregados no nível EX com salários superiores a US$100.000

A análise dos resultados nos trouxe uma estatística de teste $q \approx 7.8792$ e um valor-p ≈ 0.005, nos trazendo evidências de que a proporção de trabalhadores nos níveis SE e EX com ganhos acima de US$100.000 não é a mesma.

2.2 Teste Exato de Fisher

Alternativamente ao Teste Qui-Quadrado de Independência, utilizamos o Teste Exato de Fisher, o qual não é recomendado por ser menos robusto, não poder ter seus resultados replicados para a população e devido ao fato de as condições para a aplicação do Teste Qui-Quadrado terem sido satisfeitas. Para a formulação das hipóteses, consideramos a mesma suposição do problema anterior.

Após a execução do teste, obtivemos uma razão de chances de aproximadamente 3.0106 e um valor-p 0.00236, nos levando a concluir que a proporção de trabalhadores cujo salário excede US$100.000 para os dois níveis é diferente.

2.3 Teste da Mediana

O teste da Mediana foi empregado com o intuito de compreender se as duas amostras são provenientes de populações com a mesma mediana, uma vez que existe o interesse em analisar se a renda mediana dos profissionais da área de Ciência de Dados nas duas regiões é igual, para tanto, foram formuladas as seguintes hipóteses:

- H_0: $p_1 = p_2$ (A renda mediana das duas populações é igual)
- H_1: $p_1 > p_2$ (A renda mediana nos Estados Unidos é maior que na União Europeia)

Onde:

- p_1: renda mediana nos Estados Unidos
- p_2: renda mediana na União Europeia

Após a devida aplicação do teste, obtivemos uma estatística de teste $t \approx 222.33$ e um valor-p próximo de 0, o que nos trouxe evidências de que a renda mediana nos Estados Unidos é, na verdade, maior que na União Europeia, como se suspeitava inicialmente.

2.4 Teste das Somas dos Ranks (Wilcoxon, Mann e Whitney)

Este teste foi utilizado para a verificação da renda média nas duas regiões, uma vez que existe o interesse em entender se é plausível esperar que os cidadãos europeus recebam um salário equivalente ao dos norte-americanos para esse ramo da tecnologia, para tal averiguação, foram consideradas as seguintes afirmações:

- H_0: $E(X) = E(Y)$ (A renda média nas duas regiões é a mesma)
- H_1: $E(X) > E(Y)$ (A renda média nos Estados Unidos é maior que na União Europeia)

Onde:

- $E(X)$: renda média nos Estados Unidos
- $E(Y)$: renda média na União Europeia

O Teste de Wilcoxon nos retornou uma estatística de teste $w = 5.181.839$ e um valor-p muito próximo de zero, o que nos traz evidências de que se espera que os profissionais da área de Ciência de Dados nos EUA obtenham ganhos superiores àqueles observados para os europeus.

2.5 Teste de Kolmogorov-Smirnov para Duas Amostras

Já para este teste, queremos observar se a renda per-capita nos EUA e na UE seguem a mesma distribuição probabilística. As hipóteses formuladas seguem abaixo:

- H_0: $F_X(t) = G_Y(t)$, para todo $t \in \mathbb{R}$ (As distribuições de X e Y são similares)
- H_1: $F_X(t) \neq G_Y(t)$, para pelo menos um $t \in \mathbb{R}$ (As distribuições de X e Y não são similares)

Onde:

- X: distribuição salarial nos Estados Unidos
- Y: distribuição salarial na União Europeia

A análise dos resultados nos trouxe uma estatística de teste $t \approx 0.7854$ e um valor-p próximo de zero, o que nos leva a concluir que a distribuição salarial nas duas regiões não é a mesma.

3 Considerações Finais

Neste trabalho, foram aplicados testes não-paramétricos para a comparação de duas amostras independentes. Utilizamos os testes da Mediana, das Somas dos Ranks e de Kolmogorov-Smirnov para Duas Amostras para testar hipóteses acerca da discrepância salarial para profissionais de Ciência de Dados nos Estados Unidos e na União Europeia, enquanto os testes Qui-Quadrado e Exato de Fisher foram utilizados para comparar os salários em duas categorias profissionais distintas

Os resultados nos mostraram que, ao nível de significância de 5%, as proporções de indivíduos nas classes SE e EX cuja renda ultrapassa os US$100.000 são distintas, bem como que os salários nos EUA tendem a ser consideravelmente superiores aos Europeus, diferindo quanto a distribuição e apresentando valores mais elevados para os ganhos esperado e mediano.

Tais resultados corroboram o senso comum de que os executivos tendem a receber mais que os seniores, bem como as especulações mercadológicas de que os ganhos dos trabalhadores europeus estão defasados em comparação aos americanos.

Referências

Kaggle. Data science salaries 2023, 2023. Disponível em: https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023. Acesso em: 11 de novembro de 2024.