House of excellence

Consultores Responsáveis:
Estatiano 1
Estatiano 2
...
Estatiano n

Requerente:
ESTAT

Brasília, 20 de outubro de 2024.

ESTAT
consultoria estatística

Sumário

1 Introdução

O projeto tem como objetivo auxiliar João Neves, proprietário da academia de alta performance House of Excellence, na otimização do desempenho de seus atletas de elite, com base em análises estatísticas de suas participações nas edições dos Jogos Olímpicos de 2000 a 2016. O foco das análises é identificar padrões de desempenho, características físicas e fatores relacionados às conquistas de medalhas, oferecendo insights valiosos para melhorar a preparação e a performance futura dos atletas.

Segundo parágrafo: detalhar quais são as análises que serão feitas. TODOS OS TIPOS DE ANÁLISES, se é análise descritiva , teste de hipóteses, regressão. Descrever um pouco sobre o banco de dados(quantidade de variáveis, tipos de variáveis, etc)..

As análises foram realizadas utilizando o software R, versão 4.4.1, com pacotes especializados para manipulação de dados, visualização gráfica e modelagem estatística.

2 Referencial Teórico

2.1 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:

$$100 \times f_j$$

2.2 Média

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

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

Com:

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

2.3 Mediana

Sejam as n observações de um conjunto de dados $X = X_{(1)}, X_{(2)}, \dots, X_{(n)}$ 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:

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

2.4 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.

2.5 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.

2.5.1 Variância Populacional

Para uma população, a variância é dada por:

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

Com:

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

2.5.2 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

2.6 Desvio Padrão

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

2.6.1 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

2.6.2 Desvio Padrão Amostral

Para uma amostra, o desvio padrão é dado por:

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

Com:

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

2.7 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

2.8 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

2.9 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).

2.10 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.

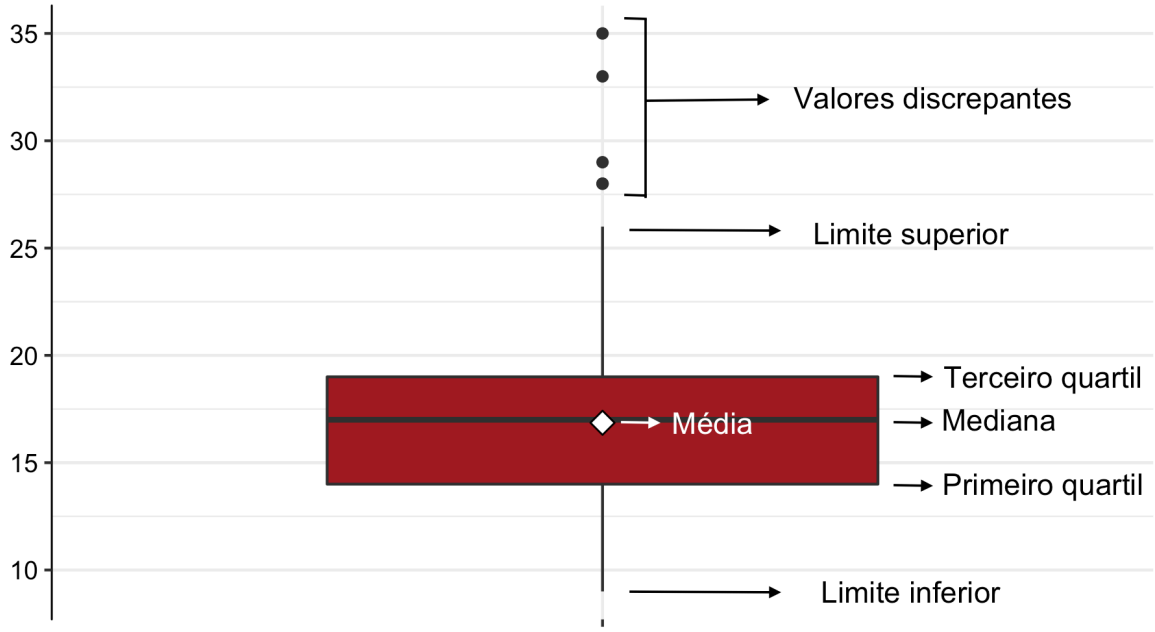

Figura 1: Exemplo de boxplot

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.

2.11 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.

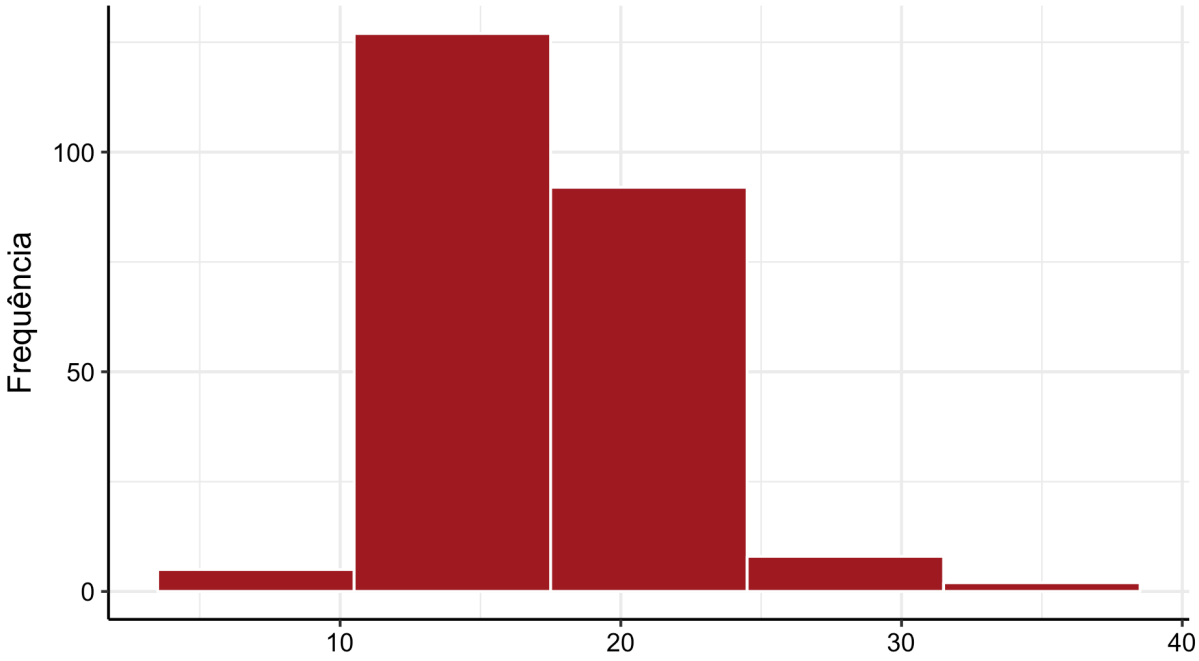

Figura 2: Exemplo de histograma

2.12 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.

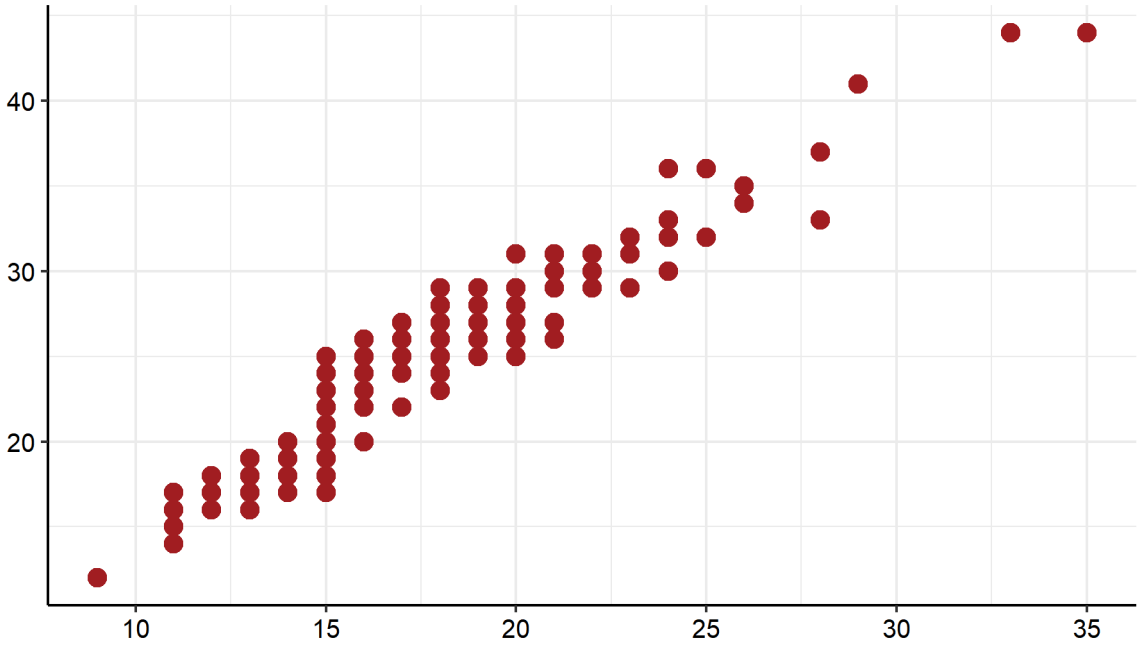

Figura 3: Exemplo de Gráfico de Dispersão

2.13 Tipos de Variáveis

2.13.1 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)

2.13.2 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)

2.14 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 $\frac{n}{2}$
- $X_{(i)} =$ 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

2.15 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:

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

Sendo que:

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

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

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 $F(k, N-k-1)$.

2.16 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:

Soma de Quadrados Total (SQTOT) = Soma de Quadrados de Tratamento (SQTRAT)
+ Soma de Quadrados de Resíduos (SQRES)

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			

2.17 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.

2.18 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.

3 Análises

3.1 Top 5 países com maior número de mulheres medalhistas

A análise, buscamos identificar os países que tiveram o maior número de mulheres medalhistas em Jogos Olímpicos de 2000 a 2016. Utilizou-se dados das edições de Sydney 2000, Atenas 2004, Pequim 2008, Londres 2012 e Rio 2016, considerando as variáveis de sexo (feminino), país de origem das atletas (Team), e a presença de uma medalha (Medal). O objetivo principal foi compreender quais nações se destacaram em termos de conquistas femininas, utilizando como métricas o número total de medalhistas por país e a frequência relativa de cada país em relação ao total de mulheres medalhistas.

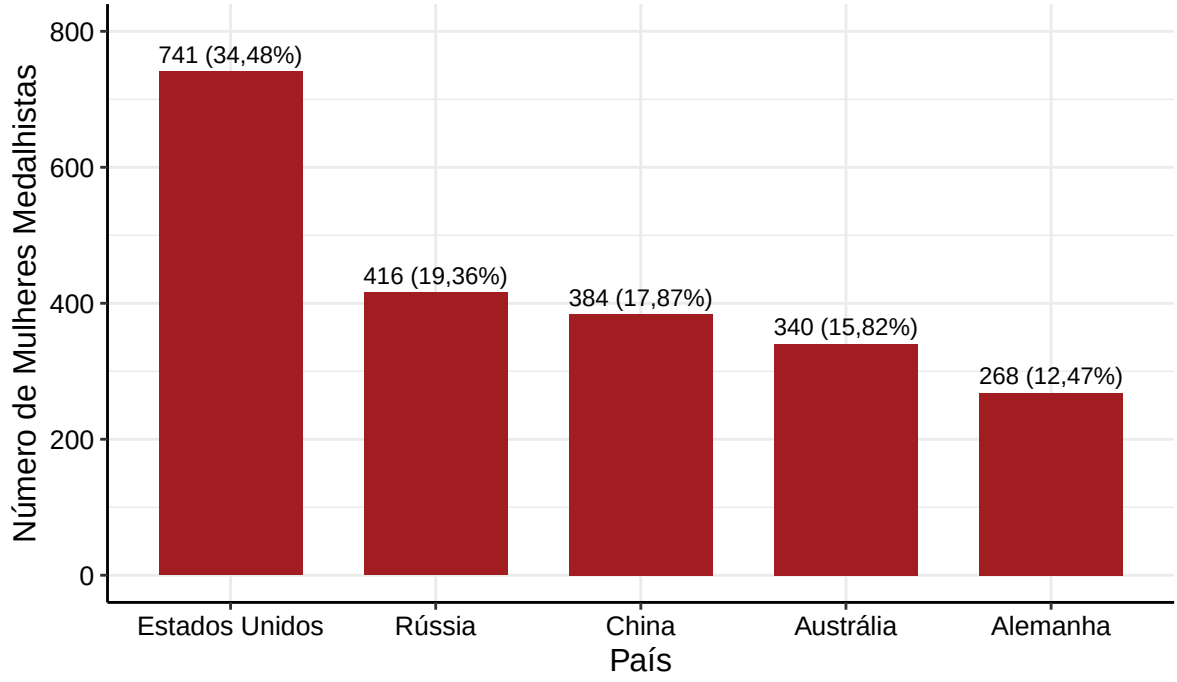

Figura 4: Gráfico de barras dos 5 países com maior número de mulheres medalhistas

Entre os anos de 2000 a 2016, os Estados Unidos se destacaram como o país com o maior número de mulheres medalhistas nas Olimpíadas, conquistando 741 medalhas, o que representa 34,48% do total. Em segundo lugar, a Rússia registrou 416 medalhas femininas, correspondendo a 19,36% do total. A China ocupa a terceira posição com 384 medalhas, o que equivale a 17,87% do total. A Austrália, com 340 medalhas (15,82%), é o quarto colocado. Por fim, a Alemanha completa o top 5 com 268 medalhas, representando 12,47% do total.

3.2 Valor IMC por esporte, estes sendo, ginástica, futebol, judô, atletismo e badminton

Nesta análise, busca-se comparar o Índice de Massa Corporal (IMC) de atletas olímpicos que competiram em diferentes modalidades esportivas, especificamente Ginástica, Judô, Atletismo e Badminton, em cinco edições dos Jogos Olímpicos: Sydney 2000, Atenas 2004, Pequim 2008, Londres 2012 e Rio 2016. O objetivo é identificar se há diferenças significativas nos valores de IMC entre esses esportes, bem como entender quais esportes tendem a ter IMCs mais altos ou mais baixos.

Tabela 2: Tabela das estatísticas descritivas por esporte

Esporte	Tamanho_amostral	Média	Mediana	Desvio_padrão	Mín	Máx
Atletismo	11673	22,03147	21,20309	3,772593	14,98079	44,37866
Badminton	968	22,39120	22,29535	1,783496	16,90101	31,14184
Ginástica	5000	21,14521	21,35929	2,298969	14,08379	30,82480
Judô	1941	25,55040	24,41404	5,118483	17,57705	63,90153

No Atletismo, a média do IMC é de 22,0, com uma mediana de 21,2 e um desvio padrão de 3,77. Isso sugere que os valores de IMC dos atletas são relativamente próximos, mas há alguns casos extremos, como evidenciado pelo valor máximo de 44,4.

Os atletas de Badminton apresentam uma média de IMC ligeiramente maior, de 22,4, e uma mediana de 22,3, com um desvio padrão baixo de 1,78. Isso indica uma menor variabilidade nos valores de IMC, com a maioria dos atletas concentrados em torno da média.

Na Ginástica, o IMC médio é de 21,1, enquanto a mediana é de 21,4, e o desvio padrão é de 2,30, indicando também uma baixa variabilidade e uma distribuição mais homogênea dos valores de IMC entre os atletas.

Já os atletas de Judô apresentam a maior média de IMC, de 25,6, com uma mediana de 24,4 e um desvio padrão mais elevado, de 5,12. Esse maior desvio padrão sugere uma variabilidade significativa no IMC dos judocas, com um valor máximo de 63,9, indicando que alguns atletas possuem um IMC substancialmente maior, o que pode estar relacionado à necessidade de uma maior massa corporal nesse esporte de combate.

De acordo com os dados tem se os judocas tendem a ter um IMC mais elevados, enquanto os ginastas apresentam os menores. As diferenças na variabilidade entre os grupos refletem os distintos requisitos físicos de cada esporte, como força e agilidade.

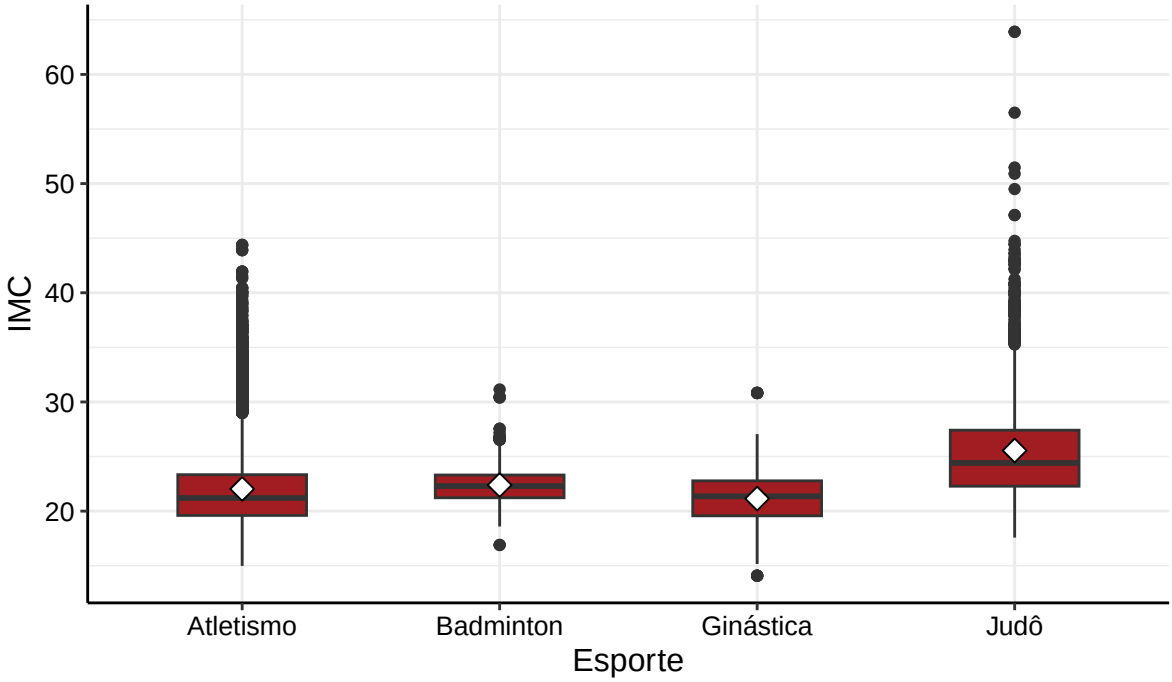

Figura 5: Boxplot da Comparação do IMC entre Esportes

Observa-se que os atletas de Atletismo e Badminton tendem a ter uma mediana de IMC menor em comparação aos de Ginástica e Judô, sugerindo que essas duas modalidades têm atletas com IMC menores. O Judô apresenta a maior mediana entre os esportes analisados, indicando uma tendência a ter um IMC mais elevados entre seus atletas.

O Judô se destaca por ter uma caixa maior, sugerindo mais variação no IMC dos seus atletas em comparação aos outros esportes. Por outro lado, os da Ginástica é menor, o que indica uma menor variação no IMC dos atletas dessa modalidade.

O boxplot também exibe varios pontos fora dos limites, também conhecidos com outliers, que representam valores atípicos. Esportes como Atletismo e Judô apresentam uma quantidade significativa de outliers, com alguns atletas tendo IMC consideravelmente mais altos que a maioria de seus pares na mesma modalidade.

3.2.1 Normalidade

O teste de Shapiro-Wilk foi aplicado para cada esporte para verificar se os dados de IMC seguem uma distribuição normal.

x
1,000000e-18
9,288618e-13
2,867683e-22

x
2,564731e-38

Os valores-p resultantes para todos os esportes foram extremamente baixos, menores que um nível de significância preestabelecido de 5%„ indicando que os dados não seguem uma distribuição normal. Mesmo assim, os grupos possuem tamanhos de amostras grandes, permitindo o uso da ANOVA (Analise de variância), pois este teste é indicado para com grandes amostras.

3.2.2 Homogeneidade

O teste de Levene foi aplicado para verificar se a variância do IMC entre os esportes é homogênea.

	Df	F value	Pr(>F)
group	3	250,2836	0
	19161	NA	NA

O resultado mostrou um valor-p muito baixo (p < 0,00000000000000022), indicando que as variâncias entre os grupos não são homogêneas. A violação da homogeneidade sugere que é preciso ter cuidado ao interpretar os resultados da ANOVA.

3.2.3 Análise de variância(ANOVA)

Apesar da violação de alguns pressupostos, aplicou-se a ANOVA para avaliar se existem diferenças significativas entre os IMCs médios dos diferentes esportes.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Esporte	3	27096,23	9032,07797	717,3436	0
Residuals	19161	241256,28	12,59101	NA	NA

O resultado indicou uma diferença significativa (p < 0,05), sugerindo que pelo menos um dos esportes possui um IMC médio diferente dos outros.

3.2.4 Teste de Tukey (Comparações Múltiplas)

Para identificar quais grupos de esportes apresentaram diferenças significativas entre si, foi realizado o teste de Tukey.

	diff	lwr	upr	p adj
Badminton-Atletismo	0,3597304	0,0530908	0,6663699	0,0137523
Ginástica-Atletismo	-0,8862583	-1,0422143	-0,7303024	0,0000000
Judô-Atletismo	3,5189291	3,2930662	3,7447920	0,0000000
Ginástica-Badminton	-1,2459887	-1,5681854	-0,9237920	0,0000000
Judô-Badminton	3,1591987	2,7979474	3,5204500	0,0000000
Judô-Ginástica	4,4051874	4,1586170	4,6517578	0,0000000

Os resultados mostraram diferenças significativas entre todos os pares de esportes analisados, com destaque para a diferença positiva entre Judô e Atletismo (3,52) e negativa entre Ginástica e Atletismo (-0,89), indicando que atletas de Judô têm, em média, IMC mais elevados em comparação aos de Atletismo, enquanto os de Ginástica tendem a ter IMC mais baixos.

3.2.5 Teste de Kruskal-Wallis

Dada a violação da normalidade e homogeneidade das variâncias, foi aplicado o teste de Kruskal-Wallis, uma alternativa não-paramétrica à ANOVA.

	Estatística	p_value	Número_de_grupos	Tamanho_da_amostra
Kruskal-Wallis chi-squared	1627,172	0	36	67474

O resultado do p-valor foi menor que 5% reforçou a conclusão da ANOVA, indicando diferenças significativas entre os grupos de esportes.

3.3 Top 3 medalhistas gerais por quantidade de cada tipo de medalha

3.4 Variação Peso por Altura

4 Conclusões