Projeto House of excellence

Consultores Responsáveis:
Maria Beatriz Cunha Barros

Requerente:
João Vitor Neves

Brasília, 4 de novembro de 2024.

ESTAT
consultoria estatística

Sumário

1 Introdução

Este relatório tem como objetivo principal a realização de análises estatísticas, a fim de ponderar e avaliar o desempenho dos atletas de elite da companhia House of Excellence ao longo de 5 olimpíadas. Através da utilização de alguns critérios estatísticos,as avaliações tendem esclarecer 4 tópicos pertinentes à empresa, os quais são descirtas como a compreensão dos países com o maior número de mulheres medalistas, a verificação e a possível correlação entre os IMCs de determinados esportes, a relação entre medalistas gerais e as medalhas conquistadas, além de entender a relação entre o peso e a altura dos atletas da equipe.

O banco de dados foi coletado e disponibilizado pela própria empresa. Trata-se de uma amostra composta por 5 variáveis qualitativas nominais, sendo essas o nome, o gênero, a idade, o time, o esporte e a modalidade. Além disso, também dispõe de 3 variáveis quantitativas contínuas, sendo essas a idade, a altura e o peso. Também conta com uma variável do tipo qualitativa ordinal descrita como o tipo da medalha conquistada pelo atleta.

Por fim, o software utilizado para o desenvolvimento da pesquisa foi a versão 4.4.1 do programa R, o qual se trata de uma linguagem de programação estatística e gráfica, além de ser uma ferramenta gratuita e específica para análise de dados.

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 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.

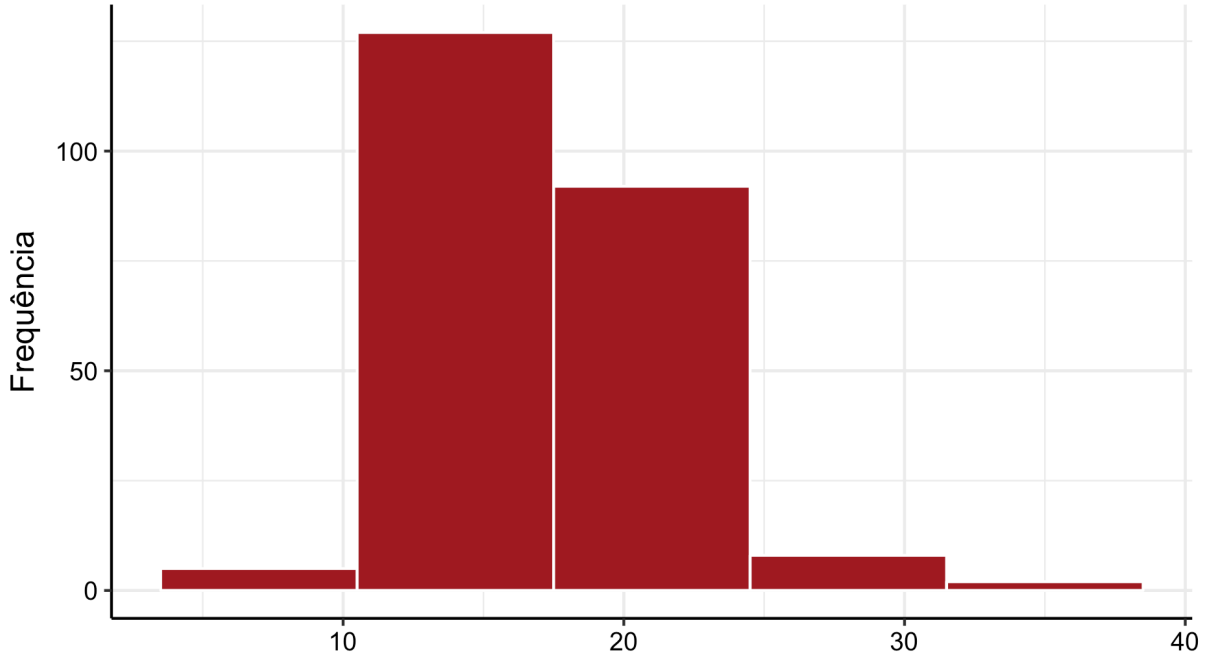


Figura 1: Exemplo de histograma

2.5 Tipos de Variáveis

2.5.1 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)

2.5.2 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)

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			

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.

2.6 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

2.7 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.

2.8 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 $y_i > y_j$.

2.9 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:

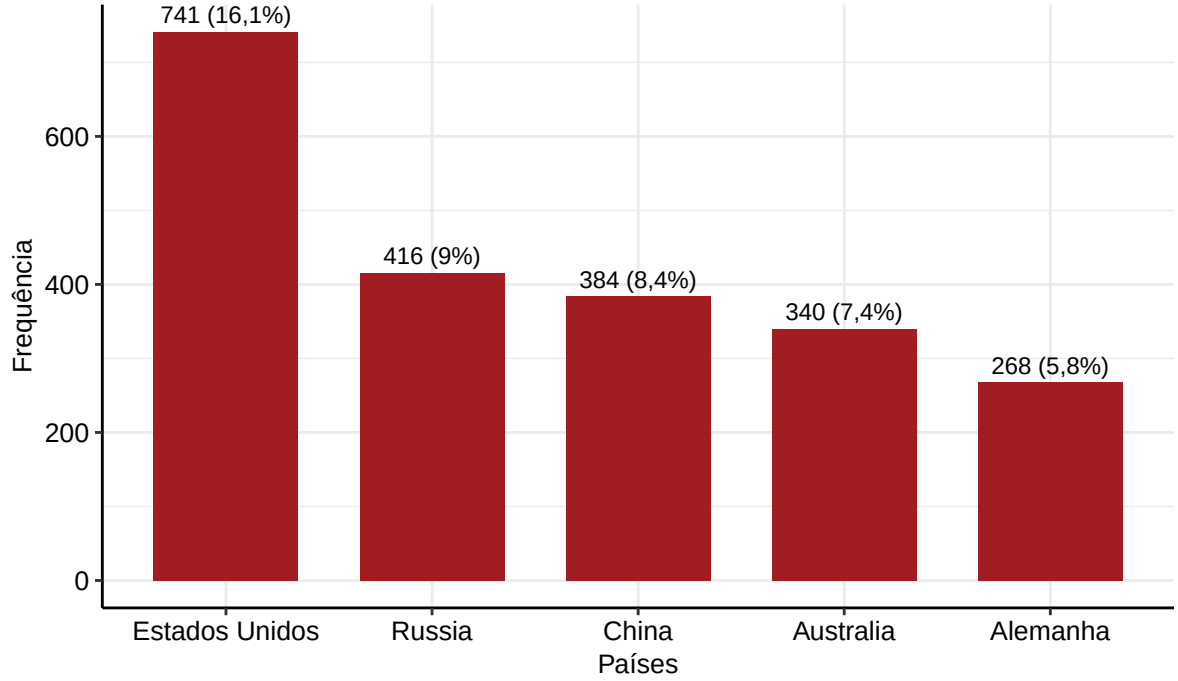
- Se $\dfrac{Y_j - Y_i}{X_j - X_i} > 0$, os pares são **concordantes** (adicione 1 a N_c)
- Se $\dfrac{Y_j - Y_i}{X_j - X_i} < 0$, os pares são **discordantes** (adicione 1 a N_d)
- Se $\dfrac{Y_j - Y_i}{X_j - X_i} = 0$, ocorreu **empate** (adicione 0,5 a N_c e a N_d)
- Se $X_j = X_i$, não há comparação

3 Análises

3.1 Top 5 países com maior número de mulheres medalistas

Para a primeira análise será utliziado as variáveis categóricas qualitativa nominais Sexo e Time, além da variável categórica qualitativa ordinal denominada Medalha.

Figura 1: Gráfico de Colunas do Top 5 países com maior número de mulheres medalistas



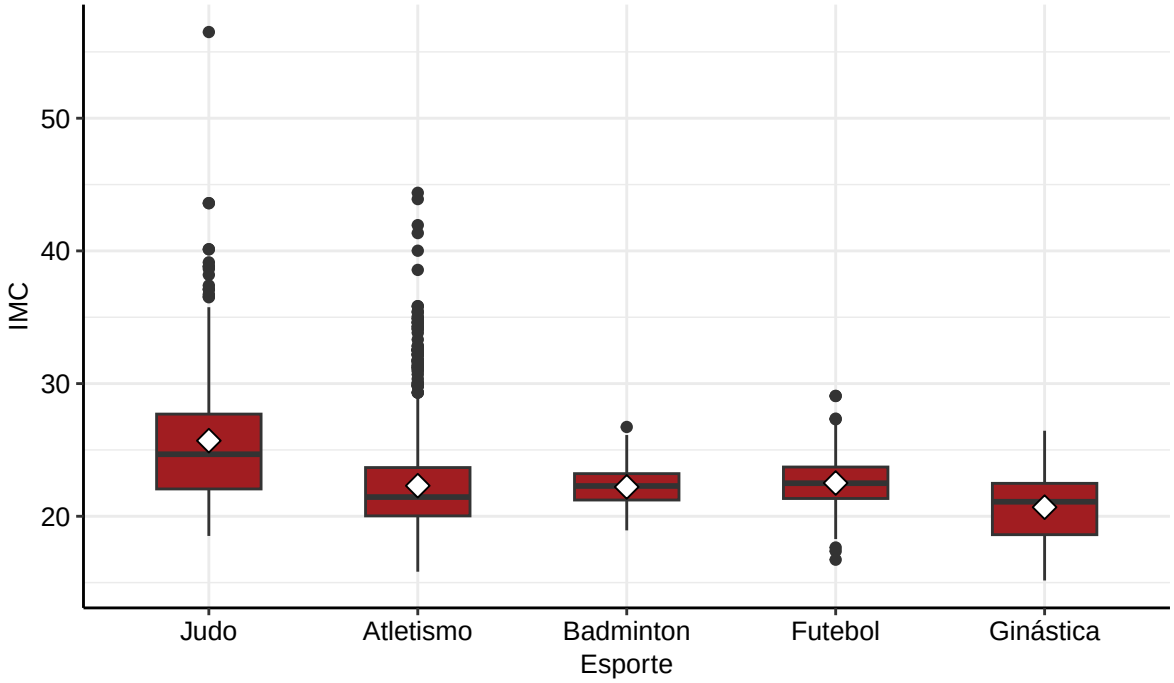
Ao selecionarmos apenas as ganhadoras mulheres dos dados disponibilizados, podemos perceber que representam 4597 ganhadores, ou seja, cerca de 45,89% do total de medalistas.

Outrossim, percebe-se por meio da Figura 1 que o país com maior número de ganhadoras mulheres são os Estados Unidos, representando cerca de 16,1% dentro desse secção denominada Top 5, seguido pela Russia (9%), pela China (8,4%), pela Austrália (7,4%) e pela Alemanha (5,8%).

3.2 Valor IMC por esporte

Para a segunda análise será utliziado a variável categórica qualitativa nominais Time, além da variável categórica qualitativa contínua denominada IMC.

Figura 2: Boxplot do IMC pelo tipo de esporte



Ao analisarmos as médias de cada esporte, podemos observar que, em comparação aos outros esportes, os atletas da ginástica possuem um valor de IMC médio mais baixo (20,7), enquanto o judo apresenta IMC médio mais alto (25,7). Segundo a figura 2, é visivel um grande número de outliners, ou seja, valores discrepantes, que estão for do intervalo esperado.

A fim de verificarmos se há diferença entre os IMC entre diferentes tipos de esporte, foi realizado primeiro o teste de normalidade Shapiro-Wilk, o qual indicou que a váriavel analisada não segue uma distribuição Normal de probabilidade($P<0,001$), ou seja, é necessário a realização de um teste não paramétrico.

$$\begin{cases} H_0 : \text{O IMC e tipo de esporte seguem uma distribuição simétrica em torno de zero} \\ H_1 : \text{O IMC e tipo de esporte não seguem uma distribuição simétrica em torno de zero} \end{cases}$$

Posteriormente foi realizado o teste Kruskal-Wallis($P<0,001$), o qual apontou que existe ao menos uma diferença significativa entre os grupos de esportes quando comparados entre si.

$$\begin{cases} H_0 : \text{A distribuição das variáveis é a mesma para todos os grupos} \\ H_1 : \text{As distribuições das variáveis não são as mesmas para todos os grupos} \end{cases}$$

Ademais, vale ressaltar que, ao comparar os esportes a partir do teste de Dunn, o judô se destacou com grandes diferenças entre sua média de IMC em comparação

com outros esportes, enquanto entre atletismo e badminton(p = 0,519) não foram encontradas diferenças significativas entre as médias.

Quadro 1: Gráfico de colunas da categoria de IMC pelo esporte praticado

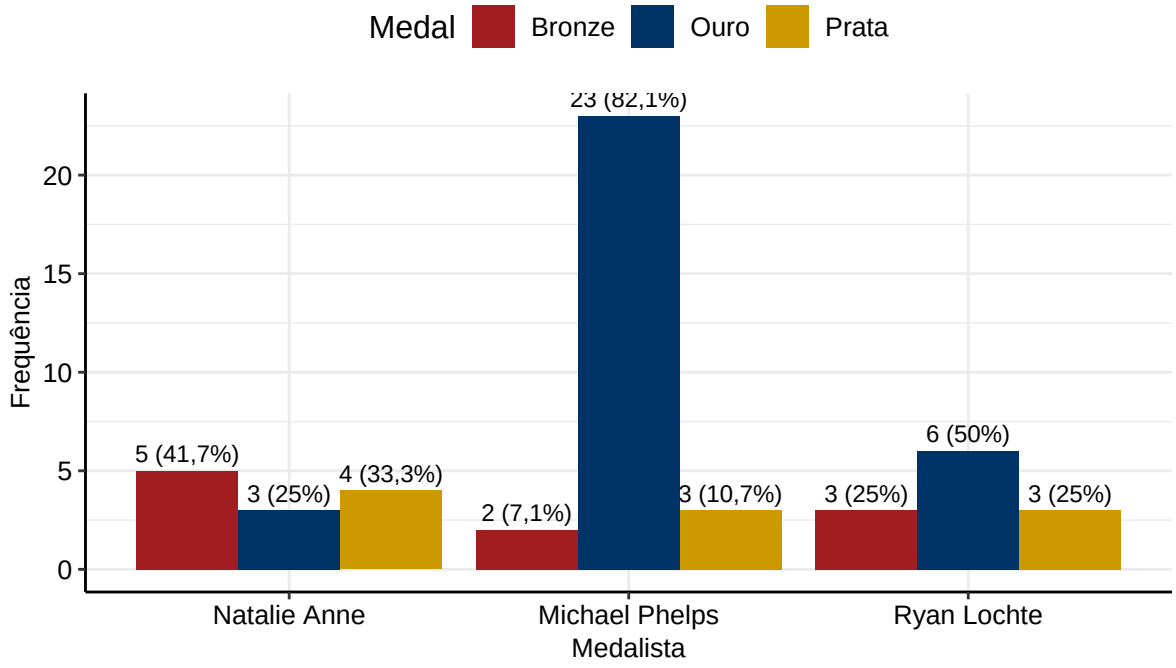
Quadro 1: Medidas resumo da(o) [nome da variável]

Percebe-se por meio da figura 3 a frequência de cada esporte dentro das categorias de IMC, as quais são classificadas como abaixo do ideal (IMC < 18.5), normal (18.5 < IMC < 25), sobrepeso (25 < IMC < 30) e obesidade (IMC > 30). Foi possível afirmar que o grupo de atletismo apresentam maior número de atletas com IMC abaixo do ideal (53,63%), enquanto os atletas do judo e badminton ficaram de fora dessa categoria.

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

Para a terceira análise, foi utilizado a variável medalha e análisado os três medalistas com maior numero de medalhas nessas últimas olimpiadas.

Figura 4: Gráfico de Colunas dos três medalistas mais pontuados e suas respectivas quantidades de medalhas.



Segundo a figura 4, foi observado uma diferença de 16 medalhas entre o primeiro colocado, o nadador Michael Phelps, e o segundo e terceiro colocados, os quais empataram em quantidade de medalhas. O valor de medalhas de ouro de Phelps chega a ser discrepante entre os outros competidores, o que indica uma maior vantagem do nadador sobre os outros atletas.

Grupo	P-valor	Decisão do teste
Nome Medalha	0.012	Rejeita H_0

Quadro 2: P-valor do teste de associação (Teste Qui-Quadrado) entre as variáveis medalista e medalha

$$\begin{cases} H_0 : \text{As variáveis medalhas e medalistas são independentes} \\ H_1 : \text{As variáveis medalhas e medalistas não são independentes} \end{cases}$$

Analisando o Quadro 2, que representa o teste de associação entre as variáveis medalista e medalha, é visível que há uma associação estatisticamente significativa entre as variáveis analisadas (P = 0,012), ou seja, é possível afirmar que a frequência de medalha parece estar associada ao medalista.

3.4 Variação peso por altura

Esta análise tem o objetivo de compreender se há uma relação entre o peso e altura dos atletas. Para tal, foram utilizadas as variáveis Peso e Altura e será utilizado um quadro de medidas resumo e um quadro indicativo do p-valor do teste que será utilizado, a fim de ilustrar as análises.

Quadro 3: P-valor do teste de coorelação entre a Altura e o Peso

Variáveis	P-valor	Decisão do teste
Peso Altura	<0,001	Rejeita H_0

Analisando o Quadro 3, que representa o teste de correlação entre as variáveis peso e altura, é visível que existe dependência entre as variáveis(P<0,001), ou seja, há uma forte evidência que o peso e a altura do atleta estão associados de forma diretamente proporcional, quando um aumenta, o outro também tende a aumentar.

$$\begin{cases} H_0 : \text{Não existe uma associação significativa entre o Peso e a Altura} \\ H_1 : \text{Existe uma associação significativa entre o Peso e a Altura} \end{cases}$$

3.5 Conclusões