consultoria estatística

Consultores Responsáveis:
Francisco Italo Rio Andrade

Requerente:
João Vitor Neves

Sumário

1 Análise Descritiva Univariada

1.1 Frequência Relativa

A frequência relativa é utilizada para a comparação entre classes de uma variável categórica com c categorias, ou para comparar uma mesma categoria em diferentes estudos.

A frequência relativa da categoria j é dada por:

$$f_j = \frac{n_j}{n}$$

Com:

- $j = 1, \ ..., \ c$
- $n_j =$ número de observações da categoria j
- $n =$ número total de observações

Geralmente, a frequência relativa é utilizada em porcentagem, dada por:

$$100 \times f_j$$

1.2 Média

A média é a soma das observações dividida pelo número total delas, dada pela fórmula:

$$\bar{X} = \frac{\sum\limits_{i=1}^{n} X_i}{n}$$

Com:

- $i = 1, \ 2, \ ..., \ n$
- $n =$ número total de observações

1.3 Mediana

Sejam as n observações de um conjunto de dados $X = X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ de determinada variável ordenadas de forma crescente. A mediana do conjunto de dados X é o valor que deixa metade das observações abaixo dela e metade dos dados acima.

Com isso, pode-se calcular a mediana da seguinte forma:

$$med(X) = \begin{cases} X_{\frac{n+1}{2}}, \text{para n ímpar} \\ \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}, \text{para n par} \end{cases}$$

1.4 Quartis

Os quartis são separatrizes que dividem o conjunto de dados em quatro partes iguais. O primeiro quartil (ou inferior) delimita os 25% menores valores, o segundo representa a mediana, e o terceiro delimita os 25% maiores valores. Inicialmente deve-se calcular a posição do quartil:

- Posição do primeiro quartil P_1:

$$P_1 = \frac{n+1}{4}$$

- Posição da mediana (segundo quartil) P_2:

$$P_2 = \frac{n+1}{2}$$

- Posição do terceiro quartil P_3:

$$P_3 = \frac{3 \times (n+1)}{4}$$

Com n sendo o tamanho da amostra. Dessa forma, $X_{(P_i)}$ é o valor do i-ésimo quartil, onde $X_{(j)}$ representa a j-ésima observação dos dados ordenados.

Se o cálculo da posição resultar em uma fração, deve-se fazer a média entre o valor que está na posição do inteiro anterior e do seguinte ao da posição.

1.5 Variância

A variância é uma medida que avalia o quanto os dados estão dispersos em relação à média, em uma escala ao quadrado da escala dos dados.

1.5.1 Variância Populacional

Para uma população, a variância é dada por:

$$\sigma^2 = \frac{\sum\limits_{i=1}^{N} (X_i - \mu)^2}{N}$$

Com:

- $X_i = i$-ésima observação da população
- $\mu =$ média populacional
- $N =$ tamanho da população

1.5.2 Variância Amostral

Para uma amostra, a variância é dada por:

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}$$

Com:

- $X_i =$ i-ésima observação da amostra
- $\bar{X} =$ média amostral
- $n =$ tamanho da amostra

1.6 Desvio Padrão

O desvio padrão é a raiz quadrada da variância. Ele avalia o quanto os dados estão dispersos em relação à média.

1.6.1 Desvio Padrão Populacional

Para uma população, o desvio padrão é dado por:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} \left(X_i - \mu\right)^2}{N}}$$

Com:

- $X_i =$ i-ésima observação da população
- $\mu =$ média populacional
- $N =$ tamanho da população

1.6.2 Desvio Padrão Amostral

Para uma amostra, o desvio padrão é dado por:

$$S = \sqrt{\frac{\sum\limits_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}}$$

Com:

- X_i = i-ésima observação da amostra
- $\bar{X}$ = média amostral
- n = tamanho da amostra

1.7 Coeficiente de Variação

O coeficiente de variação fornece a dispersão dos dados em relação à média. Quanto menor for o seu valor, mais homogêneos serão os dados. O coeficiente de variação é considerado baixo (apontando um conjunto de dados homogêneo) quando for menor ou igual a 25%. Ele é dado pela fórmula:

$$C_V = \frac{S}{\bar{X}} \times 100$$

Com:

- S = desvio padrão amostral
- $\bar{X}$ = média amostral

1.8 Coeficiente de Assimetria

O coeficiente de assimetria quantifica a simetria dos dados. Um valor positivo indica que os dados estão concentrados à esquerda em sua função de distribuição, enquanto um valor negativo indica maior concentração à direita. A fórmula é:

$$C_{Assimetria} = \frac{1}{n} \times \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S}\right)^3$$

Com:

- X_i = i-ésima observação da amostra
- $\bar{X}$ = média amostral

- $S =$ desvio padrão amostral
- $n =$ tamanho da amostra

1.9 Curtose

O coeficiente de curtose quantifica o achatamento da função de distribuição em relação à distribuição Normal e é dado por:

$$Curtose = \frac{1}{n} \times \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S} \right)^4 - 3$$

Com:

- $X_i =$ i-ésima observação da amostra
- $\bar{X} =$ média amostral
- $S =$ desvio padrão amostral
- $n =$ tamanho da amostra

Uma distribuição é dita mesocúrtica quando possui curtose nula. Quando a curtose é positiva, a distribuição é leptocúrtica (mais afunilada e com pico). Valores negativos indicam uma distribuição platicúrtica (mais achatada).

1.10 Boxplot

O boxplot é uma representação gráfica na qual se pode perceber de forma mais clara como os dados estão distribuídos. A figura abaixo ilustra um exemplo de boxplot.

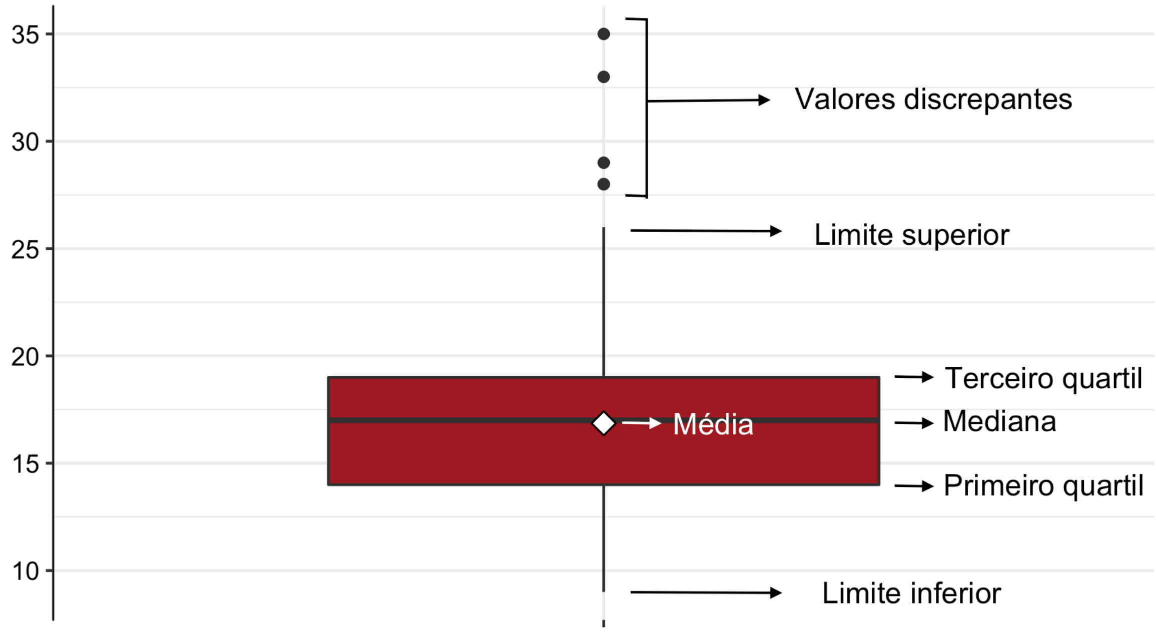

Figura 1: Exemplo de boxplot

A porção inferior do retângulo diz respeito ao primeiro quartil, enquanto a superior indica o terceiro quartil. Já o traço no interior do retângulo representa a mediana do conjunto de dados, ou seja, o valor em que o conjunto de dados é dividido em dois subconjuntos de mesmo tamanho. A média é representada pelo losango branco e os pontos são *outliers*. Os *outliers* são valores discrepantes da série de dados, ou seja, valores que não demonstram a realidade de um conjunto de dados.

1.11 Histograma

O histograma é uma representação gráfica utilizada para a visualização da distribuição dos dados e pode ser construído por valores absolutos, frequência relativa ou densidade. A figura abaixo ilustra um exemplo de histograma.

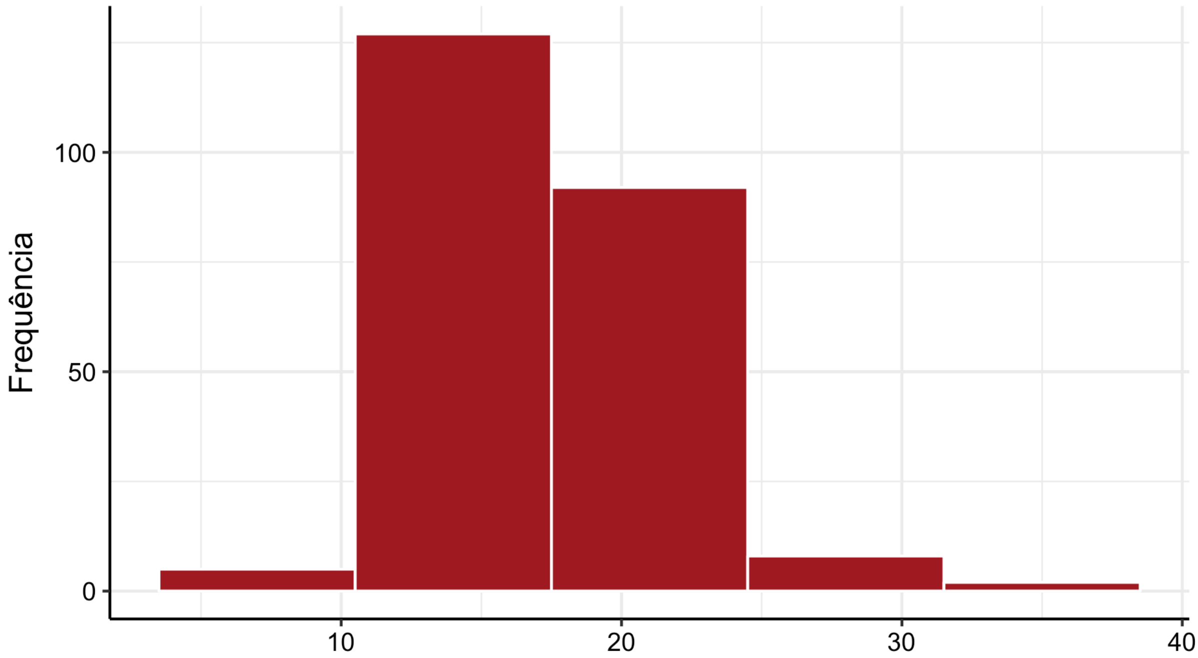

Figura 2: Exemplo de histograma

1.12 Gráfico de Dispersão

O gráfico de dispersão é uma representação gráfica utilizada para ilustrar o comportamento conjunto de duas variáveis quantitativas. A figura abaixo ilustra um exemplo de gráfico de dispersão, onde cada ponto representa uma observação do banco de dados.

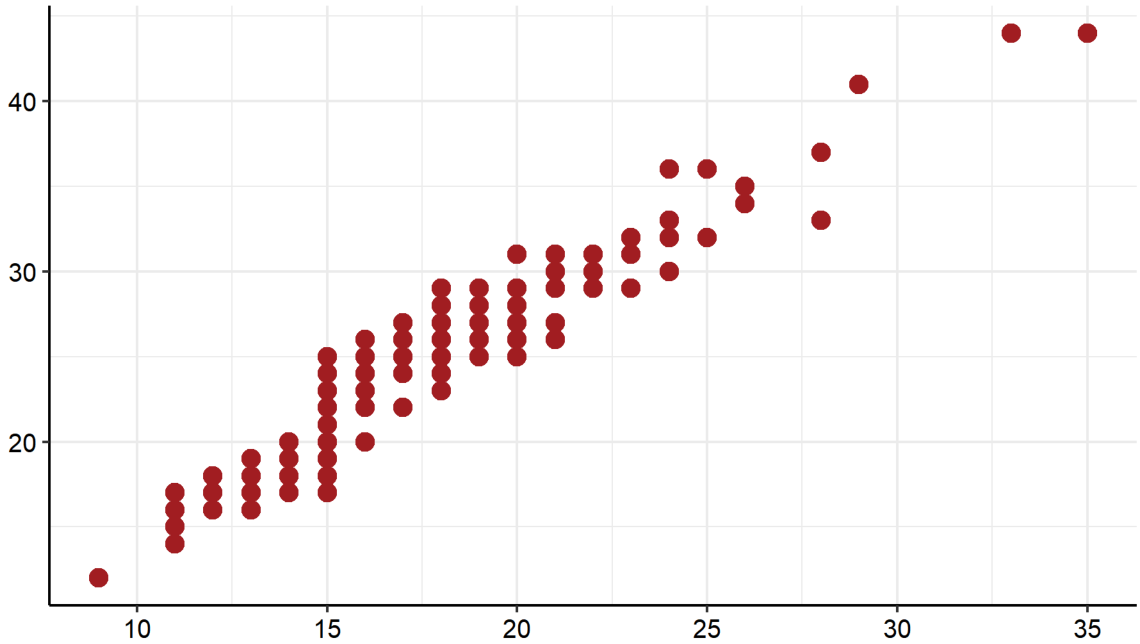

Figura 3: Exemplo de Gráfico de Dispersão

1.13 Tipos de Variáveis

1.13.1 Qualitativas

As variáveis qualitativas são as variáveis não numéricas, que representam categorias ou características da população. Estas subdividem-se em:

- **Nominais**: quando não existe uma ordem entre as categorias da variável (exemplos: sexo, cor dos olhos, fumante ou não, etc)
- **Ordinais**: quando existe uma ordem entre as categorias da variável (exemplos: nível de escolaridade, mês, estágio de doença, etc)

1.13.2 Quantitativas

As variáveis quantitativas são as variáveis numéricas, que representam características numéricas da população, ou seja, quantidades. Estas subdividem-se em:

- **Discretas**: quando os possíveis valores são enumeráveis (exemplos: número de filhos, número de cigarros fumados, etc)
- **Contínuas**: quando os possíveis valores são resultado de medições (exemplos: massa, altura, tempo, etc)

1.14 Coeficiente de Correlação de Pearson

O coeficiente de correlação de Pearson é uma medida que verifica o grau de relação linear entre duas variáveis quantitativas. Este coeficiente varia entre os valores -1 e 1. O valor zero significa que não há relação linear entre as variáveis. Quando o valor do coeficiente r é negativo, diz-se existir uma relação de grandeza inversamente proporcional entre as variáveis. Analogamente, quando r é positivo, diz-se que as duas variáveis são diretamente proporcionais.

O coeficiente de correlação de Pearson é normalmente representado pela letra r e a sua fórmula de cálculo é:

$$r_{Pearson} = \frac{\sum_{i=1}^{n} \left[(x_i - \bar{x}) \left(y_i - \bar{y} \right) \right]}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \times \sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

Onde:

- $x_i =$ i-ésimo valor da variável X
- $y_i =$ i-ésimo valor da variável Y
- $\bar{x} =$ média dos valores da variável X
- $\bar{y} =$ média dos valores da variável Y

Vale ressaltar que o coeficiente de Pearson é paramétrico e, portanto, sensível quanto à normalidade (simetria) dos dados.

1.15 Coeficiente de Correlação de Spearman

O coeficiente de correlação de Spearman é uma medida não paramétrica que verifica, através de postos de variáveis quantitativas ou qualitativas ordinais, o grau de relação linear entre duas variáveis. Este coeficiente varia entre os valores -1 e 1. O valor zero significa que não há relação linear entre as variáveis. Quando o valor do coeficiente ρ é negativo, diz-se existir uma relação de grandeza inversamente proporcional entre as variáveis. Analogamente, quando ρ é positivo, diz-se que as duas variáveis são diretamente proporcionais.

O coeficiente é calculado da seguinte maneira:

$$\rho_{Spearman} = \frac{\sum_{i=1}^{n}\left[\left(R(x_i) - \frac{n+1}{2}\right)\left(R(y_i) - \frac{n+1}{2}\right)\right]}{\sqrt{\sum_{i=1}^{n}(R(x_i)^2) - n\left(\frac{n+1}{2}\right)^2} \times \sqrt{\sum_{i=1}^{n}(R(y_i)^2) - n\left(\frac{n+1}{2}\right)^2}}$$

Onde:

- $x_i =$ i-ésimo valor da variável X
- $y_i =$ i-ésimo valor da variável Y
- $R(x_i) =$ posto relativo à observação i de X
- $R(y_i) =$ posto relativo à observação i de Y
- $n =$ número total de observações na amostra

1.16 Coeficiente de Correlação de Kendall

O coeficiente de correlação de Kendall é uma medida não paramétrica que verifica o grau de relação linear entre duas variáveis. Este coeficiente varia entre os valores -1 e 1 e utiliza observações pareadas. O valor zero significa que não há relação linear entre as variáveis. Quando o valor do coeficiente τ é negativo, diz-se existir uma relação de grandeza inversamente proporcional entre as variáveis. Analogamente, quando τ é positivo, diz-se que as duas variáveis são diretamente proporcionais.

O coeficiente de correlação de Kendall é normalmente representado pela letra τ, e sua fórmula de cálculo é:

$$\tau = \frac{C - D}{\frac{n(n-1)}{2}}$$

Onde:

- $C =$ número de pares concordantes
- $D =$ número de pares discordantes
- $n =$ tamanho da amostra

Os pares (x_i, y_i) e (x_j, y_j) são considerados concordantes se ambas as partes concordam, ou seja, se $x_i > x_j$ e $y_i > y_j$ ou se $x_i < x_j$ e $y_i < y_j$.

Já os pares (x_i, y_i) e (x_j, y_j) são discordantes se as partes discordam, ou seja, se $x_i > x_j$ e $y_i < y_j$ ou se $x_i < x_j$ e $y_i > y_j$.

1.17 Coeficiente de Goodman-Kruskal

O λ de Goodman-Kruskal mede a associação para tabulações cruzadas de variáveis qualitativas nominais. Ele mede a melhoria percentual da probabilidade da variável dependente dado o valor de outras variáveis.

O coeficiente de Goodman-Kruskal é normalmente representado pela letra λ, e sua fórmula de cálculo é:

$$\lambda = \frac{S - R}{N - R}$$

Onde:

- $S =$ a soma da maior frequência das células para cada linha
- $R =$ o maior total de linha
- $N =$ o total de todas as frequências das células

1.18 Coeficiente de Determinação (R^2)

O coeficiente R^2 de determinação utiliza a variância dentro de cada grupo para explicar a variância global dos dados. Uma forma de quantificar essa medida é utilizar a média das variâncias em cada categoria, dada por:

$$\overline{var(S)} = \frac{\sum_{i=1}^{k} n_i \times var_i(S)}{n}$$

Onde:

- $n =$ tamanho total da amostra
- $var_i(S) =$ variância dentro da categoria i
- $n_i =$ tamanho da amostra i

Assim, o coeficiente de determinação é dado por:

$$R^2 = 1 - \frac{\overline{var(S)}}{var(S)}$$

Com \$ 0 ≤R^2 ≤1\$. O valor 1 indica que a variável categórica explica 100% da variação da variável quantitativa, enquanto o valor 0 indica ausência de impacto.

1.19 Qui-Quadrado

A estatística Qui-Quadrado é uma medida de divergência entre a distribuição dos dados e uma distribuição esperada ou hipotética. Também pode ser usada para verificar independência ou associação entre variáveis categóricas. É calculada por:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Onde:

- O_i = frequência observada
- E_i = frequência esperada

1.20 Coeficiente de Contingência

O coeficiente de contingência é derivado do Qui-Quadrado e ajusta seu valor para fornecer um referencial de comparação. Seu cálculo é:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Onde:

- χ^2 = valor da estatística Qui-Quadrado
- n = tamanho da amostra

1.21 Coeficiente de Contingência Corrigido

O coeficiente de contingência corrigido ajusta o coeficiente de contingência, permitindo uma padronização entre 0 e 1. É calculado da seguinte forma:

$$C_{corr} = \sqrt{\frac{k}{k-1}} \times \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Onde:

- k = mínimo entre o número de linhas e colunas
- χ^2 = valor da estatística Qui-Quadrado
- n = tamanho da amostra

1.22 Coeficiente de Contingência V de Cramer

O coeficiente V de Cramer é utilizado para relacionar variáveis qualitativas nominais e/ou ordinais, assumindo valores entre 0 e 1. O valor 0 indica ausência de associação, e valores próximos a 1 indicam associação mais forte. O cálculo é feito por:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

Onde:

- $\chi^2 =$ valor da estatística Qui-Quadrado
- $n =$ tamanho da amostra
- $k =$ mínimo entre o número de linhas e colunas

2 Definição para Testes

2.1 Teste de Hipóteses

O teste de hipóteses tem como objetivo fornecer uma metodologia para verificar se os dados das amostras possuem indicativos que comprovem, ou não, uma hipótese previamente formulada. Ele é composto por duas hipóteses:

$$\begin{cases} H_0 : \text{hipótese a ser testada (chamada de hipótese nula)} \\ H_1 : \text{hipótese alternativa que será aceita caso a hipótese nula seja rejeitada} \end{cases}$$

Essa decisão é tomada por meio da construção de uma região crítica, ou seja, região de rejeição do teste.

2.2 Tipos de teste: bilateral e unilateral

Para a formulação de um teste, deve-se definir as hipóteses de interesse. Em geral, a hipótese nula é composta por uma igualdade (por exemplo, $H_0 : \theta = \theta_0$). Já a hipótese alternativa depende do grau de conhecimento que se tem do problema em estudo. Assim, tem-se três formas de elaborar H_1 que classificam os testes em duas categorias:

- **Teste Bilateral**:
  Esse é o teste mais geral, em que a hipótese alternativa consiste em verificar se existe diferença entre os parâmetros de interesse, independentemente de um ser maior ou menor que o outro. Dessa forma, tem-se:

$$H_1 : \theta \neq \theta_0$$

- **Teste Unilateral**:
  Dependendo das informações que o pesquisador possui a respeito do problema e os questionamentos que possui, a hipótese alternativa pode ser feita de forma a verificar se existe diferença entre os parâmetros em um dos sentidos. Ou seja:

$$H_1 : \theta < \theta_0$$

  ou

$$H_1 : \theta > \theta_0$$

Tipos de Erros Ao realizar um teste de hipóteses, existem dois erros associados: **Erro do Tipo I** e **Erro do Tipo II**.

- **Erro do Tipo I**:
  Esse erro é caracterizado por rejeitar a hipótese nula (H_0) quando essa é verdadeira. A probabilidade associada a esse erro é denotada por α, também conhecido como nível de significância do teste.

- **Erro do Tipo II**:
  Ao não rejeitar H_0 quando, na verdade, é falsa, está sendo cometido o **Erro do Tipo II**. A probabilidade de se cometer este erro é denotada por β.

2.3 Nível de significância (α)

O nível de significância do teste é o nome dado à probabilidade de se rejeitar a hipótese nula quando essa é verdadeira; essa rejeição é chamada de **erro do tipo I**. O valor de α é fixado antes da extração da amostra e, usualmente, assume 5%, 1% ou 0,1%.

Por exemplo, um nível de significância de $\alpha = 0,05$ (5%) significa que, se for tomada uma grande quantidade de amostras, em 5% delas a hipótese nula será rejeitada quando não havia evidências para essa rejeição, isto é, a probabilidade de se tomar a decisão correta é de 95%.

2.4 Estatística do Teste

A estatística do teste é o estimador que será utilizado para testar se a hipótese nula (H_0) é verdadeira ou não. Ela é escolhida por meio das teorias estatísticas.

2.5 P-valor

O **P-valor**, ou nível descritivo, é uma medida utilizada para sintetizar o resultado de um teste de hipóteses. Ele também pode ser chamado de *probabilidade de significância* do teste e indica a probabilidade de se obter um resultado da estatística de teste mais extremo do que o observado na presente amostra, considerando que a hipótese nula é verdadeira. Dessa forma, rejeita-se H_0 quando P-valor $< \alpha$, porque a chance de uma nova amostra possuir valores tão extremos quanto o encontrado é baixa, ou seja, há evidências para a rejeição da hipótese nula.

2.6 Intervalo de Confiança

Quando calcula-se um estimador pontual para o parâmetro, não é possível definir qual a possível magnitude do erro que se está cometendo. Com o objetivo de associar um erro à estimativa, são construídos os intervalos de confiança que se baseiam na distribuição amostral do estimador pontual.

Dessa forma, considere T um estimador pontual para θ e que a distribuição amostral de T é conhecida. O intervalo de confiança para o parâmetro θ será dado por t_1 e t_2, tal que:

$$P(t_1 < \theta < t_2) = \gamma$$

A probabilidade γ é estabelecida no início do estudo e representa o nível de confiança do intervalo. A interpretação desse resultado é que, se forem tiradas várias amostras de mesmo tamanho e forem calculados intervalos de confiança para cada uma, $100 \times \gamma\%$ dos intervalos irão conter o parâmetro θ. Assim, ao calcular um intervalo, pode-se dizer que há $100 \times \gamma\%$ de confiança de que o intervalo contém o parâmetro de interesse.

3 Teste de Normalidade

Os testes de normalidade são utilizados para verificar se uma variável aleatória segue um distribuição Normal de probabilidade ou não. Eles são muito importantes, pois impactam em qual teste deve ser utilizado em uma análise futura. Se o resultado do teste confirmar que a variável segue uma distribuição normal, procedimentos paramétricos podem e devem ser utilizados. Caso contrário, os métodos não paramétricos são mais recomendados.

3.1 Teste de Normalidade de Shapiro-Wilk

O **Teste de Shapiro-Wilk** é utilizado para verificar a aderência de uma variável quantitativa ao modelo da Distribuição Normal, sendo mais recomendado para amostras pequenas. A suposição de normalidade é importante para a determinação do teste a ser utilizado. As hipóteses a serem testadas são:

$$\begin{cases} H_0 : \text{A variável segue uma distribuição Normal} \\ H_1 : \text{A variável segue outro modelo} \end{cases}$$

A amostra deve ser ordenada de forma crescente para que seja possível obter as estatísticas de ordem. A estatística do teste é dada por:

$$W = \frac{1}{D}\left[\sum_{i=1}^{k} a_i \left(X_{(n-i+1)} - X_{(i)}\right)\right]$$

Com:

- K aproximadamente $\dfrac{n}{2}$
- $X_{(i)} =$ estatística de ordem *i*
- $D = \sum_{i=1}^{n}(X_i - \bar{X})^2$, em que $\bar{X}$ é a média amostral
- \$a_i = \$ constantes que apresentam valores tabelados

3.2 Teste de Normalidade de Kolmogorov-Smirnov

O **teste de Kolmogorov-Smirnov** é usado para determinar se duas distribuições de probabilidade diferem uma da outra. É baseado na diferença entre a função de distribuição acumulada teórica $F_0(x)$ e a função de distribuição acumulada da amostra $S_n(x)$. A função $S_n(x)$ é definida como a proporção das observações da amostra que são menores ou iguais a x.

O teste possui as seguintes hipóteses:

$$\begin{cases} H_0 : \text{A variável segue o modelo proposto} \\ H_1 : \text{A variável segue outro modelo} \end{cases}$$

Se a hipótese nula é verdadeira, espera-se que as diferenças entre $F_0(x)$ e $S_n(x)$ sejam pequenas e estejam dentro dos limites dos erros aleatórios. O teste de Kolmogorov-Smirnov focaliza a maior dessas diferenças. No caso do teste de normalidade de Kolmogorov-Smirnov, a função de distribuição acumulada teórica $F_0(x)$ é a função de distribuição acumulada da normal, com média e variância estimadas pela amostra. Este teste é mais recomendado para amostras grandes sem *outliers*.

3.3 Teste de Normalidade de Lilliefors

Assim como o teste de Kolmogorov-Smirnov, o **teste de Lilliefors** é utilizado para verificar se um conjunto de dados $X_1, X_2, ..., X_n$ de tamanho n segue determinada distribuição. A estatística de teste para este teste é dada por:

$$T_1 = sup_x|F^*(x) - S(X)|$$

A diferença entre a estatística T de Kolmogorov e T_1 de Lilliefors é que a função de distribuição acumulada $S_n(x)$ é obtida através dos dados padronizados da amostra, ou seja, $Z_1, Z_2, ..., Z_n$, com $Z_i = \dfrac{X_i - \bar{X}}{S}$. O teste acima é recomendado para amostras grandes com presença de valores discrepantes (*outliers*).

O teste possui as seguintes hipóteses:

$$\begin{cases} H_0 : \text{A variável segue o modelo proposto} \\ H_1 : \text{A variável segue outro modelo} \end{cases}$$

Se a hipótese nula é verdadeira, espera-se que as diferenças entre $F_0(x)$ e $S_n(x)$ sejam pequenas e estejam dentro dos limites dos erros aleatórios.

3.4 Teste de Normalidade de Anderson-Darling

O **teste de Normalidade de Anderson-Darling** é utilizado para verificar se uma amostra aleatória $X_1, X_2, ..., X_n$ de uma variável quantitativa segue uma distribuição Normal de probabilidade ou não. O teste possui as seguintes hipóteses:

$$\begin{cases} H_0 : \text{A variável segue uma distribuição Normal} \\ H_1 : \text{A variável segue outro modelo} \end{cases}$$

Se a hipótese nula for verdadeira, espera-se que o p-valor esteja acima do nível de significância α.

4 Teste de Homogeneidade de Variância

Existem diversos métodos estatísticos que possuem o pressuposto de que as variâncias de uma variável quantitativa entre 2 ou mais grupos são constantes. Para verificar essa suposição, são utilizados testes de homogeneidade de variância.

4.1 Teste de Homogeneidade de Variância de Bartlett

O **teste de Bartlett** é utilizado para testar a igualdade de três ou mais variâncias de determinadas populações. O teste é sensível à normalidade dos dados, não sendo indicado caso esse pressuposto de normalidade não seja satisfeito.

A estatística de teste é dada por:

$$B_0 = \frac{q}{c} \approx \chi^2_{k-1}$$

Com: - $c = 1 + \frac{1}{3(3k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{N-k}\right)$ - $q = (N - k)\ ln\left(S_p^2\right)\sum_{i=1}^{k}\left[(n_i - 1)\ ln\left(S_i^2\right)\right]$

- $S_p^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)S_i^2$
- $S_i^2 = \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})$
- E, para cada $i = 1, 2, \ldots, k$ amostra, tem-se:
  - $n_i =$ tamanho da amostra i
  - $S_i^2 =$ variância da amostra i
  - $N =$ soma do tamanho das amostras

As hipóteses do teste são:

$$\begin{cases} H_0 : \text{Todas as populações possuem mesma variância} \\ H_1 : \text{Ao menos uma população possui variância diferente das demais} \end{cases}$$

Sob H_0, rejeita-se a hipótese nula de igualdade de variâncias das k populações a um nível α de significância se a estatística do teste assumir valor superior ao quantil crítico respectivo da distribuição Qui-Quadrado com $k-1$ graus de liberdade.

4.2 Teste de Homogeneidade de Variância de Breusch-Pagan

O **teste de Breusch-Pagan** é utilizado para testar se a variância do erro de um modelo de regressão é constante. É indicado para grandes amostras e sensível quanto à normalidade dos resíduos.

Para o teste, ajusta-se um modelo de regressão e obtêm-se os valores preditos $\hat{y}$ e resíduos padronizados dados por:

$$u_i = \frac{e_i^2}{\sum_{i=1}^{n} \frac{e_i^2}{n}}$$

Para cada $i = 1, 2, \ldots, k$ observação, tem-se:

- $e_i =$ resíduo i
- $n = $ tamanho amostral

Em seguida, ajusta-se um modelo de regressão dos valores preditos $\hat{y}$ como variável resposta e os resíduos ajustados u_i como variável explicativa. A partir disso, obtém-se a estatística do teste:

$$\chi^2_{BP} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{2} \approx \chi^2_1$$

onde $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ é a Soma de Quadrados Explicada pelo modelo.

O teste possui as seguintes hipóteses:

$$\begin{cases} H_0: \text{As variâncias dos erros são iguais} \\ H_1: \text{As variâncias dos erros são diferentes e função multiplicativa de outras variáveis} \end{cases}$$

Sob H_0, rejeita-se a hipótese nula de igualdade de variâncias dos erros a um nível α de significância se a estatística do teste assumir valor superior ao quantil crítico respectivo da distribuição Qui-Quadrado com 1 grau de liberdade.

4.3 Teste de Homogeneidade de Variância de Levene

O **teste de Levene** consiste em fazer uma transformação nos dados originais. Para essa transformação, utiliza-se a técnica estatística de análise de variância (ANOVA).

Diferentemente de outros testes de homogeneidade de variância, o teste de Levene é não-paramétrico, ou seja, não possui pressuposto de normalidade.

A transformação dos dados é dada por:

$$z_{ij} = |x_{ij} - med(x_i)|$$

para $i = 1, 2, ..., k$ e $j = 1, 2, ..., n_i$ com k sendo o número de subgrupos, em que:

- $med(x_i) = $ mediana do subgrupo i
- $z_{ij} =$ representa a transformação nos dados
- $n_i = $ tamanho da amostra do subgrupo i

Com isso, tem-se a estatística do teste:

$$F^* = \frac{\displaystyle\sum_{i=1}^{k} \frac{n_i(\bar{z}_{i.} - \bar{z}_{..})^2}{(k-1)}}{\dfrac{\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_{i.}^2)}{\displaystyle\sum_{i=1}^{k}(n_i - 1)}}$$

Sendo que:

$$\bar{z}_{i.} = \sum_{i=1}^{k} \frac{z_{ij}}{n_i}$$

$$\bar{z}_{..} = \frac{\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{n_i} z_{ij}}{\displaystyle\sum_{i=1}^{k} n_i}$$

Sabe-se que $F^* \approx F(k, N-k-1)$.

Após a transformação dos dados originais, aplica-se o teste da ANOVA nos dados transformados. Assim, testa-se as seguintes hipóteses:

$$\begin{cases} H_0 : \text{Todas as populações possuem mesma variância} \\ H_1 : \text{Ao menos uma população possui variância diferente das demais} \end{cases}$$

Sob H_0, rejeita-se a hipótese nula de igualdade de variâncias a um nível α de significância se a estatística do teste F^* assumir valor superior ao quantil crítico respectivo da distribuição $F(k, N-k-1)$.

4.4 Teste F de Igualdade de Variância

O **teste F de igualdade de variância** é utilizado para verificar se duas populações possuem a mesma variância a um nível α de significância. O teste a seguir pressupõe normalidade dos dados.

Considere duas populações, X e Y, com médias μ_X, μ_Y, variâncias amostrais S_X^2, S_Y^2 e tamanhos de amostras n e m, respectivamente. Tem-se a estatística do teste:

$$\frac{S_X^2}{S_Y^2} \approx F(n-1, m-1)$$

Testa-se as seguintes hipóteses:

$$\begin{cases} H_0 : X \text{ e } Y \text{ possuem mesma variância populacional} \\ H_1 : X \text{ e } Y \text{ não possuem mesma variância populacional} \end{cases}$$

Caso os valores de S_X^2 e S_Y^2 sejam próximos, é esperado que o valor da estatística do teste esteja próximo de um e, assim, não deve-se rejeitar H_0. Caso esses valores estejam distantes, a estatística do teste se distanciará de um, levando à rejeição da hipótese nula de variâncias iguais.

4.5 Teste de Variância Constante de White

O **teste de White** permite verificar se a variância é constante em um modelo matemático. Sua metodologia baseia-se em ajustar um modelo de regressão dos resíduos do modelo original ao quadrado, tendo como variáveis explicativas um polinômio de 2° grau de cada variável explicativa X_i do modelo original e suas interações. O ponto fraco desse teste é a perda de muitos graus de liberdade, levando à diminuição do poder do teste. Caso a amostra seja suficientemente grande, essa perda de graus de liberdade não trará consequências relevantes.

As hipóteses do teste são:

$$\begin{cases} H_0 & : \text{Variância é constante} \\ H_1 & : \text{Variância não é constante} \end{cases}$$

4.6 Teste de Homogeneidade de Variância de Brown-Forsythe

O **teste de Brown-Forsythe** é um teste utilizado para verificar se a variância dos erros de um modelo de regressão linear simples é constante para diferentes valores da variável explicativa X. É uma modificação do teste de Levene, dessa forma, não depende da normalidade dos erros do modelo. Isto é, ele é um teste robusto para afastamentos sérios da normalidade dos erros. O tamanho da amostra deve ser suficientemente grande de modo que a dependência entre os resíduos possa ser ignorada.

As hipóteses do teste são:

$$\begin{cases} H_0 & : \text{A variância dos erros é constante} \\ H_1 & : \text{A variância dos erros não é constante} \end{cases}$$

O primeiro passo é dividir a amostra $X_1, X_2, \ldots, X_n$ em dois grupos de acordo com os níveis da variável explicativa X:

- O grupo 1 é formado pelas observações de X com nível baixo: e_{i1} é o i-ésimo resíduo do grupo 1, $i = 1, \ldots, n_1$.
- O grupo 2 é formado pelas observações de X com nível alto: e_{i2} é o i-ésimo resíduo do grupo 2, $i = 1, \ldots, n_2$.

Sejam $\tilde{e_1}$ e $\tilde{e_2}$ as medianas dos resíduos dos grupos 1 e 2, respectivamente, e os desvios em valor absoluto:

$$d_{i1} = |e_{i1} - \tilde{e_1}|$$
$$d_{i2} = |e_{i2} - \tilde{e_2}|$$

A estatística do teste é dada por:

$$t^*_{BF} = \frac{\bar{d_1} - \bar{d_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

em que $\bar{d_1}$ e $\bar{d_2}$ são as médias de d_{i1} e d_{i2}, respectivamente. A variância agrupada é definida por:

$$s^2 = \frac{\sum_{i=1}^{n_1}(d_{i1} - \bar{d_1})^2 + \sum_{i=1}^{n_2}(d_{i2} - \bar{d_2})^2}{n-2}$$

Sob a hipótese nula (H_0) verdadeira, ou seja, se a variância dos erros é constante e as amostras n_1 e n_2 não são muito pequenas, t^*_{BF} tem aproximadamente distribuição *t-Student* com $n-2$ graus de liberdade.

5 Teste de Comparação de Médias

5.1 Teste t Pareado

Considere duas amostras dependentes $x_1, \ldots, x_n$ e $y_1, \ldots, y_n$, em que as observações são pareadas, ou seja, $(x_1, y_1), \ldots, (x_n, y_n)$. Seja $D_i = x_i - y_i$, para $i = 1, \ldots, n$. Então, a amostra $D_1, \ldots, D_n$ é obtida a partir das diferenças entre os valores de cada par. A suposição é de que a população das diferenças segue distribuição Normal com média μ_D e variância σ_D^2.

As hipóteses do teste são:

$$\begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D \neq 0 \end{cases}$$

Em que μ_D é a média populacional das diferenças e é obtida por $\mu_D = \mu_X - \mu_Y$, sendo μ_X e μ_Y as médias correspondentes às populações de X e Y, respectivamente. Ou seja, está-se testando se a média da diferença é 0 ou não.

Os parâmetros μ_D e σ_D^2 são estimados pela média e variância amostrais:

$$\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n} \sim N\left(\mu_D, \frac{\sigma_D^2}{n}\right)$$

$$S_D^2 = \frac{\sum_{i=1}^{n} (D_i - \bar{D})^2}{n-1}$$

Assim, a estatística do teste é dada por:

$$T = \frac{\sqrt{n}\,(\bar{D} - \mu_D)}{S_D} \sim t_{n-1}$$

O critério de decisão utilizado para se rejeitar ou não a hipótese nula é a comparação do p-valor do teste com o nível α de significância adotado para a realização do teste. A um nível de significância α de erro, rejeita-se a hipótese H_0 se o p-valor for menor que α.

5.2 Teste de Wilcoxon Pareado

Considere observações pareadas $(X_1, Y_1), \ldots, (X_n, Y_n)$. Existe interesse em comparar se as medidas de posição de duas amostras são iguais no caso em que as amostras são dependentes. Não há necessidade de haver normalidade nos dados para este teste. Seja $D_i = x_i - y_i$, para $i = 1, \ldots, n$. Então, a amostra $D_1, \ldots, D_n$ é obtida a partir das diferenças entre os valores de cada par. As hipóteses deste teste são:

$$\begin{cases} H_0 : D \text{ segue uma distribuição simétrica em torno de zero} \\ H_1 : D \text{ não segue uma distribuição simétrica em torno de zero} \end{cases}$$

O teste é feito a partir da ordenação da variável D_i e postos (ou ranks) são atribuídos a cada observação. Algumas observações podem receber a mesma posição na ordenação. Esse fenômeno é denominado empate. Os postos são calculados da seguinte maneira:

$$R_i = \begin{cases} R(x_i, y_i), & \text{se} \quad D_i > 0 \\ -R(x_i, y_i), & \text{se} \quad D_i < 0 \end{cases}$$

com $R(x_i, y_i)$ sendo o posto associado a (x_i, y_i). A estatística de teste é dada pela soma dos postos com sinais positivos:

$$T^+ = \sum_{i=1}^{n} R_i$$

Em caso de empates ou se $n > 50$, utiliza-se a aproximação normal:

$$T^+ = \frac{\sum_{i=1}^{n} R_i}{\sum_{i=1}^{n} R_i^2}$$

Rejeita-se a hipótese nula se T^+ está fora da região de aceitação, a um determinado nível de significância α previamente estabelecido, para a distribuição de probabilidade w_i conhecida para T^+.

5.3 Teste t de Comparação de Médias para Variâncias Populacionais Conhecidas

Considere duas amostras independentes $(x_1, \ldots, x_n)$ e $(y_1, \ldots, y_m)$. O objetivo é comparar as médias dessas populações, verificando se podem ser consideradas iguais ou não. Sabendo que as variâncias populacionais são conhecidas e sob a suposição de normalidade nos dados em ambas as populações (simetria nos dados), testa-se as seguintes hipóteses:

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

Sendo μ_X e μ_Y as médias correspondentes às populações de X e Y, respectivamente. Os parâmetros μ_X e μ_Y são estimados pelas médias amostrais:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\bar{Y} = \frac{\sum_{i=1}^{m} Y_i}{m}$$

Assim, tem-se a estatística de teste:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0, 1)$$

- n, m = tamanho da amostra de X e Y, respectivamente
- σ_X^2, σ_Y^2 = variância de X e Y, respectivamente

O critério de decisão utilizado para se rejeitar ou não a hipótese nula é a comparação do p-valor do teste com o nível α de significância adotado para a realização do teste. A um nível de significância α de erro, rejeita-se a hipótese H_0 se o p-valor for menor que α.

5.4 Teste t de Comparação de Médias para Variâncias Populacionais Desconhecidas e Iguais

Considere duas amostras independentes $x_1, \ldots, x_n$ e $y_1, \ldots, y_m$. Existe interesse em comparar as médias dessas populações, verificando se podem ser consideradas iguais ou não. Sob a suposição de normalidade nos dados em ambas as populações (simetria nos dados) e igualdade entre suas variâncias, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, testa-se as seguintes hipóteses:

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

Sendo μ_X e μ_Y as médias correspondentes às populações de X e Y, respectivamente. Os parâmetros μ_X e μ_Y são estimados pelas médias amostrais:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\bar{Y} = \frac{\sum_{i=1}^{m} Y_i}{m}$$

E os parâmetros σ_X^2 e σ_Y^2 pelas variâncias amostrais:

$$S_X^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

$$S_Y^2 = \frac{\sum_{i=1}^{m} (Y_i - \bar{Y})^2}{m-1}$$

Assim, é construída a estatística de teste:

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

- n, m = tamanho da amostra de X e Y, respectivamente
- $n + m - 2$ = número de graus de liberdade da distribuição t-Student
- S_p^2 = variância combinada de S_X^2 e S_Y^2

Com S_p^2 dada por:

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

O critério de decisão utilizado para se rejeitar ou não a hipótese nula é a comparação do p-valor do teste com o nível α de significância adotado para a realização do teste. A um nível de significância α de erro, rejeita-se a hipótese H_0 se o p-valor for menor que α.

5.5 Teste t de Comparação de Médias para Variâncias Populacionais Desconhecidas e Diferentes

Considere duas amostras independentes $(x_1, \ldots, x_n)$ e $(y_1, \ldots, y_m)$. Existe interesse em comparar as médias dessas populações, verificando se podem ser consideradas iguais ou não. Sob a suposição de normalidade nos dados em ambas as populações (simetria nos dados) e diferença entre suas variâncias, $\sigma^2_X \neq \sigma^2_Y$, testa-se as seguintes hipóteses:

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

Sendo μ_X e μ_Y as médias correspondentes às populações de X e Y, respectivamente. Os parâmetros μ_X e μ_Y são estimados pelas médias amostrais:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\bar{Y} = \frac{\sum_{i=1}^{m} Y_i}{m}$$

e os parâmetros σ^2_X e σ^2_Y pelas variâncias amostrais:

$$S^2_X = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

$$S^2_Y = \frac{\sum_{i=1}^{m}(Y_i - \bar{Y})^2}{m-1}$$

Assim, é construída a estatística de teste:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S^2_X}{n} + \frac{S^2_Y}{m}}} \sim t_v$$

- n, m = tamanho da amostra de X e Y, respectivamente
- v = número de graus de liberdade da distribuição t-Student

Com v dado por:

$$v = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{\left(\frac{S_X^2}{n}\right)^2}{n-1} + \frac{\left(\frac{S_Y^2}{m}\right)^2}{m-1}}$$

A um nível de significância α de erro, sob a hipótese H_0 verdadeira, afirma-se que há igualdade de médias da população.

5.6 Teste de Wilcoxon-Mann-Whitney

O teste de Wilcoxon-Mann-Whitney ou apenas Mann-Whitney é utilizado para comparar dois grupos independentes sem supor nenhuma distribuição. Isso ocorre pois o teste baseia-se em postos atribuídos a cada observação da variável quantitativa após serem ordenadas. O teste considera as hipóteses:

$$\begin{cases} H_0 : \text{As populações têm a mesma distribuição} \\ H_1 : \text{As populações têm distribuições distintas} \end{cases}$$

Para cada caso a seguir, a estatística do teste se diferencia:

a) Com nenhum ou poucos empates:

$$T = \sum_{i=1}^{n} R(X_i)$$

Com:

- $R(X_i)$ o posto atribuído ao i-ésimo elemento da amostra
- n o tamanho da amostra

b) Com grandes amostras:

$$Z = \frac{T - E(T)}{\sqrt{V(T)}}$$

Com:

- $E(T) = \dfrac{n(N+1)}{2}$
- $V(T) = \dfrac{nm(N+1)}{13}$

c) **Com muitos empates:**

$$Z = \frac{T - E(T)}{\sqrt{V_c(T)}}$$

Com:

– $V_c(T) = \frac{nm}{N(N-1)} \sum_{i=1}^{N} R_i^2 - \frac{nm(N+1)^2}{4(N-1)}$

– $R_i =$ posto das N observações.

5.7 Análise de Variância (ANOVA)

A Análise de Variância, mais conhecida por ANOVA, consiste em um teste de hipótese, em que é testado se as médias dos tratamentos (ou grupos) são iguais. Os dados são descritos pelo seguinte modelo:

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, ..., a \quad e \quad j = 1, ..., N$$

Em que:

- i é o número de tratamentos
- j é o número de observações
- y_{ij} é a j-ésima observação do i-ésimo tratamento

No modelo, μ é a média geral dos dados e α_i é o efeito do tratamento i na variável resposta. Já e_{ij} é a variável aleatória correspondente ao erro. Supõe-se que tal variável tem distribuição de probabilidade Normal com média zero e variância σ^2. Mais precisamente, $e_{ij} \sim N(0, \sigma^2)$.

A variabilidade total pode ser decomposta na variabilidade devida aos diferentes tratamentos somada à variabilidade dentro de cada tratamento:

$$\begin{aligned}\text{Soma de Quadrados Total (SQTOT)} &= \text{Soma de Quadrados de Tratamento (SQTRAT)} \\ &+ \text{Soma de Quadrados de Resíduos (SQRES)}\end{aligned}$$

Sendo o estudo não balanceado, ou seja, quando os tratamentos possuem tamanhos de amostra distintos:

$$SQTOT = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N}$$

$$SQTRAT = \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N}$$

$$SQRES = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i}$$

Em que:

- n_i é o número de observações do i-ésimo tratamento
- N é o número total de observações
- $y_{..} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}$
- $y_{i.} = \sum_{j=1}^{n_i} y_{ij}$

As hipóteses do teste são:

$$\begin{cases} H_0 : \text{As médias dos } a \text{ tratamentos são iguais} \\ H_1 : \text{Existe pelo menos um par de médias diferente} \end{cases}$$

A estatística do teste é composta pelo Quadrado Médio de Tratamento (QMTRAT) e Quadrado Médio de Resíduos (QMRES), sendo a definição de Quadrado Médio a divisão da Soma de Quadrados pelos seus graus de liberdade. Por conta da suposição de Normalidade dos erros no modelo, a estatística do teste, F, tem distribuição F de Snedecor com $(a-1)$ e $(\sum_{i=1}^{a} n_i - a)$ graus de liberdade.

$$F_{obs} = \frac{QMTRAT}{QMRES} = \frac{\frac{SQTRAT}{(a-1)}}{\frac{SQRES}{(\sum_{i=1}^{a} n_i - a)}}$$

A hipótese nula é rejeitada caso o p-valor seja menor que o nível de significância pré-fixado. A Tabela **??** abaixo resume as informações anteriores:

Tabela 1: Tabela de Análise de Variância

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Estatística F	P-valor
Tratamento	$(a-1)$	SQTRAT	$\frac{SQTRAT}{(a-1)}$	$\frac{QMTRAT}{QMRES}$	$P(F > F_{obs})$
Resíduos	$(\sum_{i=1}^{a} n_i - a)$	SQRES	$\frac{SQRES}{(\sum_{i=1}^{a} n_i - a)}$		
Total	$(\sum_{i=1}^{a} n_i - 1)$	SQTOT			

5.8 Teste de Fisher LSD

Após a rejeição da hipótese nula da Análise de Variância (ANOVA), deve-se identificar quais médias diferem. Para isso, é utilizado o teste de Fisher LSD, tendo como objetivo comparar as médias duas a duas. Consiste em realizar múltiplos testes t de comparação de médias, cada um com nível de significância α. As hipóteses são:

$$\begin{cases} H_0 : \mu_i = \mu_j \\ H_1 : \mu_i \neq \mu_j \end{cases}$$

A estatística do teste é dada por:

$$T = t_{(\alpha/2;\, GL_{res})} \sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) QM_{res}}$$

Em que:

- $t_{(\alpha/2;\, GL_{res})}$ é o valor da distribuição *t*-Student com número de graus de liberdade do resíduo
- n é o número de observações do tratamento/grupo i
- m é o número de observações do tratamento/grupo j
- QM_{res} é o Quadrado Médio do Resíduo obtido da tabela de Análise de Variância

Rejeita-se a hipótese nula caso o módulo da diferença entre as médias ($|\bar{y}_i - \bar{y}_j|$) seja maior ou igual a T. Caso contrário, não se pode afirmar que as médias diferem.

5.9 Teste de Tukey HSD

Após a rejeição da hipótese nula da Análise de Variância (ANOVA), deve-se identificar quais médias diferem. Para isso, é utilizado o teste de Tukey HSD, tendo como objetivo comparar as médias duas a duas. Diferentemente de outros testes, ele controle o erro global do teste. Ou seja, a probabilidade de se cometer pelo menos um erro do tipo I é igual a α. As hipóteses são:

$$\begin{cases} H_0 : \mu_i = \mu_j \\ H_1 : \mu_i \neq \mu_j \end{cases}$$

A estatística do teste é dada por:

$$T = Tukey_{(\alpha;\, a;\, N-a)} \sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) \frac{QM_{res}}{2}}$$

Em que:

- α é o nível de significância global do teste
- a é o número de tratamentos/grupos
- N é o número total de observações
- $Tukey_{(\alpha;\, a;\, N-a)}$ é o quantil da distribuição de *Tukey* com esses parâmetros
- QM_{res} é o Quadrado Médio do Resíduo obtido da tabela de Análise de Variância
- n é o número de observações do tratamento/grupo i
- m é o número de observações do tratamento/grupo j

Rejeita-se a hipótese nula caso o módulo da diferença entre as médias ($|\bar{y}_i - \bar{y}_j|$) seja maior ou igual a T. Caso contrário, não se pode afirmar que as médias diferem.

5.10 Teste de Dunnett

Após a rejeição da hipótese nula da Análise de Variância (ANOVA), deve-se identificar quais médias diferem. Para isso, é utilizado o teste de Dunnett, tendo como objetivo comparar o controle com as demais médias, não sendo aplicável para casos em que não se deseja comparar um grupo controle com os tratamentos. As hipóteses são:

$$\begin{cases} H_0 : \mu_i = \mu_c \\ H_1 : \mu_i \neq \mu_c \end{cases}$$

A estatística do teste é dada por:

$$\Delta = d_{(\alpha;a-1;GLres)}\sqrt{\Big(\frac{1}{b_c}+\frac{1}{b_i}\Big)QMRes}$$

Em que:

- μ_i é a média do tratamento/grupo i
- μ_c é a média do tratamento/grupo controle
- $d_{(\alpha;a-1;GLres)}$ é uma constante da tabela de Dunnet, que depende do número de tratamentos sem o controle (a − 1) e do número de graus de liberdade do resíduo (GLres)
- b_i é o número de observações do tratamento i
- b_c é o número de observações do grupo controle
- $QMres$ é o Quadrado Médio do Resíduo obtido da tabela de Análise de Variância

Rejeita-se a hipótese nula caso o módulo da diferença entre as médias ($|\bar{y_i} - \bar{y_c}|$) seja maior ou igual a Δ. Caso contrário, não se pode afirmar que as médias diferem.

5.11 Teste de Duncan

Após a rejeição da hipótese nula da Análise de Variância (ANOVA), deve-se identificar quais médias diferem. Para isso, é utilizado o teste de Duncan, tendo como objetivo comparar a amplitude de um conjunto de médias amostrais com uma amplitude mínima significante calculada. As hipóteses são:

$$\begin{cases} H_0 : \mu_i = \mu_j \\ H_1 : \mu_i \neq \mu_j \end{cases}$$

A estatística do teste é dada por:

$$R_m = r_m \frac{QMRes}{n}$$

Em que:

- μ_i é a média do tratamento/grupo i
- μ_c é a média do tratamento/grupo j

- r_m é o valor da amplitude mínima studentizada significante ao nível de significância α, encontrado em tabelas, dependente do número m de médias abrangidas na amplitude em comparação e de $GLres$
- n é o número de observações em cada grupo
- QM_{res} é o Quadrado Médio do Resíduo obtido da tabela de Análise de Variância

Rejeita-se a hipótese nula caso o módulo da diferença entre as médias ($|\bar{y_i} - \bar{y_j}|$) seja maior ou igual a Δ. Caso contrário, não se pode afirmar que as médias diferem.

5.12 Teste de Kruskal-Wallis

O teste de Kruskal-Wallis é utilizado para comparar dois ou mais grupos independentes sem supor nenhuma distribuição. É um método baseado na comparação de postos, os quais são atribuídos a cada observação de uma variável quantitativa após serem ordenadas.

As hipóteses do teste de Kruskal-Wallis são formuladas da seguinte maneira:

$$\begin{cases} H_0 : \text{Não existe diferença entre os grupos} \\ H_1 : \text{Pelo menos um grupo difere dos demais} \end{cases}$$

A estatística do teste de Krukal-Waliis é definida da seguinte maneira:

$$H_{Kruskal-Wallis} = \frac{\left[\frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right] - 3(n+1)}{1 - \left[\frac{\sum_j (t_j^3 - t_j)}{n^3 - n}\right]} \approx \chi^2_{(k-1)}$$

Com: - $k =$ número de grupos

- $R_i =$ soma dos postos do grupo i
- $n_i =$ número de elementos do grupo i
- $n =$ tamanho total da amostra
- $t_j =$ número de elementos no j-ésimo empate (se houver)

Se o p-valor for menor que o nível de significância α, rejeita-se a hipótese nula.

5.13 Teste de Comparação Múltipla de Conover

Se rejeitarmos a hipótese nula no teste de Kruskal-Wallis, é necessário realizar comparações múltiplas para detectar quais pares de populações podem ser considerados diferentes.

As populações i e j são consideradas diferentes se a seguinte inequação é satisfeita:

$$\left|\frac{R_{i.}}{n_i} - \frac{R_{j.}}{n_j}\right| > t_{1-\alpha/2}\left(S^2\frac{N-1-T}{N-k}\right)^{1/2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)^{1/2}$$

em que $R_{i.}$ e $R_{j.}$ são as somas dos postos das amostras i e j, respectivamente, $t_{1-\alpha/2}$ é o quantil (1 − α/2) da distribuição t−Student com (N − k) graus de liberdade, que é equivalente ao intervalo:

$$\left[\left(\frac{R_{i.}}{n_i} - \frac{R_{j.}}{n_j}\right) \pm t_{1-\alpha/2}\left(S^2\frac{N-1-T}{N-k}\right)^{1/2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)^{1/2}\right]$$

não conter o zero.

5.14 Teste de Dunn

Após a rejeição da hipótese nula do teste de Kruskall-Wallis, deve-se identificar quais médias diferem. Para isso, é utilizado o teste de Dunn, tendo como objetivo comparar as médias dos grupos 2 a 2, controlando o erro global dos testes. Ah hipóteses são:

$$\begin{cases} H_0 : \mu_i = \mu_j \\ H_1 : \mu_i \neq \mu_j \end{cases}$$

Seja W_i a soma dos ranks do i-ésimo grupo e n_i o número de observações do i-ésimo grupo, e seja $\hat{W}_i$ a média dos ranks do i-ésimo grupo, então a estatística do teste para os grupos A e B é:

$$z_{A,B} = \frac{\hat{W}_A - \hat{W}_B}{\sigma_{A,B}}$$

Em que:

- $$\sigma_{A,B} = \sqrt{\left[\frac{N(N+1)}{12} - \frac{\sum_{s=1}^{r}\tau_s^3 - \tau_s}{12(N-1)}\right]\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}$$

- N é o tamanho total da amostra de todos os grupos;
- r é o número de empates nos ranks entre todos os grupos;
- τ_s é o número de observações entre todos os grupos com o s-ésimo rank empatado.

Se o p-valor for menor que o nível de significância fixado α, rejeita-se a hipótese nula.

5.15 Teste de Friedman

O Teste de Friedman é utilizado para comparar dois ou mais grupos dependentes, sendo que a suposição de Normalidade não precisa ser atendida. As hipóteses do teste são:

$$\begin{cases} H_0 : \text{A característica em estudo é igual em todos os grupos} \\ H_1 : \text{A característica em estudo difere em pelo menos um grupo} \end{cases}$$

Em termos estatísticos, tem-se que a média da característica em estudo é igual em todos os grupos. A estatística do teste é dada por:

$$Q^2 = \left[\frac{12}{nk(n+1)} \sum_{i=1}^{k} R_i^2 \right] - 3n(k+1)$$

Com:

- $k =$ número de grupos
- $R_i =$ soma dos postos do grupo i
- $n =$ número de elementos nos grupos (igual em todos os grupos)

Os postos são obtidos após a ordenação dos dados dentro de cada grupo e a estatística do teste segue a distribuição Qui-Quadrado com $(k-1)$ graus de liberdade. Se o p-valor for menor que o nível de significância α, rejeita-se a hipótese nula.

6 Teste de Associação

6.1 Testes Qui-Quadrado

Os testes a seguir utilizam como base a estatística χ^2, apresentando mudanças nos graus de liberdade da sua distribuição de acordo com o teste que será utilizado. No geral,

$$\chi_v^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

em que v expressa os graus de liberdade, o_i é a frequência observada e e_i é chamado de valor esperado e representa a frequência que seria observada se H_0 fosse verdadeira.

6.1.1 Teste de Aderência

O Teste de Aderência Qui-Quadrado tem como objetivo verificar se uma variável, qualitativa ou quantitativa, segue determinada distribuição com probabilidades especificadas. Para os dois casos, será formada uma tabela de contingência com uma linha e s colunas; a linha terá a frequência observada em cada categoria presente nas colunas.

As hipóteses para este teste podem ser escritas como:

$$\begin{cases} H_0 : p_i = p_{i0} \\ H_1 : p_i \neq p_{i0} \end{cases}$$

em que p_i é a frequência relativa observada na categoria i e p_{i0} é a probabilidade que deseja-se testar para cada categoria.

Para esse teste, utiliza-se a seguinte estatística:

$$\chi_v^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

na qual $v = s-1$ representa o número de graus de liberdade, s é o total de colunas da tabela de contingência, o_i é o valor observado na amostra e e_i é o valor que seria observado caso a hipótese nula (H_0) fosse verdadeira.

Então, sob a hipótese de H_0 verdadeira, a estatística do teste seguirá a distribuição χ_v^2.

6.1.2 Teste de Homogeneidade

O Teste de Homogeneidade tem como objetivo verificar se uma variável (seja ela qualitativa, seja quantitativa) se comporta de forma similar, ou homogênea, em várias subpopulações. Nesse teste, o tamanho da amostra em cada subpopulação é fixado e seleciona-se uma amostra dentro de cada uma. Para realizar o teste, serão comparadas se as proporções de cada evento são semelhantes (frequências observadas). Então, as hipóteses podem ser escritas como:

$$\begin{cases} H_0 : \text{O comportamento da variável é homogêneo nas subpopulações} \\ H_1 : \text{O comportamento da variável não é homogêneo nas subpopulações} \end{cases}$$

ou

$$\begin{cases} H_0 : p_1 = p_2 = ... = p_n \\ H_1 : p_i \neq p_j, \text{ para algum } i \neq j \end{cases}$$

em que p_i é a proporção em cada evento.

A estatística utilizada nesse teste é a estatística Qui-Quadrado:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

em que:

- $e_{ij} = $ valor esperado na i-ésima linha e na j-ésima coluna e é dado por:

$$\frac{(total\ da\ linha\ i) \times (total\ da\ coluna\ j)}{total\ geral}$$

- $o_{ij} =$ valor observado na i-ésima linha e na j-ésima coluna

Então, considerando a hipótese nula (H_0) verdadeira, a estatística χ^2 seguirá uma distribuição Qui-Quadrado com $v = (r-1)(s-1)$ graus de liberdade, em que r representa o número total de linhas da tabela e s o número total de colunas.

6.1.3 Teste de Independência

Esse teste tem como objetivo verificar se existe associação entre duas variáveis, sendo mais recomendado para variáveis qualitativas (principalmente nominais). O princípio básico deste método é comparar proporções, ou seja, as possíveis divergências entre as frequências observadas e esperadas para um certo evento. Para esse teste, as hipóteses podem ser escritas como:

$$\begin{cases} H_0 : \text{A variável X é independente da variável Y} \\ H_1 : \text{A variável X depende da variável Y} \end{cases}$$

Este teste é baseado no cálculo dos valores esperados. Os valores esperados são os valores que seriam observados caso a hipótese nula fosse verdadeira:

$$e_{ij} = \frac{(total\ da\ linha\ i) \times (total\ da\ coluna\ j)}{total\ geral}$$

Para isso, utiliza-se a seguinte estatística:

$$\chi^2_v = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

em que:

- e_{ij} = valor esperado na i-ésima linha e na j-ésima coluna
- o_{ij} = valor observado na i-ésima linha e na j-ésima coluna
- $v = (r-1)(s-1)$ representa o número de graus de liberdade
- r = número total de linhas
- s = número total de colunas

Então, sob a hipótese de H_0 ser verdadeira, a estatística do teste seguirá a distribuição χ^2_v.

Para que a aproximação Qui-Quadrado seja satisfatória, é preciso que a amostra seja relativamente grande, com todos os valores esperados maiores ou iguais a 5 ou no máximo 20% deles seja menor que 5 com todos maiores que 1. Caso isso não ocorra, utiliza-se a correção de **Yates**.

6.1.4 Teste de Correlação de McNemar

O teste de correlação de McNemar tem como objetivo verificar se existem mudanças nas respostas em diferentes períodos de tempo para determinada variável em estudo. Esse teste pode ser feito com variáveis qualitativas nominais ou ordinais, é um teste não-paramétrico (não depende da suposição de normalidade) e, por avaliar diferenças entre períodos, é feito com amostras pareadas, ou seja, um grupo de indivíduos é avaliado em um determinado período e, após algum tempo, esses mesmos indivíduos são avaliados novamente.

Para a realização do teste, uma tabela 2x2 é feita que auxilia a testar a significância de qualquer mudança.

	Depois	
Antes	**+**	**-**
+	A	B
-	C	D

Por meio dessa tabela, verifica-se que o número total de elementos que tiveram alguma mudança é a soma das células B e C. Então, essas são as células de interesse do teste.

As hipóteses podem ser escritas como:

$$\begin{cases} H_0 : \text{Não houve mudanças no período de tempo em estudo} \\ H_1 : \text{Houve mudanças no período em estudo} \end{cases}$$

Assim, sob H_0, espera-se que as frequências de B e C sejam iguais, ou seja, o número de elementos em cada uma das células que tiveram mudanças deve ser aproximadamente $\dfrac{(B+C)}{2}$. Dessa forma, a estatística do teste é baseada na estatística Qui-Quadrado e, aplicada a esse problema, é expressa por:

$$\chi^2 = \frac{(B-C)^2}{B+C}$$

Como a distribuição Qui-Quadrado é contínua, aplica-se uma correção na fórmula para que possa ser utilizada neste teste, que é específico para variáveis nominais e ordinais. Sendo assim, a estatística do teste com a correção é:

$$\chi^2 = \frac{(|B-C|-1)^2}{B+C}$$

Portanto, a estatística do teste seguirá uma distribuição Qui-Quadrado com 1 grau de liberdade (χ^2_1).

6.2 Teste Exato de Fisher

Esse teste é um caso particular do testes qui-quadrado de homogeneidade e do teste de independência. Ele tem como objetivo comparar as proporções de cada evento, bem como a independência entre as variáveis em estudo, porém, os totais de linha e os totais de coluna na tabela de contingência são fixos e esse teste só pode ser realizado em tabelas de tamanho 2x2. Além disso, ele é utilizado para analisar dados qualitativos (nominais ou ordinais) quando o tamanho de amostra é pequeno.

As hipóteses para esse teste são:

$$\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{cases}$$

Observação: as hipóteses também podem testar a diferença para algum dos sentidos: $p_1 < p_2$ ou $p_1 > p_2$, em que p_1 é a probabilidade da coluna 1 dado a linha 1, e p_2 é a probabilidade da coluna 1 dado a linha 2.

A estatística utilizada nesse teste é dada pela número de observações presentes na casela (1, 1) da tabela de contingência e seu valor máximo é o total da coluna 1.

A estatística T do teste, considerando H_0 verdadeira, tem distribuição:

a) Exata dada pela distribuição hipergeométrica:

$$P(T = O_{11}) = \frac{\binom{n_1}{O_{11}}\binom{N-n_1}{c_1-O_{11}}}{\binom{N}{c_1}}$$

em que:

$O_{11} =$ valor observado na casela 1x1 (linha 1 e coluna 1)

$n_1 =$ total da linha 1

$N =$ total geral

$c_1 =$ total na coluna 1

b) Aproximada, para grandes amostras, para a distribuição Qui-Quadrado com 1 grau de liberdade se o teste for bilateral, ou distribuição Normal se o teste for unilateral.

6.3 Teste de Correlação de Pearson

O coeficiente de correlação linear de Pearson indica a força e a direção do relacionamento linear entre duas variáveis quantitativas. É um índice adimensional com valores situados entre -1 e 1, no qual o valor -1 representa total correlação linear negativa entre as variáveis (quando o valor de uma variável cresce, o valor da outra diminui) e o valor 1 representa total correlação linear positiva entre elas (ambas crescem simultaneamente). Esse coeficiente é obtido por meio da fórmula:

$$r_{Pearson} = \frac{\sum_{i=1}^{n} [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

em que

- x_i = i-ésimo valor da variável X
- y_i = i-ésimo valor da variável Y
- $\bar{x}$ = média dos valores da variável X
- $\bar{y}$ = média dos valores da variável Y
- $r_{Pearson}$ = coeficiente de correlação linear de Pearson amostral

Para o teste de correlação de Pearson, tem-se as seguintes hipóteses:

$$\begin{cases} H_0 : \text{Não há correlação linear entre as variáveis } X \text{ e } Y \ (\rho_{Pearson} = 0) \\ H_1 : \text{Há correlação linear entre as variáveis } X \text{ e } Y \ (\rho_{Pearson} \neq 0) \end{cases}$$

em que $\rho_{Pearson}$ é o parâmetro a ser testado: coeficiente de correlação linear populacional.

Se X e Y tem distribuição normal, tem-se que a estatística do teste é dada por:

$$t_{Pearson} = \frac{r_{Pearson}\sqrt{n-2}}{\sqrt{1 - r_{Pearson}^2}} \sim t_{n-2}$$

Assim, sob H_0, $t_{Pearson}$ segue uma distribuição *t*-Student com $(n-2)$ graus de liberdade.

6.4 Teste de Correlação de Postos de Spearman

O coeficiente de correlação de Spearman é uma medida não paramétrica que verifica, por meio de postos de variáveis quantitativas ou qualitativas ordinais, o grau de correlação linear entre duas variáveis. Esse coeficiente varia entre os valores -1 e 1. O valor zero significa que não há relação linear entre as variáveis. Quando o valor do coeficiente r é negativo, diz-se ter uma relação de grandeza inversamente proporcional entre as variáveis. Analogamente, quando r é positivo, afirma-se que as duas variáveis são diretamente proporcionais.

O coeficiente é calculado da seguinte maneira:

$$r_{Spearman} = \frac{\sum_{i=1}^{n}\left[\left(R(x_i) - \frac{n+1}{2}\right)\left(R(y_i) - \frac{n+1}{2}\right)\right]}{\sqrt{\sum_{i=1}^{n}(R(x_i)^2) - n\left(\frac{n+1}{2}\right)^2} \times \sqrt{\sum_{i=1}^{n}(R(y_i)^2) - n\left(\frac{n+1}{2}\right)^2}}$$

no qual

- $x_i =$ i-ésimo valor da variável X
- $y_i =$ i-ésimo valor da variável Y
- $R(x_i) =$ posto atribuído a x_i, quando comparado a outros valores de x
- $R(y_i) =$ posto atribuído a y_i, quando comparado a outros valores de y
- $n =$ número total de observações na amostra
- $r_{Spearman} =$ coeficiente de correlação de postos de Spearman amostral

Observação: ao ordenar de forma crescente as observações $x_1, x_2, ..., x_n, y_1, y_2, ..., y_n$ afirma-se que $R(X_1) = 1$ se x_1 é o menor valor da amostra, $R(X_3) = 2$ se x_3 é o segundo menor valor da amostra, $R(X_4) = 3$ se x_4 é o terceiro menor valor, e assim sucessivamente. Quando há empates nas observações, o posto atribuído a elas é a média dos postos que teriam se não houvesse empate. Por exemplo, se X assume os valores 9, 5, 6 e 9, tem-se duas observações com mesmo valor e, assim, seus postos serão obtidos por meio da média entre 3 e 4, que seriam seus postos se não houvesse empate.

Por ser um método não paramétrico, não há suposições para o teste.

Para a realização do teste, são feitas as seguintes hipóteses:

$$\begin{cases} H_0: \text{Não há correlação de postos entre as variáveis } X \text{ e } Y \ (\rho_{Spearman} = 0) \\ H_1: \text{Há correlação de postos entre as variáveis } X \text{ e } Y \ (\rho_{Spearman} \neq 0) \end{cases}$$

em que $\rho_{Spearman}$ é o coeficiente de correlação de postos populacional (parâmetro a ser testado com base em $r_{Spearman}$).

Além disso, a hipótese alternativa também pode ser escrita para testar se a correlação dos postos é positiva ($\rho_{Spearman} > 0$) ou negativa ($\rho_{Spearman} < 0$).

A estatística T do teste ($T = r_{Spearman}$), se H_0 é verdadeira, tem distribuição:

- **a) Pequenas amostras e sem/poucos empates:** exata com os valores apresentados em uma tabela.
- **b) Grandes amostras ou muitos empates:** aproximada pela Normal Padrão (Normal com média 0 e variância 1), tal que:

$$w_p = \frac{z_p}{\sqrt{(n-1)}}$$

no qual

- $w_p =$ quantil de ordem p da distribuição que a estatística T segue
- $z_p =$ quantil de ordem p da distribuição Normal padrão
- $n =$ número total de observações na amostra

6.5 Teste de Correlação de Postos de Kendall

Esse teste tem como objetivo verificar, por meio da comparação de postos, se existe independência entre as variáveis, avaliando a concordância e discordância dos pares. As variáveis em estudo podem ser qualitativas ordinais ou quantitativas. Assim, o total de pares é $\binom{n}{2}$, em que n é o tamanho da amostra e $\binom{n}{2}$ representa a combinação das n observações da amostra tomadas de duas a duas. Considere, então, que N_c representa o número de pares concordantes e N_d é o número de pares discordantes. Os pares são concordantes se ambos os valores de X e Y de uma observação (um par) são maiores que os valores de X e Y de outra observação; os pares são discordantes se os valores das variáveis de uma observação diferem os valores de outra observação em direções opostas (por exemplo, $X_1 > X_2$ e $Y_1 < Y_2$).

As hipóteses para esse teste podem ser escritas como:

$$\begin{cases} H_0 : X \text{ e } Y \text{ são independentes (não há correlação entre elas)} \\ H_1 : \text{Há correlação de Kendall entre } X \text{ e } Y \end{cases}$$

A estatística do teste pode ter duas formas que variam conforme a presença de empates entre os pares:

a) Sem empates: $\tau = N_c - N_d$

Considerando H_0 verdadeira, essa estatística tem:

i) Distribuição exata apresentada em um tabela se o tamanho da amostra n for menor que 60.

ii) Aproximada pela Normal Padrão em caso de n grande:

$$w_p = z_p \frac{\sqrt{n(n-1)(2n+5)}}{18}$$

b) Com empates: $\tau = \dfrac{N_c - N_d}{N_c + N_d}$

Considerando H_0 verdadeira, essa estatística tem:

i) Distribuição exata apresentada em um tabela se o tamanho da amostra n for menor que 60.

ii) Aproximada pela Normal Padrão em caso de n grande:

$$w_p = z_p \frac{\sqrt{n(n-1)(2n+5)}}{18}$$

Para realizar a comparação dos pares e concluir se serão concordantes ou discordantes, pode-se utilizar as seguintes regras de decisão:

- Se $\dfrac{Y_j - Y_i}{X_j - X_i} > 0$, os pares são **concordantes** (adicione 1 a N_c)
- Se $\dfrac{Y_j - Y_i}{X_j - X_i} < 0$, os pares são **discordantes** (adicione 1 a N_d)
- Se $\dfrac{Y_j - Y_i}{X_j - X_i} = 0$, ocorreu **empate** (adicione 0,5 a N_c e a N_d)
- Se $X_j = X_i$, não há comparação

6.6 Correlação Parcial de Kendall

Em alguns casos, quando se observa a correlação entre duas variáveis, há possibilidade de que essa relação ocorra devido à associação de cada uma dessas variáveis com uma terceira. Na correlação parcial, os efeitos desta terceira variável (Z) sobre as outras duas variáveis (X e Y) são controlados, mantendo-a constante.

Para o cálculo desse coeficiente, comparam-se os postos atribuídos às observações em cada variável. Os postos representam a posição que um determinado valor da variável ocuparia se os dados estivessem ordenados de forma crescente. Assim, esses postos são atribuídos aos valores das três variáveis de forma independente do resultado das outras. Em seguida, são formados pares para os postos e esses pares são colocados de forma que as observações de Z estejam em ordem crescente. Tomando