



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
CURSO DE GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO
PROBABILIDADE E ESTATÍSTICA (INE5405)

Leonardo de Sousa Marques
Rafael Veronezi Ribeiro
Ruan Alboni Ferreira
Thayse Estevo Teixeira

**Inferência Estatística de Dados do Exame Nacional de Desempenho dos Estudantes
(ENADE)**

Florianópolis
2025

1 Introdução

O **Exame Nacional de Desempenho dos Estudantes** (ENADE) foi instituído em 2004 pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) com o objetivo de avaliar o desempenho dos concluintes de cursos de graduação no Brasil. A cada ciclo trienal, determinados cursos em todo o país realizam a prova, que avalia o conhecimento dos estudantes sobre os conteúdos programáticos definidos nas diretrizes curriculares de cada curso (SEAVI, 2020).

Nesse contexto, o ENADE fornece um conjunto de métricas destinadas à avaliação dos cursos e das instituições de ensino superior, tais como o Conceito Enade (CE), o Conceito Preliminar de Curso (CPC), o Indicador de Diferença entre os Desempenhos Observados e Esperado (IDD) e o Índice Geral de Cursos (IGC). Esses indicadores, juntamente com as variáveis Categoria Administrativa e Modalidade de Ensino dos cursos avaliados, serão detalhados na seção de Materiais e Métodos.

O presente trabalho dá continuidade à análise exploratória realizada no Trabalho 1 da disciplina, ampliando o estudo por meio da aplicação de técnicas de inferência estatística – a partir dos dados de uma amostra, criamos testes de hipóteses para realizar conclusões e fazer generalizações sobre a população. Assim como na análise anterior, a população considerada corresponde aos cursos do estado de Santa Catarina. A partir desse conjunto de dados, será selecionada uma amostra aleatória, permitindo a realização de inferências estatísticas sobre a população com base nos resultados obtidos anteriormente, apresentados na Tabela ??.

A partir dos indicadores apresentados pelo ENADE, serão formuladas e testadas hipóteses relacionadas às médias, proporções, correlações e regressões, com o objetivo de aprofundar a compreensão acerca das variáveis quantitativas e qualitativas previamente examinadas.

Tabela 1: Estatísticas Descritivas dos Indicadores Contínuos (SC)

Indicador	Média	Mediana	Desvio Padrão	Variância	CV (%)	Mínimo	Máximo	Amplitude
Conceito ENADE (Contínuo)	2.478	2.488	0.971	0.943	39.196	0	4.972	4.972
IDD (Contínuo)	2.667	2.631	0.983	0.965	36.838	0	5	5
CPC (Contínuo)	2.885	2.857	0.636	0.404	22.029	0.977	4.603	3.626
IGC (Contínuo)	2.703	2.726	0.565	0.319	20.909	0.926	4.416	3.49

1.1 Objetivos

O objetivo principal do trabalho é tomar os indicadores apresentados anteriormente e que também foram utilizados no trabalho anterior e testar hipóteses levantadas sobre eles, sendo essas hipóteses aplicadas sobre a média e proporção dos dados de indicadores, além de hipóteses sobre a correlação e regressão entre duas dessas variáveis.

Com base nos dados dos indicadores, foram definidas as hipóteses a serem analisadas no presente trabalho:

1. Hipótese sobre a média:

H_0 : A média do CPC em cursos com ensino presencial em Santa Catarina é igual a média do CPC em cursos com ensino a distância.

2. Hipótese sobre a proporção:

H_0 : A proporção de instituições federais em Santa Catarina é de 25%.

3. Hipótese sobre a Correlação:

H_0 : Não há correlação entre o IDD e o Conceito Enade.

4. Hipóteses sobre a Regressão:

H_0 : O IDD não tem influência significativa sobre o Conceito Enade.

2 Materiais e Métodos

Nesta seção são apresentadas as variáveis quantitativas e qualitativas selecionadas para as análises estatísticas, as quais foram previamente exploradas no trabalho anterior e agora serão utilizadas na etapa de inferência estatística. A base de dados considerada compreende os cursos participantes do ENADE no ciclo trienal de 2021 a 2023, pertencentes ao estado de Santa Catarina, que constitui a população de interesse deste estudo.

Para a realização das inferências estatísticas, será determinado o tamanho de amostra necessário para cada tipo de variável, conforme as expressões apresentadas nas Equações 1a e 1b, aplicadas respectivamente às variáveis qualitativas e quantitativas:

$$\left. \begin{array}{l} (a) \quad n_0 = \frac{z^2 \cdot p \cdot (1-p)}{d^2} \\ (b) \quad n_0 = \frac{z^2 \cdot \sigma^2}{d^2} \end{array} \right\} n = \frac{n_0}{1 + \frac{n_0}{N}} \quad (1)$$

Nas equações, n_0 representa o tamanho amostral inicial e n o tamanho amostral corrigido para populações finitas. O parâmetro z corresponde ao valor crítico da distribuição normal padrão associado ao nível de confiança adotado. Para obtermos um nível de confiança de 95%, iremos considerar $z = 1,96$. A variância populacional σ^2 é conhecida através da Tabela ???. O parâmetro p indica a proporção esperada de ocorrência de um determinado evento. Para nossa amostragem, consideraremos $p = 0,5$, a fim de maximizar a variância $p \cdot (1-p)$. Já a variável d expressa o erro amostral tolerável, frequentemente fixado em 5%. E, por fim, N representa o tamanho da população.

2.1 Variáveis Quantitativas

1. Conceito Enade Contínuo (CE) - Variável Quantitativa Contínua

O Conceito Enade (CE) foi uma variável escolhida por avaliar o desempenho dos cursos de uma instituição de ensino. Sua métrica é baseada nos resultados obtidos por seus estudantes no Enade em relação ao desempenho geral da respectiva área acadêmica em todo o território nacional. Para fins de cálculo, um curso tem que ter pelo menos 2 participantes com resultados válidos no exame para gerar o Conceito Enade, um valor entre 0 e 5. (INEP, 2024b).

Conforme do anexo 1, que sintetiza as fórmulas supracitadas, o tamanho da amostra aleatória para essa variável será $n = 699$.

2. Indicador de Diferença entre os Desempenhos Observados e Esperado Contínuo (IDD) - Variável Quantitativa Contínua:

O IDD foi escolhido por avaliar o quanto um curso de graduação impactou no desenvolvimento de seus estudantes. Ele leva em conta o desempenho dos estudantes no Enade, sua nota de ingresso no ensino superior, a qualidade do curso, entre outras medidas de desempenho, para então comparar o resultado aos dos demais cursos do país. Tudo isso é então compactado em uma variável entre 0 e 5.

Conforme o script do anexo 1, o tamanho da amostra aleatória para essa variável será $n = 707$.

3. Conceito Preliminar de Curso Contínuo (CPC) - Variável Contínua

Por outro lado, o CPC foi uma escolha por avaliar a qualidade dos cursos de graduação. As métricas levadas em conta para o seu cálculo são o desempenho dos estudantes no Enade; a contribuição do curso ao desenvolvimento do aluno (IDD); a titulação e regime de trabalho do corpo docente (Censo da Educação Superior); e a opinião dos estudantes sobre a didática, infraestrutura e oportunidades de evolução acadêmica e profissional que o curso proporciona (Questionário do Estudante) (INEP, 2024a). Essa medida é relativa ao resultado médio da área de avaliação em todo o país, gerando uma faixa contínua definida de 1 a 5.

Conforme o script do anexo 1, o tamanho da amostra aleatória para essa variável será $n = 426$.

2.2 Variáveis Qualitativas

1. Categoria Administrativa - Variável Nominal

Categoria Administrativa aborda o órgão responsável por gerir a instituição de ensino de determinado curso, sendo as 7 categorias possíveis: Pública Municipal, Pública Estadual, Pública Federal, Comunitária/Confessional, Privada sem Fins Lucrativos, Privada com Fins Lucrativos e Especial. Sua escolha vem do interesse de comparar a presença e o desempenho das instituições públicas e privadas no exame. Conforme o script do anexo 1, o tamanho da amostra aleatória para essa variável será $n = 150$.

2. Modalidade de Ensino - Variável Nominal

A Modalidade de Ensino representa como as aulas do curso são administradas aos estudantes, tendo 2 modalidades possíveis: Educação Presencial e Educação a Distância. Por ser um tema recente, essa variável foi incluída para analisar o impacto dessas modalidades no desempenho e qualidade das instituições. Conforme o script do anexo 1, o tamanho da amostra aleatória para essa variável será $n = 300$.

2.2.1 Base de Dados

A base de dados escolhida para o trabalho foi a mesma já apresentada no primeiro trabalho: os dados do Conceito Enade de 2021, 2022 e 2023 disponível no site do ENADE em formato XLSX, contendo além do indicador de Conceito Enade, o CPC e o IDD.

Novamente, os dados foram tomados por um intervalo suficiente para abranger um ciclo trienal completo e filtrados com um script em Python, para formatação das tabelas e conversão do arquivo XLSX para CSV, e um script em R, para conter somente as entradas referentes ao estado de Santa Catarina.

3 Resultados e Discussões

Nesta seção iremos apresentar as análises estatísticas realizadas com base nas variáveis selecionadas na seção de Materiais e Métodos. Inicialmente, serão discutidas as estatísticas descritivas das variáveis quantitativas e qualitativas, seguido pelos testes de hipóteses levantados na seção de Objetivos.

3.1 Análise das Estatísticas Descritivas

Observando a base de dados selecionada e suas medições descritivas calculadas sobre as amostras feitas, é possível extrair algumas características das variáveis de interesse analisadas. As variáveis quantitativas, por exemplo, tem uma distribuição perceptivelmente semelhante à normal, com uma concentração de valores em torno da média e gradativa redução em direção às extremidades, porém cada uma possui suas particularidades.

O Conceito Enade, com média 2,4753 e mediana 2,4753, apresenta certa inclinação a ter valores na parte inferior do intervalo (abaixo de 2,5), com seu desvio padrão de 0,9464 indicando a presença de valores afastados da média influenciando no seu cálculo. O IDD, por sua vez, tem um comportamento semelhante, porém invertido, com média 2.6829, mediana 2.661 e desvio padrão 1.0202 indicando a sua inclinação à parte superior do intervalo.

O CPC possui uma leve inclinação para a parte de baixo do intervalo (menor que 3), com média 2.8335 e mediana 2.8249, e também concentra seus valores mais ao redor da média, como indicado por seu desvio padrão de 0.6381, indicando menores valores nas extremidades. O IGC apresenta um comportamento muito semelhante, com média 2.6834 e mediana 2.6845, com desvio padrão ainda menor de 0.5985.

Essas inclinações, porém, são muito leves para terem uma influência significativa nos intervalos de confiança de $\gamma = 95\%$ para a média dessas variáveis, sendo que os de Conceito Enade (2.405, 2.5456) e IDD (2.6076, 2.7582) compartilham uma extensão semelhante aos de CPC (2.7727, 2.8942) e IGC (2.5908, 2.776), mostrando sua semelhança em aproximar o valor estimado a partir de uma amostra de tamanho particular para cada variável.

Já ao observarmos as variáveis qualitativas, a modalidade de ensino presencial domina fortemente sobre a educação a distância, com cerca de 77% dos cursos. Enquanto isso, as categorias administrativas públicas concentram por volta de 5.6% das instituições de ensino. Nessa variáveis é possível perceber que os intervalos de confiança de 95% para estimar suas proporções são mais extensos que os das quantitativas, sendo (72.5957, 82.071) para modalidade de ensino Presencial e (0.834, 10.4019) para categorias administrativas Públicas.

3.2 Testes de Hipótese e Resultados

3.2.1 Hipótese sobre a Média

O objetivo da análise foi comparar a média do Conceito Preliminar de Curso (CPC) entre cursos com Ensino Presencial e cursos com Ensino a Distância (EaD) em Santa Catarina. A variável utilizada foi o CPC contínuo, enquanto a modalidade de ensino identificou o tipo de curso. As hipóteses formuladas foram:

- Hipótese nula (H_0): A média do CPC em cursos com ensino presencial em Santa Catarina é igual a média do CPC em cursos com ensino a distância (i.e., $\mu_{\text{Presencial}} = \mu_{\text{EaD}}$).
- Hipótese alternativa (H_1): A média do CPC em cursos com ensino presencial em Santa Catarina é maior que a média do CPC em cursos com ensino a distância (i.e., $\mu_{\text{Presencial}} > \mu_{\text{EaD}}$).

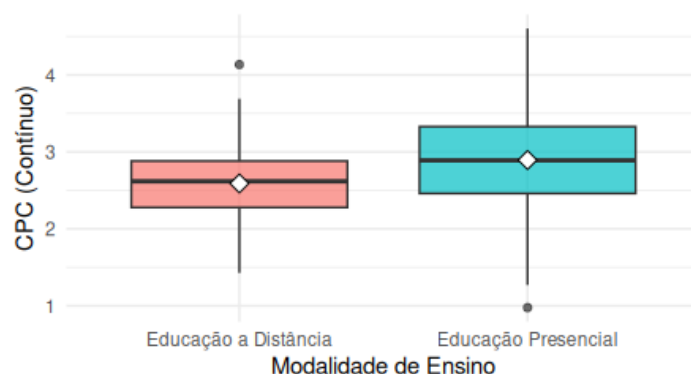
Para a análise, foi selecionada uma amostra aleatória de $n = 426$ cursos pertencentes às duas modalidades de interesse. Em seguida, foram extraídos os valores do CPC contínuo para cada

grupo, e aplicou-se o teste t de Welch, que não pressupõe igualdade de variâncias entre as amostras. Para isso, utilizou-se a função `t.test` do R, com o parâmetro `var.equal = FALSE`, nível de confiança de 95% (`conf.level = 0.95`) – assumindo um nível de significância de 5% – e `alternative = "greater"`, aplicando assim um teste unilateral à direita, pois o interesse era verificar se as médias dos cursos presenciais eram significativamente maiores. O *script* em R utilizado pode ser consultado no Anexo 2.

Os resultados do teste indicaram $t = 4.6738$, graus de liberdade aproximados $df = 162.24$ e p-valor $= 3,09 \times 10^{-6}$. As médias amostrais foram 2,8942 para os cursos presenciais e 2,5898 para os cursos a distância. O intervalo de confiança unilateral de 95% para a diferença das médias foi $[0,1967, +\infty)$, o que significa que, com 95% de confiança, a média dos cursos presenciais excede a média dos cursos EaD em pelo menos aproximadamente 0,197 pontos. O gráfico da Figura ?? ilustra essa diferença, mostrando uma distribuição de CPCs mais elevada entre os cursos presenciais.

Diante desses resultados, **rejeita-se a hipótese nula**, pois o p-valor é muito inferior ao nível de significância de 5%, indicando que a diferença observada não é resultado do acaso. Conclui-se, portanto, que os cursos com ensino presencial apresentam CPC médio significativamente superior aos cursos com ensino a distância em Santa Catarina. Esse resultado sugere um desempenho médio mais elevado dos cursos presenciais segundo o indicador CPC.

Figura 1: Distribuição do Conceito Preliminar de Curso (CPC) por Modalidade de Ensino.



3.2.2 Hipótese sobre a Proporção

O objetivo desta análise foi verificar se a proporção de instituições federais em Santa Catarina é diferente de 25%, buscando contrariar a hipótese nula. A variável analisada foi a *categoria administrativa*, sendo considerado “sucesso” cada instituição classificada como Pública Federal. As hipóteses formuladas foram:

- Hipótese nula (H_0): A proporção de instituições federais em Santa Catarina é igual a 25% (i.e., $p = 0,25$).
- Hipótese alternativa (H_1): A proporção de instituições federais em Santa Catarina é diferente de 25% (i.e., $p \neq 0,25$).

Foi selecionada uma amostra aleatória simples de $n = 150$ instituições, das quais 8 foram identificadas como federais, resultando em uma proporção amostral de $\hat{p} = 0,0533$. Aplicou-se o teste de proporção única (`prop.test` no R) sem correção de continuidade e com nível de confiança de 95%.

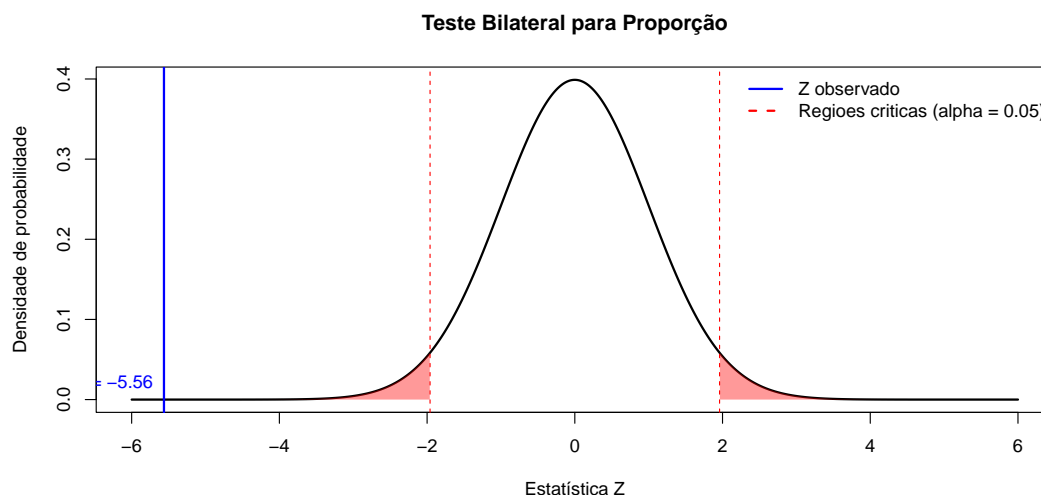
O resultado indicou uma estatística qui-quadrado de $\chi^2 = 30,942$, equivalente a uma estatística padronizada $z = -5,56$, com p -valor de aproximadamente $2,66 \times 10^{-8}$. O intervalo de confiança de 95% para a proporção verdadeira foi estimado em $[0,027; 0,102]$.

Como o valor calculado de $z = -5,56$ está muito além do limite crítico bilateral de $z_{\text{crítico}} = \pm 1,96$, o ponto amostral situa-se claramente dentro da região de rejeição de H_0 . Essa relação pode ser visualizada na Figura ??, que mostra a distribuição normal padrão utilizada no teste, destacando as regiões críticas em vermelho e o valor observado de z em azul.

Dado que o p -valor é muito inferior ao nível de significância de 5%, **rejeita-se a hipótese nula**, concluindo que há evidências estatísticas de que a proporção de instituições federais em Santa Catarina é significativamente diferente de 25%. Observa-se, contudo, que a proporção observada (5,3%) é consideravelmente menor do que o valor hipotetizado, indicando uma sub-representação de instituições federais no estado. Em termos absolutos, a diferença entre a proporção observada e a esperada é de aproximadamente 0,20, ou seja, 19,7 pontos percentuais a menos.

Do ponto de vista prático, esse resultado sugere que a participação das instituições federais em Santa Catarina é inferior ao quarto esperado, o que pode refletir características estruturais do sistema de ensino superior estadual, marcado por forte presença de instituições privadas. Apesar da robustez estatística do teste, é importante considerar que o número de instituições federais observadas (apenas oito) é pequeno, o que pode reduzir a precisão das estimativas. Ainda assim, o intervalo de confiança exclui claramente o valor de 25%, reforçando a conclusão de que a proporção real é significativamente menor do que a hipotetizada.

Figura 2: Distribuição normal padrão do teste bilateral para a proporção de instituições federais ($p = 0,25$), destacando as regiões críticas ($\alpha = 0,05$) e o valor observado de $z = -5,56$.



3.2.3 Hipótese sobre a Correlação

- Hipótese nula (H_0): Não há correlação entre o IDD e o Conceito Enade
- Hipótese Alternativa (H_1): Existe correlação entre IDD e Conceito Enade

Para testar a Hipótese, primeiro foram tomadas as tabelas de IDD e CPC filtradas para o estado de Santa Catarina e criadas amostras independentes de $n_{\text{semente}} = 123$ e $n_{\text{CE}} = 698$ e $n_{\text{IDD}} = 706$.

As entradas foram associadas por meio do código da instituição, código do curso e ano em uma única tabela. Observou-se a presença de valores nulos, o que precisaram ser descartados já que não poderiam ser trabalhados no teste. Ao final, obteve-se uma tabela resultante dos valores de IDD e Conceito Enade relacionados para cada entrada, com um total de 388 entradas.

Realizando o teste de correlação de Pearson em R considerando 5% de significância (α), obteve-se os seguintes resultados: O Coeficiente de Correlação de Pearson (r) = 0.6945254 e o p-valor $< 2.2 \times 10^{-16}$, com grau de liberdade de 386, valor de t calculado igual a 18.966 e intervalo de confiança de 95% igual a [0.6391626, 0.7427272].

Observou-se uma forte correlação positiva do coeficiente de correlação entre IDD e Conceito Enade ao estar próximo de $r = 1$, além de estar dentro do intervalo de confiança calculado. O p-valor se mostrou muito inferior ao nível de significância adotado (0.05), o que leva a rejeição da hipótese nula H_0 , ou seja, existe sim uma correlação significativa entre os valores dos indicadores de IDD e o Conceito Enade.

3.2.4 Hipótese sobre a Regressão

- Hipótese nula (H_0): O IDD não tem influência significativa sobre o Conceito Enade
- Hipótese Alternativa (H_1): O IDD tem influência significativa sobre o Conceito Enade

Para testar a hipótese, deve-se primeiro definir o modelo. O modelo de regressão linear ajustado é:

$$\text{Conceito Enade} = \beta_0 + \beta_1 \times (\text{IDD}) + \varepsilon$$

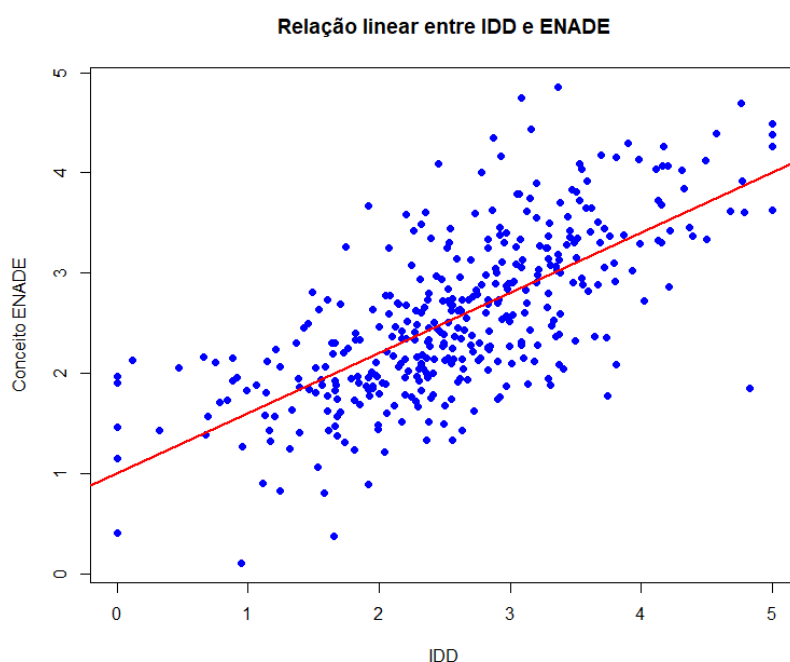
A hipótese nula será aceita caso o valor de β_1 seja igual a 0, ou seja, o valor do IDD não tem influência sobre o valor do Conceito Enade.

Utilizando o R, analisa-se a regressão e calcula-se os parâmetros utilizando os dados presentes na tabela resultante das amostras de IDD e Conceito Enade. Os resultados obtidos foram: intercepto ($\beta_0 = 1,00395$) e coeficiente angular ($\beta_1 = 0,59984$), erro padrão residual de 0,5829 e *p-valor* $< 2,2 \times 10^{-16}$, com o coeficiente de determinação ajustado (R^2) igual a 0,481. Na Figura ??, observa-se o gráfico da relação entre as amostras de IDD e Conceito Enade.

Percebe-se que o coeficiente angular estimado é de 0.59984, valor significativamente diferente de 0 como propõe a hipótese, além disso, o p-valor calculado se manteve abaixo do nível de significância de 5%, portanto pode-se rejeitar a hipótese nula H_0 , concluindo que existe sim a influência de IDD sobre o valor do Conceito Enade.

Além disso, pode-se também tirar conclusões adicionais quanto ao coeficiente de determinação ajustado (0.481), indicando que aproximadamente 48,1% da variação no valor do Conceito Enade é explicada pelo IDD, sugerindo que cursos de valores maiores de IDD tendem a apresentar melhores valores do Conceito Enade.

Figura 3: Relação de regressão linear entre IDD e Conceito Enade.



4 Considerações Finais

Este trabalho avançou na análise exploratória dos dados do ENADE de Santa Catarina, aplicando métodos de inferência estatística para testar hipóteses formuladas sobre os indicadores de qualidade. O objetivo central foi validar suposições sobre a média do Conceito Preliminar de Curso (CPC) entre modalidades, a proporção de instituições federais e a relação de dependência entre o IDD e o Conceito Enade.

A aplicação dos testes revelou um cenário de conclusões distintas. Para a **Hipótese 1**, encontrou-se uma diferença estatisticamente robusta ($p\text{-valor} \approx 3,09 \times 10^{-6}$), confirmando que os cursos presenciais possuem, em média, um CPC superior aos da modalidade EaD em Santa Catarina. Em contrapartida, a **Hipótese 2** não pôde ser corroborada; o teste não forneceu evidências para rejeitar a hipótese nula, indicando que a proporção de instituições federais não é significativamente maior que os 25% hipotetizados.

As análises de correlação e regressão (**Hipóteses 3 e 4**) foram particularmente elucidativas. Confirmou-se uma forte correlação positiva ($r \approx 0,694$) e uma influência estatisticamente significativa do IDD sobre o Conceito Enade. No entanto, é crucial ponderar a magnitude desse efeito. O coeficiente de determinação ($R^2 \approx 0,481$) demonstra que, embora o IDD seja um preditor relevante, ele explica menos da metade da variabilidade no Conceito Enade. Isso sugere que a significância estatística não se traduz, neste caso, em uma influência prática dominante.

A constatação de que 51,9% da variação do Conceito Enade não é explicada pelo IDD aponta para a principal limitação deste modelo de regressão. Fica evidente a necessidade de investigações futuras que incluam outras variáveis (como as que compõem o próprio CPC: infraestrutura, qualificação docente ou a percepção discente) para construir um modelo preditivo mais completo.

Em suma, este estudo reforça a natureza probabilística da inferência estatística. A rejeição de uma hipótese nula não representa uma prova irrefutável, mas sim uma forte evidência contra H_0 ,

baseada no nível de significância adotado de 5%. Os resultados aqui obtidos, portanto, devem ser interpretados como indicadores de tendências fortes, e não como regras determinísticas, sublinhando a importância de não equiparar significância estatística com causalidade ou relevância prática absoluta.

Referências

INDICADORES DA EDUCAÇÃO SUPERIOR (CGGI/DAES/INEP), COORDENAÇÃO-GERAL DE. **Nota Técnica CEI/CGGI/DAES nº 19/2024 – Metodologia utilizada no cálculo do CPC referente ao ano de 2023.** [S.l.], 2024. Acessado em: 18 set. 2025. Disponível em: <https://download.inep.gov.br/educacao_superior/enade/notas_tecnicas/2023/nota_tecnica_n_19_2024_cei_cggi_daes_inep_metodologia_utilizada_no_calculo_do_cpc_referente_ao_ano_de_2023.pdf>.

_____. **Nota Técnica CEI/CGGI/DAES nº 6/2024 – Metodologia utilizada no cálculo do Conceito Enade referente ao ano de 2023.** [S.l.], 2024. Acessado em: 18 set. 2025. Disponível em: <https://download.inep.gov.br/educacao_superior/enade/notas_tecnicas/2024/nota_tecnica_6_metodologia_calculo_conceito_enade_2023.pdf>.

SECRETARIA ESPECIAL DE AVALIAÇÃO INSTITUCIONAL. **CONCEITO ENADE (CE) E CONCEITO PRELIMINAR DE CURSO (CPC) DOS CURSOS DE GRADUAÇÃO DO INMA.** [S.l.]: Universidade Federal de Mato Grosso do Sul (UFMS), 2020. Acessado em: 30 ago. 2025. Disponível em: <<https://seavi.ufms.br/files/2020/05/Relat%C3%B3rio-1-INMA-2020.pdf>>.

5 Anexos - Códigos Fonte

Todos os códigos, bases de dados, gráficos e tabelas podem ser observados em detalhes no repositório: <https://github.com/leonardosm14/ENADE-Estatisticas>

Listing 5.1: Anexo A: Amostragem

```
# --- Diretório ---
setwd("~/Documentos/UFSC/ENADE-Estatisticas")

# Carrega os dados filtrados de Santa Catarina
source(file = "src/script_sc.r")

# --- Parâmetros fixos padrão ---
z <- 1.96 # nível de confiança de 95%
d <- 0.05 # erro amostral tolerável (5%)

# --- Variáveis Quantitativas ---

# --- Funções ---
n0 <- function(variancia) {
  return((z^2 * variancia) / (d^2))
}

n <- function(n0, N) {
  return(n0 / (1 + (n0 / N)))
}

# --- 1. Conceito ENADE (Contínuo) ---
variancia_enade <- var(data_CPC_SC$conceito_enade_.continuo., na.rm = TRUE)
N_enade <- nrow(data_CPC_SC)

n0_enade <- n0(variancia_enade)
n_enade <- n(n0_enade, N_enade)

cat("\n--- Conceito ENADE (Contínuo) ---\n")
cat("Variância:", round(variancia_enade, 4), "\n")
cat("N:", N_enade, "\n")
cat("n0:", round(n0_enade, 2), "\n")
cat("n:", ceiling(n_enade), "\n")

# --- 2. IDD (Contínuo) ---
variancia_idd <- var(data_IDD_SC$idd_.continuo., na.rm = TRUE)
N_idd <- nrow(data_IDD_SC)
```

```

n0_idd <- n0(variancia_idd)
n_idd <- n(n0_idd, N_idd)

cat("\n--- IDD (Contínuo) ---\n")
cat("Variância:", round(variancia_idd, 4), "\n")
cat("N:", N_idd, "\n")
cat("n0:", round(n0_idd, 2), "\n")
cat("n:", ceiling(n_idd), "\n")

# --- 3. CPC (Contínuo) ---
variancia_cpc <- var(data_CPC_SC$cpc_.continuo., na.rm = TRUE)
N_cpc <- nrow(data_CPC_SC)

n0_cpc <- n0(variancia_cpc)
n_cpc <- n(n0_cpc, N_cpc)

cat("\n--- CPC (Contínuo) ---\n")
cat("Variância:", round(variancia_cpc, 4), "\n")
cat("N:", N_cpc, "\n")
cat("n0:", round(n0_cpc, 2), "\n")
cat("n:", ceiling(n_cpc), "\n")

# --- 4. IGC (Contínuo) ---
variancia_igc <- var(data_IGC_SC$igc_.continuo., na.rm = TRUE)
N_igc <- nrow(data_IGC_SC)

n0_igc <- n0(variancia_igc)
n_igc <- n(n0_igc, N_igc)

cat("\n--- IGC (Contínuo) ---\n")
cat("Variância:", round(variancia_igc, 4), "\n")
cat("N:", N_igc, "\n")
cat("n0:", round(n0_igc, 2), "\n")
cat("n:", ceiling(n_igc), "\n")

# --- Variáveis Qualitativas ---

p <- 0.5

# n0 é comum a todos

n0 <- z^2 * p * (1-p) / d^2

```

```

# --- 1. Categoria Administrativa ---
N_cat <- nrow(data_IGC_SC)
n_cat <- n(n0, N_cat)

cat("\n--- Categoria Administrativa ---\n")
cat("N:", N_cat, "\n")
cat("n:", ceiling(n_cat), "\n")

# --- 2. Modalidade de Ensino ---

N_mod <- nrow(data_CPC_SC) # total de cursos (cada curso tem uma modalidade)
n_mod <- n(n0, N_mod)

cat("\n--- Modalidade de Ensino ---\n")
cat("N:", N_mod, "\n")
cat("n:", ceiling(n_mod), "\n")

```

Listing 5.2: Anexo B: Análise Descritiva

```

# --- Pacotes ---
library("dplyr")

# Carrega os dados filtrados de Santa Catarina
source(file = "script_sc.r")

# --- Filtrando os dados ---

# ENADE
dados_enade <- data_CPC_SC %>%
  select(conceito_enade_.continuo.) %>%
  filter(!is.na(conceito_enade_.continuo.))

# CPC
dados_cpc <- data_CPC_SC %>%
  select(cpc_.continuo.) %>%
  filter(!is.na(cpc_.continuo.))

# IDD
dados_idd <- data_IDD_SC %>%
  select(idd_.continuo.) %>%
  filter(!is.na(idd_.continuo.))

```

```

# IGC
dados_igc <- data_IGC_SC %>%
  select(igc_.continuo.) %>%
  filter(!is.na(igc_.continuo.))

# Modalidade de Ensino
dados_me <- data_CPC_SC %>%
  select(modalidade_de_ensino) %>%
  filter(!is.na(modalidade_de_ensino))

# Categoria Administrativa
dados_ca <- data_IGC_SC %>%
  select(categoria_administrativa) %>%
  filter(!is.na(categoria_administrativa))

# --- Criar amostra aleatória de tamanho n = 426, conforme "amostragem.r" ---

# ENADE
set.seed(123) # para reprodutibilidade
amostra_enade <- dados_enade %>%
  sample_n(size = 699, replace = FALSE)

#CPC
set.seed(123) # para reprodutibilidade
amostra_cpc <- dados_cpc %>%
  sample_n(size = 426, replace = FALSE)

# IDD
set.seed(123) # para reprodutibilidade
amostra_idd <- dados_idd %>%
  sample_n(size = 707, replace = FALSE)

# IGC
set.seed(123) # para reprodutibilidade
amostra_igc <- dados_igc %>%
  sample_n(size = 163, replace = FALSE)

# Modalidade de Enino
set.seed(123) # para reprodutibilidade
amostra_me <- dados_me %>%
  sample_n(size = 300, replace = FALSE)

# Categoria Administrativa

```

```

set.seed(123) # para reprodutibilidade
amostra_ca <- dados_ca %>%
  sample_n(size = 89, replace = FALSE)

# --- Estatísticas Descritivas ---

# ENADE
media_enade <- mean(amostra_enade$conceito_enade_.continuo.)
mediana_enade <- median(amostra_enade$conceito_enade_.continuo.)
dp_enade <- sd(amostra_enade$conceito_enade_.continuo.)
erro_enade = qt((1+0.95)/2, 699-1) * (dp_enade/(sqrt(699)))

cat("Média Conceito Enade:", round(media_enade, 4), "\n")
cat("Mediana Conceito Enade:", round(mediana_enade, 4), "\n")
cat("Desvio Padrão Conceito Enade:", round(dp_enade, 4), "\n")
cat("Intervalo de Confiança da Média do Conceito Enade:", round(media_enade-
  erro_enade, 4), "-", round(media_enade+erro_enade, 4), "\n")

# CPC
media_cpc <- mean(amostra_cpc$cpc_.continuo.)
mediana_cpc <- median(amostra_cpc$cpc_.continuo.)
dp_cpc <- sd(amostra_cpc$cpc_.continuo.)
erro_cpc = qt((1+0.95)/2, 426-1) * (dp_cpc/(sqrt(426)))

cat("Média CPC:", round(media_cpc, 4), "\n")
cat("Mediana CPC:", round(mediana_cpc, 4), "\n")
cat("Desvio Padrão CPC:", round(dp_cpc, 4), "\n")
cat("Intervalo de Confiança da Média do CPC:", round(media_cpc-erro_cpc, 4),
  "-", round(media_cpc+erro_cpc, 4), "\n")

# IDD
media_idd <- mean(amostra_idd$idd_.continuo.)
mediana_idd <- median(amostra_idd$idd_.continuo.)
dp_idd <- sd(amostra_idd$idd_.continuo.)
erro_idd = qt((1+0.95)/2, 707-1) * (dp_idd/(sqrt(707)))

cat("Média IDD:", round(media_idd, 4), "\n")
cat("Mediana IDD:", round(mediana_idd, 4), "\n")
cat("Desvio Padrão IDD:", round(dp_idd, 4), "\n")
cat("Intervalo de Confiança da Média do IDD:", round(media_idd-erro_idd, 4),
  "-", round(media_idd+erro_idd, 4), "\n")

# IGC

```



```

media_igc <- mean(amostra_igc$igc_.continuo.)
mediana_igc <- median(amostra_igc$igc_.continuo.)
dp_igc <- sd(amostra_igc$igc_.continuo.)
erro_igc = qt((1+0.95)/2, 163-1) * (dp_igc/(sqrt(163)))

cat("Média IGC:", round(media_igc, 4), "\n")
cat("Mediana IGC:", round(mediana_igc, 4), "\n")
cat("Desvio Padrão IGC:", round(dp_igc, 4), "\n")
cat("Intervalo de Confiança da Média do IGC:", round(media_igc-erro_igc, 4),
    "-", round(media_igc+erro_igc, 4), "\n")

# Modalidade de ensino
presencial <- sum(amostra_me$modalidade_de_ensino == "Educação Presencial")
total <- nrow(amostra_me)
prop_me <- (presencial / total)
erro_me <- qnorm((1+0.95)/2)*(sqrt(prop_me*(1-prop_me)/300))

cat("Proporção de Cursos Presenciais:", round(prop_me*100, 4), "\n")
cat("Intervalo de Confiança da Proporção de Cursos Presenciais:", round((
    prop_me-erro_me)*100, 4), "-", round((prop_me+erro_me)*100, 4), "\n")

# Categoria Adiministrativa
publicas <- sum(amostra_ca$categoria_administrativa == "Pública Federal",
    amostra_ca$categoria_administrativa == "Pública Estadual",
    amostra_ca$categoria_administrativa == "Pública Munucipal")
total <- nrow(amostra_ca)
prop_ca <- (publicas / total)
erro_ca <- qnorm((1+0.95)/2)*(sqrt(prop_ca*(1-prop_ca)/89))

cat("Proporção de Instituições Públicas:", round(prop_ca*100, 4), "\n")
cat("Intervalo de Confiança da Proporção de Instituições Públicas:", round((
    prop_ca-erro_ca)*100, 4), "-", round((prop_ca+erro_ca)*100, 4), "\n")

```

Listing 5.3: Anexo C: Hipótese sobre a Média

```

setwd("~/Documents/ENADE-Estatisticas")

# --- Pacotes ---
library(dplyr)
library(ggplot2)

# --- Filtrando os dados ---
dados_modalidade <- data_CPC_SC %>%

```

```

filter(modalidade_de_ensino %in% c("Educação Presencial", "Educação a Distâ
ncia")) %>%
select(modalidade_de_ensino, cpc_.continuo.) %>%
filter(!is.na(cpc_.continuo.))

# --- Criar amostra aleatória de tamanho n = 426, conforme "amostragem.r" ---
set.seed(123) # para reprodutibilidade
dados_amostra <- dados_modalidade %>%
  sample_n(size = 426, replace = FALSE)

# --- Separando os grupos ---
presencial <- dados_amostra %>%
  filter(modalidade_de_ensino == "Educação Presencial") %>%
  pull(cpc_.continuo.)

ead <- dados_amostra %>%
  filter(modalidade_de_ensino == "Educação a Distância") %>%
  pull(cpc_.continuo.)

# --- Teste t de Welch ---
teste_t <- t.test(presencial, ead,
                  alternative = "greater", # teste unilateral à direita
                  var.equal = FALSE) # Welch (variâncias diferentes)

# --- Exibir resultado ---
print(teste_t)

# --- Boxplot comparativo ---
dev.new()
pdf("boxplot_modalidade.pdf", width = 7, height = 5)
print(
  ggplot(dados_amostra, aes(x = modalidade_de_ensino, y = cpc_.continuo., fill =
    modalidade_de_ensino)) +
  geom_boxplot(alpha = 0.7) +
  stat_summary(fun = mean, geom = "point", shape = 23, size = 3, fill = "white
    ") +
  labs(x = "Modalidade de Ensino",
       y = "CPC (Contínuo)") +
  theme_minimal() +
  theme(legend.position = "none")
)
dev.off()

```

Listing 5.4: Anexo D: Hipótese sobre a Proporção

```
setwd("~/Documents/ENADE-Estatisticas")
source(file = "src/script_sc.r")

# --- Pacotes ---
library(dplyr)

# --- Amostra aleatória de instituições ---
set.seed(123)
amostra_n <- 150
amostra_instituicoes <- data_IGC_SC %>%
  sample_n(amostra_n)

# --- Contagem de instituições federais ---
sucesso <- sum(amostra_instituicoes$categoria_administrativa == "Pública
  Federal")
total <- nrow(amostra_instituicoes)
prop_observada <- sucesso / total

cat("Tamanho da amostra:", total, "\n")
cat("Número de instituições federais na amostra:", sucesso, "\n")
cat("Proporção observada:", round(prop_observada, 3), "\n\n")

# --- Teste de hipótese para proporção única (bilateral) ---
# H0: p = 0.25
# H1: p 0.25
teste_prop <- prop.test(x = sucesso, n = total, p = 0.25, alternative = "two.
  sided", correct = FALSE)
print(teste_prop)

# --- Interpretação ---
cat("\nInterpretação:\n")
if (teste_prop$p.value < 0.05) {
  cat("Rejeita-se H0: A proporção de instituições federais ésignificativamente
    diferente de 25%.\n")
} else {
  cat("Não se rejeita H0: Não há evidências de que a proporção de instituições
    federais seja diferente de 25%.\n")
}

# --- Gráfico do teste bilateral e salvamento em PDF ---
p0 <- 0.25
phat <- prop_observada
```

```

n <- total
z <- (phat - p0) / sqrt(p0 * (1 - p0) / n)
print(z)
alpha <- 0.05
zcrit <- qnorm(1 - alpha / 2)
print(zcrit)

# Caminho de saída
pdf("teste_bilateral.pdf", width = 10, height = 5)

# Dados da distribuição normal padrão
x <- seq(-6, 6, length = 1000)
y <- dnorm(x)

# Plot principal
plot(x, y, type = "l", lwd = 2, col = "black",
     main = "Teste Bilateral para Proporção",
     xlab = "Estatística Z", ylab = "Densidade de probabilidade")

# Regiões críticas
polygon(c(x[x >= zcrit], rev(x[x >= zcrit])),
       c(y[x >= zcrit], rep(0, sum(x >= zcrit))),
       col = rgb(1, 0, 0, 0.4), border = NA)
polygon(c(x[x <= -zcrit], rev(x[x <= -zcrit])),
       c(y[x <= -zcrit], rep(0, sum(x <= -zcrit))),
       col = rgb(1, 0, 0, 0.4), border = NA)

# Linha do valor observado
abline(v = z, col = "blue", lwd = 2)
abline(v = c(-zcrit, zcrit), col = "red", lty = 2)

# Legenda e texto
legend("topright",
     legend = c("Z observado", "Regioes criticas (alpha = 0.05)"),
     col = c("blue", "red"), lwd = 2, lty = c(1, 2), bty = "n")

text(z, dnorm(z) + 0.02, labels = sprintf("z = %.2f", z),
     col = "blue", pos = ifelse(z < 0, 2, 4))

dev.off()

```

Listing 5.5: Anexo E: Hipótese sobre a Correlação e Regressão

```
# Diretório da base de dados - Talvez precisa alterar, dependendo de onde o
# repositório estiver clonado.
setwd("~/Documents/ENADE-Estatisticas")

# ---- Filtragem das tabelas ----

# Retira eventuais entradas nulas
data_CPC_SC_limpo <- na.omit(data_CPC_SC[, c("ano", "codigo_da_ies", "
      nome_da_ies", "codigo_do_curso", "area_de_avaliacao", "conceito_enade_.
      continuo.")])
data_IDD_SC_limpo <- na.omit(data_IDD_SC[, c("ano", "codigo_da_ies", "
      nome_da_ies", "codigo_do_curso", "area_de_avaliacao", "idd_.continuo.")])

# Faz a amostragem de acordo com os valores de n calculados em "amostragem.r"
# Note que as amostras são independentes
set.seed(123)
amostra_CPC <- data_CPC_SC_limpo[sample(1:nrow(data_CPC_SC_limpo), 699, replace
      = FALSE), ]
rownames(amostra_CPC) <- NULL
amostra_IDD <- data_IDD_SC_limpo[sample(1:nrow(data_IDD_SC_limpo), 707, replace
      = FALSE), ]
rownames(amostra_IDD) <- NULL

# Faz o merge das tabelas pelas colunas-chave
data_merged <- merge(
  x = amostra_CPC,
  y = amostra_IDD,
  by = c("ano", "codigo_da_ies", "codigo_do_curso"),
  suffixes = c("_CPC", "_IDD")
)

# Seleciona as colunas principais para análise
data_corre_reg <- data.frame(
  ano = data_merged$ano,
  codigo_da_ies = data_merged$codigo_da_ies,
  codigo_do_curso = data_merged$codigo_do_curso,
  nome_da_ies = data_merged$nome_da_ies_CPC,
  nome_do_curso = data_merged$area_de_avaliacao_CPC,
  conceito_enade = data_merged$conceito_enade_.continuo.,
  idd = data_merged$idd_.continuo.
)
```

```

# Remove linhas com valores faltantes (NA) em Enade ou IDD
data_corre_reg <- subset(data_corre_reg, !is.na(conceito_enade) & !is.na(idd))
rownames(data_corre_reg) <- NULL

# ---- Teste de hipótese sobre correlação ----

# Teste de correlação entre ENADE e IDD
cor.test(data_corre_reg$conceito_enade, data_corre_reg$idd, method = "pearson")

# ---- Teste de hipótese sobre regressão ----

# Modelo de regressão linear (ENADE dependente de IDD)
modelo <- lm(conceito_enade ~ idd, data = data_corre_reg)

# Sumário com teste t dos coeficientes
summary(modelo)

# Gráfico da Relação ENADE e IDD contínuos
dev.new()
plot(data_corre_reg$idd, data_corre_reg$conceito_enade,
      xlab = "IDD", ylab = "Conceito ENADE",
      main = "Relação linear entre IDD e ENADE",
      pch = 19, col = "blue")

#Plot da Linha de regressão
abline(modelo, col = "red", lwd = 2)

```