SME 0810 - Métodos Não Paramétricos

Trabalho Prático 1/3

Testes para Duas Amostras Independentes

Aluno(a) 1: Gabriel Felipe Machado de Oliveira Nº USP: 11908695
Aluno(a) 2: Karen Neves Perciliano Nº USP:12559357
Aluno(a) 3: Thalita Sousa Pereira Nº USP: 12559187

1 Introdução

O estudo de métodos não paramétricos para duas amostras independentes é de grande importância para a Estatística. Esses métodos permitem investigar relações, diferenças ou associações entre duas amostras distintas sem a necessidade de pressupor distribuições específicas para os dados. Assim, torna-se possível avaliar a independência ou a existência de padrões nas amostras com maior flexibilidade.

Neste estudo, foram aplicados de forma prática os seguintes testes para duas amostras independentes: (1) Teste Qui-Quadrado de Independência; (2) Teste Exato de Fisher; (3) Teste da Mediana; (4) Teste de Soma dos Ranks de Wilcoxon, Menn e Whitney; (5) Teste Kolmogorov-Smirnov para duas amostras. Para cada um deles, foram selecionados datasets adequados, sobre os quais foram definidas hipóteses específicas a serem testadas. As análises foram realizadas utilizando o software estatístico R.

Ao final de cada aplicação, são apresentadas as conclusões e comentários pertinentes, destacando a adequabilidade dos métodos às hipóteses propostas e avaliando suas implicações práticas.

2 Testes Não Paramétricos para Duas Amostras Independentes

2.1 Teste Qui-Quadrado (χ^2) de Indepenência

O Teste Qui-Quadrado de Independência possui como enfoque principal contextos em que há duas amostras independentes, e possui como finalidade determinar a significância de diferenças entre dois grupos dado a aplicação de um determinado tratamento aos dados da amostra.

2.1.1 Descrição dos Dados

Para a aplicação do Qui-Quadrado, foi utilizado um Dataset da plataforma Kaggle, que traz inforações de rendas de homens e mulheres. Na base, há 15 colunas e mas apenas duas utilizadas: `gender` (variável qualitativa nominal) e `income` (variável dicotômica podendo ser $>$ 50k de renda ou $\leq$50k), e além disso, um total de 48.842 observações.

2.1.2 Objetivos e Hipóteses

O objetivo principal foi investigar a relação da renda e dos sexo das pessoas da amostra. Declaração das hipóteses:

H_0 : Não há relação entre o sexo e a renda das pessoas

x

H_1 : Há relação entre o sexo e a renda das pessoas

Como tabela de contingência, ocorreu da seguinte forma:

	Menor ou igual a 50k	Maior que 50k
Female	9592	1179
Male	15128	6662

2.1.3 Resultados

Trata-se então de um teste bilateral. Após a aplicação do método obteve-se um uma estatística de teste $q - squared = 1517,9$ e ainda $p - valor < 2,2e - 16$.

2.1.4 Conclusões e Comentários

Ao final, aplicando o Teste Qui-Quadrado na amostra de homens e mulheres e suas respectivas rendas, obteve-se que $p - valor < 2,2e - 16 < \alpha = 0,05$, logo, rejeita-se H_0.

Portanto, ao nível de significância de 5% os dados deram evidências estatisticamente suficientes de que há uma relação significativa entras as varáveis analisadas, sexo e renda, uma vez que a hipótese nula de indepencência foi rejeitada.

2.2 Teste Exato de Fisher

O Teste dos Sinais possui como enfoque principal contextos em que há duas amostras independentes, e possui como finalidade examinar se duas populações diferem entre si, comparando as proporções das unidades amostrais nas duas classificações ou grupos.

2.2.1 Descrição dos Dados

Para a aplicação do Teste Exato de Fischer, foi utilizado o dataset que fornce informações de pacientes com ou sem hipertensão. Na base há ao todo 5110 observações e 12 colunas.Foi utilizado somente as colunas `hypertension` (coluna dicotômica qualitativa nominal que indicas 1 para presença de hipertensão e 0 para ausência) e `gender` (coluna também qualitativa nominal dicotômica que indica o gênero.)

2.2.2 Objetivos e Hipóteses

O objetivo principal foi investigar a diferença das proporções de pacientes com hipertensão observados nos grupos masculinos e femininos. Declaração das hipóteses:

$$H_0 : p_1 = p_2$$
$$\text{x}$$
$$H_1 : p_1 \neq p_2$$

A table de contingência ocorreu da seguinte forma:

	Hipertensão	Não Hipertensão
Feminino	276	2718
Masculino	222	1893

2.2.3 Resultados

Trata-se então de um teste bilateral. Utilizando um nível de significância $\alpha = 0,05 = 5\%$ para a verificação das hipóteses, após a aplicação do método obteve-se um $p - valor = 0,1377$.

2.2.4 Conclusões e Comentários

Assim, ao final, aplicando o Exato de Fisher na amostra dos pacientes passíveis de hipertensão, verificou-se que $p - valor = 0,1377 > \alpha = 0,05$, logo, não rejeita-se H_0.

Portanto, ao nível de significância de 5% os dados deram evidências estatisticamente suficientes de que os grupos não possuem diferenças significativas com relação à hipertensão sendo do sexo masculino ou sexo feminino, uma vez que não rejeitamos a hipótese nula.

2.3 Teste da Mediana

O Teste da Mediana é utilizado em contextos em que há duas amostras Independentes. A ideia principal é verificar se c amostras independentes diferem pelas locações (tendências centrais), ou seja, se c grupos são provenientes de populações com a mesma mediana.

2.3.1 Descrição dos Dados

Para a aplicação do Teste da Mediana, foi utilizado um Dataset fornecido pelo sotware R chamado *PlantGrowth*. A base fornece resultados de um experimento que comparou os rendimentos (medidos pelo peso plantas secas) obtidos sob uma condição de controle e duas condições de tratamento diferentes. Isto é, há duas colunas `weight` (variável contínua) e `group`

(variável qualitativa nominal) que compõem a base de dados, e 30 linhas no total. Abaixo algumas medidas estatísticas dos dados:

Medida	**Valor**
Min.	3,590
1st Qu.	4,550
Mediana	5,155
Média	5,073
3rd Qu.	5,530
Max.	6,310

Tabela 2.3.1: Sumário estatísco da variável `weight`

2.3.2 Objetivos e Hipóteses

O objetivo principal foi investigar se as amostras de controle e tratamento são provenientes de populações com medianas iguais ou distintas.

Declaração das hipóteses:

H_0 : as amostras são provenientes de populações com mesma mediana;

x

H_1 : pelo menos duas amostras são provenientes de populações com medianas diferentes.

A table de contingência ocorreu da seguinte forma:

	Abaixo da Mediana	Acima da Mediana
Controle	4	6
Tratamento	6	4

2.3.3 Resultados

Trata-se então de um teste bilateral. Utilizando um nível de significância $\alpha = 0,05 = 5\%$ para a verificação das hipóteses, após a aplicação do método obteve-se um $p - valor = 0,655$ bem como uma estatística de teste $X = 0,2$.

2.3.4 Conclusões e Comentários

Ou seja, ao final, aplicando o Teste da Mediana na amostras de plantas submetidas ao tratamento de solo, obteve-se que $p - valor = 0,655 > \alpha = 0,05$, logo, não rejeita-se H_0.

Portanto, ao nível de significância de 5% os dados deram evidências estatisticamente suficientes de que as amostras são provenientes de populações com a mesma mediana, isto é, não houve diferença nas medianas das amostras de controle e de tratamento, dado que não rejeitamos a hipótese nula.

2.4 Teste da Soma dos Ranks

O Teste da Soma dos Ranks é utilizado em contextos em que há duas amostras Independentes. A ideia principal é comprovar se dois grupos independentes foram ou não extraídos de uma mesma população.

2.4.1 Descrição dos Dados

Para a aplicação do teste, também foi utilizado o dataset que fornce informações de pacientes com ou sem hipertensão. Como visto, na base há ao todo 5110 observações e 12 colunas. Aqui, utilizou-se somente as colunas `hypertension` (coluna dicotômica qualitativa nominal que indicas 1 para presença de hipertensão e 0 para ausência) e `avg-glucose-level` (coluna quantitaiva contínua que traz a média dos níveis de clicose do paciente). Abaixo um breve resumo estatístico sobre a variável contínua:

Medida	**Valor**
Min.	55,120
1st Qu.	77,245
Mediana	91,885
Média	106,147
3rd Qu.	114,090
Max.	271,740

Tabela 2.4.1: Sumário estatísco da variável `avg-glucose-level`

2.4.2 Objetivos e Hipóteses

O objetivo principal foi investigar se houve diferença dos níveis médios de glicose no sange das amostras dos pacientes com hipertensão masculino e feminino, e se essa diferença foi significativa.

Declaração das hipóteses:

$$H_0 : \mathrm{p}_1 = p_2$$

(Não existe diferença entre o nível médio de glicose entre os dois grupos, masculino e feminino.)

x

$H_1 : p_1 \neq p_2$

(Existe diferença entre o nível médio de glicose entre os dois grupos, masculino e feminino.)

2.4.3 Resultados

Trata-se então de um teste bilateral. Utilizando um nível de significância $\alpha = 0,05 = 5\%$ para a verificação das hipóteses, após a aplicação do teste da Soma dos Ranks obteve-se um $p - valor = 0,0004816$ bem como uma estatística de teste $W = 2984886$.

2.4.4 Conclusões e Comentários

Ao final, ao aplicar o Teste da Soma dos Ranks na amostra dos níveis médios de glicose para pacientes do sexo masculino e feminino, obteve-se que $p-valor = 0,0004816 < \alpha = 0,05$, Logo, rejeita-se H_0.

Ao nível de significância de 5% os dados deram evidências estatisticamente suficientes de que houve diferença nos níveis de glicose entre pacientes do sexo masculino e feminino, uma vez que a hipótes nula foi rejeitada.

2.5 Teste de Kolmogorv Smirnov

O Teste da Soma dos Ranks é utilizado em contextos em que há duas amostras Independentes. Ao final busca-se verificar se duas amostras independentes pertencem à mesma população ou se provêm de populações com a mesma distribuição.

2.5.1 Descrição dos Dados

Para a aplicação do teste, também foi utilizado o dataset que fornece informações de eficiência de motores em carros automáticos e carros manuais. Os dados foram extraídos da revista Motor Trend US de 1974 e compreendem o consumo de combustível e 10 aspectos do design e desempenho do automóvel para 32 automóveis (modelos de 1973-74), isto é, há 32 observações na amostra.

Para o teste, utilizou as colunas `mpg` (variável numérica dicotômica, indica se os carros são automáticos com 1 ou manuais com 0) e `am` (variável quantitativa contínua que indica a eficiência/gasto do motor). Abaixo uma sumarização da variável `am`:

Medida	Valor
Min.	10,400
1st Qu.	15,430
Mediana	19,200
Média	20,09
3rd Qu.	22,800
Max.	33,900

Tabela 2.5.1: Sumário estatísco da variável `am`

2.5.2 Objetivos e Hipóteses

O objtivo, ao final, foi investigar se os carros automáticos possuem uma eficiência menor ou igual aos manuais (hipóte nula), assim, atestar se essa diferença é estatisticamente significativa. Ou seja:

H_0 : Carros automáticos têm em média, uma eficiência menor ou igual à de carros manais.

x

H_1 : Carros automáticos têm em média, uma eficiência maior à de carros manais.

2.5.3 Resultados

Trata-se então de um teste unilateral à direita. Utilizando um nível de significância $\alpha = 0,05 = 5\%$ para a verificação das hipóteses, após a aplicação do teste de Kolmogorov-Smirnov obteve-se um $p - valor = 0,0,0009701$ bem como uma estatística de teste $D^+ = 0,63563$.

2.5.4 Conclusões e Comentários

Ao final, ao aplicar o Teste de Kolmogorov-Smirnov para amostras independentes de carros automáticos e manuais, obteve-se que $p - valor = 0,0009701 < \alpha = 0,05$, logo, rejeita-se H_0.

Portanto, ao nível de significância de 5% os dados deram evidências estatisticamente suficientes de que carros automáticos possuem, em média, eficiência em seu funcionamento maior que carros manuais, dado que rejeitamos a hipótes nula.

3 Referências

- Plataforma Kaggle, dataset adult.csv. URL:https://www.kaggle.com/datasets/wenruliu/adult-income-dataset. Acesso em 17 de novembro de 2024.

- Plataforma Kaggle, dataset. URL:https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset. Acesso em 17 de novembro de 2024.
- Software R. Data PlantGrowth. URL: https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/PlantGrowth. Acesso em 17 de novembro de 2024.
- Software R. Data mtcars. URL: https://search.r-project.org/CRAN/refmans/explore/html/use-data-mtcars.html. Acesso em 17 de novembro de 2024.