UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO

DEPARTAMENTO DE ESTATÍSTICA

SME0807 - Técnicas de Amostragem: Análise utilizando dados do Enem

Ariel Bor Cheng Chen (11877260)
Douglas Sudré Souza (10733820)
Josiel Ramalho da Rosa (11735139)
Lucas Roberto de Oliveira Lopes (10850460)
Rafael Fragoso Marin (11809318)
Vinicius Loureiro Siqueira (10857119)

São Carlos, SP

1 Introdução

Em uma pesquisa, a determinação de uma amostra é essencial ao querermos entender, seja uma característica ou um comportamento, acerca de uma população, mas não conseguimos examinar o todo.

O uso inadequado de um procedimento amostral pode levar a um viés de interpretação do resultado.(Bolfarine, Bussab, 2004). Através de uma amostra representativa, suficiente e aleatória podemos inferir sobre uma população a partir de parâmetros estimados garantindo que a interpretação do resultado estará livre de viéses originários da escolha da amostra.

O Exame Nacional do Ensino Médio (ENEM) foi criado em 1998 com o objetivo de avaliar as competências básicas para o exercício pleno da cidadania e como uma "modalidade alternativa ou complementar aos exames de acesso aos cursos profissionalizantes pós-médios e ao ensino superior" (Brasil. Inep, 1998, p. 2). A partir de 2009 o exame também começou a ser utilizado como forma de acesso ao ensino superior no Brasil. O Sistema de Seleção Unificada (Sisu) passou a operar em larga escala no processo de alocação dos candidatos às vagas. (Silveira, Barbosa, Silva, 2015). A prova é dividida em cinco áreas do conhecimento, correspondentes às Ciências Naturais e suas Tecnologias, Linguagens, Códigos e suas Tecnologias, Matemática e suas Tecnologias, Ciências Humanas e suas Tecnologias e Redação.

Recentemente, durante a reunião da Coordenação do Curso de Estatística e Ciência de Dados, do dia 07/06/2022, estabeleceram o aumento das notas mínimas do ENEM para o curso a fim de equiparar com as notas mínimas de outros cursos do Instituto De Ciências Matemáticas e de Computação, como a de Bacharelado em Ciência da Computação, Ciência de Dados e também os de Licenciatura/Bacharelado em Matemática. Nessa reunião foi estabelecida as notas mínimas para: Matemática e suas Tecnologias (450), Ciências da Natureza e suas tecnologias (350), Ciências Humanas e suas Tecnologias (350), Linguagens, Códigos e suas Tecnologias (400) e Redação (400). Esta nova nota estabelecida gerou algumas repercussões entre os graduandos do curso sobre o quanto isso pode afetar futuramente a entrada de novos estudantes e isso nos motivou a ver como é o desempenho das pessoas que prestam o vestibular ENEM e comparar com as novas notas estabelecidas.

A partir de todo contexto apresentado, este trabalho objetiva-se a analisar o desempenho dos vestibulandos, calculando, através de técnicas de amostragem e inferência estatística, uma estimativa para a nota média da área de Matemática e suas Tecnologias e também para as áreas de Ciências da Natureza e suas Tecnologias, Ciências Humanas e suas Tecnologias, Linguagens, Códigos e suas Tecnologias e Redação, com um intervalo de confiança de 95%. Além disso, pretende-se analisar se a média geral de Matemática é superior a 450 e também se a média

geral em Ciências da Natureza e Humanas é superior a 300, assim como para Linguagens e Redação é superior a 400, que são as novas notas mínimas estabelecidas pela Coordenação do Curso de Estatística e Ciência de Dados.

2 Fundamentação Teórica

Publicado em 26 de dezembro de 2018, o artigo "Plano de amostragem: um estudo de caso utilizando dados do Enem"de Fernando Antonio de Melo Pereira, Raquel Alves Basílio, Maria Jessiane Alexandre da Silva e Roberto Junior J. Oliveira apresenta técnicas e procedimentos de amostragem para pesquisas em educação, que busquem generalizar resultados. Foi-se utilizado a base de dados da prova do ENEM aplicada em 2014 para descrever procedimentos computacionais e aspectos teóricos sobre a seleção de uma amostra, com base no desempenho dos alunos no ENEM. As escolas privadas da cidade de Natal, no estado de Rio Grande do Norte, compõem a população e uma amostra de escolas nesta região foi obtida através de estratificação e agrupamento. Também utilizou-se amostragem proporcional para o tamanho. (Pereira, Basílio, Silva, Oliveira, 2018).

Como esse artigo desenvolveu um tema semelhante ao nosso resolvemos usar ele como suporte teórico para a pesquisa, onde utilizamos o conjunto de dados da prova do ENEM aplicada em 2019.

3 Metodologia

3.1 Análise Descritiva

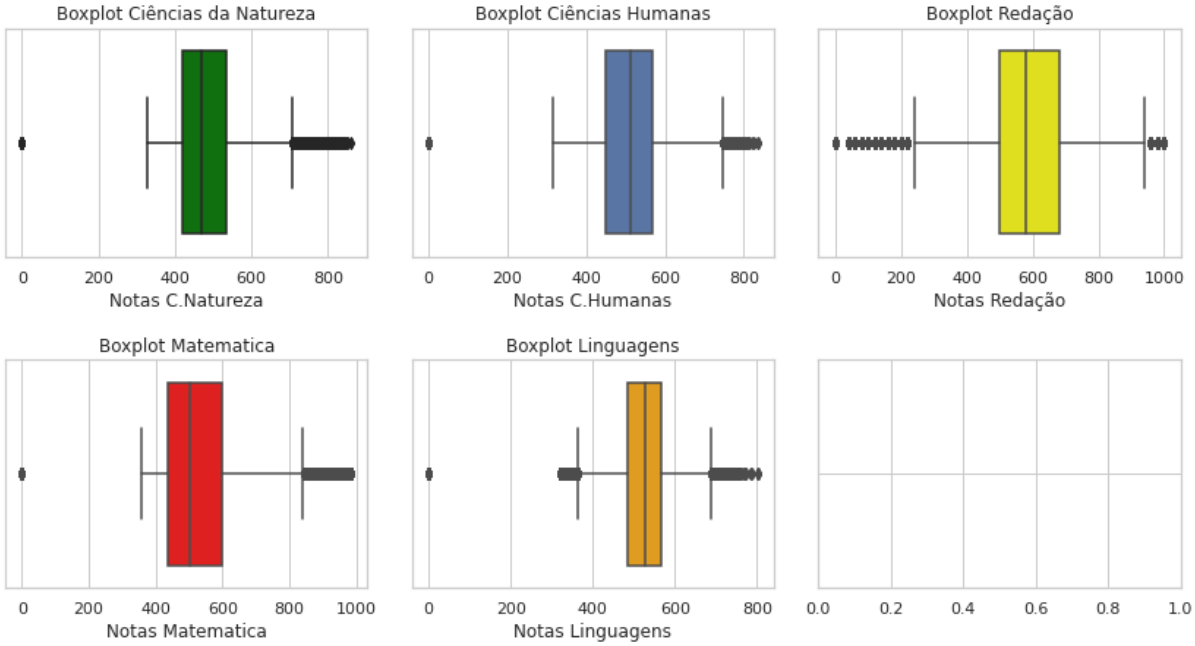

Figura 3.1.1: Boxplot das notas de toda população

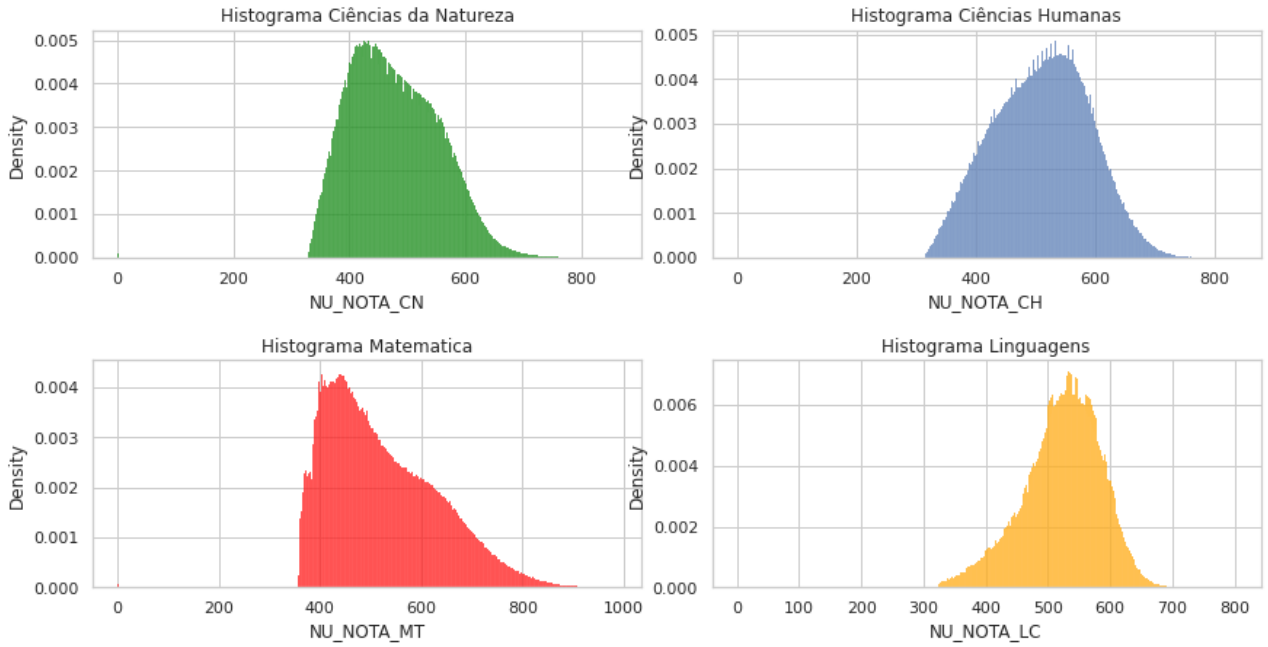

Figura 3.1.2: Histograma das notas de toda população

Com relação aos dados de toda amostra, o que podemos notar é que a prova de matemática apresenta o melhor desempenho. Nela foram onde os alunos conseguiram as maiores notas em comparação as outras, e tiveram menos notas baixas. Já na prova de linguagens, podemos ver que o oposto ocorre, nela tiveram as notas mais baixas e foi a matéria que teve o menor número de notas altas.

Nas provas de ciências da natureza e ciências humanas, tivemos um desempenho parecido nas notas mais altas, e um índice um pouco maior de notas mais baixas em ciências humanas.

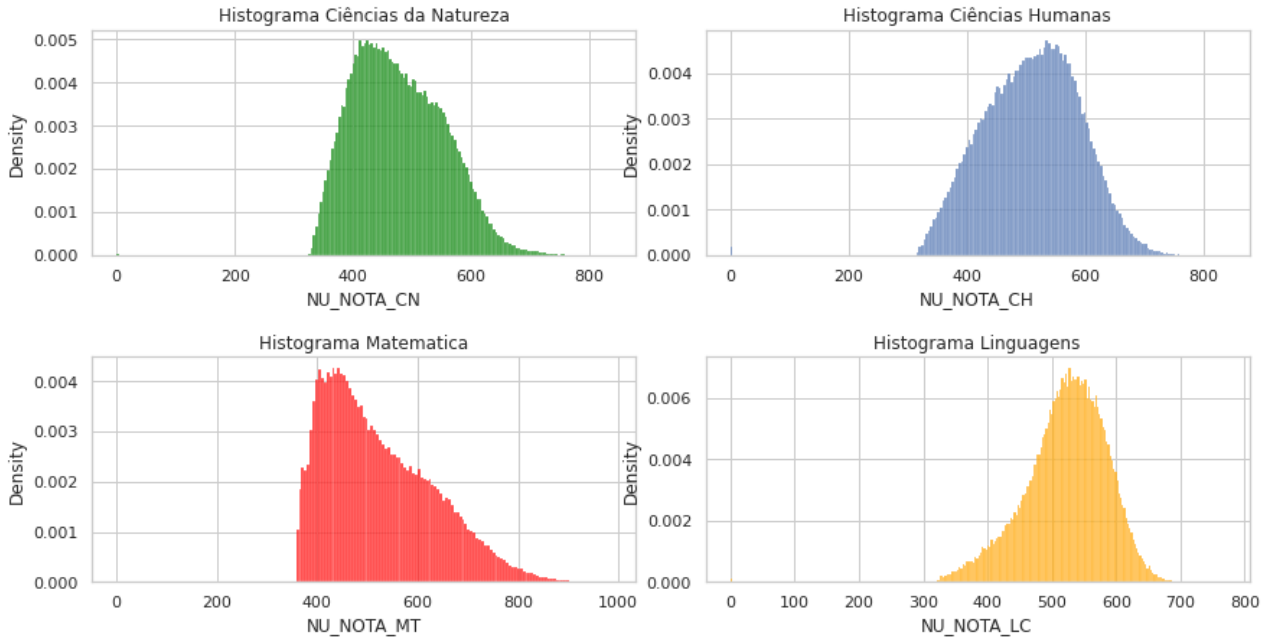

Figura 3.1.3: Histograma da amostra

Em relação a nossa amostra, temos indício de que é representativa, já que os dados seguem quase o mesmo padrão do que foi visto para os dados populacionais.

3.2 Amostragem

O nosso objetivo principal é avaliar o desempenho na área de Matemática e suas Tecnologias, então iremos calcular o tamanho da amostra para o estudo a fim de obter boas estimativas para as notas dessa área.

Para definir o tamanho da amostra, vamos utilizar uma Amostragem Aleatória Simples sem Reposição (AAS_s), com isso, para um erro máximo de esimação B definido, temos a seguinte equação para o cálculo do tamanho de amostra:

$$n = \frac{1}{(\frac{B}{z_{\alpha/2} \cdot s})^2 + \frac{1}{N}}$$

O nosso conjunto de dados possui um total de 3701909 observações e uma variância estimada para a nota em Matemática de 11886.3323, então definindo um erro máximo de estimação de 50 e nível de significância 5%, temos o seguinte cálculo para o tamanho da amostra:

$$n = \frac{1}{(\frac{B}{z_{\alpha/2} \cdot s})^2 + \frac{1}{N}} = \frac{1}{(\frac{50}{z_{0.025} \cdot 11886.3323})^2 + \frac{1}{3701909}} = \frac{1}{(\frac{50}{1.96 \cdot 11886.3323})^2 + \frac{1}{3701909}} \approx 205076.9743$$

Com isso, temos:

$$n = 205077$$

Portanto, fixamos uma semente para pegar uma amostra de 20577 observações e fazer as análises de interesse.

3.3 Estimativas para os objetivos

O nosso interesse de estudo é estimar a nota média em Matemática, assim como para as outras áreas de Ciências da Natureza, Ciências Humanas, Linguagens e Redação. Calculando também um intervalo com 95% de confiança para as médias estimadas.

Desempenho na área de Matemática e suas Tecnologias

Vamos calcular uma estimativa para a nota média em Matemática utilizando a nossa amostra selecionada, assim como um intervalo de confiança de 95%, para isso, vamos calcular primeiro a estimativa para a média assim como para sua variância, sabendo que a variância das notas (s^2) é de 11825.1205:

$$\bar{y} = \frac{t(\mathbf{s})}{n} \approx 523.0787$$

$$var(\bar{y}) = (1 - \frac{n}{N}) \cdot \frac{s^2}{n} = (1 - \frac{205077}{3701909}) \cdot \frac{11825.1205}{205077} = 0.0545$$

Com a estimativa para a média e sua respectiva estimativa de variância, podemos calcular o intervalo de confiança de 95%:

$$IC(\bar{y}, 95\%) = \bar{y} \pm z_{\alpha/2} \cdot \sqrt{var(\bar{y})} = 523.0787 \pm 1.96 \cdot \sqrt{0.0545}$$

$$IC(\bar{y}, 95\%) = [522.6213; 523.5361]$$

Com isso, temos uma estimativa para a nota média em Matemática de 523.0797 e um intervalo com 95% de confiança para a média de [522.6213 ; 523.5361].

Desempenho na área de Ciências da Natureza e suas Tecnologias

Vamos calcular uma estimativa para a nota média em Ciências da Natureza utilizando a nossa amostra selecionada, assim como um intervalo de confiança de 95%, para isso, vamos calcular primeiro a estimativa para a média assim como para sua variância, sabendo que a variância das notas (s^2) é de 5765.1925:

$$\bar{y} = \frac{t(\mathbf{s})}{n} \approx 477.7713$$

$$var(\bar{y}) = (1 - \frac{n}{N}) \cdot \frac{s^2}{n} = (1 - \frac{205077}{3701909}) \cdot \frac{5765.1925}{205077} = 0.0266$$

Com a estimativa para a média e sua respectiva estimativa de variância, podemos calcular o intervalo de confiança de 95%:

$$IC(\bar{y}, 95\%) = \bar{y} \pm z_{\alpha/2} \cdot \sqrt{var(\bar{y})} = 509.7053 \pm 1.96 \cdot \sqrt{0.0266}$$

$$IC(\bar{y}, 95\%) = [477.4519; 478.0907]$$

Com isso, temos uma estimativa para a nota média em Ciências da Natureza de 477.7713 e um intervalo com 95% de confiança para a média de [477.4519 ; 478.0907].

Desempenho na área de Ciências Humanas e suas Tecnologias

Vamos calcular uma estimativa para a nota média em Ciências Humanas utilizando a nossa amostra selecionada, assim como um intervalo de confiança de 95%, para isso, vamos calcular primeiro a estimativa para a média assim como para sua variância, sabendo que a variância das notas (s^2) é de 6535.6547:

$$\bar{y} = \frac{t(\mathbf{s})}{n} \approx 509.7053$$

$$var(\bar{y}) = (1 - \frac{n}{N}) \cdot \frac{s^2}{n} = (1 - \frac{205077}{3701909}) \cdot \frac{6535.6547}{205077} = 0.0301$$

Com a estimativa para a média e sua respectiva estimativa de variância, podemos calcular o intervalo de confiança de 95%:

$$IC(\bar{y}, 95\%) = \bar{y} \pm z_{\alpha/2} \cdot \sqrt{var(\bar{y})} = 509.7053 \pm 1.96 \cdot \sqrt{0.0301}$$

$$IC(\bar{y}, 95\%) = [509.3653; 510.0454]$$

Com isso, temos uma estimativa para a nota média em Ciências da Humanas de 509.7053 e um intervalo com 95% de confiança para a média de [509.3653 ; 510.0454].

Desempenho na área de Linguagens, Códigos e suas Tecnologias

Vamos calcular uma estimativa para a nota média em Linguagens utilizando a nossa amostra selecionada, assim como um intervalo de confiança de 95%, para isso, vamos calcular primeiro a estimativa para a média assim como para sua variância, sabendo que a variância das notas (s^2) é de 3939.2381:

$$\bar{y} = \frac{t(\mathbf{s})}{n} \approx 522.2624$$

$$var(\bar{y}) = (1 - \frac{n}{N}) \cdot \frac{s^2}{n} = (1 - \frac{205077}{3701909}) \cdot \frac{3939.2381}{205077} = 0.0181$$

Com a estimativa para a média e sua respectiva estimativa de variância, podemos calcular o intervalo de confiança de 95%:

$$IC(\bar{y}, 95\%) = \bar{y} \pm z_{\alpha/2} \cdot \sqrt{var(\bar{y})} = 522.2624 \pm 1.96 \cdot \sqrt{0.0181}$$

$$IC(\bar{y}, 95\%) = [521.9983; 522.5264]$$

Com isso, temos uma estimativa para a nota média em Linguagens de 522.2624 e um intervalo com 95% de confiança para a média de [521.9983 ; 522.5264].

Desempenho na área de Redação

Vamos calcular uma estimativa para a nota média em Redação utilizando a nossa amostra selecionada, assim como um intervalo de confiança de 95%, para isso, vamos calcular primeiro a estimativa para a média assim como para sua variância, sabendo que a variância das notas (s^2) é de 33340.5283:

$$\bar{y} = \frac{t(\mathbf{s})}{n} \approx 579.6213$$

$$var(\bar{y}) = (1 - \frac{n}{N}) \cdot \frac{s^2}{n} = (1 - \frac{205077}{3701909}) \cdot \frac{33340.5283}{205077} = 0.1536$$

Com a estimativa para a média e sua respectiva estimativa de variância, podemos calcular o intervalo de confiança de 95%:

$$IC(\bar{y}, 95\%) = \bar{y} \pm z_{\alpha/2} \cdot \sqrt{var(\bar{y})} = 579.6213 \pm 1.96 \cdot \sqrt{0.1536}$$

$$IC(\bar{y}, 95\%) = [578.8532; 580.3894]$$

Com isso, temos uma estimativa para a nota média em Linguagens de 579.6213 e um intervalo com 95% de confiança para a média de [578.8532 ; 580.3894].

3.4 Teste de Hipótese

O teste de hipóteses nos auxilia na tomada de decisões sobre uma população a partir da amostra, nos permitindo verificar se os dados amostrais trazem evidências que apoiem ou não suposicões acerca dos parâmetros investigados. Chamamos estas suposições de hipóteses estatísticas.

A hipótese estatística se refere a qualquer afirmação acerca da distribuição de probabilidades de uma ou mais variáveis aleatórias.(Bolfarine, Bussab, 2004).

A hipótese de interesse é chamada de hipótese nula H_0. Caso esta hipótese seja rejeitada, tomamos como verdadeira a hipótese alternativa H_1/H_A.

Sendo a variável aleatória X distribuída de acordo com a função de densidade (ou de probabilidade) $f(x|\theta)$, com $\theta \in \Theta$, dizemos que a distribuição de X está totalmente especificada quando conhecemos $f(x|\theta)$ e θ. (Bolfarine, Bussab, 2004).

A distribuição de X será dita estar parcialmente especificada quando conhecemos a função de densidade (ou de probabilidade) $f(x|\theta)$, mas não θ. Associados às hipóteses H_0 e H_1, definimos os conjuntos Θ_0 e Θ_1, ou seja, H_0 afirma que $\theta \in \Theta_0$ (notação: $H_0 : \theta \in \Theta_0$) e H_1 afirma que $\theta \in \Theta_1$ (notação: H_1: $\theta \in \Theta_1$). (Bolfarine, Bussab, 2004).

No caso em que $\Theta_0 = \theta_0$ consideramos que H_0 é simples. Caso contrário, dizemos que H_0 é composta. O mesmo vale para a hipótese alternativa H_1. (Bolfarine, Bussab, 2004).

No nosso caso, nossa hipótese básica corresponde à analisar se a média geral de matemática é superior a 450. Deste modo, temos como hipóteses nula e alternativa:

$$\begin{cases} H_0 : \text{média das notas em matematica} \leq 450; \\ H_1 : \text{média das notas em matematica} > 450. \end{cases}$$

Além da hipótese básica, apresentaremos também algumas hipóteses secundárias, que correspondem à analisar se a média geral em Ciências da Natureza e suas Tecnologias e em Ciências Humanas e suas Tecnologias é superior a 350, assim como analisar se a média geral em Linguagens, Códigos e suas Tecnologias e suas Tecnologias e em Redação e suas Tecnologias é superior a 400.

Para realizar os testes de hipótese, vamos utilizar a estatística de teste:

$$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{var(\hat{\theta}}} \sim N(0,1)$$

Através dessa estatística, vamos calcular o nível descritivo (p-valor) e para ver se é significativo a nível de significância 5%. Como o nosso caso é unilateral à direta, temos que:

$$p - valor = P(Z \geq z|H_0) = 1 - P(Z \leq z|H_0)$$

Se p-valor < 0.05, com base nessa amostra coletada, rejeitados H_0 a nível de significância 5%.

Hipótese básica: Matemática

Vamos analisar se a média geral de Matemática é superior a 450. Deste modo, temos como hipóteses nula e alternativa:

$$\begin{cases} H_0 : \text{média das notas em matematica} \leq 450; \\ H_1 : \text{média das notas em matematica} > 450. \end{cases}$$

Temos:

$$z = \frac{\hat{\theta} - \theta_0}{\sqrt{var(\hat{\theta})}} = \frac{523.0786 - 450}{\sqrt{0.0545}} \approx 313.128$$

e,

$$p - valor = 1 - P(Z \leq z|H_0) = 1 - P(Z \leq 313.128|H_0) = 1 - 1 = 0.0$$

Como nosso p-valor deu 0.0, com base nessa amostra e ao nível de significância 5%, temos evidências para rejeitar H_0, e ficar com a hipótese de que a média das notas em Matemárica é superior a 450.

Hipótese secundária: Ciência da Natureza

Vamos analisar se a média geral de Ciência da Natureza é superior a 350. Deste modo, temos como hipóteses nula e alternativa:

$$\begin{cases} H_0 : \text{média das notas em ciência da natureza} \leq 350; \\ H_1 : \text{média das notas em ciência da natureza} > 350. \end{cases}$$

Temos:

$$z = \frac{\hat{\theta} - \theta_0}{\sqrt{var(\hat{\theta})}} = \frac{477.7713 - 350}{\sqrt{0.0266}} \approx 784.0797$$

e,

$$p - valor = 1 - P(Z \leq z|H_0) = 1 - P(Z \leq 784.0797|H_0) = 1 - 1 = 0.0$$

Como nosso p-valor deu 0.0, com base nessa amostra e ao nível de significância 5%, temos evidências para rejeitar H_0, e ficar com a hipótese de que a média das notas em Ciência da Natureza é superior a 350.

Hipótese secundária: Ciência Humanas

Vamos analisar se a média geral de Ciência Humanas é superior a 350. Deste modo, temos como hipóteses nula e alternativa:

$$\begin{cases} H_0 : \text{média das notas em ciência humanas} \leq 350; \\ H_1 : \text{média das notas em ciência humanas} > 350. \end{cases}$$

Temos:

$$z = \frac{\hat{\theta} - \theta_0}{\sqrt{var(\hat{\theta})}} = \frac{509.7053 - 350}{\sqrt{0.0301}} \approx 920.4682$$

e,

$$p - valor = 1 - P(Z \leq z|H_0) = 1 - P(Z \leq 920.4682|H_0) = 1 - 1 = 0.0$$

Como nosso p-valor deu 0.0, com base nessa amostra e ao nível de significância 5%, temos evidências para rejeitar H_0, e ficar com a hipótese de que a média das notas em Ciência Humanas é superior a 350.

Hipótese secundária: Linguagens

Vamos analisar se a média geral de Linguagens é superior a 400. Deste modo, temos como hipóteses nula e alternativa:

$$\begin{cases} H_0 : \text{média das notas em linguagens} \leq 400; \\ H_1 : \text{média das notas em linguagens} > 400. \end{cases}$$

Temos:

$$z = \frac{\hat{\theta} - \theta_0}{\sqrt{var(\hat{\theta})}} = \frac{522.2624 - 400}{\sqrt{0.0181}} \approx 907.6546$$

e,

$$p - valor = 1 - P(Z \leq z|H_0) = 1 - P(Z \leq 907.6546|H_0) = 1 - 1 = 0.0$$

Como nosso p-valor deu 0.0, com base nessa amostra e ao nível de significância 5%, temos evidências para rejeitar H_0, e ficar com a hipótese de que a média das notas em Linguagens é

superior a 400.

Hipótese secundária: Redação

Vamos analisar se a média geral de Redação é superior a 400. Deste modo, temos como hipóteses nula e alternativa:

$$\begin{cases} H_0 : \text{média das notas em redação} \leq 400; \\ H_1 : \text{média das notas em redação} > 400. \end{cases}$$

Temos:

$$z = \frac{\hat{\theta} - \theta_0}{\sqrt{var(\hat{\theta})}} = \frac{579.6213 - 400}{\sqrt{0.1536}} \approx 458.3588$$

e,

$$p - valor = 1 - P(Z \leq z|H_0) = 1 - P(Z \leq 458.3588|H_0) = 1 - 1 = 0.0$$

Como nosso p-valor deu 0.0, com base nessa amostra e ao nível de significância 5%, temos evidências para rejeitar H_0, e ficar com a hipótese de que a média das notas em Redação é superior a 400.

4 Conclusão

Como em todos os testes de hipóteses ficamos com a hipótese alternativa, ou seja, as médias são superiores que a nova nota mínima estabelecida, acreditamos que essa mudança não irá afetar a entrada de novos estudantes.

Além disso, por se tratar do curso Bacharelado em Estatística e Ciência de Dados, na Universidade de São Paulo, é natural que as notas de corte sejam maiores que as mínimas estabelecidas.

Portanto, mesmo que futuramente essa nota venha a afetar a entrada de um estudante, o problema não está nas notas mínimas, e sim, na qualidade de ensino que antecede os vestibulares, como é o caso do ENEM.

5 Referências Bibliográficas

BOLFARINE, H.; BUSSAB, W. O. Elementos de Amostragem. São Paulo: ABE - Projeto Fisher, Edgard Blücher, 2005.

BOLFARINE, H.; SANDOVAL, M.C. Introdução à Inferência Estatística. SBM. 2002.

PEREIRA, F. A. M.; BASÍLIO, R. A.; SILVA, M. J. A.; OLIVEIRA, R. J. J. Plano de amostragem: um estudo de caso utilizando dados do Enem. . Revista Gestão e Tecnologia. V. 14, n. 3, p. 177-202, 26 de dez de 2018.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Enem: documento básico. Brasília, 1998.

UFSC - ine - Depto. de Informática e Estatística: Teste de Hipóteses. Disponível em: https://www.inf.ufsc.br/ andre.zibetti/probabilidade/teste-de-hipoteses.html: :text=Nos

MINDMINERS: Como definir amostragem de uma pesquisa? - Frankenthal, R., 15 Feb 2022. Disponível em: https://mindminers.com/blog/como-definir-amostragem-de-pesquisa/: :text=