Universidade de Brasília
Departamento de Estatística

Análise do desempenho acadêmico nas disciplinas de Probabilidade e Estatística e Bioestatística do Departamento de Estatística

Ananda Almeida de Sá

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Ananda Almeida de Sá

Análise do desempenho acadêmico nas disciplinas de Probabilidade e Estatística e Bioestatística do Departamento de Estatística

Orientadora: Maria Teresa Leão Costa

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Resumo

Atualmente, a Estatística tem se tornado cada vez mais relevante no processo decisório em diversas áreas de atuação, impulsionada pelo aprimoramento dos resultados obtidos por meio dos modelos estatísticos. Sendo assim, em vários cursos de graduação estão incluídas uma disciplina de Estatística nos seus currículos. Nesse contexto, esse estudo tem como objetivo identificar os fatores que influenciam o desempenho acadêmico dos estudantes nas disciplinas de Probabilidade e Estatística e Bioestatística oferecidas para diversos cursos de graduação da Universidade de Brasília. Os dados foram obtidos dos bancos de dados dos sistemas acadêmicos da UnB (SIGAA e SIGRA) referentes ao período de 1994 a 2019. Para a modelagem dos dados foi utilizada a técnica de Regressão Logística Multinível, devido à natureza binária da variável resposta (aprovação ou reprovação) e à estrutura hierárquica dos dados, considerando alunos e turmas. A análise foi realizada no período de 2017 a 2019. Os resultados indicaram que fatores como a modalidade da disciplina, o tempo de permanência do aluno na universidade e a porcentagem de faltas estão associados ao desempenho dos alunos.

Palavras-chaves: Estatística, desempenho acadêmico, Probabilidade e Estatística, Bioestatística, Universidade de Brasília, Regressão Logística Multinível.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

Atualmente, no processo de tomada de decisão, independente da área de atuação, a Estatística vem se tornando uma ferramenta muito utilizada. Tal crescimento é justificado pelo aprimoramento dos resultados obtidos através dos modelos estatísticos. Para Ignácio (2010, p. 177), os equipamentos e softwares permitem a manipulação de grande quantidade de dados, o que veio a dinamizar o emprego dos métodos estatísticos.

Apesar da relevância da matéria para as áreas de conhecimento, no atual sistema de ensino, a Estatística aparece de forma discreta no ensino médio e obrigatória em apenas alguns cursos superiores (ARA, 2006). Para além disso, nem todos os cursos de graduação têm a disciplina como obrigatória e poucos são os cursos que ela é ofertada como uma optativa. Desse modo, a pouca familiaridade da disciplina tem como consequência a falta de motivação para a aprendizagem e um alto índice de reprovação.

"Considera-se como necessária a ampliação do conhecimento da universidade sobre si mesma e sobre seus estudantes, de forma a garantir o cumprimento adequado de suas funções científicas e sociais"(VENDRAMINI et al., 2004). Consequentemente, é de suma importância ao Departamento de Estatística da Universidade de Brasília (UnB) compreender o perfil dos estudantes e identificar quais são os fatores associados ao desempenho dos que cursam as matérias da Estatística.

O Departamento de Estatística, além de ofertar as disciplinas que comportam o fluxo da graduação de Estatística, ofertam matérias para os alunos dos outros departamentos da universidade, podendo ser obrigatória ou não para a formação. São elas: Estatística Aplicada, Probabilidade e Estatística; e Bioestatística. Para o estudo proposto, o desempenho será analisado para as duas últimas.

As duas disciplinas, tanto Probabilidade e Estatística quanto Bioestatística possuem carga horária de 60 horas e têm como pré-requisito Cálculo 1 ou Matemática 1. Por isso, o período letivo mínimo que o aluno precisa estar para cursá-las é o segundo. A primeira está inclusa no fluxo curricular como obrigatória para os cursos de Ciências da Computação, Ciências Contábeis, Ciências Econômicas, Engenharia de Produçao, Engenharia de Redes e Comunicação, Engenharia Elétrica e Engenharia Mecatrônica. Já Bioestatística, para os cursos de Engenharia Florestal e Agronomia.

Ademais, uma das formas de identificar a aprendizagem ou não do aluno é por meio da avaliação (BRUM; LISKA, 2020). Portanto, utilizando os dados disponibilizados tanto pelo Departamento de Estatística quanto pelo os bancos dos sistemas da univer-

sidade, Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA) e Sistema de Informações Acadêmicas (SIGRA), a análise do rendimento será realizada através dos resultados das avaliações dos alunos que cursaram as disciplinas de interesse.

Desta forma, o referido estudo tem como objetivo identificar as características dos estudantes, bem como gerar insumos para detectar quais são os fatores que estão associados ao desempenho acadêmico dos alunos que cursam as disciplinas, Probabilidade e Estatística; e Bioestatística, ofertadas pelo Departamento de Estatística na Universidade de Brasília (UnB).

A estrutura do referido estudo iniciou com a introdução, seguida pelo referencial teórico utilizado para elaboração do estudo. A metodologia detalhou todos os procedimentos de limpeza e tratamento de dados, assim como suas limitações. No capítulo referente à análise exploratória, todas as variáveis obtidas no banco de dados foram analisadas. A modelagem utilizada para o estudo foi a técnica de Regressão Logística Multinível, devido à natureza binária da variável resposta (aprovação ou reprovação) e à estrutura hierárquica dos dados, considerando alunos e turmas. Por fim, na conclusão, foram apresentados os fatores que influenciaram diretamente o desempenho dos alunos, que incluem a modalidade da disciplina, o tempo de permanência do aluno na universidade e o percentual de faltas.

2 Referencial Teórico

2.1 Regressão Logística

A regressão logística é uma técnica estatística que permite a predição dos valores determinados por uma variável de interesse de estudo, nomeada como reposta ou dependente, classificada como categórica ou qualitativa, binárias ou não, relacionadas a uma ou mais variáveis que afetariam a resposta, nomeadas como variáveis explicativas, independentes ou preditoras. Esse modelo de regressão logística é um caso específico dos modelos lineares generalizados, que são uma classe de modelos que generalizam a regressão linear clássica para lidar com diferentes tipos de variáveis de resposta.

A variável resposta é denominada como binária quando apresenta apenas dois resultados possíveis, em que a ocorrência de um determinado evento é definido como o "sucesso", caso o contrário, o "fracasso". Portanto, levando em consideração apenas um único evento observado, a variável resposta, Y_i, segue distribuição Bernoulli:

$$Y_i = \begin{cases} 1, \text{ se o evento ocorreu, ou seja, sucesso;} \\ 0, \text{ caso contrário, fracasso.} \end{cases}$$

A probabilidade de sucesso é denominada por π_i e ainda $0 \leq \pi_i \leq 1$. Já a probabilidade de fracasso é complementar a de sucesso e é dado por $1 - \pi_i$. Ou ainda, $P(Y_i = 1) = \pi_i$ e $P(Y_i = 0) = 1 - \pi_i$. Além disso, a média da distribuição é igual a $E(Y_i) = \pi_i$, consequentemente varia de 0 a 1 e sua variância é dada por $V(Y_i) = \pi_i(1-\pi_i)$.

2.2 Regressão Logística Binária

O modelo de regressão logística é representado pela seguinte função:

$$\pi_{(x_i)} = \frac{exp(\beta_0 + \beta_1 X_i)}{1 + exp(\beta_0 + \beta_1 X_i)},$$

em que, β_0 e β_1 são os parâmetros da regressão e X_i é a única variável explicativa, por ser uma regressão logística simples. E ainda,

$$\pi_{(x_i)} = E[Y_i]. \tag{2.2.1}$$

A curva que representa a função de regressão logística está descrita na Figura 1.

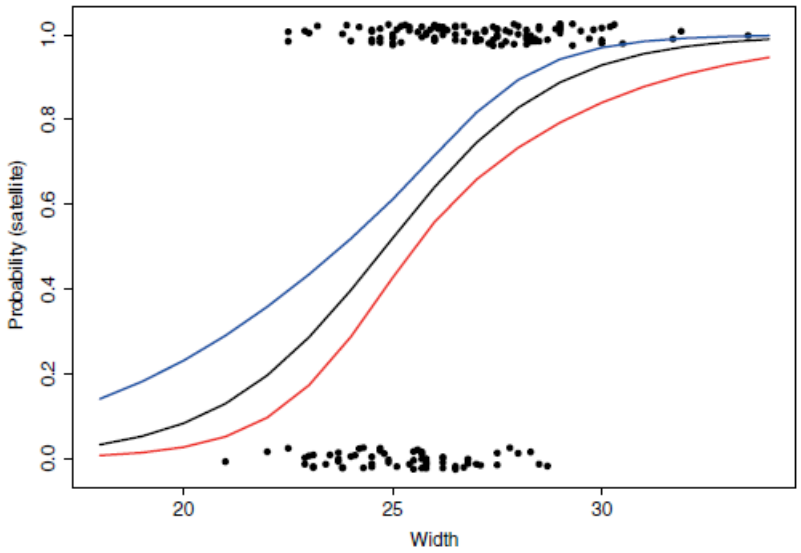


Figura 1: Exemplos da curva da função de regressão logística

Fonte: Agresti, An Introduction to Categorical Data Analysis (p. 97).

Ao aplicarmos a transformação denominada *logito*, representada por $\pi^*_{(x_i)}$, a função de regressão logística para $E[Y_i]$ pode ser linearizada. Essa transformação é o logatirimo natural da *odds*, que pode ser interpretada como a razão entre as chances de um evento ocorrer e as chances de ele não ocorrer.

$$\pi^*_{(x_i)} = ln(odds) = ln\left(\frac{\pi_{(x_i)}}{1-\pi_{(x_i)}}\right) = \beta_0 + \beta_1 X_i. \tag{2.2.2}$$

2.3 Estimação dos parâmetros do modelo de regressão

Para realizar o ajuste do modelo de regressão, faz-se necessário estimar os parâmetros β_0 e β_1. Esses estimadores serão obtidos através dos valores que maximizam a função de log-verossimilhança:

$$ln(L(\beta_0, \beta_1)) = \sum_{i=1}^{n} Y_i \left(\beta_0 + \beta_1 X_i\right) - \sum_{i=1}^{n} ln[1 + exp\left(\beta_0 + \beta_1 X_i\right)].$$

Não há uma forma algébrica fechada para determinar os estimadores que maximizam tal função. Portanto, através do auxílio de programas estatísticos, utilizando métodos iterativos como o de Newton-Raphson ou do Score, encontra-se uma solução para o sistema. Com isso, representados por b_0 e b_1 os estimadores de β_0 e β_1, respectivamente,

tem-se que:

$$\hat{\pi}_i = \frac{exp(b_0 + b_1 X_i)}{1 + exp(b_0 + b_1 X_i)}. \tag{2.3.1}$$

Levando em consideração uma variável aleatória determinada através das n_i repetições do valor de X_i, em que i está no intervalo de 1 a c, nesse caso deve-se considerar a seguinte função de log-verossimilhança:

$$ln(L(\beta_0, \beta_1)) = \sum_{j=1}^{c} \{\ln \binom{n_i}{Y_j} + Y_j(\beta_0 + \beta_1 X_j) - n_j ln[1 + exp(\beta_0 + \beta_1 X_j)]\}.$$

2.4 Interpretação do b_1

Considerando $x_i = x$, a partir da fórmula dada por 2.3.1, tem-se:

$$\hat{\pi} = \frac{exp(b_0 + b_1 x)}{1 + exp(b_0 + b_1 x)}.$$

sendo, a $odds_x$

$$\frac{\pi_i(x)}{1 - \pi_i(x)} = exp(b_0 + b_1 x) = e^{b_0}(e^{b_1})^x,$$

e a $odds_{x+1}$,

$$\frac{\pi(x+1)}{1 - \pi(x+1)} = exp(b_0 + b_1(x+1)) = e^{b_0}(e^{b_1})^x e^{b_1}.$$

A razão de chances entre $x + 1$ e x é dada por:

$$\frac{odds_{x+1}}{odds_x} = \frac{e^{b_0}(e^{b_1})^x e^{b_1}}{e^{b_0}(e^{b_1})^x} = e^{b_1},$$

Logo, aplicando a transformação de *logito*,

$$ln\left(\frac{odds_{x+1}}{odds_x}\right) = ln(e^{b_1}) = b_1.$$

Portanto, a chance estimada de sucesso (*odds*) para cada unidade que aumenta em x é multiplicada por e^{b_1}. Isto é, a chance estimada de sucesso no nível $x + 1$ é igual a no nível x multiplicada por e^{b_1}.

2.5 Regressão Logística Múltipla

Quando existe mais de uma variável explicativa, o modelo de regressão logística utilizado é representado pela extensão do modelo de regressão logística simples. Então, utilizando a função 2.2.1, tem-se:

$$\pi_i = \frac{exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})},\ i = 1, ...c.$$

E ainda, a tranformação de *logito* é análoga a apresentada a função 2.2.2:

$$\pi_i^* = ln(odds) = ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}.$$

A estimação dos parâmetros para a regressão múltipla é análoga a estimação na regressão simples. Porém, como há mais variáveis preditoras, o estimador dos b_j será dado pelo conjunto de todos os b_j até o $p - 1$. Portanto,

$$\boldsymbol{b_{p\times 1}} = \begin{bmatrix} b_0 \\ b_1 \\ : \\ b_{p-1} \end{bmatrix}.$$

2.6 Inferência sobre os parâmetros do modelo

Para grandes amostras, uma propriedade interessante dos estimadores de máxima verossimilhança é a distribuição assintótica para uma distribuição normal, que posteriormente pode convergir para uma distribuição normal padrão. Considerando Z uma variável aleatória que tem distribuição normal padrão, tem-se:

$$\frac{b_k - \beta_k}{s(b_k)} \sim Z,\ para\ k = 0, 1, ...p - 1,$$

em que $s(b_k)$ é o estimador de máxima verossimilhança do desvio padrão de b_k.

2.6.1 Teste de Wald

Para verificar se uma determinada variável preditora possui uma relação estatisticamente significativa em relação a variável resposta, aplica-se o teste de Wald. Logo,

considerando as seguintes hipóteses:

$$\begin{cases} H_0)\ \beta_k = 0, \\ H_1)\ \beta_k \neq 0. \end{cases}$$

A estatística do teste é dada por:

$$Z^* = \frac{b_k}{s(b_k)} \sim N(0,1).$$

O estatística do teste de Wald pode ser apresentada de uma outra forma, quando elevada ao quadrado e neste caso sua distribuição amostral será a distribuição χ^2 com 1 grau de liberdade.

2.6.2 Teste da Razão de Verossimilhança

Análogo ao Teste de Wald, o Teste da Razão de Verossimilhança é utilizado para verificar se um subconjunto de variáveis preditoras possuem uma relação estatisticamente significativa em relação a variável resposta. Logo, as hipóteses a serem consideradas para esse teste são:

$$\begin{cases} H_0)\ \beta_1 = \beta_2 = ... = \beta_m = 0, \\ H_1) \text{ Existe pelo menos um } \beta_k \text{ em } H_0 \text{ diferente de 0, em que } k = 1, ..., m. \end{cases}$$

A estatística do teste é dada por:

$$G^2 = -2(L_0 - L_1) \sim \chi^2, \tag{2.6.1}$$

em que, sob H_0, a estatística G^2 recebe uma distribuição amostral de χ^2 com m graus de liberdade. Os valores que são representados por L_0 e L_1 são os máximos das funções de log-verossimilhança do modelo reduzido, isto é, quando H_0 é verdadeira, e do modelo completo, isto é, quando H_1 é verdadeira, respectivamente.

2.7 Adequabilidade do modelo

A análise feita sobre o ajustamento do modelo de regressão logística é realizada através de testes de adequação em que as hipóteses testadas são:

$$\begin{cases} H_0) \text{ O modelo de Regressão Logística ajusta-se aos dados,} \\ H_1) \text{ O modelo não se ajusta.} \end{cases}$$

2.7.1 Teste de adequação qui-quadrado de Pearson

Assumindo que Y_{ij} recebe as observações independentes e repetidas de um ou mais níveis da variável explicativa, tem-se a estatística do teste:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=0}^{1} \frac{(f_{ji} - fe_{ji})^2}{fe_{ji}}, \tag{2.7.1}$$

em que, f_{ji} é o número de sucessos para n_i repetições do i-ésimo conjunto de valores das variáveis explicativas e fe_{ji} é o número de sucessos ajustado ou esperado obtido por $n_i * \hat{\pi}_i$, $\hat{\pi}_i$ é a probabilidade de sucesso predita no modelo ajustado.

Sob H_0, a estatística do teste tem distribuição aproximadamente qui-quadrado com $c-k$ graus de liberdade quando n grande e $k < c$, em que k é o número de parâmetros do modelo e c é o número de conjuntos de valores distintos das variáveis explicativas.

2.7.2 Estatística G^2 do Teste da Razão de Verossimilhança

Considerando a estatística do Teste da Razão de Verossimilhança, indicada pela Fórmula 2.6.1 e o modelo completo é determinado $E[Y_{ij}] = \pi_j$, tal que $j = 1, 2, ..., c$, e a probabilidade e estimativa de π_i são dadas $p_{ij} = \frac{Y_j}{n_j}$ e $\hat{\pi}_{ij}$, respectivamente. Com isso, a estatística do teste é dada por:

$$G^2 = -2 \sum_{i=1}^{c} \left[Y_j \; ln\left(\frac{\widehat{\pi}_i}{p_j}\right) + (n_j - Y_j) \; ln\left(\frac{1-\widehat{\pi}_i}{1-p_j}\right) \right],$$

sob H_0, a estatística do teste tem distribuição aproximadamente qui-quadrado com $c - k$ graus de liberdade quando n_j grande e $k < c$, em que k é o número de parâmetros do modelo e c é o número de conjuntos de valores distintos das variáveis explicativas.

2.7.3 Teste de Hosmer–Lemeshow

Para realizar o teste de Hosmer-Lemeshow, os valores obtidos pelas variáveis explicativas no modelo devem estar ordenados. Feito isso, são construídas classes com base nas probabilidades estimadas, ou nos percentis ou com base em valores fixados.

A estatística do teste será a mesma definida em 2.7.1, com mesma distribuição amostral de qui-quadrado, mas com o $c - 2$ graus de liberdade.

2.8 Seleção dos modelos

Após a verificação da adequabilidade do modelo, o próximo passo é definir qual deles é o que mais representa o fenômeno em análise. Essa verificação é feita utilizando critérios de seleção, são eles: critério de informação de Akaike (AIC_p), critério de informação Bayesiano (SBC_p) e o *pseudo* R^2.

2.8.1 Critério de informação de Akaike (AIC_p)

A estatística *deviance* é definida por $-2ln(L(\hat{\mathbf{b}}))$, o critério de informação de Akaike é dado por:

$$AIC_p = deviance + 2p = -2ln(L(\hat{\mathbf{b}})) + 2p,$$

quanto menor o valor de AIC_p, melhor é o modelo. A análise gráfica desse critério é feita da seguinte maneira:

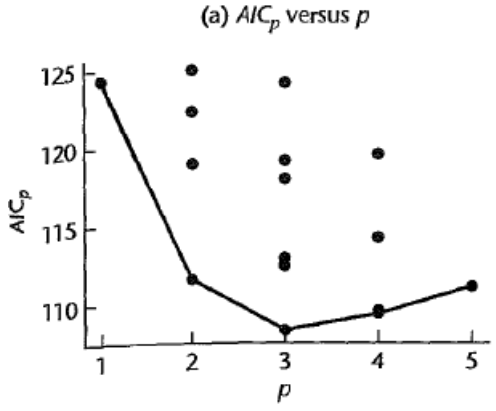


Figura 2: Exemplo do gráfico do AIC_p

Fonte: Neter et al. Applied Linear Statistical Models (p. 585).

De acordo com a figura 2, o critério de seleção de modelos determinou que os que apresentam três parâmetros são os mais adequados para o fenômeno do exemplo. Dentre esses, a seleção final é feita pelo o que apresenta o menor valor de AIC_p.

2.8.2 Critério de informação Bayesiano (SBC_p)

Também conhecido por BIC, esse critério é dado por:

$$SBC_p = deviance + p.ln(n) = -2ln(L(\hat{\mathbf{b}})) + p.ln(n),$$

quanto menor o valor de SBC_p, melhor é o modelo. A análise gráfica desse critério é feita da seguinte maneira:

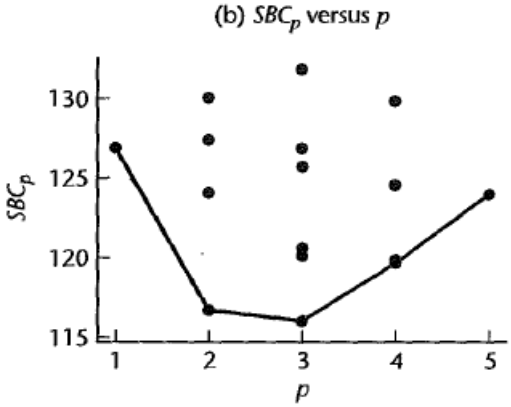


Figura 3: Exemplo do gráfico do SBC_p

Fonte: Neter et al. Applied Linear Statistical Models (p. 585).

De acordo com a figura 3, o critério de seleção de modelos determinou que os que apresentam três parâmetros são os mais adequados para o fenômeno do exemplo. Dentre esses, a seleção final é feita pelo o que apresenta o menor valor de SBC_p.

2.8.3 Pseudo $\mathbf{R}^2$

Análogo ao utilizado em regressão linear, o critério *pseudo* R^2 determina a quantidade ideal de parâmetros através do R_p^2:

$$R_p^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y_i})^2},$$

em que, $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ é a medida da variação total de quando a variável expli-

cativa é considerada e $\sum_{i=1}^{n}(Y_i - \overline{Y_i})^2$ é a medida da varição total de Y_i. O modelo que apresentar o maior valor do *pseudo* R^2 é o mais adequado.

Além desses três critérios de seleção, existem outros métodos de seleção de modelos automatizados utilizando programação computacional para selecionar o modelo mais parcimonioso, ou seja, com o menor número de variáveis e que se ajuste bem ao fenômeno estudado. Os métodos são *Backward*, *Forward* e *Stepwise*, e eles também utilizam os critérios mencionados anteriormente, como AIC, BIC e até mesmo a *deviance* para selecionar qual o modelo que melhor se ajusta aos dados.

2.8.4 Métodos de Seleção Automática

Backward

Esse método é realizado através da regressão passo a passo, em que o modelo inicial utiliza todas as variáveis explicativas e a cada etapa verifica se há a possibilidade de retirar alguma variável explicativa até encontrar o modelo mais adequado.

Forward

Esse método é realizado através da regressão passo a passo, em que o modelo inicial é dado por nenhuma das variáveis explicativas e a cada etapa verifica a possibilidade de incluir as variáveis explicativas até encontrar o modelo mais adequado.

Stepwise

Esse método é realizado primeiramente pela análise exploratória das variáveis explicativas e identifica quais delas são mais significantes para explicar a variável resposta. Feito isso, o método adiciona ou remove de forma sistematizada as variáveis que melhor predizem o fenômeno estudado. O método de *Stepwise* é uma junção dos métodos *Forward* e do *Backward*.

2.9 Análise dos resíduos

Como mencionado anteriormente, a variável resposta é binária, então seus resíduos irão assumir valores quando $Y_i = 1$ e quando $Y_i = 0$. Para cada um dos casos, o i-ésimo

resíduo, e_i, assumirá os seguintes valores:

$$e_i = \begin{cases} 1 - \hat{\pi}_i, \text{ para } Y_i = 1, \\ -\hat{\pi}_i, \text{ para } Y_i = 0. \end{cases}$$

Diferente de como é na regressão linear, os resíduos apresentam apenas duas alternativas de resposta. Por isso, não serão normalmente distribuídos e a análise sobre os resíduos na regressão logística é feita com o auxílio de outras medidas. Dentre elas, o Resíduos de Pearson e o Resíduos de Pearson Semistudentizados.

2.9.1 Resíduos de Pearson

$$rp_i = \frac{Y_I - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}.$$

2.9.2 Resíduos de Pearson Semistudentizados

$$rsp_i = \frac{rp_i}{\sqrt{1 - h_{ii}}},$$

em que h_{ii} é o i-ésimo elemento da matriz, $\hat{W}^{1/2}X(X'\hat{W}X)^{-1}X'\hat{W}^{1/2}$, sendo $\hat{W}$ a matriz (n x n) com elementos $\hat{\pi}_i(1 - \hat{\pi}_i)$.

2.9.3 Gráficos de resíduos

A análise de maneira gráfica dos resíduos é utilizado apenas para identificar a adequação do ajuste do modelo.

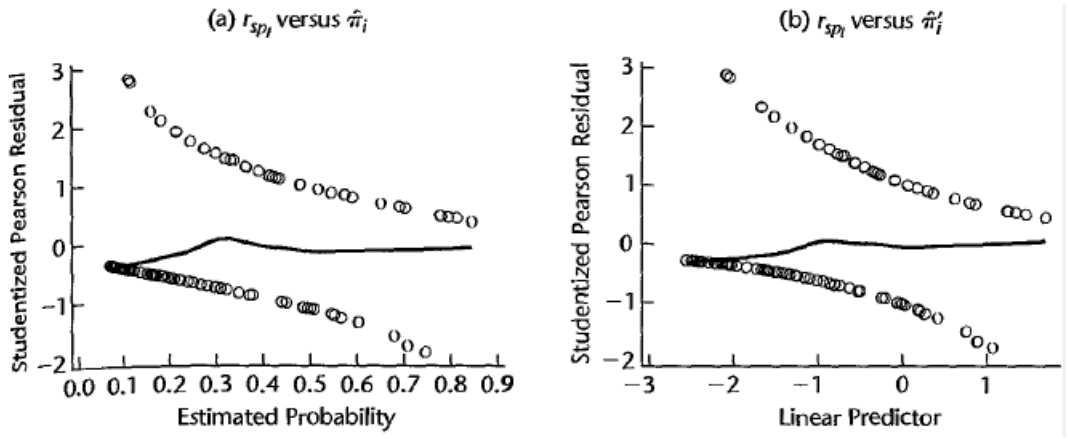


Figura 4: Exemplo do gráfico de resíduos

Fonte: Neter et al. Applied Linear Statistical Models (p. 595).

Pela figura 4, o gráfico dos resíduos do modelo de regressão logística que apresenta um bom ajuste apresentará uma linha horizontal que vai oscilar em torno de zero, qualquer desvio nessa linha determina a inequabilidade do modelo.

2.10 Detecção de observações influentes

2.10.1 Influência na estatística qui-quadrado de Pearson

Como visto anteriormente, χ^2 é dado pela estatística de Pearson do conjunto de dados completo. E ainda, o χ_i^2 é o valor da estatística para o i-ésimo caso desconsiderado. Então, o valor de delta qui-quadrado é definida por:

$$\Delta\chi_i^2 = \chi^2 - \chi_i^2.$$

Esse valor identifica a influência do i-ésimo caso no modelo, mas a determinação de qual variável é mais influente é feita através de auxílio de gráficos.

2.11 Previsão de uma nova observação

2.11.1 Escolha da regra de previsão

Na regressão logística, algumas regras são determinadas para fazer previsões para novas observações a partir da estimativa $\hat{\pi}_h$.

1- Usar 0,5 como ponto de corte

Utilizar 0,5 como um ponto de corte é uma boa regra quando as probabilidade de ocorrer 0 ou 1 são parecidos, ou quando o custo de fazer uma observação errada para cada um dos casos for próximo.

2 - Escolher o melhor ponto de corte para o modelo

Descartada a possibilidade de utilizar a Regra 1, a escolha do melhor ponto de corte para o modelo é determinado através da menor proporção de previsões incorretas. Essa regra é sugerida quando o conjunto de dados é extraído de maneira aleatória e as proporções de 0 e de 1 são adequadas.

3 - Usar probabilidades a *priori* e custos de previsões incorretas para determinar o corte

2.11.2 Estimativa da taxa de erro de previsão

Após a escolha do conjunto de dados de validação, definido através das regras de previsão, determina-se qual é a taxa de erro dessa previsão. Com isso, se essa taxa de erro é aproximadamente a mesma para o conjunto de dados de construção, então há uma indicação confiável da capacidade preditiva do modelo em análise. Se não, o modelo não é indicado para predizer as novas observações.

2.12 Curva de ROC

Uma outra forma de validar a confiabilidade do modelo é analisando a curva de ROC (Curva de característica de operação do receptor). Essa curva mostra a sensibilidade e a especificidade das previsões para todos os π_0. A sensibilidade é a probabilidade de prever $\hat{Y} = 1$ quando $Y = 1$ e a especificidade é a probabilidade de prever $\hat{Y} = 0$ quando $Y = 0$.

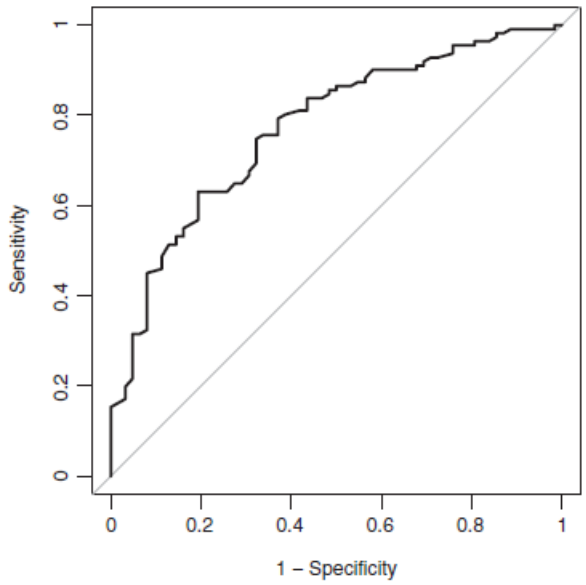


Figura 5: Exemplo do gráfico da curva de ROC

Fonte: Agresti, An Introduction to Categorical Data Analysis (p. 112).

Portanto, pela figura 5, quanto maior for a área sob a curva, melhor o poder preditivo do modelo.

2.13 Regressão Multinível

A análise multinível, também conhecida como análise de regressão hierárquica, é uma técnica estatística que investiga relações entre variáveis em diferentes níveis de agregação. Essa abordagem é útil para examinar como os níveis diferentes de um fenômeno estão relacionados com o resultado, pois ela permite ao pesquisador controlar variáveis que estão em níveis mais altos ou mais baixos do que o nível de interesse. Por exemplo, no estudo de relações entre educação e desempenho, a análise multinível pode levar em consideração as variáveis da escola, da turma e dos estudantes, assim como representado pela Figura 6 .

Figura 6: Exemplo dos níveis que podem ser analisados pela regressão multinível

Essa abordagem é útil porque permite ao pesquisador examinar relações entre variáveis que existem em diferentes níveis de agregação, tornando possível determinar o quanto cada nível contribui para o resultado. Outra vantagem da análise multinível é que ela leva em consideração os erros padrão, os intervalos de confiança e os testes de hipótese apropriados, dado que leva em consideração a correlação entre os níveis.

2.13.1 Método para dois níveis

A Regressão Multinível considera que exista apenas uma variável resposta no mais baixo nível. Assim, considerando Y_{ij} a i-ésima observação do nível primeiro nível e a j-ésima do segundo nível, o modelo multinível de regressão linear para dois níveis pode ser expressado da seguinte forma:

$$Y_{ij} = \gamma_{00} + \gamma_{po}X_{pij} + \gamma_{oq}Z_{qj} + \gamma_{pq}Z_{qj}X_{pij} + \mu_{pj}X_{pij} + \mu_{0j} + e_{ij},$$

em que γ_{00} é o intercepto da regressão, X_{pij} são as p variáveis explicativas do primeiro nível definido, chamadas de efeitos fixos, Z_{qj} são as q variáveis explicativas do outro nível, que por sua vez são definidas como efeitos aleatórios. O μ_{0j} e e_{ij} são os erros usuais desses níveis, respectivamente. Por fim, μ_{pj} são os resíduos do nível 2 dos coeficientes das variáveis explicativas X_{pij} do nível 1.

2.13.2 Passos para a construção do modelo

- **Passo 1**

Primeiramente, analisa-se o modelo sem nenhuma variável explicativa:

$$Y_{ij} = \gamma_{00} + \mu_{0j} + e_{ij}, \tag{2.13.1}$$

esse modelo é relevante pois proporciona a estimativa de correlação intraclasse ρ:

$$\rho = \frac{\sigma^2_{\mu 0}}{\left(\sigma^2_{\mu 0} + \sigma^2_e\right)}, \tag{2.13.2}$$

onde $\sigma^2_{\mu 0}$ é a variância dos resíduos $\mu 0$ e σ^2_e é a variância dos resíduos e_{ij}. Segundo Hox (2010, p. 17), 39% para essa correlação é um valor consideravelmente elevado. Além disso, o modelo sem nenhuma variável proporciona também o *deviance*, já mencionado anteriormente, utilizado para comparar modelos.

- **Passo 2**

Incluindo todas as variáveis explicativas fixas do nível mais baixo, obtém-se o modelo desse passo:

$$Y_{ij} = \gamma_{00} + \gamma_{po} X_{pij} + \mu_{0j} + e_{ij}. \tag{2.13.3}$$

Neste passo, as variáveis explicativas desse nível são analisadas com mais acurácia.

- **Passo 3**

Dessa vez, acrescentam-se as variáveis explicativas do nível seguinte:

$$Y_{ij} = \gamma_{00} + \gamma_{po} X_{pij} + \gamma_{oq} Z_{qj} + \mu_{0j} + e_{ij}, \tag{2.13.4}$$

Os modelos dos passos 2.13.3 e 2.13.4 são chamados modelos de componentes de variância justamente por decomporem a variância do intercepto em níveis distintos de variância para cada nível hierárquico.

- **Passo 4**

O modelo desse penúltimo passo é chamado de modelo de coeficientes randômicos. Já nesse modelo o nível de avaliação passa a ser micro e tem um componente significativo de variância.

$$Y_{ij} = \gamma_{00} + \gamma_{po}X_{pij} + \gamma_{oq}Z_{qj} + \mu_{pj}X_{pij} + \mu_{0j} + \varepsilon_{ij},$$

- **Passo 5**

Finalmente, são acrescentadas todas as variáveis explicativas dos níveis que tiveram variância significativa de no passo anterior:

$$Y_{ij} = \gamma_{00} + \gamma_{po}X_{pij} + \gamma_{oq}Z_{qj} + \gamma_{pq}Z_{qj}X_{pij} + \mu_{pj}X_{pij} + \mu_{0j} + \varepsilon_{ij}.$$

2.13.3 Estimação dos parâmetros

A estimação de parâmetros na modelagem de nível múltiplo é, principalmente, feita pelo método de máxima verossimilhança. Dois tipos diferentes de função de verossimilhança são usados nessa modelagem: *full maximum likelihood* (FML) e *restricted maximum likelihood* (RML). No primeiro método, os coeficientes de regressão e os componentes de variância são incluídos na função de verossimilhança. No outro, apenas os componentes de variância são incluídos na função de verossimilhança, e os coeficientes de regressão são estimados em uma segunda etapa.

A estimação de máxima verossimilhança produz estimativas de parâmetros e erros padrão correspondentes. Esses podem ser usados para realizar um teste de teste de Wald. Logo, considerando as seguintes hipóteses:

$$\begin{cases} H_0)\ \gamma_k = 0, \\ H_1)\ \gamma_k \neq 0. \end{cases}$$

A estatística do teste é dada por:

$$Z = \frac{\hat{\gamma}}{\hat{\sigma_\gamma}} \sim N(0,1).$$

O estatística do teste de Wald pode receber uma outra distribuição amostral conhecida quando elevado ao quadrado, χ^2 com 1 grau de liberdade.

2.13.4 Escolha e comparação dos modelos

A análise do *deviance* é necessária para ajustar, avaliar e comparar modelos estatísticos. Assim como na regressão logística múltipla, o critério utilizado é quanto menor a diferença, melhor o ajuste. O processo começa com o ajuste inicial, seguido da adição de variáveis explicativas. Além disso, essa medida pode ser usada como critério para determinar a necessidade de adicionar variáveis ao modelo de diferentes níveis.

2.13.5 Modelos lineares generalizados multiníveis

Segundo Hox (2010, p. 117), "Wong e Mason (1985), Gibbons e Bock (1987), Longford (1993), Goldstein (1991, 2003) e Raudenbush e Bryk (2002) descrevem a extensão multivariada dos modelos lineares generalizados". Então, o modelo de dois níveis para proporções é escrito da seguinte forma, no caso das variáveis respostas binárias:

$$Y_{ij} = \gamma_{00} + \gamma_{po}X_{pij} + \gamma_{oq}Z_{qj} + \gamma_{pq}Z_{qj}X_{pij} + \mu_{pj}X_{pij} + \mu_{0j}.$$

A variância residual usual de menor nível e_{ij} não aparece na equação do modelo, pois faz parte da especificação da distribuição de erro. Se a distribuição de erro for binomial, a variância é uma função da proporção populacional π_{ij} : $\sigma^2 = (\pi_{ij}/(1 - \pi_{ij}))$ e não precisa ser estimada separadamente. Em alguns softwares é possível estimar um fator de escala para a menor variância de nível.

3 Metodologia

3.1 Conjunto de dados

Os dados do estudo exposto foram obtidos através dos bancos de dados extraídos dos sistemas de informações acadêmicas da Universidade de Brasília (SIGRA e SIGAA, atualmente) para as disciplinas Bioestatística e Probabilidade e Estatística no período de 1994 até 2019. Os bancos continham para cada estudante as seguintes informações: Curso do estudante, Código da disciplina, Nome da disciplina, Período no qual a disciplina foi cursada, Turma, Menção, Faltas, Hora de início da aula, Hora do término da aula, ID do estudante e ID do professor que ministrou a disciplina (ID resultantes de codificação aleatória).

A ferramenta computacional que foi utilizada para a limpeza, análise descritiva e modelagem dos dados foi o programa R (versão 4.2.1).

3.2 Limpeza e tratamento dos dados

O processo de limpeza dos dados tanto para a base inicial de Bioestatística como a de Probabilidade e Estatística foi padronizado. A otimização do tratamento de dados é justificado principalmente pelo banco inicial de ambas disciplinas apresentarem as mesmas variáveis.

Na análise prévia dos dados, foram identificados alguns problemas, como mais de uma hora de início e término da aula em um período, mais de um código de professor na mesma turma e semestre, entre outros. Para resolver os horários duplicados, foi criada uma coluna chamada "Horário", na qual as turmas com horários cruzados foram identificadas, e a coluna "Turno"derivada da dessa nova variável. Para as turmas com mais de um código de professor, foi criada uma nova coluna que identifica as substituições.

Apenas os casos em que os estudantes concluíram a disciplina foram levados em consideração, ignorando trancamentos e créditos concedidos. O desempenho dos alunos (aprovado, reprovado com ou sem a menção SR) foi registrado através da coluna "Menção". Se o estudante obteve "SR", "II"ou "MI", foi apontado como reprovado, enquanto que se sua menção foi "MM", "MS"ou "SS", foi considerado como aprovado. Além disso, apenas a primeira tentativa de cada aluno na disciplina foi considerada para o banco de dados final.

3.3 Análise exploratória

Após a limpeza e tratamento dos dados, foram realizadas análises exploratórias de dados, incluindo análises univariadas e bivariadas das principais variáveis presentes no banco de dados. Para essa análise, foram criados gráficos e medidas estatísticas para visualizar e compreender as relações entre as variáveis.

Em seguida, as análises entre as variáveis explicativas consideradas mais importantes foram realizadas e a variável resposta, apenas dos últimos três períodos mais recentes, de 2017 a 2019. Além disso, essas análises foram duplicadas para os casos em que o aluno obteve uma menção final diferente de SR e descartando os que abandonaram as disciplinas. Finalmente, foram executados alguns testes qui-quadrado para avaliar as associações entre as variáveis explicativas e o desempenho do estudante.

3.4 Modelagem dos dados

A modelagem dos dados foi realizada através da técnica de Regressão Logística Multinível, pois a variável resposta é binária, o "sucesso"é determinado pela aprovação do aluno, e existe uma estrutura hierárquica de dados formada por dois níveis principais: os alunos e as turmas. Essa abordagem nos permite levar em consideração as particularidades de cada nível hierárquico para prever o desempenho acadêmico dos alunos. Os efeitos fixos, que são as características referentes ao nível dos alunos, são: o tempo na UnB, o curso e as faltas. Já os efeitos aleatórios, que são as características referentes ao nível da turma, foi considerado apenas o turno e o professor que ministrou a disciplina. Por fim, optou-se por analisar os casos com a base completa, ou seja, considerando as menções SR, e desenvolver um outro modelo desconsiderando essa menção.

Ao trabalhar com estatísticas e análise de dados usando o programa R, inicialmente foi criado um modelo completo de todas as variáveis relevantes, determinadas pelo teste Qui-quadrado. Posteriormente, foi elaborada a validação da base completa com as amostras de teste e de validação. Em seguida, foram analisadas as razões de chances considerando o modelo final validado. Para finalizar, vários diagnósticos foram realizados no modelo com o uso de gráficos de resíduos, curva de ROC e matriz de confusão.

4 Análise exploratória

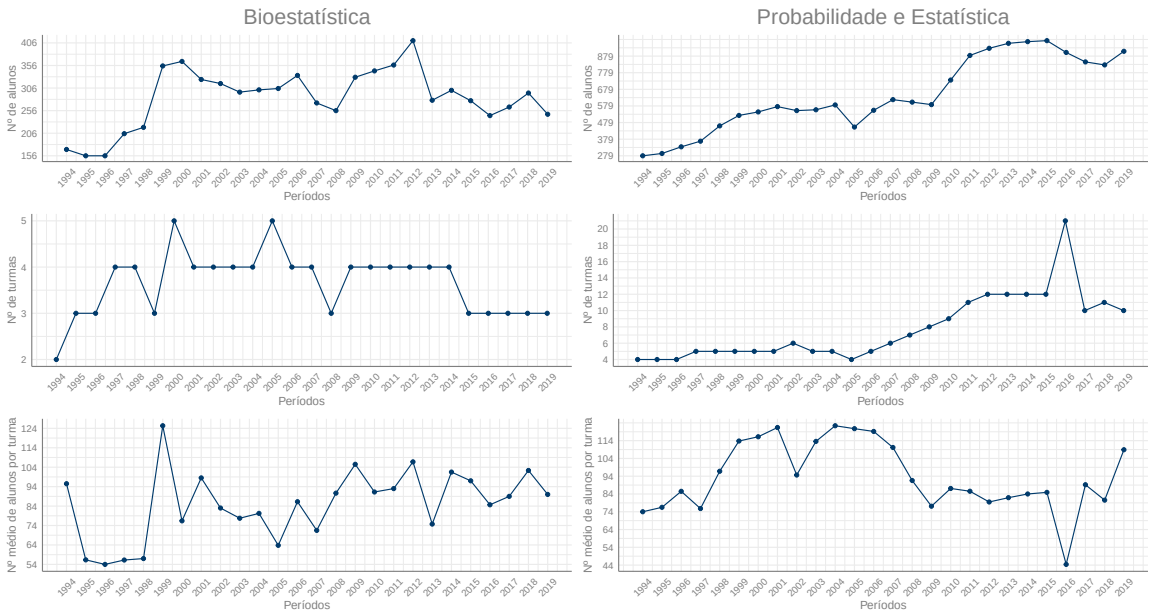
4.1 Análise das características das disciplinas Bioestatística e Probabilidade e Estatística

Nessa parte do estudo, busca-se estudar as características da oferta das disciplinas Bioestatística e Probabilidade e Estatística e sua evolução ao longo do período de 1994 e 2019. O banco de dados é composto por 7993 e 17689 estudantes de Bioestatística e Probabilidade e Estatística, respectivamente.

O Gráfico 7 aponta para um aumento no número de estudantes nas duas disciplinas durante o período analisado. Em relação à Bioestatística, o número de estudantes acompanhava o número de turmas até 2009, após esse período o número de turmas foi estabilizado até 2015 em níveis mais baixos. Apesar da oscilação no número médio de estudantes por turma, esse comportamento pode ser observado pelo pico de estudantes em 2012, quando o número médio de alunos por turma também aumentou.

Já para a disciplina de Probabilidade e Estatística, o número de estudantes também acompanhava o número de turmas até 2016, quando houve um pico no número de turmas e, consequentemente, o número médio de alunos por turma atingiu seu ponto mais baixo.

Gráfico 7: Número de alunos por período, número de turmas por período e número médio de alunos por turma em relação aos períodos.

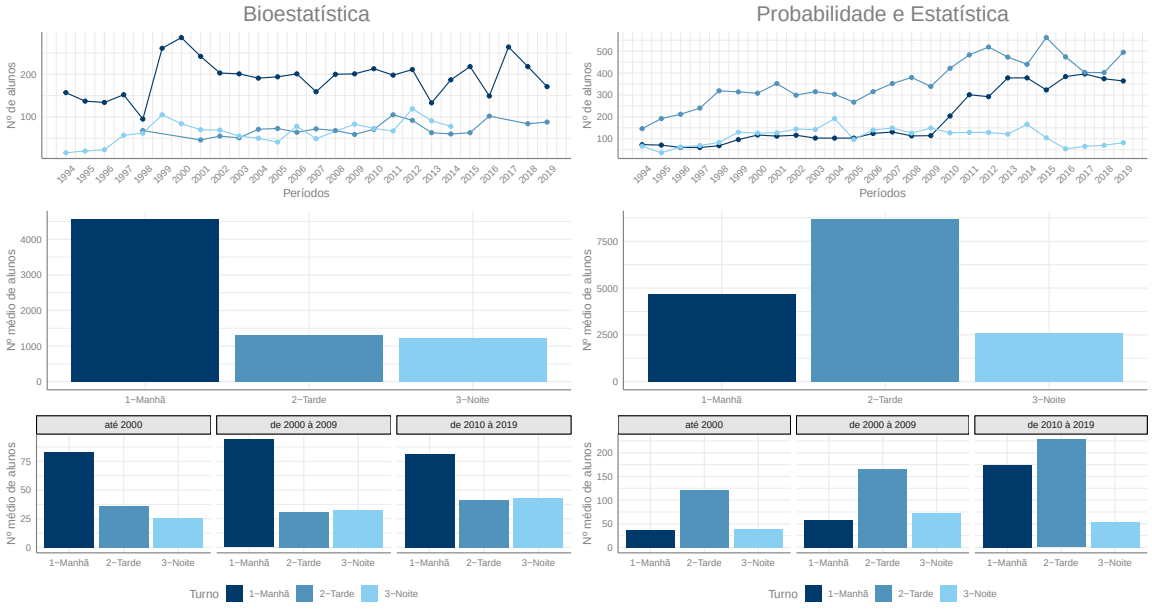


4.1.1 Turnos

Para a disciplina Bioestatística, o Gráfico 8, indica que a maioria dos alunos cursaram nos horários do turno da manhã, especificamente o de 8h às 9h50, visto é o único que foi ofertado em todos os períodos. Embora o turno da noite seja o segundo mais frequente, o horário de 14h às 15h50 é o segundo mais procurado e as turmas noturnas com os horários cruzados para essa disciplina foram ofertadas até 2005.[1]

Analisando o Gráfico 8 para Probabilidade e Estatística, a maioria dos alunos cursou no turno da tarde. Apesar de o turno da manhã e da noite não tenham apresentado muita diferença no número de alunos até 2009, depois desse período houve um aumento no noturno. As turmas noturnas com os horários cruzados para essa disciplina foram ofertadas até 2018.[2]

Gráfico 8: Número de alunos por turno em relação aos períodos, número médio de alunos por período em relação aos turnos e períodos.

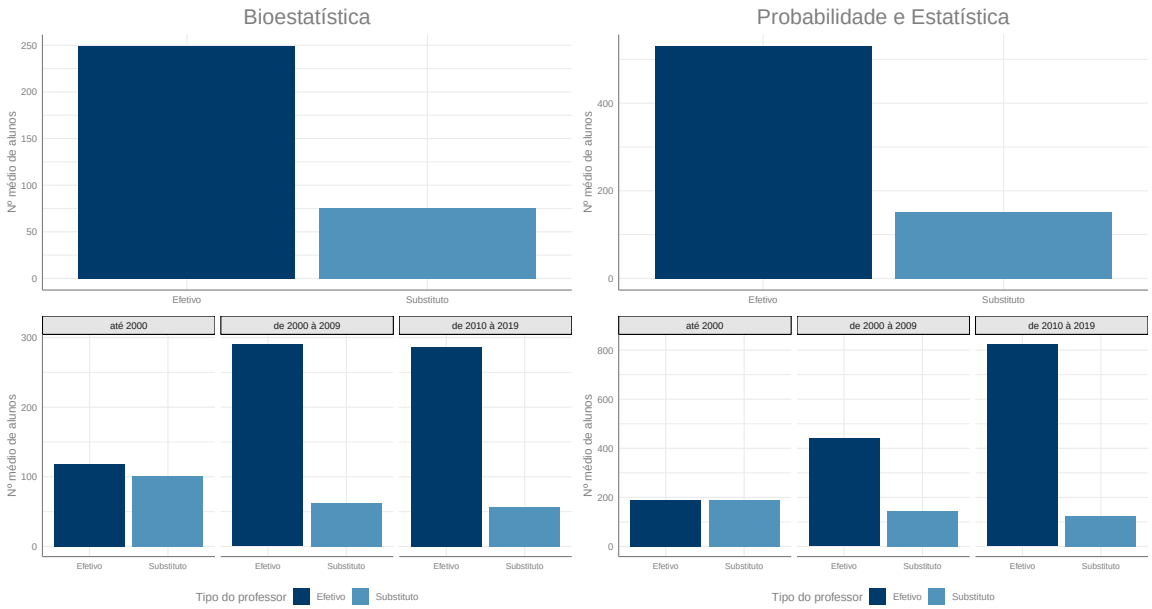


4.1.2 Professores

Segundo o Gráfico 9, é evidente que, para ambas as disciplinas, mais alunos cursaram pela primeira vez com professores efetivos do que com professores substitutos. Contudo, ao analisar por período, nota-se que até o ano de 2000, o número médio de alunos que cursaram pela primeira vez com professores efetivos e com professores substitutos é bastante semelhante.

[1] Consulte a tabela no apêndice.
[2] Consulte a tabela no apêndice.

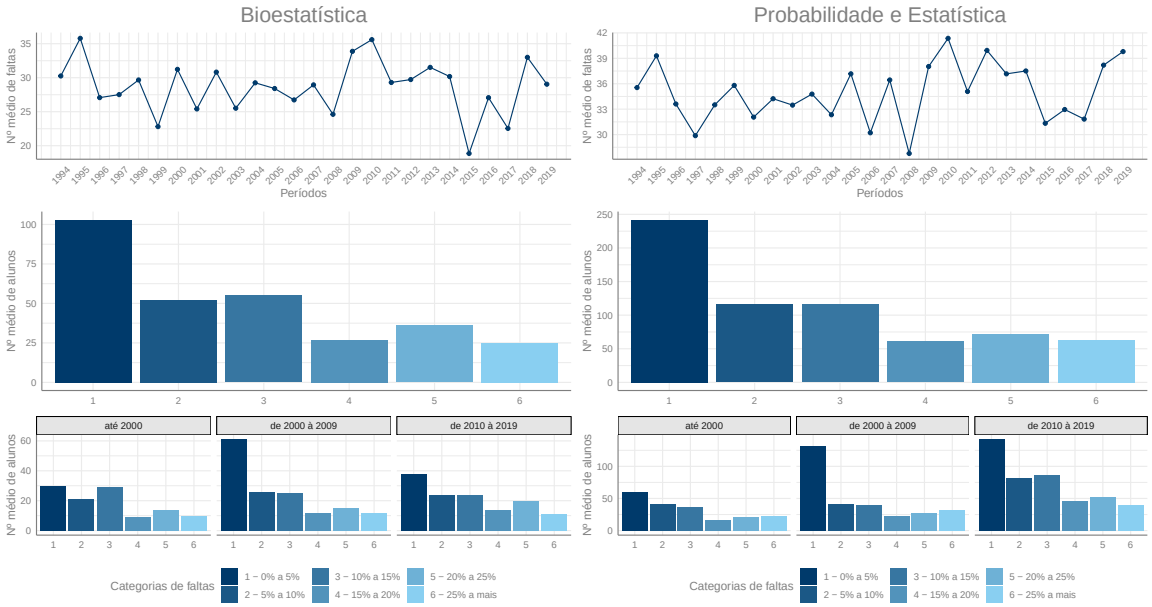
Gráfico 9: Número médio de alunos em relação ao tipo do professor que ministrou as disciplinas e os períodos.



4.1.3 Faltas

De acordo com o Gráfico 10, a média de faltas para ambas as disciplinas varia amplamente de acordo com os períodos examinados. No entanto, a maioria dos alunos, em média, tende a ter faltas entre 0% e 5%. Ao analisar por período, percebe-se ainda que o intervalo com a menor média de alunos, tanto para Bioestatística como para Probabilidade e Estatística, é entre 15% e 20%.

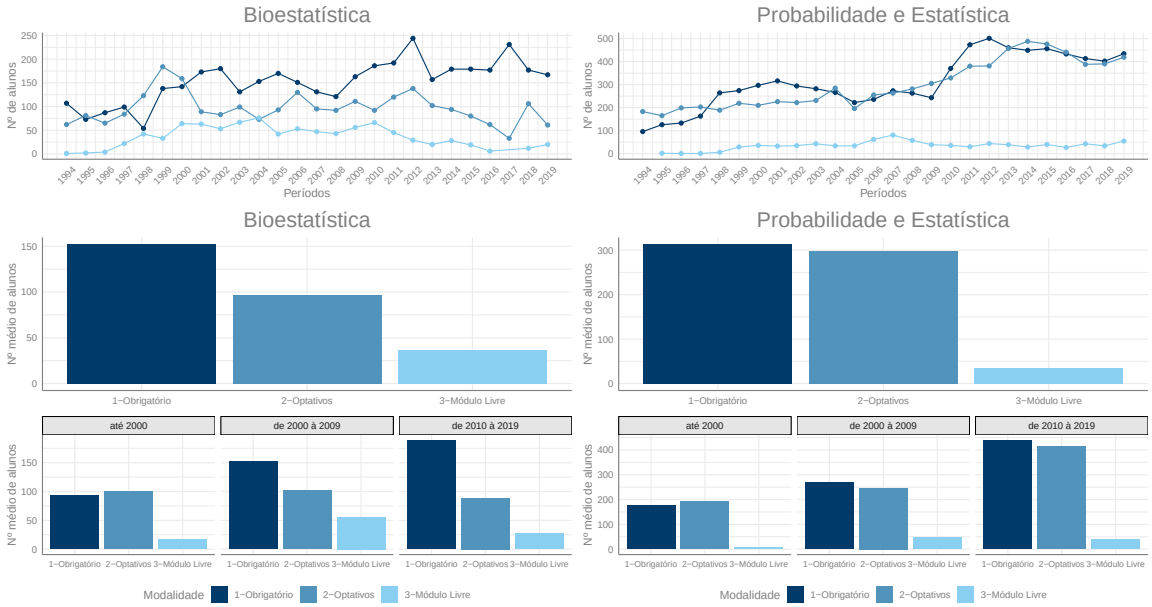
Gráfico 10: Percentual de faltas em relação aos períodos, número médio de alunos por período em relação as percentual de faltas e aos períodos.



4.1.4 Modalidade das disciplinas

É evidente, pelo Gráfico 11, que para as duas matérias em análise, o número médio de estudantes que possuem elas no currículo obrigatório é maior do que o das outras modalidades. No entanto, quando analisado o período detalhadamente, verifica-se que para Bioestatística, a partir do ano 2000, o número de alunos que cursam a disciplina como optativa diminui em relação aos que cursam como obrigatória. Já para a outra disciplina, esse número é bastante próximo ao longo de todo o período observado, inclusive nos três anos mais recentes.

Gráfico 11: Número de alunos por modalidade da disciplina em relação aos períodos, número médio de alunos por período em relação as modalidades da disciplina e aos períodos.



4.1.5 Cursos

Conforme o gráfico 12, para a Bioestatística, nota-se que os dois principais cursos que possuem o maior número médio de alunos que fizeram a disciplina pela primeira vez são aqueles em que ela é considerada obrigatória. No entanto, quando observado o gráfico 13, percebe-se que até 2000, o curso que apresentava o maior número de alunos médios era o de Ciências Biológicas.

Para Probabilidade e Estatística, conforme o gráfico 12, o comportamento geral do número médio de alunos não é o mesmo da disciplina anterior, já que Engenharia Civil, que é o curso que aparece com maior frequência, não inclui a disciplina como obrigatória. Quando observamos o gráfico 13, notamos que o número médio de alunos tende a ser mais bem distribuído no decorrer do tempo, com relação a Bioestatística. Vale destacar também que a categoria "Outros"[3] apresentou crescimento muito significativo no último período analisado, apesar de representar cursos com quantidade média de estudantes muito baixa.

[3]Consulte o apêndice.

Gráfico 12: Cursos dos estudantes com os maiores números médios de alunos.

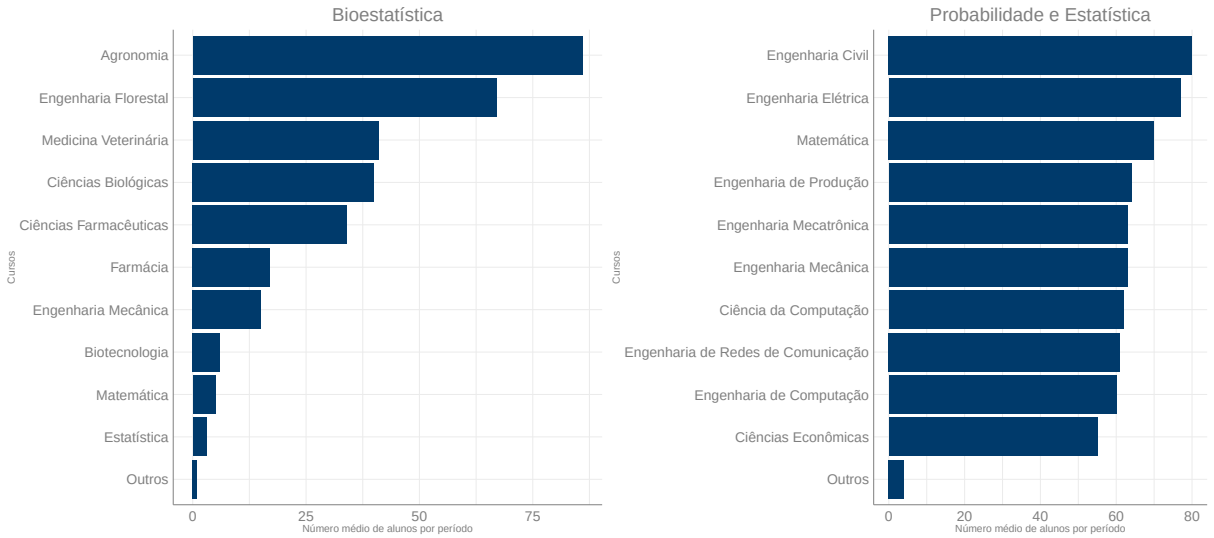
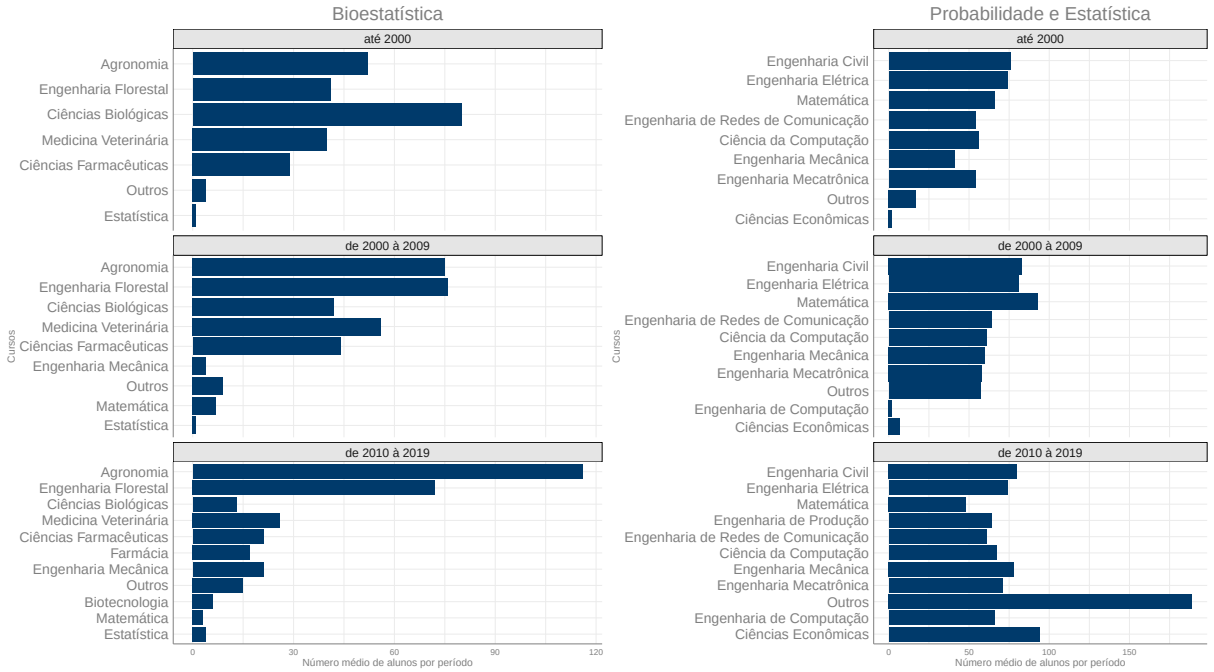


Gráfico 13: Cursos dos estudantes com os maiores números médios de alunos em relação aos períodos.

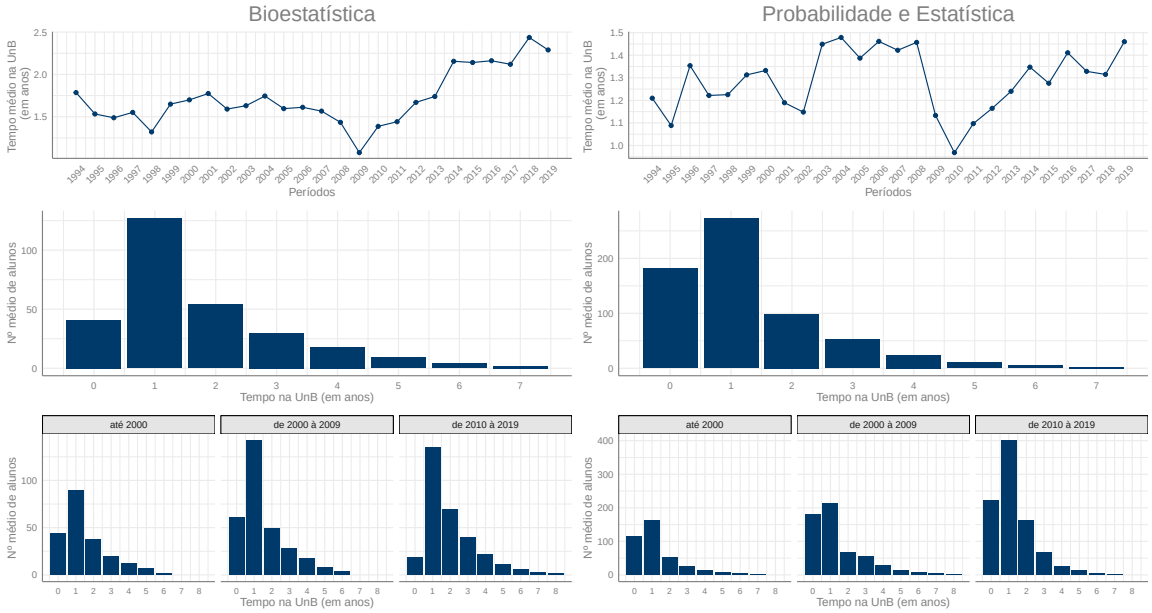


4.1.6 Tempo dos estudantes na UnB até cursar pela primeira vez as disciplinas, em anos completos

Ao verificar o Gráfico 14, identifica-se que o tempo médio em anos inteiros dos universitários varia bastante conforme os períodos para as duas disciplinas. No entanto, para Bioestatística, verifica-se que a maioria dos alunos, em média, estão entre 1 e 2 anos

na Universidade de Brasília. Para Probabilidade e Estatística, a maioria se encontra no intervalo de até 1 ano. De 2000 a 2009, o comportamento da maioria dos alunos, em média, de Bioestatística era semelhante ao de Probabilidade e Estatística, mas a partir de 2010 esse comportamento mudou.

Gráfico 14: Tempo médio em anos dos estudantes na UnB até tentar pela primeira vez as disciplinas.



4.1.7 Menção dos estudantes

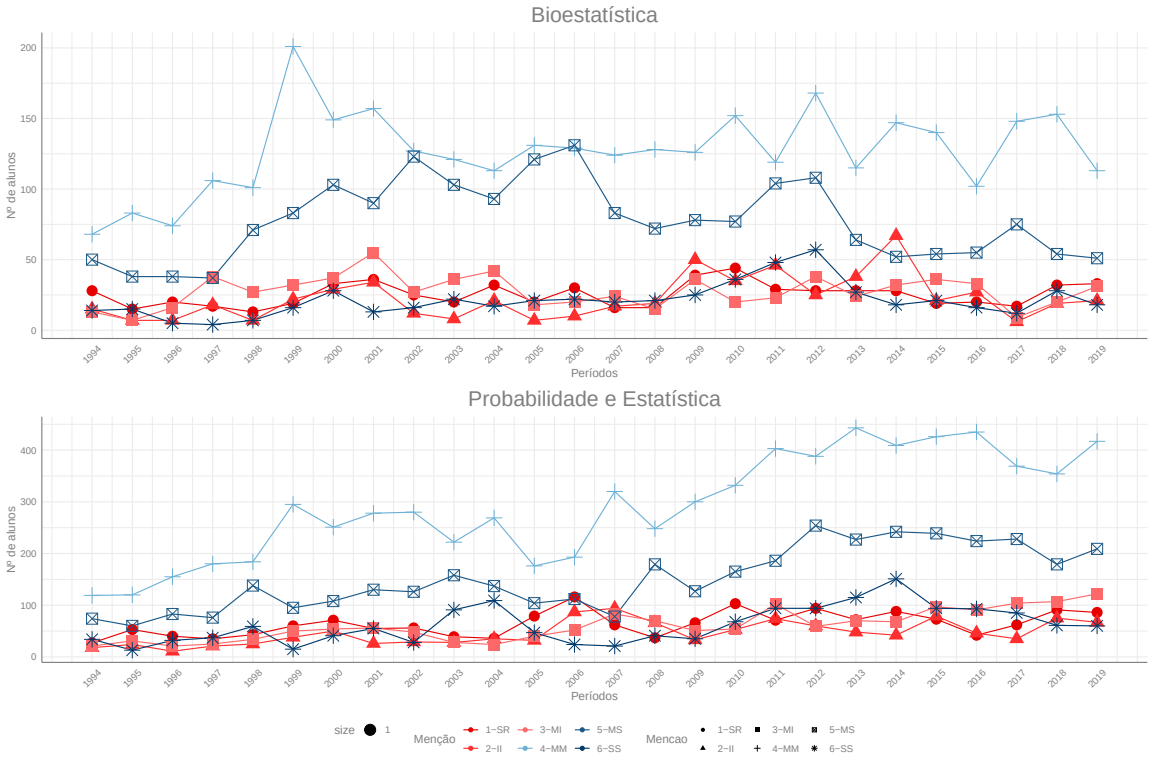
Referente as menções[4] dos estudantes, para as duas disciplinas, observa-se que a menção MM (menção mínima para aprovação) é a que apresenta a maior frequência ao decorrer dos anos analisados, seguida pela MS, como visto no Gráfico 15. Outra informação relevante é que em ambas as disciplinas, a média de alunos com a menção SR[5] é considerável em comparação com as demais menções.

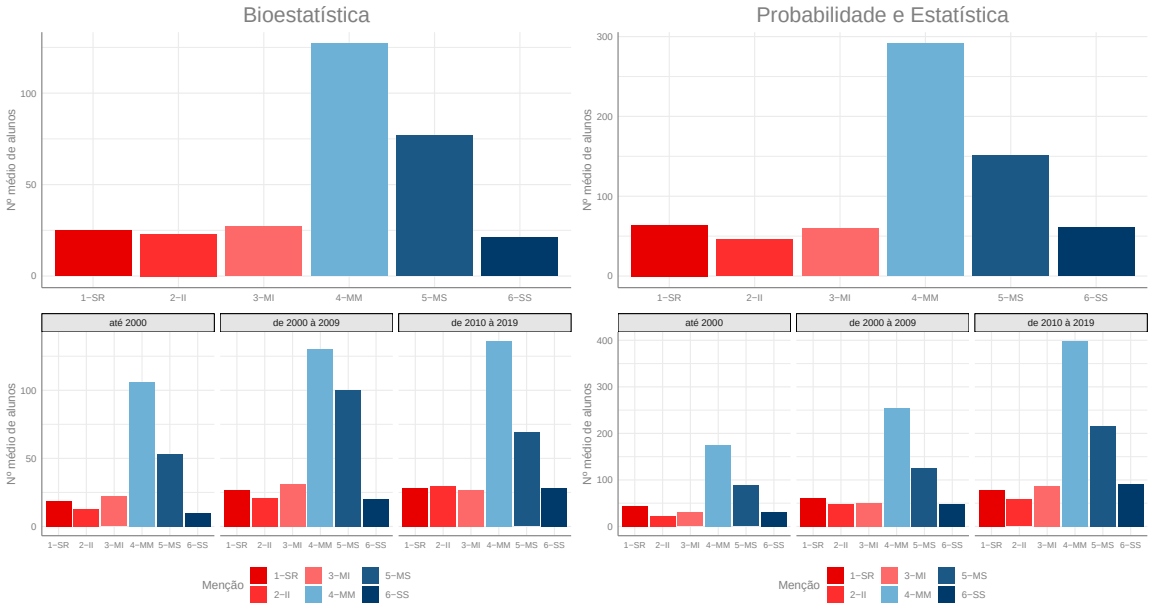
Importante destacar que em Bioestatística, até 2009, as menções que resultam em reprovação dos alunos, ou seja, SR, II e MI, são mais frequentes do que a menção SS. Depois desse período, a menção SS tende a se igualar à MI. Já em Probabilidade e Estatística, também até 2009, a média de alunos reprovados com a menção SR é maior do que a média de alunos aprovados com a menção SS. Após esse período, essa tendência se inverte e os valores para a menção SS aumentam.

[4]Neste estudo, as menções avaliadas são, em ordem crescente: SR, II, MI, MM, MS e SS, sendo que as três primeiras correspondem a reprovados e as três últimas representam menções de aprovação dos estudantes.

[5]A menção SR pode ser interpretada como uma forma de desistência do estudante, já que é concedida quando o aluno tem mais de 25% de faltas ou obtém nota zero em todas as avaliações da matéria.

Gráfico 15: Número de alunos por menções da disciplina em relação aos períodos, número médio de alunos por período em relação as menções da disciplina e aos períodos.





4.1.8 Desempenho dos estudantes

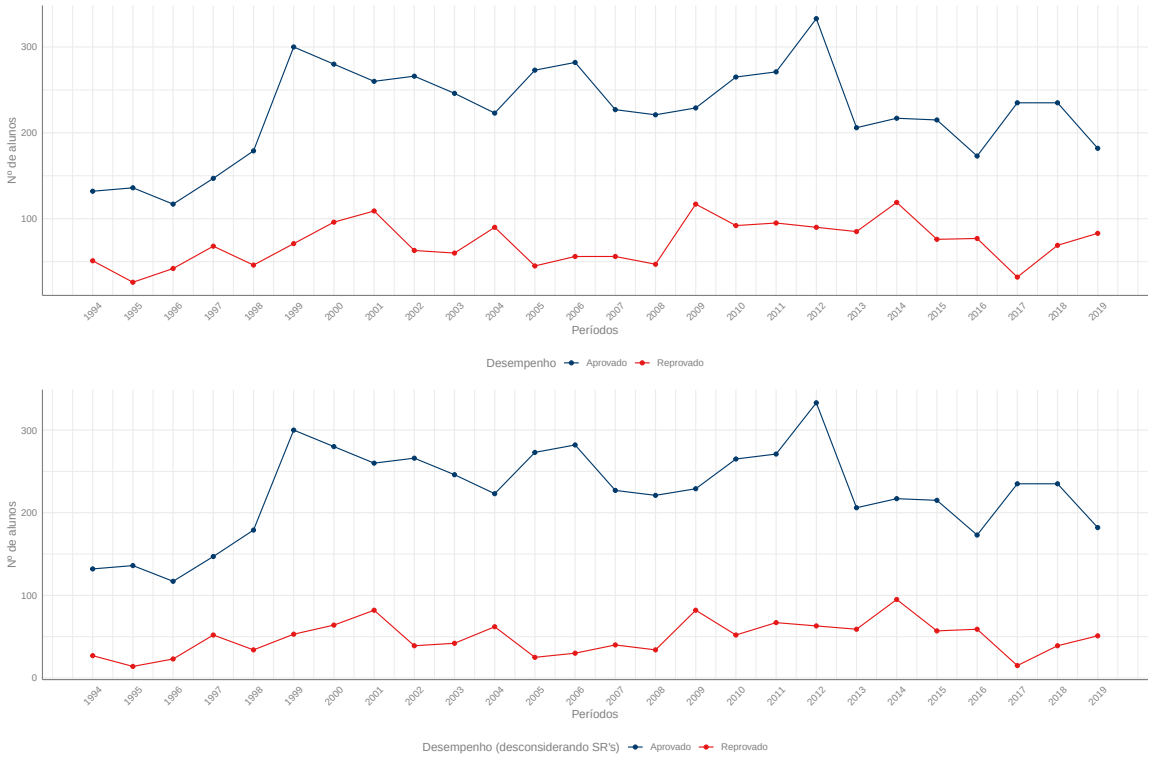
- **Bioestatística**

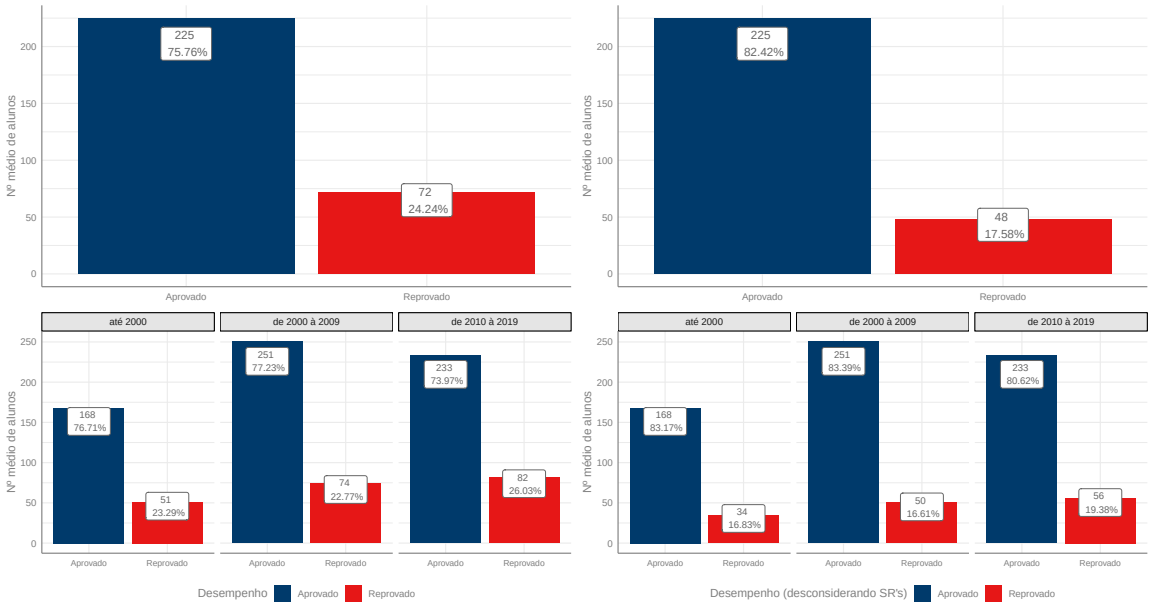
Pelo Gráfico 16, é evidente observar que, mesmo desconsiderando os estudantes

com a menção SR, o número de alunos aprovados é significativamente maior do que o de reprovados em todos os anos considerados no estudo. Além disso, a média de universitários reprovados cresceu ao longo dos períodos, em ambas as análises.

Considerando as menções SR, em média, de 297 estudantes avaliados, 72 deles foram reprovados. Isso significa que a taxa de aprovação representa aproximadamente 76%, enquanto a proporção complementar é em torno de 24%. Em contrapartida, quando a menção é desconsiderada, de 273 estudantes analisados, 48 deles foram reprovados. Então, aproximadamente 82% dos estudantes foram aprovados, enquanto cerca de 17% não foram. Dessa forma, desconsiderando a menção de SR, a taxa de reprovação decai cerca de 7%. Esse declínio na taxa de reprovação é identificado ao longo dos períodos também.

Gráfico 16: Número de alunos por período em relação ao desempenho em Bioestatística, número médio de alunos em relação ao desempenho e o período na disciplina.



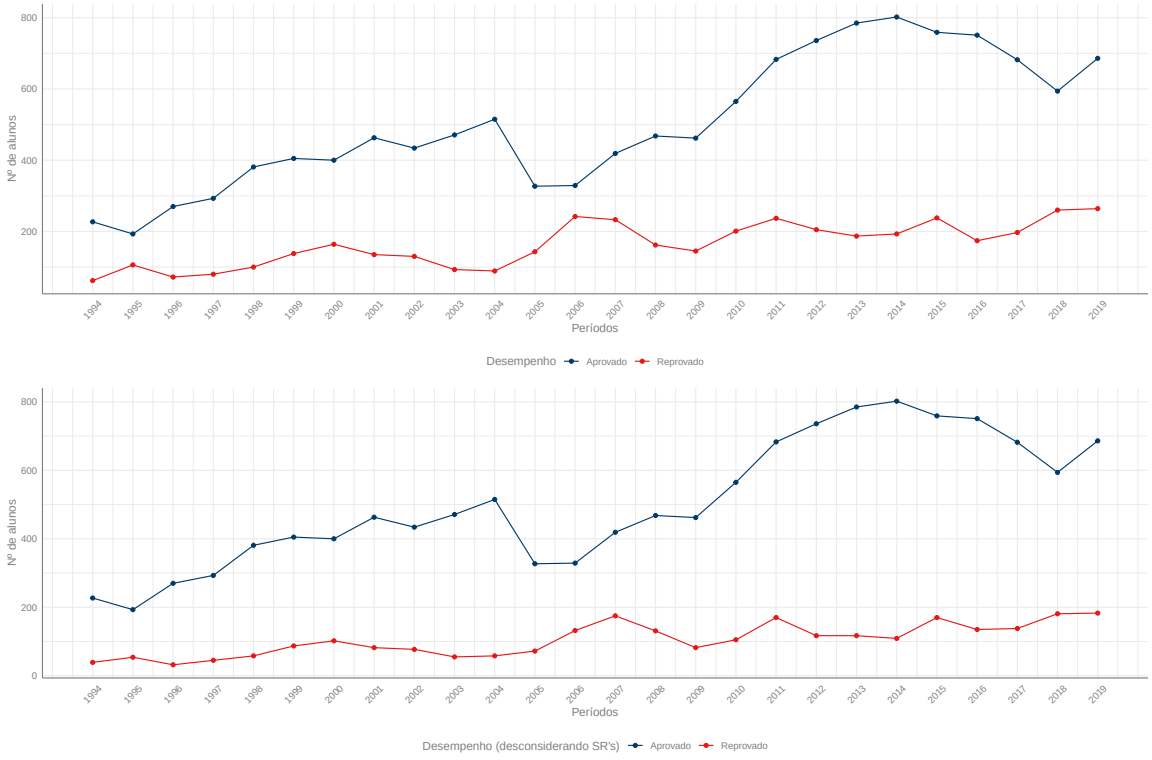

Nº médio de alunos
225
75.76%
72
24.24%
Aprovado
Reprovado
até 2000
de 2000 à 2009
de 2010 à 2019
168
76.71%
51
23.29%
251
77.23%
74
22.77%
233
73.97%
82
26.03%
Desempenho
Aprovado
Reprovado
225
82.42%
48
17.58%
168
83.17%
34
16.83%
251
83.39%
50
16.61%
233
80.62%
56
19.38%
Desempenho (desconsiderando SR's)

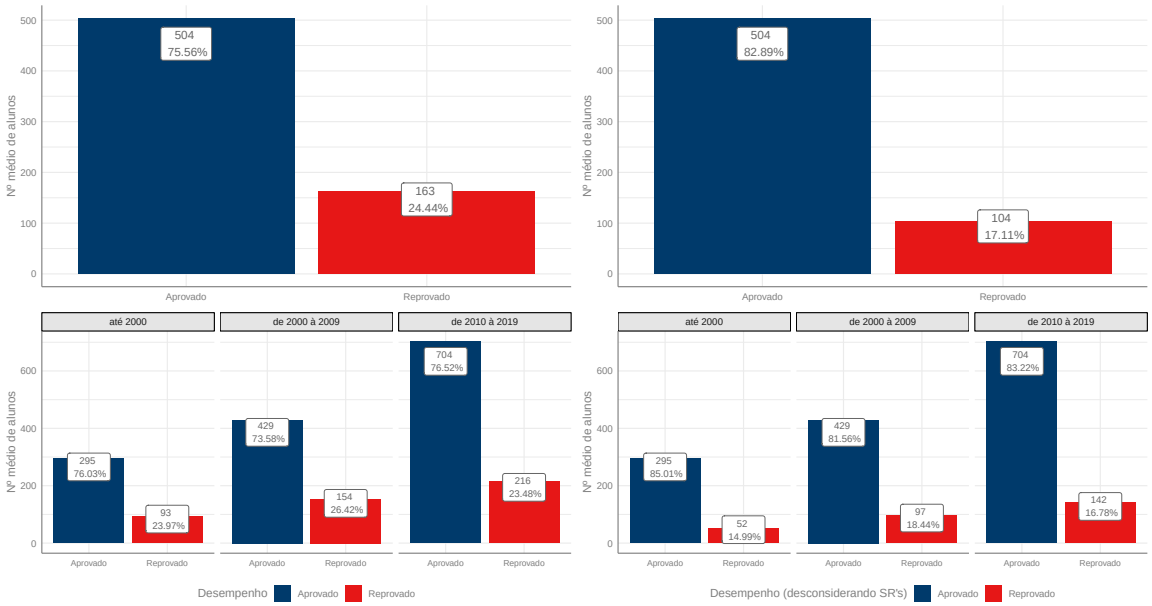
- **Probabilidade e Estatística**

Como é possível ver no Gráfico 17, assim como para Bioestatística, mesmo sem levar em conta os estudantes com a menção SR, há uma diferença entre o número de alunos aprovados e o de reprovados em todos os anos avaliados no estudo de Probabilidade e Estatística. E ainda, nota-se um aumento na média de estudantes reprovados ao longo dos períodos, tanto na primeira quanto na segunda análise.

De 667 estudantes avaliados, considerando as menções SR, aproximadamente 163 deles foram reprovados. Isso significa que a aprovação corresponde a cerca de 76% e a reprovação a 24%, assim como em Bioestatística. Já quando a menção SR é ignorada, dentre os 608 estudantes avaliados, cerca de 104 foram reprovados, o que representa uma aprovação de aproximadamente 83% e uma reprovação de cerca de 17%. Assim, desconsiderando a menção SR, a taxa de reprovação diminui em cerca de 7%, com o mesmo comportamento da disciplina anteriormente analisada, e esse declínio pode ser identificado ao longo dos períodos.

Gráfico 17: Número de alunos por período em relação ao desempenho em Probabilidade e Estatística, número médio de alunos em relação ao desempenho e o período na disciplina.





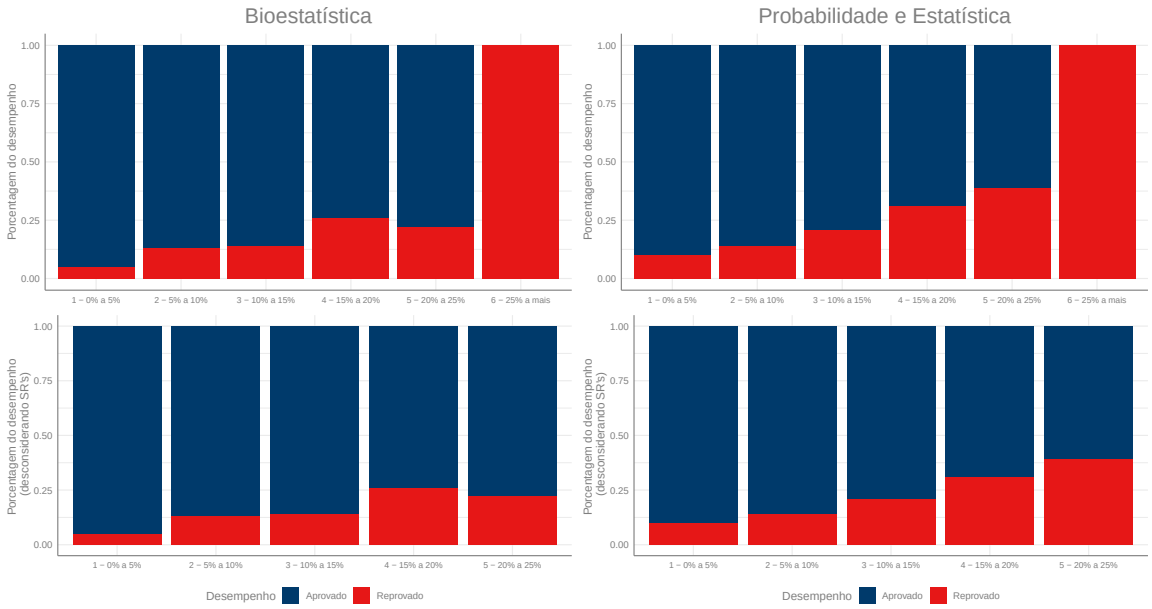
4.2 Análises entre a variável resposta e as explicativas

Na etapa da modelagem decidiu-se trabalhar somente com os resultados dos anos de 2017 e 2019, que são os mais recentes no banco de dados analisado. A razão para essa escolha é que muitos fatores mudaram na Universidade ao longo dos anos, o que torna importante avaliar quais são as variáveis atualmente que influenciam o desempenho dos estudantes em Bioestatística e Probabilidade e Estatística. Dois casos foram considerados: dados completos e dados sem menções SR (pois a menção SR pode ser interpretada como uma forma de desistência da disciplina pelo estudante). Em seguida, realizaram-se algumas análises bivariadas, para verificar a relação entre as variáveis candidatas a participar do modelo e o resultado final obtido pelos estudantes na primeira vez que cursou a disciplina.

4.2.1 Faltas

No Gráfico 18, é evidente observar que a faixa com valores maiores que 25% correspondem às proporções dos estudantes com menções SR. Por isso, na base sem essa menção, a faixa não aparece mais no Gráfico. Para Probabilidade e Estatística, quanto maior o número de faltas, maior a proporção de reprovações. Para Probabilidade e Estatística, esse comportamento é percebido até a faixa de 15% a 20%, onde há um decréscimo na taxa de reprovações para as faixas de 20% a 25%.

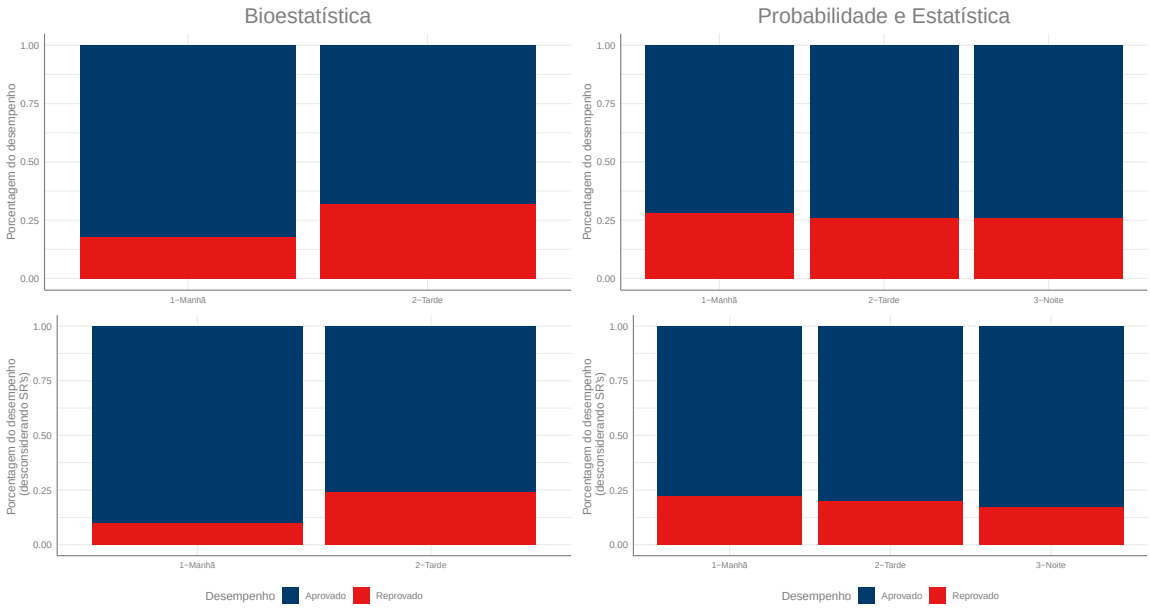
Gráfico 18: Desempenho dos estudantes segundo percentual de faltas na disciplina (2017-2019).



4.2.2 Turno de oferta das disciplinas

De acordo com o Gráfico 19, para a disciplina de Bioestatística, a análise do desempenho dos estudantes revelou que o turno da tarde teve a maior taxa de reprovação. Já para Probabilidade e Estatística, quando considerada a base completa, a taxa de reprovação foi bastante semelhante para todos os turnos, com o turno da manhã destacando com a maior taxa. Desconsiderando SR, o turno da manhã continua com a maior taxa, mas a taxa de reprovação caiu de forma significativa para o turno da noite quando comparado aos outros.

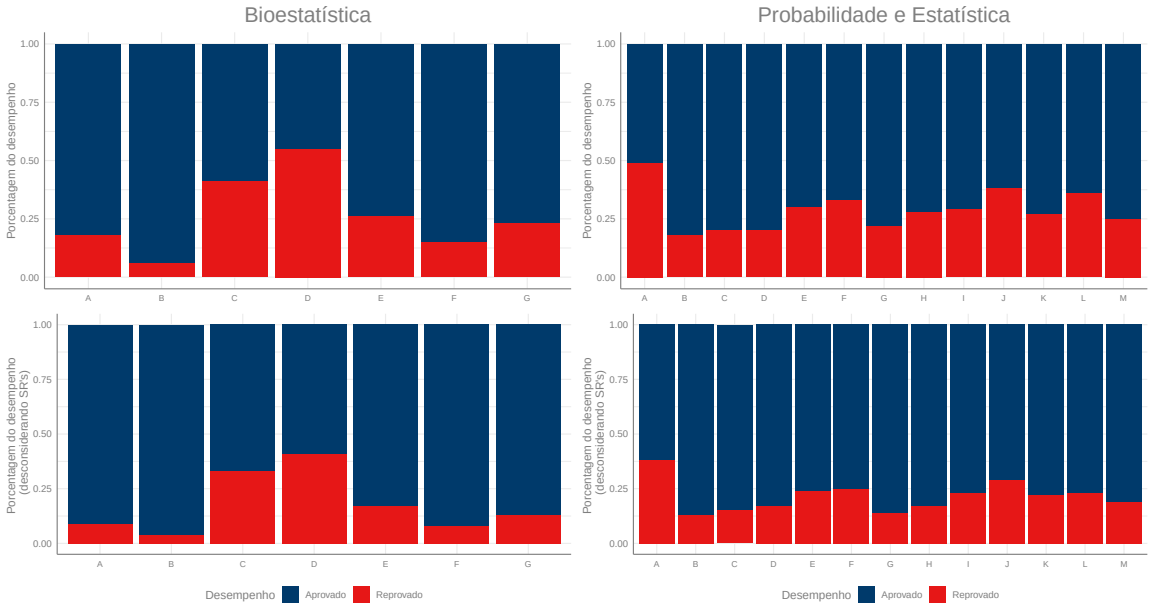
Gráfico 19: Desempenho dos estudantes segundo os turnos (2017-2019).



4.2.3 Professor

Analisando o Gráfico 20, para Bioestatística, observou-se uma tendência de maior reprovação, em média, para as turmas que foram ministradas pelo professor D, nas duas análises. Enquanto para Probabilidade e Estatística, essa tendência é evidenciada pelo professor A. Por outro lado, a taxa de aprovação média dos professores H e L decaiu mais rapidamente quando retirados casos sem SR.

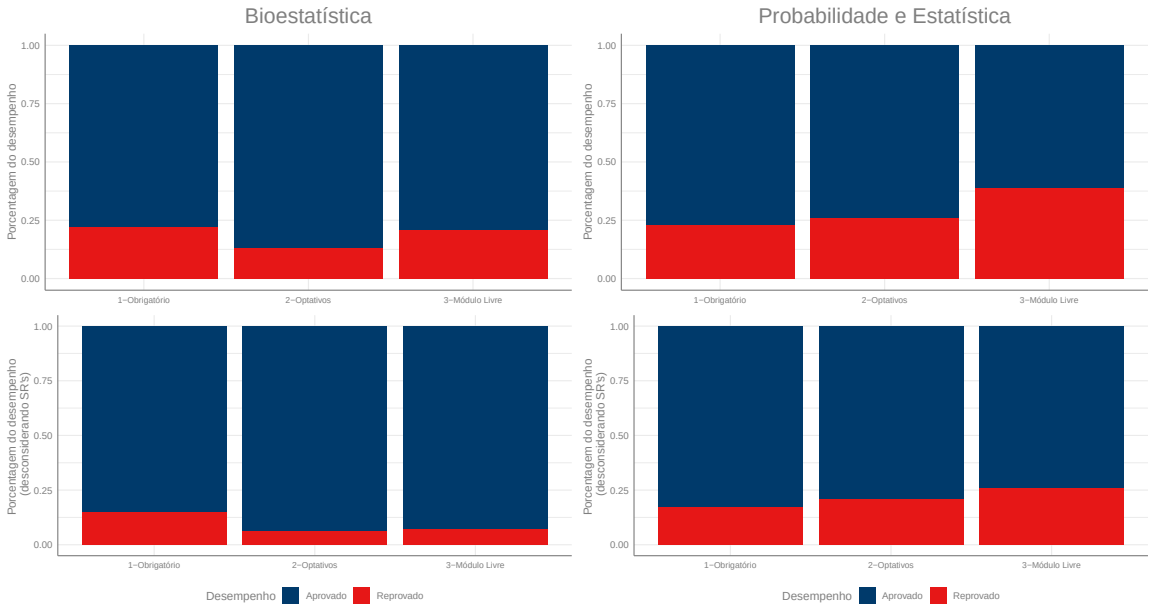
Gráfico 20: Desempenho dos estudantes segundo o professor que ministrou as disciplinas (2017-2019).



4.2.4 Modalidade das disciplinas

Para Bioestatística, segundo o Gráfico 21, tanto a análise da base completa como desconsiderando os alunos sem rendimento, a maior taxa de reprovação foi nos cursos que ofertam a disciplina como obrigatória. Entretanto, para Probabilidade e Estatística, a maior taxa de reprovação, em ambas análises, foram nos cursos onde a disciplina não é ofertada.

Gráfico 21: Desempenho dos estudantes segundo a modalidade das disciplinas (2017-2019).

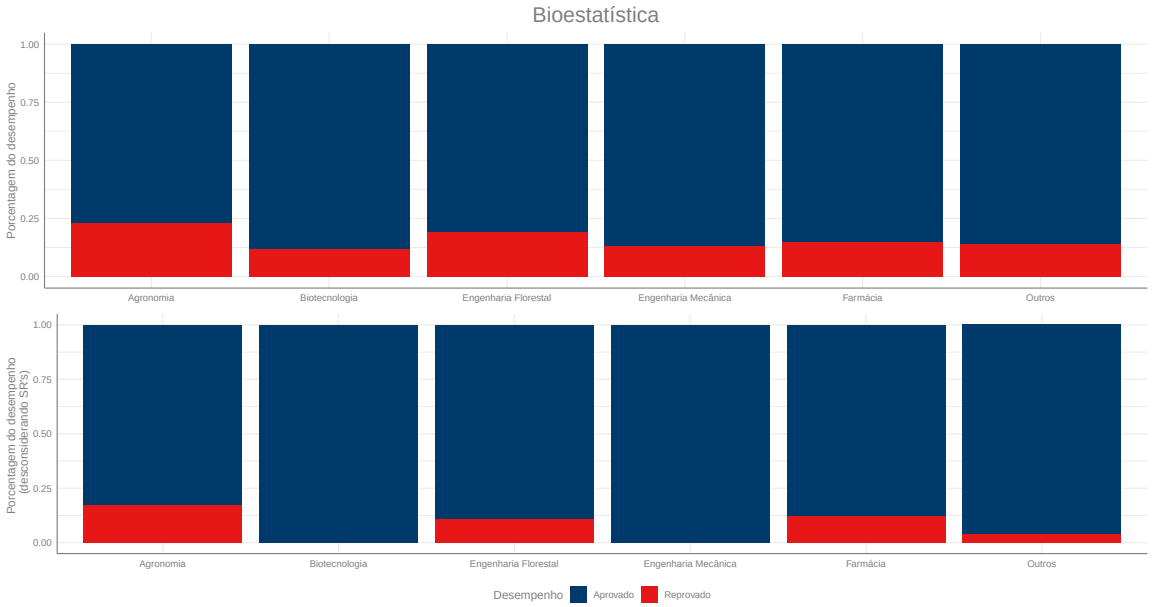


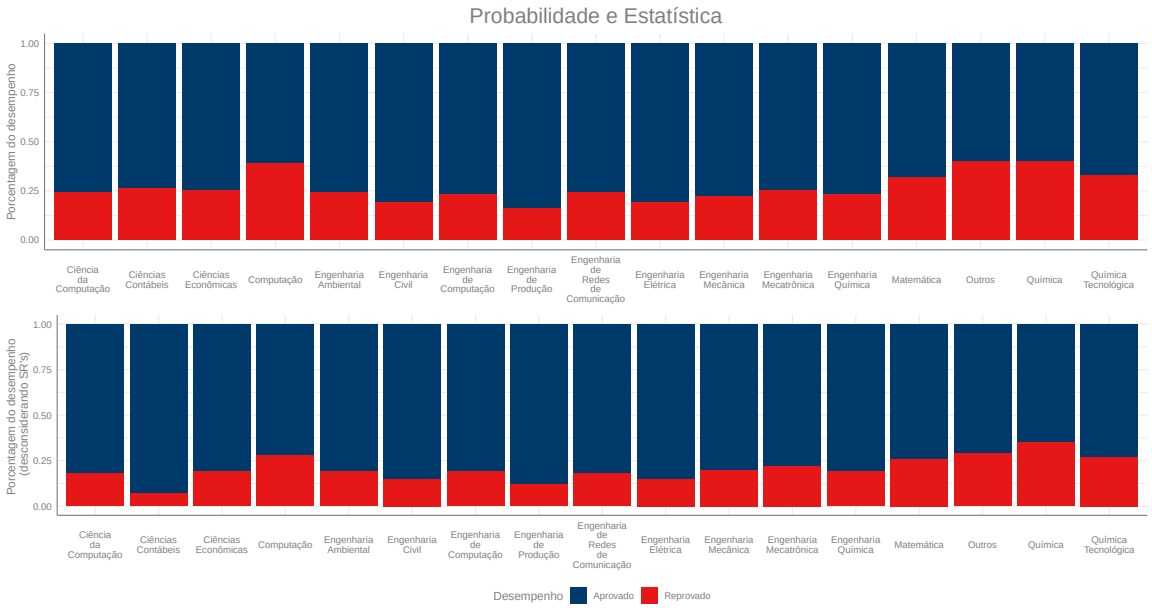
4.2.5 Cursos

As maiores proporções de reprovados para Bioestatística são encontradas nos cursos onde ela é obrigatória. Como mostra o gráfico 22, os estudantes das disciplinas Biotecnologia e Engenharia Mecânica tiveram toda a proporção de reprovados devido à aprovação de 100% quando se ignora a menção. Por outro lado, Probabilidade e Estatística apresentaram um desempenho semelhante nos dois cenários, com exceção de Ciências Contábeis que teve um aumento abrupto na aprovação. Isso sugere que a maior proporção de estudantes com menção SR está no curso de Ciências Contábeis. A categoria "Outros"[6] foi determinada para os cursos que apresentavam as quantidades médias de alunos menos relevantes.

[6]Consulte o apêndice.

Gráfico 22: Desempenho dos estudantes segundo os cursos que a quantidade média de alunos é mais relevante (2017-2019).



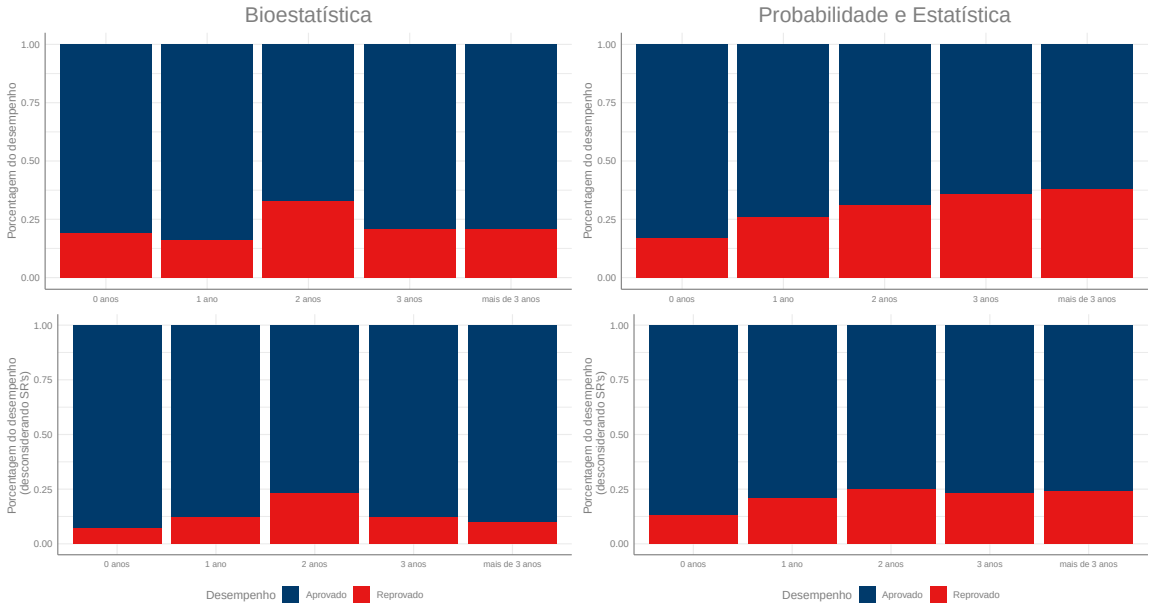


4.2.6 Tempo do estudante na UnB

Por fim, o Gráfico 23 mostra que, para a disciplina de Bioestatística, a maior taxa de reprovação, nos dois conjuntos de dados, é dos estudantes que estão há 2 anos na Universidade. Além disso, é possível verificar que o comportamento de reprovação da disciplina de Probabilidade e Estatística, considerando a menção SR, aumenta conforme o tempo de permanência na Universidade. No entanto, desconsiderando a menção SR, esse comportamento não persiste e a maior taxa de reprovação passa a ser dos estudan-

tes que ingressaram há dois anos na UnB, exatamente como acontece na disciplina de Bioestatística.

Gráfico 23: Desempenho dos estudantes segundo o tempo, em anos inteiros, que o estudante está na UnB até a primeira tentativa nas disciplinas (2017-2019).



4.3 Testes qui-quadrado

Para realizar a comparação entre a variável resposta e as variáveis explicativas das duas disciplinas no período de 2017 a 2019, consideram-se as hipóteses abaixo:

$$\begin{cases} H_0) \text{ Não há associção entre as variáveis,} \\ H_1) \text{ Há associção entre as variáveis.} \end{cases}$$

De acordo com as Tabelas 1 e 2, para ambas disciplinas, os p-valores obtidos através do teste Qui-quadrado apresentaram valores pequenos para a maioria das variáveis. Portanto, com o nível de significância de 5%, existem indícios estatísticos para rejeitar a hipótese nula, ou seja, há evidências de que existe uma associação entre as variáveis explicativas e o desempenho dos estudantes. O p-valor da variável "Turno"para Probabilidade e Estatística superou o nível de significância nas duas análises da matéria, o que indica que não há associação entre o turno em que a disciplina é ofertada e o desempenho dos alunos.

Tabela 1: Resultados dos testes qui-quadrado entre a variável resposta e as explicativas da disciplina Bioestatística.

Teste qui-quadrado entre as variáveis resposta e as variáveis explicativas								
	Caso geral				Desconsiderando SR			
Variável explicativa	**% de aprovação**	**Estatística do teste**	**df**	**p-valor**	**% de aprovação**	**Estatística do teste**	**df**	**p-valor**
Faltas								
De 0% a 5%	95,45%				95,45%			
De 5% a 10%	86,67%				86,67%			
De 10% a 15%	85,81%	336,07	5	<0,001	85,81%	33,623	4	<0,001
De 15% a 20%	73,91%				73,91%			
De 20% a 25%	76,86%				76,86%			
Mais de 25%	0%				0%			
Turno								
Manhã	80,09%	13,923	1	<0,001	88,58%	18,926	1	<0,001
Tarde	66,48%				74,52%			
Professor								
A	82,22%				91,36%			
B	94,16%				95,56%			
C	56,45%	61,975	6	<0,001	64,81%	49,124	6	<0,001
D	44,83%				59,09%			
E	73,05%				82,73%			
F	84,78%				91,76%			
G	76,97%				86,67%			
Modalidade								
Obrigatória	74,55%	9,744	2	0,0076	80,52%	14,277	2	<0,001
Optativa	85,15%				77,21%			
Módulo livre	78,79%				72,17%			
Cursos								
Agronomia	72,66%				80,17%			
Biotecnologia	81,48%				100,00%			
Engenharia Florestal	78,04%	11,825	5	0,0372	87,43%	25,054	2	<0,001
Farmácia	82,76%				100,00%			
Engenharia Mecânica	86,67%				86,75%			
Outros	85,53%				95,59%			
Tempo na UnB								
0 anos	76,47%				92,86%			
1 ano	83,33%	20,686	4	<0,001	87,96%	11,891	4	0,0181
2 anos	65,70%				76,87%			
3 anos	78,62%				87,69%			
Mais que 3 anos	75,60%				86,99%			

Tabela 2: Resultados dos testes qui-quadrado entre a variável resposta e as explicativas da disciplina Probabilidade e Estatística.

Teste qui-quadrado entre as variáveis resposta e as variáveis explicativas								
	Caso geral				Desconsiderando SR			
Variável explicativa	% de aprovação	Estatística do teste	df	p-valor	% de aprovação	Estatística do teste	df	p-valor
Faltas								
De 0% a 5%	90,00%				90,00%			
De 5% a 10%	85,45%				85,97%			
De 10% a 15%	78,52%	891,982	5	<0,001	78,38%	175,302	4	<0,001
De 15% a 20%	68,80%				68,85%			
De 20% a 25%	60,84%				61,28%			
Mais de 25%	0,00%				0,00%			
Turno								
Manhã	69,33%	4,450	2	0,108	76,82%	5,643	2	0,059
Tarde	74,43%				79,45%			
Noite	72,20%				83,16%			
Professor								
A	51,06%				61,94%			
B	81,61%				86,67%			
C	79,52%				84,62%			
D	79,88%				82,91%			
E	68,98%				76,02%			
F	65,99%	82,361	12	<0,001	74,55%	62,057	12	<0,001
G	77,86%				86,42%			
H	71,90%				82,71%			
I	71,02%				76,82%			
J	61,96%				71,25%			
K	72,00%				77,59%			
L	64,47%				76,56%			
M	75,34%				81,48%			
Modalidade								
Obrigatória	73,40%	13,400	2	0,001	80,52%	6,061	2	0,048
Optativa	70,05%				77,21%			
Módulo livre	59,29%				72,17%			
Cursos								
Ciência da Computação	70,94%				79,05%			
Ciências Contábeis	74,29%				92,86%			
Ciências Econômicas	69,85%				77,74%			
Computação	59,38%				70,37%			
Engenharia Ambiental	72,63%				78,41%			
Engenharia Civil	76,23%				84,16%			
Engenharia de Computação	71,18%				78,37%			
Engenharia de Produção	81,25%	45,315	16	<0,001	87,56%	37,27	16	0,001
Engenharia de Redes de Comunicação	72,57%				80,38%			
Engenharia Elétrica	78,01%				83,24%			
Engenharia Mecânica	74,75%				79,57%			
Engenharia Mecatrônica	70,18%				75,47%			
Engenharia Química	72,64%				79,38%			
Matemática	63,89%				73,02%			
Outros	58,97%				70,23%			
Química	60,32%				66,67%			
Química Tecnológica	65,09%				73,40%			
Tempo na UnB								
0 anos	82,23%				92,86%			
1 ano	71,24%	61,184	4	<0,001	87,96%	32,127	4	<0,001
2 anos	65,98%				76,87%			
3 anos	62,55%				87,69%			
Mais que 3 anos	60,26%				86,99%			

5 Modelagem

O processo de modelagem dos dados considera apenas as variáveis explicativas que possuem associação com a variável resposta. Portanto, para o mesmo período analisado, de 2017 até 2019, as variáveis explicativas que possuem essa condição para Bioestatística são: as faixas de faltas dos estudantes (referência: Faltas [1- 0% a 5%]), o turno em que a disciplina foi ofertada (referência: Manhã), qual foi o professor que aplicou a matéria (referência: professor A), a modalidade da disciplina (referência: Obrigatório), os cursos que a quantidade média de alunos é mais relevante (referência para Bioestatística: Agronomia e para Probabilidade e Estatística: Ciência da Computação) e o tempo na Universidade do aluno. Para Probabilidade e Estatística foi desconsiderada apenas o turno em que foi ofertada pois essa variável não apresentou associação significativa.

Foi aplicada uma Regressão Logística Multinível para a análise, pois a variável resposta é binária, o "sucesso"é determinado pela aprovação do aluno, e existe uma estrutura hierárquica de dados formada por dois níveis principais: os alunos e as turmas. Essa abordagem nos permite levar em consideração as particularidades de cada nível hierárquico para prever o desempenho acadêmico dos alunos.

Por fim, optou-se por analisar os casos com a base completa, ou seja, considerando as menções SR, e desenvolver um outro modelo desconsiderando essa menção.

5.1 Modelagem caso geral

Nessa modelagem, foi desconsiderada a variável "faixas de faltas"dos estudantes, pois apresenta uma autocorrelação com a menção SR. Isso ocorre pois a principal condição para obter a menção SR é o estudante apresentar uma porcentagem de faltas superior a 25%. Para Bioestatística, foram analisados 844 alunos e 18 turmas, e para Probabilidade e Estatística, foram analisados 2866 alunos e 62 turmas.

Os resultados obtidos pela Tabela 3 indicam que houve diferenças significativas entre o desempenho dos estudantes entre o curso Agronomia e Engenharia Florestal, os outros cursos não apresentaram diferenças significativas, apesar de não considerado pelo nível de significância, o curso Farmácia seguirá no modelo para análise. Os demais resultados das outras variáveis dos efeitos fixos foram significativos. Além disso, foram verificados as correlações das variáveis dos efeitos aleatórios e apresentaram correlações relativamente altas.

A correlação intraclasse para Bioestatística deu 17%, ou seja, levando em consideração que 36% é um nível alto, a correlação apresentou um valor intermediário. E ainda, os coeficientes de determinação R^2 marginal e condicional para esse modelo foram, 0,066 e 0,226, respectivamente. Isso significa que o modelo completo, tanto com os efeitos fixos como com os efeitos aleatórios, explica cerca de 23% da variância dos dados, enquanto que o modelo desconsiderando os efeitos aleatórios explica cerca de 7% da mesma variância. No presente trabalho, é apresentado o resultado do modelo multinível considerando o efeito aleatório.

Tabela 3: Resultados do modelo geral testado com as variáveis de referência para Bioestatística.

Bioestatística				
Efeitos fixos [aluno]	**Valor**	**Erro padrão**	**Z**	**P-valor**
Intercepto	-1,03	0,20	-5,12	0,00
Modalidade [Módulo Livre]	-1,10	0,47	-2,34	0,02
Modalidade [Optativos]	-1,93	0,56	-3,42	0,00
Tempo na UnB	0,16	0,06	2,59	0,01
Curso [Biotecnologia]	0,90	0,75	1,20	0,23
Curso [Engenharia Florestal]	-0,58	0,22	-2,62	0,01
Curso [Engenharia Mecânica]	0,18	0,71	0,25	0,80
Curso [Farmácia]	1,10	0,62	1,76	0,08
Efeitos aleatórios [turma]	**Variância**	**Desvio padrão**		
Intercepto	0,30	0,55		
Professor [B]	4,62	2,15		
Professor [C]	0,64	0,80		
Professor [D]	1,53	1,24		
Professor [E]	0,30	0,55		
Professor [F]	0,23	0,48		
Professor [G]	0,43	0,66		
Turno [Tarde]	0,14	0,38		

Já os resultados obtidos pela Tabela 4 indicam que não houve diferenças significativas entre o desempenho dos estudantes da modalidade Obrigatória em relação as outras modalidades, portanto a variável Modalidade será retirada para dar continuidade a modelagem. No entanto, houve diferenças significativas para a variável Tempo na UnB. Além disso, houve diferenças significativas entre o desempenho dos estudantes do curso de Ciência da Computação com o curso Engenharia de Produção e Engenharia Ambiental, os outros cursos não apresentaram diferenças significativas e foram desconsiderados do modelo. Foram verificados as correlações das variáveis dos efeitos aleatórios e apresentaram correlações relativamente altas.

A correlação intraclasse para Probabilidade e Estatística deu 6%, ou seja, a correlação apresentou um valor fraco. Porém, os coeficientes de determinação R^2 marginal

e condicional para esse modelo foram, 0,044 e 0,102, respectivamente. Isso significa que o modelo completo, tanto com os efeitos fixos como com os efeitos aleatórios, explica cerca de 10% da variância dos dados, enquanto que o modelo desconsiderando os efeitos aleatórios explica cerca de 4% da mesma variância. No presente trabalho, é apresentado o resultado do modelo multinível considerando o efeito aleatório.

Tabela 4: Resultados do modelo geral testado com as variáveis de referência para Probabilidade e Estatística.

Probabilidade e Estatística				
Efeitos fixos [aluno]	**Valor**	**Erro padrão**	**Z**	**P-valor**
Intercepto	-1,39	0,18	-7,60	0,00
Modalidade [Módulo Livre]	-0,06	0,46	-0,13	0,90
Modalidade [Optativos]	0,15	0,27	0,54	0,59
Tempo na UnB	0,29	0,04	6,96	0,00
Curso [Ciências Contábeis]	-0,63	0,44	-1,43	0,15
Curso [Ciências Econômicas]	0,12	0,22	0,56	0,58
Curso [Computação]	0,51	0,48	1,05	0,29
Curso [Engenharia Ambiental]	-0,72	0,32	-2,21	0,03
Curso [Engenharia Civil]	-0,12	0,28	-0,43	0,67
Curso [Engenharia de Computação]	0,06	0,27	0,23	0,82
Curso [Engenharia de Produção]	-0,51	0,25	-2,05	0,04
Curso [Engenharia de Redes de Comunicação]	0,22	0,25	0,90	0,37
Curso [Engenharia Elétrica]	-0,26	0,25	-1,04	0,30
Curso [Engenharia Mecânica]	-0,26	0,27	-0,95	0,34
Curso [Engenharia Mecatrônica]	0,29	0,23	1,27	0,20
Curso [Engenharia Química]	0,03	0,28	0,12	0,91
Curso [Matemática]	0,11	0,33	0,33	0,74
Curso [Outros]	0,24	0,30	0,80	0,42
Curso [Química]	0,04	0,34	0,12	0,90
Efeitos aleatórios [turma]	**Variância**	**Desvio padrão**		
Intercepto	0,78	0,88		
Professor [B]	0,41	0,64		
Professor [C]	0,47	0,69		
Professor [D]	0,39	0,63		
Professor [E]	0,78	0,88		
Professor [F]	0,26	0,51		
Professor [G]	0,26	0,51		
Professor [H]	0,78	0,88		
Professor [I]	0,63	0,80		
Professor [J]	0,43	0,66		
Professor [K]	0,78	0,88		
Professor [L]	0,78	0,88		
Professor [M]	0,78	0,88		

5.1.1 Validação do modelo

Para testar se os resultados obtidos no modelo geral podem ser considerados representativos, realizou-se o processo de modelagem para amostras do banco de dados, separando-as em dados de teste e de validação do modelo. A regra de decisão foi de 70% para a amostra de teste, e 30% para a validação. Então, para Bioestatística, na amostra de teste foram consideradas 590 alunos, consequentemente na de validação foram considerados os 254 restantes. Já para Probabilidade e Estatística, na amostra de teste foram consideradas 2005 alunos e na de validação os 861 alunos restantes.

Os resultados obtidos a partir dos dados de teste e de validação de ambas as tabelas 5 e 6 indicam que os valores observados certamente estabelecem padrões semelhantes entre si.

Tabela 5: Coeficientes dos modelos de teste e de validação testados com as variáveis de referência para Bioestatística.

Bioestatística								
	Amostra teste				Amostra validação			
Efeitos fixos [aluno]	**Valor**	**Erro padrão**	**Z**	**P-valor**	**Valor**	**Erro padrão**	**Z**	**P-valor**
Intercepto	-1,02	0,24	-4,27	0,00	-1,00	0,37	-2,73	0,01
Modalidade [Módulo Livre]	-0,81	0,53	-1,53	0,13	-1,69	1,17	-1,44	0,15
Modalidade [Optativos]	-1,80	0,42	-4,34	0,00	-1,60	0,59	-2,71	0,01
Tempo na UnB	0,16	0,07	2,32	0,02	0,13	0,12	1,15	0,25
Curso [Engenharia Florestal]	-0,72	0,27	-2,69	0,01	-0,47	0,40	-1,16	0,25
Curso [Farmácia]	0,84	0,52	1,62	0,11	0,57	0,82	0,70	0,48
Efeitos aleatórios [turma]	**Variância**	**Desvio padrão**			**Variância**	**Desvio padrão**		
Intercepto	0,17	0,42			0,00	0,00		
Professor [B]	4,70	2,17			11,19	3,35		
Professor [C]	0,78	0,88			0,38	0,62		
Professor [D]	0,35	0,59			6,29	2,51		
Professor [E]	0,17	0,42			0,54	0,74		
Professor [F]	0,17	0,42			3,36	1,83		
Professor [G]	0,39	0,63			0,00	0,00		
Turno [Tarde]	0,21	0,46			0,54	0,74		

Tabela 6: Coeficientes dos modelos de teste e de validação testados com as variáveis de referência para Probabilidade e Estatística.

Probabilidade e Estatística								
	Amostra teste				Amostra validação			
Efeitos fixos [aluno]	**Valor**	**Erro padrão**	**Z**	**P-valor**	**Valor**	**Erro padrão**	**Z**	**P-valor**
Intercepto	-1,46	0,21	-6,89	0,00	-1,28	0,30	-4,22	0,00
Tempo na UnB	0,33	0,05	7,00	0,00	0,19	0,06	3,05	0,00
Curso [Engenharia Ambiental]	-0,50	0,35	-1,43	0,15	-0,75	0,60	-1,26	0,21
Curso [Engenharia de Produção]	-0,57	0,29	-1,92	0,05	-0,51	0,44	-1,16	0,24
Efeitos aleatórios [turma]	**Variância**	**Desvio padrão**			**Variância**	**Desvio padrão**		
Intercepto	1,35	1,16			0,27	0,52		
Professor [B]	0,57	0,75			0,27	0,52		
Professor [C]	0,67	0,82			0,08	0,29		
Professor [D]	0,35	0,59			0,07	0,27		
Professor [E]	0,92	0,96			0,12	0,34		
Professor [F]	0,73	0,86			0,06	0,25		
Professor [G]	0,68	0,83			0,29	0,54		
Professor [H]	1,35	1,16			0,19	0,44		
Professor [I]	0,80	0,89			0,26	0,51		
Professor [J]	1,02	1,01			1,05	1,02		
Professor [K]	1,35	1,16			0,27	0,52		
Professor [L]	0,67	0,82			0,27	0,52		
Professor [M]	1,35	1,16			0,27	0,52		

O modelo de predição por regressão logística foi aplicado à amostra de teste, de ambas disciplinas, gerando coeficientes que foram então comparados aos resultados verdadeiros obtidos na amostra de validação. Essa comparação revelou que os resultados obtidos são estatisticamente significativos, indicando que o modelo é válido para a amostra. Além disso, isto também significa que as variáveis preditoras do modelo são significativas para a previsão dos resultados.

De acordo com as Tabelas 7 e 8, verificou-se que a taxa de acerto de Bioestatística e de Probabilidade e Estatística foi de 73,05% e 62,64%, respectivamente. Essas porcentagens indicaram que os valores observados parece ser relativamente satisfatório. Como resultado, o banco de dados completo foi usado para realizar a modelagem final dos dados.

Tabela 7: Resultados da validação do modelo de teste para Bioestatística.

Validação do modelo teste			
Estimado/ Observado	**Aprovado**	**Reprovado**	**Total**
Aprovado	375	78	453
Reprovado	81	56	137
Total	456	134	590

Tabela 8: Resultados da validação do modelo de teste para Probabilidade e Estatística.

Validação do modelo teste			
Estimado/ Observado	Aprovado	Reprovado	Total
Aprovado	987	305	1292
Reprovado	444	269	713
Total	1.431	574	2.005

5.1.2 Razão de chances

Analisando a Tabela 9, se o estudante cursar a matéria como Optativas e Módulo Livre a chance de ser aprovado em sua primeira tentativa são reduzidas para 0,34 e 0,20 vezes, respectivamente a chance de aprovação de alunos que cursam a disciplina como Obrigatória. Além isso, o aumento de um ano do aluno na Universidade leva ao um aumento de 1,16 na chance de aprovação na sua vez cursando. Por fim, a chance do aluno de Engenharia Florestal ser aprovado é 0,56 vezes a chance do aluno de Agronomia ser aprovado.

Tabela 9: Resultados da razão de chances do modelo geral para Bioestatística.

Bioestatística			
Efeitos fixos [aluno]	Razão de chances	IC [95%]	P-valor
Intercepto	0,36	0,25 – 0,54	<0,001
Modalidade [Módulo Livre]	0,34	0,13 – 0,85	0,02
Modalidade [Optativos]	0,20	0,10 – 0,38	<0,001
Tempo na UnB	1,16	1,04 – 1,31	0,01
Curso [Engenharia Florestal]	0,56	0,36 – 0,86	0,01
Curso [Farmácia]	2,17	0,94 – 5,03	0,07

E ainda, analisando a Tabela 10, o aumento de um ano do aluno na Universidade leva ao um aumento de 1,34 na chance de aprovação na sua primeira vez cursando. Por fim, a chance do aluno de Engenharia Ambiental e Engenharia de Produção ser aprovado é 0,57 e 0,58 vezes, respectivamente, a chance do aluno de Ciência da Computação ser aprovado.

Tabela 10: Resultados da razão de chances do modelo geral para Probabilidade e Estatística.

Probabilidade e Estatística			
Efeitos fixos	**Razão de chances**	**IC [95%]**	**P-valor**
Intercepto	0,25	0,18 – 0,35	<0,001
Tempo na UnB	1,34	1,24 – 1,44	<0,001
Curso [Engenharia Ambiental]	0,57	0,31 – 1,02	0,059
Curso [Engenharia de Produção]	0,58	0,36 – 0,95	0,031

5.1.3 Diagnóstico do modelo

- Resíduos

Ao analisar ambas as disciplinas, observou-se nos Gráficos 24 e 25 que o gráfico dos resíduos oscila entre o zero, o que indica que o modelo está ajustado de forma adequada aos dados.

Gráfico 24: Gráfico de resíduos e valores ajustados do modelo geral para Bioestatística.

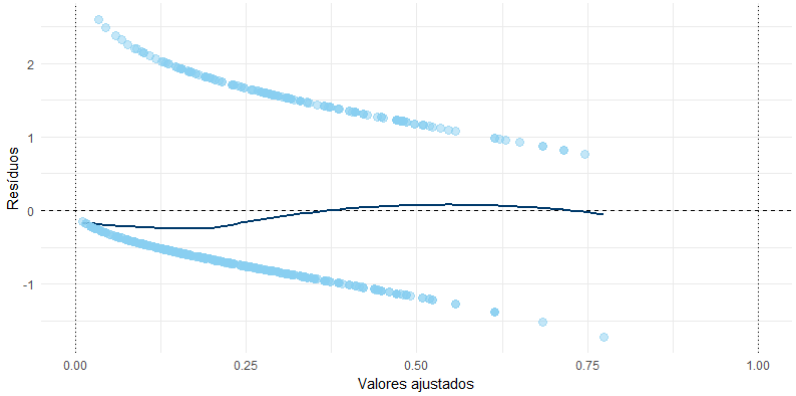
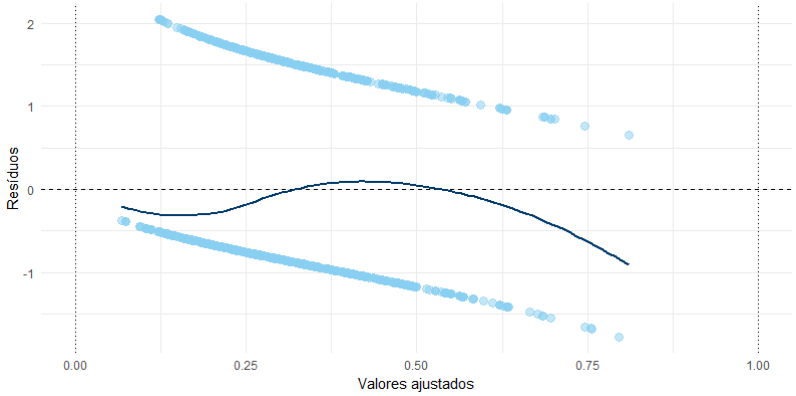


Gráfico 25: Gráfico de resíduos e valores ajustados do modelo geral para Probabilidade e Estatística.



- Curva ROC

Os gráficos 26 e 27 mostram que o modelo preditivo é bastante confiável, como a área sob a curva é relativamente grande para ambas disciplinas. Valores de 0,742 e 0,682 foram considerados como um ponto de corte apropriado na seleção da regra para obter resultados mais precisos.

Gráfico 26: Gráfico da curva de ROC do modelo geral para Bioestatística.

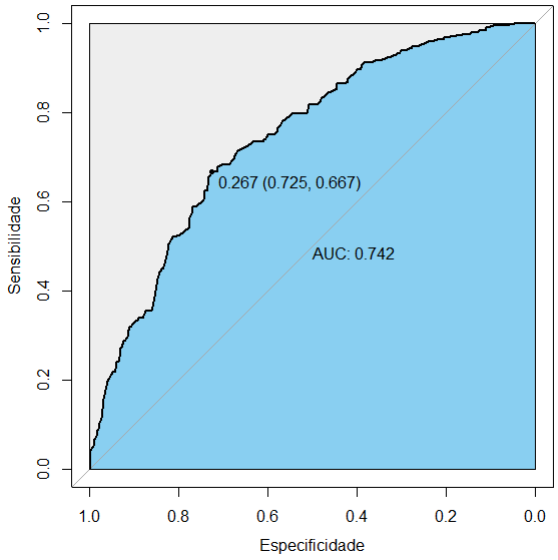
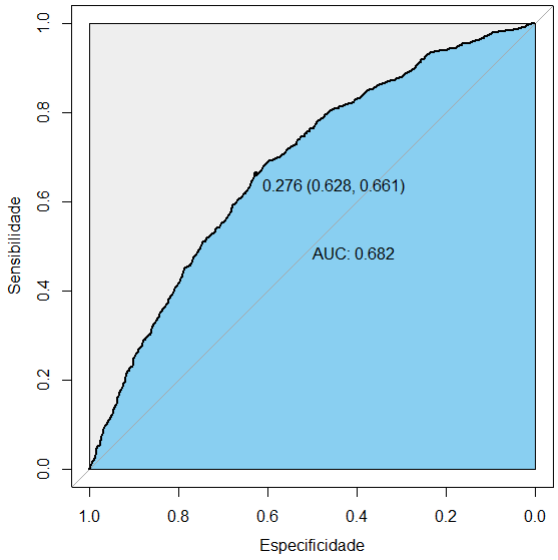


Gráfico 27: Gráfico da curva de ROC do modelo geral para Probabilidade e Estatística.



Por fim, analisando os resultados da matriz de confusão, tem-se que os resultados obtidos nas Tabelas 11 e 12 indicam que os modelos de Bioestatística e Probabilidade e

Estatística têm uma boa acurácia, aproximadamente 74% e 67%, respectivamente. Além disso, ambos também apresentaram sensibilidades consideravelmente elevadas, 85% e 79%, respectivamente, e especificidades aproximadas de 44% e 43%, respectivamente.

Tabela 11: Resultados da matriz de confusão do modelo geral para Bioestatística.

Matriz de confusão			
Estimado/ Observado	**Aprovado**	**Reprovado**	**Total**
Aprovado	527	125	652
Reprovado	92	100	192
Total	619	225	844

Medidas	**Valores**
Acurácia	0,7429
IC [95% para acurácia]	(0,712, 0,7721)
Sensibilidade	0,8514
Especificidade	0,4444

Tabela 12: Resultados da matriz de confusão do modelo geral para Probabilidade e Estatística.

Matriz de confusão			
Estimado/ Observado	**Aprovado**	**Reprovado**	**Total**
Aprovado	1467	578	2045
Reprovado	381	440	821
Total	1.848	1.018	2.866

Medidas	**Valores**
Acurácia	0,6654
IC [95% para acurácia]	(0,6478, 0,6827)
Sensibilidade	0,7938
Especificidade	0,4322

5.2 Modelagem desconsiderando menção SR

Nessa modelagem, foram retirados os casos dos alunos que receberam "faixas de faltas"dos estudantes. Para Bioestatística, foram analisados 761 alunos e 18 turmas, e para Probabilidade e Estatística, foram analisados 2594 alunos e 62 turmas.

Os resultados obtidos pela Tabela 13 indicam que houve diferenças significativas entre o desempenho dos estudantes entre o curso Agronomia e Engenharia Florestal, os outros cursos não apresentaram diferenças significativas, apesar de não considerado pelo nível de significância, o curso Farmácia seguirá no modelo para análise. Além disso, não houve diferenças significativas para a variável Tempo na UnB, portanto a variável será retirada do modelo para dar seguimento a modelagem. Os demais resultados das outras variáveis dos efeitos fixos foram significativos. Foram verificados as correlações das variáveis dos efeitos aleatórios e apresentaram correlações relativamente altas.

A correlação intraclasse para Bioestatística deu 12%, ou seja, a correlação apresentou um valor intermediário. E ainda, os coeficientes de determinação R^2 marginal e condicional para esse modelo foram, 0,854 e 0,872, respectivamente. Isso significa que o modelo completo, tanto com os efeitos fixos como com os efeitos aleatórios, explica cerca de 87% da variância dos dados, enquanto que o modelo desconsiderando os efeitos aleatórios explica cerca de 85% da mesma variância. No presente trabalho, é apresentado o resultado do modelo multinível considerando o efeito aleatório.

Tabela 13: Resultados do modelo geral testado com as variáveis de referência para Bioestatística.

Bioestatística				
Efeitos fixos [aluno]	**Valor**	**Erro padrão**	**Z**	**P-valor**
Intercepto	-2,72	0,43	-6,27	0,00
Modalidade [Módulo Livre]	-1,85	0,79	-2,34	0,02
Modalidade [Optativos]	-2,77	1,06	-2,62	0,01
Tempo na UnB	0,02	0,08	0,23	0,82
Curso [Biotecnologia]	-14,58	1281,25	-0,01	0,99
Curso [Engenharia Florestal]	-0,74	0,28	-2,64	0,01
Curso [Engenharia Mecânica]	-14,99	940,62	-0,02	0,99
Curso [Farmácia]	2,11	1,11	1,90	0,06
Faltas [2 - 5% a 10%]	1,22	0,44	2,80	0,01
Faltas [3 - 10% a 15%]	1,39	0,46	3,02	0,00
Faltas [4 - 15% a 20%]	2,12	0,48	4,44	0,00
Faltas [5 - 20% a 25%]	2,04	0,46	4,42	0,00
Efeitos aleatórios [turma]	**Variância**	**Desvio padrão**		
Intercepto	0,00	0,00		
Professor [B]	0,72	0,85		
Professor [C]	0,67	0,82		
Professor [D]	4,33	2,08		
Professor [E]	0,20	0,44		
Professor [F]	0,36	0,60		
Professor [G]	0,00	0,00		
Turno [Tarde]	0,26	0,51		

Pela Tabela 14 indicam que não houve diferenças significativas entre o desempenho dos estudantes da modalidade Obrigatória em relação as outras modalidades, então a variável será retirada do modelo. No entanto, houve diferenças significativas para a variável Tempo na UnB. Além disso, considerando o nível de significância de 10%, houve diferenças significativas entre o desempenho dos estudantes do curso de Ciência da Computação apenas com o curso de Ciências Contábeis, Computação e Engenharia de Produção, os outros cursos não apresentaram diferenças significativas. Foram verificados as correlações das variáveis dos efeitos aleatórios e apresentaram correlações relativamente altas.

A correlação intraclasse para Probabilidade e Estatística deu 5%, ou seja, a correlação apresentou um valor fraco. Porém, os coeficientes de determinação R^2 marginal e condicional para esse modelo foram, 0,036 e 0,082, respectivamente.Isso significa que o modelo completo, tanto com os efeitos fixos como com os efeitos aleatórios, explica cerca de 8% da variância dos dados, enquanto que o modelo desconsiderando os efeitos aleatórios explica cerca de 4% da mesma variância. No presente trabalho, é apresentado o resultado do modelo multinível considerando o efeito aleatório.

Tabela 14: Resultados do modelo geral testado com as variáveis de referência para Probabilidade e Estatística.

Probabilidade e Estatística				
Efeitos fixos [aluno]	**Valor**	**Erro padrão**	**Z**	**P-valor**
Intercepto	-2,62	0,24	-10,83	0,00
Modalidade [Módulo Livre]	-0,43	0,57	-0,76	0,45
Modalidade [Optativos]	0,18	0,32	0,55	0,59
Tempo na UnB	0,12	0,05	2,35	0,02
Curso [Ciências Contábeis]	-1,36	0,78	-1,75	0,08
Curso [Ciências Econômicas]	0,20	0,26	0,76	0,45
Curso [Computação]	1,03	0,60	1,73	0,08
Curso [Engenharia Ambiental]	-0,48	0,37	-1,27	0,20
Curso [Engenharia Civil]	-0,16	0,33	-0,47	0,64
Curso [Engenharia de Computação]	0,13	0,32	0,40	0,69
Curso [Engenharia de Produção]	-0,55	0,31	-1,80	0,07
Curso [Engenharia de Redes de Comunicação]	0,17	0,29	0,57	0,57
Curso [Engenharia Elétrica]	-0,11	0,30	-0,36	0,72
Curso [Engenharia Mecânica]	-0,22	0,32	-0,68	0,50
Curso [Engenharia Mecatrônica]	0,40	0,27	1,49	0,14
Curso [Engenharia Química]	0,05	0,33	0,16	0,87
Curso [Matemática]	0,12	0,40	0,31	0,76
Curso [Outros]	0,39	0,34	1,13	0,26
Curso [Química]	0,36	0,39	0,92	0,36
Falta [2 - 5% a 10%]	0,35	0,19	1,83	0,07
Falta [3 - 10% a 15%]	1,02	0,17	6,19	0,00
Falta [4 - 15% a 20%]	1,51	0,17	8,72	0,00
Falta [5 - 20% a 25%]	1,85	0,17	11,16	0,00
Efeitos aleatórios [turma]	**Variância**	**Desvio padrão**		
Intercepto	0,47	0,69		
Professor [B]	0,47	0,69		
Professor [C]	0,31	0,56		
Professor [D]	0,52	0,72		
Professor [E]	0,53	0,73		
Professor [F]	0,57	0,75		
Professor [G]	0,36	0,60		
Professor [H]	0,47	0,69		
Professor [I]	0,94	0,97		
Professor [J]	0,27	0,52		
Professor [K]	0,47	0,69		
Professor [L]	0,47	0,69		
Professor [M]	0,47	0,69		

5.2.1 Validação do modelo

Para testar se os resultados obtidos no modelo geral podem ser considerados representativos, realizou-se o processo de modelagem para amostras do banco de dados, separando-as em dados de teste e de validação do modelo. A regra de decisão foi de 75% para a amostra de teste, e 25% para a validação. Então, para Bioestatística, na amostra de teste foram consideradas 570 alunos, consequentemente na de validação foram

considerados os 191 restantes. Já para Probabilidade e Estatística, na amostra de teste foram consideradas 1556 alunos e na de validação os 589 alunos restantes.

Os resultados obtidos a partir dos dados de teste e de validação de ambas as tabelas 15 e 16 indicam que os valores observados certamente estabelecem padrões semelhantes entre si.

Tabela 15: Coeficientes dos modelos de teste e de validação testados com as variáveis de referência para Bioestatística.

Bioestatística								
	Amostra teste				Amostra validação			
Efeitos fixos [aluno]	**Valor**	**Erro padrão**	**Z**	**P-valor**	**Valor**	**Erro padrão**	**Z**	**P-valor**
Intercepto	-2,75	0,44	-6,30	0,00	-3,67	0,95	-3,87	0,00
Modalidade [Módulo Livre]	-1,44	0,80	-1,79	0,07	-19,61	5621,22	0,00	1,00
Modalidade [Optativos]	-3,42	1,04	-3,29	0,00	-19,14	2834,12	-0,01	0,99
Curso [Engenharia Florestal]	-0,97	0,34	-2,85	0,00	-0,06	0,54	-0,11	0,91
Curso [Farmácia]	2,91	1,09	2,66	0,01	18,48	2834,12	0,01	0,99
Faltas [2 - 5% a 10%]	1,22	0,48	2,54	0,01	1,91	1,01	1,88	0,06
Faltas [3 - 10% a 15%]	1,45	0,51	2,85	0,00	2,12	1,07	1,97	0,05
Faltas [4 - 15% a 20%]	2,24	0,52	4,30	0,00	2,65	1,10	2,41	0,02
Faltas [5 - 20% a 25%]	1,81	0,51	3,55	0,00	3,54	1,04	3,40	0,00
Efeitos aleatórios [turma]	**Variância**	**Desvio padrão**			**Variância**	**Desvio padrão**		
Intercepto	0,00	0,00			0,00	0,00		
Professor [B]	0,26	0,51			0,00	0,00		
Professor [C]	0,56	0,75			0,00	0,00		
Professor [D]	3,32	1,82			12,02	3,47		
Professor [E]	0,00	0,00			0,00	0,00		
Professor [F]	0,55	0,74			0,00	0,00		
Professor [G]	0,00	0,00			0,00	0,00		
Turno [Tarde]	0,52	0,72			0,00	0,00		

Tabela 16: Coeficientes dos modelos de teste e de validação testados com as variáveis de referência para Probabilidade e Estatística.

Probabilidade e Estatística								
	Amostra teste				Amostra validação			
Efeitos fixos [aluno]	**Valor**	**Erro padrão**	**Z**	**P-valor**	**Valor**	**Erro padrão**	**Z**	**P-valor**
Intercepto	-2,28	0,29	-7,92	0,00	-3,01	0,41	-7,40	0,00
Tempo na UnB	0,12	0,06	1,92	0,06	0,09	0,07	1,34	0,18
Curso [Ciências Econômicas]	-15,61	617,53	-0,03	0,98	-0,06	0,86	-0,07	0,94
Curso [Computação]	0,28	0,41	0,69	0,49	1,13	0,57	1,97	0,05
Curso [Engenharia de Produção]	-0,87	0,37	-2,38	0,02	-0,11	0,54	-0,20	0,84
Faltas [2 - 5% a 10%]	0,17	0,25	0,65	0,51	0,51	0,28	1,81	0,07
Faltas [3 - 10% a 15%]	0,94	0,21	4,48	0,00	1,06	0,26	4,13	0,00
Faltas [4 - 15% a 20%]	1,67	0,22	7,58	0,00	1,20	0,27	4,48	0,00
Faltas [5 - 20% a 25%]	1,88	0,21	8,85	0,00	1,74	0,25	6,86	0,00
Efeitos aleatórios [turma]	**Variância**	**Desvio padrão**			**Variância**	**Desvio padrão**		
Intercepto	0,44	0,66			0,48	0,69		
Professor [B]	0,44	0,66			0,48	0,69		
Professor [C]	0,16	0,40			0,25	0,50		
Professor [D]	2,57	1,60			0,48	0,69		
Professor [E]	0,15	0,39			0,35	0,59		
Professor [F]	0,51	0,72			0,85	0,92		
Professor [G]	0,44	0,67			0,50	0,71		
Professor [H]	0,44	0,66			0,48	0,69		
Professor [I]	0,20	0,45			0,48	0,69		
Professor [J]	0,44	0,66			0,30	0,55		
Professor [K]	0,44	0,66			0,48	0,69		
Professor [L]	0,44	0,66			0,48	0,69		
Professor [M]	0,44	0,66			0,48	0,69		

O modelo de predição por regressão logística foi aplicado à amostra de teste, de ambas disciplinas, gerando coeficientes que foram então comparados aos resultados verdadeiros obtidos na amostra de validação. Essa comparação revelou que os resultados obtidos são estatisticamente significativos, indicando que o modelo é válido para a amostra. Além disso, isto também significa que as variáveis preditoras do modelo são significativas para a previsão dos resultados.

De acordo com as Tabelas 17 e 18, verificou-se que a taxa de acerto de Bioestatística e de Probabilidade e Estatística foi de 78,07% e 73,52%, respectivamente. Essas porcentagens indicaram que os valores observados parece ser relativamente satisfatório. Como resultado, o banco de dados completo foi usado para realizar a modelagem final dos dados.

Tabela 17: Resultados da validação do modelo de teste para Bioestatística.

Validação do modelo teste			
Estimado/ Observado	**Aprovado**	**Reprovado**	**Total**
Aprovado	407	43	450
Reprovado	82	38	120
Total	489	81	570

Tabela 18: Resultados da validação do modelo de teste para Probabilidade e Estatística.

Validação do modelo teste			
Estimado/ Observado	**Aprovado**	**Reprovado**	**Total**
Aprovado	1025	210	1235
Reprovado	202	119	321
Total	1.227	329	1.556

5.2.2 Razão de chances

Analisando a Tabela 19, se o estudante cursar a matéria como Optativas e Módulo Livre a chance de ser aprovado em sua primeira tentativa são reduzidas para 0,16 e 0,02 vezes, respectivamente a chance de aprovação de alunos que cursam a disciplina como Obrigatória. A chance do aluno de Engenharia Florestal e Farmácia ser aprovado é 0,48 e 22,45 vezes a chance do aluno de Agronomia ser aprovado. Por fim, o modelo revelou um resultado inesperado com relação a chance de aprovação segundo as categorias de percentual de faltas: a chance do aluno que faltou entre 5% a 10%, 10% a 15%, 15% a 20% e 20% a 25% ser aprovado é 3,41, 4,07, 8,28 e 7,85 vezes, respectivamente, a chance

do aluno que teve de 0% a 5% faltas ser aprovado.

Tabela 19: Resultados da razão de chances do modelo geral para Bioestatística.

Bioestatística			
Efeitos fixos [aluno]	**Razão de chances**	**IC [95%]**	**P-valor**
Intercepto	0,07	0,03 – 0,15	<0,001
Modalidade [Módulo Livre]	0,16	0,03 – 0,75	0,02
Modalidade [Optativos]	0,02	0,00 – 0,17	<0,001
Curso [Engenharia Florestal]	0,48	0,28 – 0,83	0,01
Curso [Farmácia]	22,45	2,73 – 185,02	0,00
Faltas [2 - 5% a 10%]	3,41	1,45 – 7,98	0,01
Faltas [3 - 10% a 15%]	4,07	1,66 – 10,01	0,00
Faltas [4 - 15% a 20%]	8,28	3,28 – 20,95	<0,001
Faltas [5 - 20% a 25%]	7,85	3,21 – 19,23	<0,001

E ainda, analisando a Tabela 20, o aumento de um ano do aluno na Universidade leva ao um aumento de 1,12 na chance de aprovação na sua primeira vez cursando. A chance do aluno de Ciências Contábeis, Computação e Engenharia de Produção ser aprovado é 0,26, 1,81 e 0,58, respectivamente, vezes a chance do aluno de Ciência da Computação ser aprovado. Por fim, assim como em Bioestatística, o modelo apresentou um resultado não esperado: a chance do aluno que faltou entre 5% a 10%, 10% a 15%, 15% a 20% e 20% a 25% ser aprovado é 1,39, 2,72, 4,52 e 6,26 vezes, respectivamente, a chance do aluno que teve de 0% a 5% faltas ser aprovado.

Tabela 20: Resultados da razão de chances do modelo geral para Probabilidade e Estatística.

Probabilidade e Estatística			
Efeitos fixos	**Razão de chances**	**IC [95%]**	**P-valor**
Intercepto	0,08	0,05 – 0,12	<0,001
Tempo na UnB	1,12	1,02 – 1,22	0,016
Curso [Ciências Econômicas]	0,26	0,06 – 1,19	0,083
Curso [Computação]	1,81	0,94 – 3,50	0,076
Curso [Engenharia de Produção]	0,58	0,32 – 1,05	0,073
Faltas [2 - 5% a 10%]	1,39	0,96 – 2,00	0,082
Faltas [3 - 10% a 15%]	2,72	1,97 – 3,75	<0,001
Faltas [4 - 15% a 20%]	4,52	3,23 – 6,33	<0,001
Faltas [5 - 20% a 25%]	6,26	4,54 – 8,64	<0,001

5.2.3 Diagnóstico do modelo

- Resíduos

Ao analisar ambas as disciplinas, observou-se nos Gráficos 28 e 29 que o gráfico

dos resíduos oscila entre o zero, o que indica que o modelo está ajustado de forma adequada aos dados.

Gráfico 28: Gráfico de resíduos e valores ajustados do modelo geral para Bioestatística.

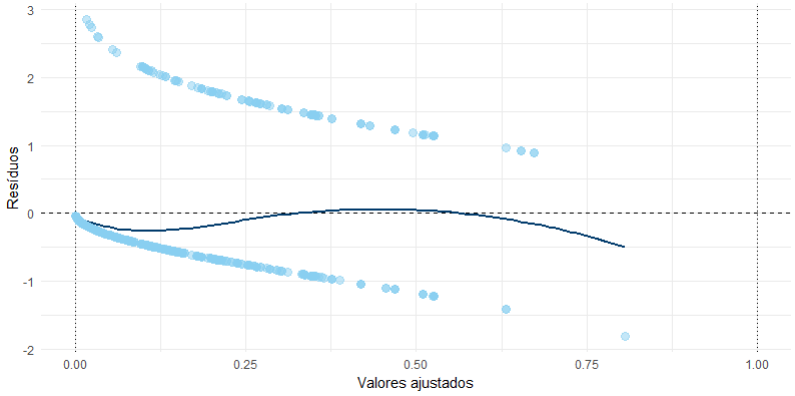
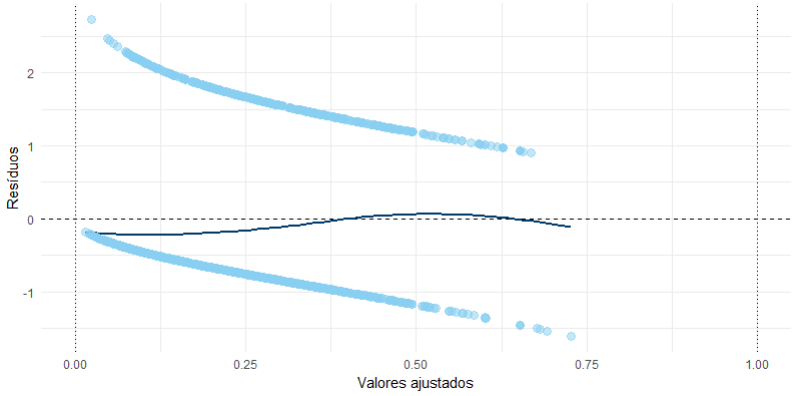


Gráfico 29: Gráfico de resíduos e valores ajustados do modelo geral para Probabilidade e Estatística.



- Curva ROC

Os gráficos 30 e 31 mostram que o modelo preditivo é bastante confiável, como a área sob a curva é relativamente grande para ambas disciplinas. Valores de 0,809 e 0,735 foram considerados como um ponto de corte apropriado na seleção da regra para obter resultados mais precisos.

Gráfico 30: Gráfico da curva de ROC do modelo geral para Bioestatística.

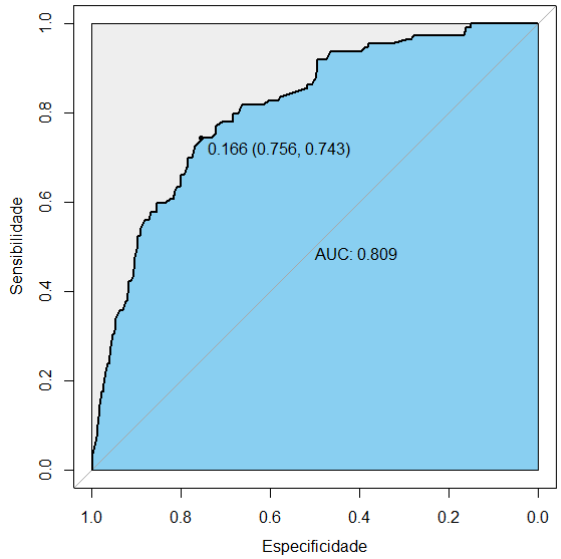
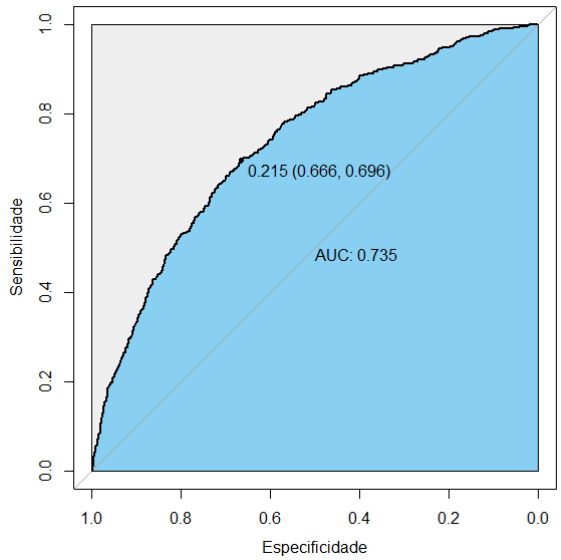


Gráfico 31: Gráfico da curva de ROC do modelo geral para Probabilidade e Estatística.



Por fim, analisando os resultados da matriz de confusão, tem-se os resultados obtidos nas Tabelas 21 e 22 indicam que os modelos de Bioestatística e Probabilidade e Estatística têm uma boa acurácia, aproximadamente 77% e 75%, respectivamente. Além disso, ambos também apresentaram sensibilidades consideravelmente elevadas, 94% e 86%, respectivamente, e especificidades aproximadas de 35% e 42%, respectivamente.

Tabela 21: Resultados da matriz de confusão do modelo geral para Bioestatística.

Matriz de confusão			
Estimado/ Observado	**Aprovado**	**Reprovado**	**Total**
Aprovado	517	135	652
Reprovado	37	72	109
Total	554	207	761

Medidas	**Valores**
Acurácia	0,774
IC [95% para acurácia]	(0,7426, 0,8032)
Sensibilidade	0,9332
Especificidade	0,3478

Tabela 22: Resultados da matriz de confusão do modelo geral para Probabilidade e Estatística.

Matriz de confusão			
Estimado/ Observado	**Aprovado**	**Reprovado**	**Total**
Aprovado	1675	370	2045
Reprovado	277	272	549
Total	1.952	642	2.594

Medidas	**Valores**
Acurácia	0,7506
IC [95% para acurácia]	(0,7335, 0,7671)
Sensibilidade	0,8581
Especificidade	0,4237

6 Conclusão

Essa pesquisa avaliou os fatores associados ao desempenho dos estudantes da Universidade de Brasília nas disciplinas Bioestatística e Probabilidade e Estatística, que introduzem conceitos básicos de Estatística para estudantes de diversos cursos da universidade. Foram analisados dois modelos, incluindo a menção SR (caso geral) e desconsiderando essa menção.

Nos dois modelos multiníveis para Bioestatística, as variáveis que apresentaram associação com o desempenho acadêmico foram: a modalidade da disciplina, o tempo na UnB em anos completos, o curso e o percentual de faltas dos estudantes. Os resultados revelaram que se o aluno cursar a disciplina como Optativas ou Módulo Livre a chance de ser aprovado é reduzida em relação a modalidade Obrigatória, a cada um ano do aluno na universidade, há um aumento na chance de aprovação na disciplina, os alunos que são dos cursos de Agronomia e Engenharia Florestal apresentaram melhor desempenho em comparação aos outros cursos, e, por fim, os modelos revelaram um resultado inesperado com relação a chance de aprovação segundo as categorias de percentual de faltas, a proporção de aprovação aumenta à medida que o número de faltas aumenta.

Nos dois modelos multiníveis para Probabilidade e Estatística, as variáveis que apresentaram associação com o desempenho acadêmico foram as mesmas para Bioestatística. Os resultados apresentaram as mesma inferências, com exceção da variável curso do estudante. No caso de Probabilidade e Estatística, os estudantes de Ciência da Computação, Engenharia Ambiental e Engenharia de Produção apresentaram uma taxa de aprovação mais alta em relação aos outros cursos.

No caso de Probabilidade e Estatística, os efeitos aleatórios não se mostraram tão significativos na utilização da regressão multinível. Uma sugestão seria utilizar a regressão logística como técnica de modelagem para prever o desempenho acadêmico.

Os resultados da análise indicaram que os modelos que excluíram a menção SR apresentaram uma acurácia significativamente maior em comparação ao modelo que incluiu essa menção. Diante disso, seria interessante realizar uma análise mais aprofundada para compreender os motivos pelos quais ocorrem tantos casos de SR e quais características apresentam os alunos com esta menção.

Por fim, é importante destacar que os modelos utilizados para identificar as disciplinas se basearam principalmente nas informações da oferta da disciplina e em um número limitado de variáveis relacionadas às características dos alunos. Portanto, para

complementar esse estudo, seria interessante incorporar características mais detalhadas dos professores e dos alunos, incluindo informações sociodemográficas e acadêmicas. Isso poderia ajudar na investigação do resultado contraditório encontrado com relação a influência do percentual de faltas no rendimento dos estudantes. Além de aprimorar a precisão dos modelos e a compreender melhor os fatores que influenciam o desempenho dos estudantes nas disciplinas Bioestatística e Probabilidade e Estatística.

Referências

AGRESTI, A. *Categorical data analysis*. [S.l.]: John Wiley & Sons, 2003.

AGRESTI, A. *An introduction to categorical data analysis*. [S.l.]: John Wiley & Sons, 2018.

ARA, A. B. *O ensino de Estatística e a busca do equilíbrio entre os aspectos determinísticos e aleatórios da realidade*. Tese (Doutorado) — Universidade de São Paulo, 2006.

BRUM, E. D.; LISKA, G. R. A avaliação de rendimento de alunos em disciplinas relacionados a estatística. *I Simpósio Sul-Americano de Pesquisa em Ensino de Ciências*, n. 1, 2020.

HOX, J. J. *Multilevel analysis: Techniques and applications*. [S.l.]: Routledge, 2010.

IGNáCIO, S. A. Importância da estatística para o processo de conhecimento e tomada de decisão. *Revista Paranaense de Desenvolvimento-RPD*, n. 118, p. 175–192, 2010.

LAROS, J. A.; MARCIANO, J. L. P. Análise multinível aplicada aos dados do nels: 88. *Estudos em Avaliação Educacional*, v. 19, n. 40, p. 263–278, 2008.

MAIA, J. A. et al. Modelação hierárquica ou multinível: Uma metodologia estatística e um instrumento Útil de pensamento na investigação em ciências do desporto. *Revista Portuguesa de Ciências do Desporto*, Universidade do Porto. Faculdade de Ciências do Desporto e de Educação Física, p. 92–107, 2003.

NETER, J.; KUTNER, M. H.; NACHTSHEIM, C. *Applied linear statistical models*. [S.l.]: Irwin Chicago, 1996.

VENDRAMINI, C. M. M. et al. Construção e validação de uma escala sobre avaliação da vida acadêmica (eava). *Estudos de Psicologia (Natal)*, SciELO Brasil, v. 9, p. 259–268, 2004.

Apêndice

Número médio de alunos por periodo em relação aos horários e períodos:

Número médio de alunos								
	Disciplinas							
	Bioestatística				Probabilidade e Estatística			
Efeitos fixos [aluno]	Geral	Até 2000	De 2000 à 2009	De 2010 à 2019	Geral	Até 2000	De 2000 à 2009	De 2010 à 2019
1 - 08:00/09:50	3966	83	82	69	1338	0	0	79
2 - 10:00/11:50	717	0	39	55	3411	37	58	105
3 - 14:00/15:50	1133	36	31	40	5073	62	89	158
4 - 16:00/17:50	167	0	0	44	3893	97	77	74
5 - 18:00/19:50	567	26	35	43	59	39	0	20
6 - 19:00/20:50	78	1	40	0	478	0	32	42
8 - 19:00/21:00 - 21:00/22:50	585	24	32	0	2105	0	70	42

Cursos classificados como “Outros” para a análise descritiva da variável “Curso”:

- **Bioestatística**

“Administração”
“Ciência da Computação”
“Ciências Contábeis”
“Ciências Econômicas”
“Ciências Naturais”
“Ciências Sociais”
“Computação”
“Comunicação Social”
“Direito”
“Educação Física”
“Enfermagem”
“Enfermagem e Obstetrícia”
“Engenharia”
“Engenharia Aeroespacial”
“Engenharia Ambiental”
“Engenharia Automotiva”
“Engenharia Civil”
“Engenharia de Computação”
“Engenharia de Energia”
“Engenharia de Produção”
“Engenharia de Redes de Comunicação”
“Engenharia Elétrica”
“Engenharia Mecatrônica”
“Engenharia Química”
“Filosofia”
“Física”
“Fisioterapia”
“Geofísica”
“Geografia”
“Geologia”
“Gestão Ambiental”
“Gestão do Agronegócio”
“História”
“Letras”
“Medicina”
“Nutrição”
“Odontologia”
“Psicologia”
“Química”
“Química Tecnológica”
“Saúde Coletiva”

- **Probabilidade e Estatística**

"Administração"
"Agronomia"
"Arquitetura e Urbanismo"
"Arquivologia"
"Biotecnologia"
"Ciência Política"
"Ciências Ambientais"
"Ciências Biológicas"
"Ciências Contábeis"
"Ciências Farmacêuticas"
"Ciências Naturais"
"Ciências Sociais"
"Computação"
"Comunicação"
"Comunicação Social"
"Design/Desenho Industrial"
"Direito"
"Educação Física"
"Enfermagem"
"Engenharia"
"Engenharia Aeroespacial"
"Engenharia Ambiental"
"Engenharia Automotiva"
"Engenharia de Energia"
"Engenharia de Software"
"Engenharia Eletrônica"
"Engenharia Florestal"
"Engenharia Química"
"Estatística"
"Farmácia"
"Filosofia"
"Física"
"Geofísica"
"Geografia"
"Geologia"
"Gestão Ambiental"
"Gestão de Agronegócios"
"Gestão de Políticas Públicas"
"Gestão do Agronegócio"
"História"
"Informática"
"Letras"
"Línguas Estrangeiras Aplicadas - MSI"
"Medicina Veterinária"
"Nutrição"
"Pedagogia"
"Psicologia"
"Química"
"Química Tecnológica"
"Relações Internacionais"
"Serviço Social"
"Teoria, Crítica e História da Arte"
"Turismo"

Cursos classificados como “Outros” para a análise da variável resposta e a variável explicativa “Cursos”:

- **Bioestatística**

“Administração”
“Ciência da Computação”
“Ciências Econômicas”
“Ciências Naturais”
“Computação”
“Comunicação Social”
“Direito”
“Educação Física”
“Enfermagem”
“Engenharia”
“Engenharia Aeroespacial”
“Engenharia Ambiental”
“Engenharia Automotiva”
“Engenharia Civil”
“Engenharia de Computação”
“Engenharia de Energia”
“Engenharia de Produção”
“Engenharia de Redes de Comunicação”
“Engenharia Elétrica”
“Engenharia Mecatrônica”
“Engenharia Química”
“Filosofia”
“Geofísica”
“Geografia”
“Geologia”
“Gestão do Agronegócio”
“Letras”
“Nutrição”
“Odontologia”
“Psicologia”
“Química Tecnológica”
“Saúde Coletiva”

- **Probabilidade e Estatística**

"Agronomia"
"Arquitetura e Urbanismo"
"Biotecnologia"
"Ciência Política"
"Ciências Ambientais"
"Ciências Biológicas"
"Ciências Naturais"
"Ciências Sociais"
"Comunicação"
"Comunicação Social"
"Design/Desenho Industrial"
"Direito"
"Educação Física"
"Enfermagem"
"Engenharia"
"Engenharia Aeroespacial"
"Engenharia Automotiva"
"Engenharia de Energia"
"Engenharia de Software"
"Engenharia Eletrônica"
"Engenharia Florestal"
"Farmácia"
"Filosofia"
"Gestão de Agronegócios"
"Gestão de Políticas Públicas"
"Gestão do Agronegócio"
"História"
"Letras"
"Línguas Estrangeiras Aplicadas - MSI"
"Medicina Veterinária"
"Nutrição"
"Pedagogia"
"Psicologia"
"Relações Internacionais"
"Serviço Social"
"Teoria, Crítica e História da Arte"