

Instituto Federal de Santa Catarina
Engenharia de Telecomunicações

JAMILLY DA SILVA PINHEIRO

Identificação Comportamental, via Análise Estatística,
de Progressão e Abandono junto aos Cursos do IFSC
Campus São José para Tomada de Providências de
Maneira Prévia.

São José, SC
2024

Instituto Federal de Santa Catarina
Engenharia de Telecomunicações

JAMILLY DA SILVA PINHEIRO

Identificação Comportamental, via Análise Estatística,
de Progressão e Abandono junto aos Cursos do IFSC
Campus São José para Tomada de Providências de
Maneira Prévia.

Prof. Ramon Hugo de Souza, Coordenador

São José, SC
2024

Índice

1	Introdução	4
1.1	Considerações Iniciais sobre Dados e Dificuldades Encontradas	4
2	Testes Estatísticos	4
2.1	Teste Qui-Quadrado (χ^2) de Pearson	6
3	Dados de Abandono e Análise Estatística	7
4	Conclusões	8

1 Introdução

Este trabalho objetiva a análise de dados de abandono relativos ao curso de Engenharia de Telecomunicações, junto ao campus São José, do Instituto Federal de Santa Catarina (IFSC). A partir de tal análise, apresenta-se então dados probabilísticos para tomada de decisão quanto ao incremento de vagas para ingresso junto a instituição. Para que tal seja possível, inicia-se com a definição de dados de abandono, base para este trabalho:

Definição 1: *Dados de abandono se referem a alunos quem estavam matriculados em um semestre e não se encontram mais matriculados no semestre posterior ao primeiro.*

De maneira que a busca pela computação de informação probabilística, com base em dados estatísticos, vem a estabelecer então a hipótese padrão a qual busca-se obter comprovação por meio deste trabalho:

Hipótese 1: *É possível ser pré-calculado o índice de abandono antes do término de um semestre, com base em dados anteriores, e, a partir de tal cálculo, é possível se definir um número maior de vagas considerando-se uma probabilidade pré-definida relativa a abandono.*

Porém é preciso que seja estabelecido que este trabalho fará análise apenas de informações relativas a primeira e segunda fases, e então relativas a abandono na primeira fase, de maneira a ser possível se calcular então um incremento de vagas com fundo probabilístico. Ou seja, parte-se da seguinte premissa, fortemente embasada em estatística:

Premissa 1: *Com base em tratamento de dados, para obtenção de média e desvio padrão, relativos a abandono, é possível se lidar probabilisticamente com o abandono, permitindo-se então que seja possível definir-se margem de probabilidade esperada quanto a tais dados.*

Apresenta-se então, na seção 2, conceitos sobre testes estatísticos e o teste qui-quadrado de Pearson, para contextualização; na seção 3 são apresentados os dados estatísticos obtidos sobre o abandono considerado para então aplicar-se a análise estatística sobre tais dados, com base no teste de Pearson apresentado na seção 2; e, por fim, as conclusões são apresentadas na seção 4.

1.1 Considerações Iniciais sobre Dados e Dificuldades Encontradas

2 Testes Estatísticos

Para modelo de testes fora selecionada a utilização de testes *chi-squared*, também referidos como testes qui-quadrado, *chi-square* ou χ^2 . Sendo tal modelo um teste estatístico de hipótese

(definição 2) utilizado na análise de tabelas de contingência (definição 3) quando as amostras são grandes o suficiente. De maneira simplificada, este tipo de teste é utilizado primariamente para examinar-se se duas variáveis categóricas (duas dimensões da tabela de contingência) são independentes e influenciam no teste estatístico [14]. O teste é tido como válido quando o teste estatístico apresenta distribuição qui-quadrado sob a hipótese nula, especificamente o teste qui-quadrado de Pearson [4] ou variantes do mesmo [16, 9, 11, 2].

O teste *chi-squared* de Pearson é utilizado para determinar-se se existe diferença estatística significativa entre as frequências esperadas e as observadas em uma ou mais categorias de uma tabela de contingência. Para tabelas de contingência com amostras menores utiliza-se do teste exato de Fisher [5, 6, 1].

Definição 2: *Um teste estatístico de hipótese é um método estatístico de inferência utilizado para decidir-se se os dados são suficientes para embasar uma hipótese em particular. Tal tipo de teste se firma então no cálculo de um teste com base estatística. A decisão é então realizada ou pela comparação do teste estatístico a um valor crítico ou, de maneira equivalente, pela avaliação de um p-value calculado a partir do teste estatístico. Existem mais de 100 testes estatísticos especializados definidos [10, 7].*

Definição 3: *Em estatística, uma tabela de contingência, também referida como tabulação cruzada ou crosstab, é um tipo de tabela em formato de matriz a qual apresenta a distribuição de frequência multivariada de variáveis. Sendo tais utilizadas amplamente em pesquisas de opinião, business intelligence, engenharia e em pesquisas científicas. Tais proveem uma ideia base da inter-relação entre duas variáveis e podem auxiliar a encontrar-se interações entre tais. O termo tabela de contingência foi utilizado pela primeira vez por Karl Pearson em “On the Theory of Contingency and Its Relation to Association and Normal Correlation” [8], parte da série de memórias de pesquisa da Drapers’ Company Biometric Memoirs I publicada em 1904.*

Nos testes padrão de aplicações as observações são classificadas em classes mutuamente exclusivas. Se a hipótese nula de que não há diferenças entre classes na população for verdadeira, o teste estatístico calculado das observações segue a frequência de distribuição qui-quadrado. O propósito deste teste é avaliar o quão provável as frequências observadas estariam assumindo que a hipótese nula é verdadeira.

Testes estatísticos que acompanham uma distribuição qui-quadrado ocorrem quando as observações são independentes. Existem também testes qui-quadrado para se testar a hipótese nula de independência de um par de variáveis aleatórias baseada em observações dos pares.

Os testes qui-quadrado geralmente se referem a testes para os quais a distribuição do teste estatístico se aproxima da distribuição chi-quadrada assintoticamente, significando que a distribuição da amostra (se a hipótese nula for verdadeira) do teste estatístico se aproxima de uma distribuição qui-quadrado cada vez mais perto à medida que o tamanho das amostras aumenta.

2.1 Teste Qui-Quadrado (χ^2) de Pearson

O teste qui-quadrado de Pearson é um teste estatístico aplicado a conjuntos de dados categóricos para avaliar qual a probabilidade de qualquer diferença observada entre os conjuntos que tenha surgido por acaso. Este teste é o mais utilizado amplamente dentre os muitos testes qui-quadrado (como Yates, razão de verossimilhança, teste de Portmanteau em séries temporais, etc) - procedimentos estatísticos cujos resultados são avaliados por referência à distribuição qui-quadrado. Suas propriedades foram estudadas pela primeira vez por Karl Pearson em 1900 [4]. Em contextos onde é importante a distinção entre a estatística do teste e sua distribuição, variantes do teste qui-quadrado de Pearson ou estatística são utilizados.

É um teste de *p-value* com a seguinte configuração [12, 13]:

- Antes do experimento, o experimentador fixa um certo número N de amostras a serem coletadas;
- Os dados observados são definidos como $(O_1, O_2, \dots, O_n)$, com a contagem do número de amostras de um conjunto finito de categorias dadas satisfazendo $\sum_i O_i = N$;
- A hipótese nula é que os números de contagem são amostrados de uma distribuição multinomial $Multinomial(N; p_1, \dots, p_n)$. Ou seja, os dados subjacentes são amostrados independentes e distribuídos de forma idêntica (IID) [3, 15] de uma distribuição categórica $Categorical(p_1, \dots, p_n)$ sobre as categorias fornecidas;
- O teste estatístico qui-quadrado de Pearson é definido como $\chi^2 = \sum_i \frac{(O_i - Np_i)^2}{Np_i}$. O *p-value* do teste estatístico é calculado numericamente ou consultando-o em uma tabela;
- Se o *p-value* for pequeno o suficiente (geralmente $p < 0.05$ por convenção), a hipótese nula é rejeitada e conclui-se que os dados observados não seguem a distribuição multinomial.

Um exemplo simples é testar a hipótese de que um dado comum de seis lados é “justo” (ou seja, todos os seis resultados têm a mesma probabilidade de ocorrer). Neste caso, os dados observados são $(O_1, O_2, \dots, O_6)$, o número de vezes que os dados caíram em cada número. A

hipótese nula é $Multinomial(N; \frac{1}{6}, \dots, \frac{1}{6})$ e $\chi^2 = \sum_{i=1}^6 \frac{(O_i - \frac{n}{6})^2}{\frac{n}{6}}$. Como detalhado a seguir, se $\chi^2 > 11.07$ então a imparcialidade dos dados pode ser rejeitada no nível de $p < 0.05$.

3 Dados de Abandono e Análise Estatística

Entrada	Com prova	Sem prova	SiSU/ENEM	Vagas rem.	Total	Disp.
2018.1	33	2	1	0	36	36
2018.2	24	13	0	0	37	36
2019.1	0	32	2	0	34	36
2019.2	0	1	33	1	35	36
2020.1	0	8	27	0	35	36
2020.2	0	2	34	0	36	36
2021.1	0	0	41	0	41	36
2021.2	0	0	32	18	50	36
2022.1	0	0	29	10	39	36
2022.2	0	2	18	22	42	36
2023.1	14	3	15	0	32	36
2023.2	1	23	15	0	39	40
2024.1	22	0	23	0	45	40

Tabela 1: Dados relativos a entrada de alunos no curso de Engenharia de Telecomunicações do semestre 2018.1 até o semestre 2024.1 segundo modalidade de entrada. Apresentado em negrito as vagas totais quando não preenchendo o número de vagas disponíveis.

Período	Com prova	Sem prova	SiSU/ENEM	Vagas rem.	Total
2018.1-2018.2	7+3	0+1	0+0	0+0	7+1
2018.2-2019.1	7+4	7+2	0+0	0+0	14+6
2019.1-2019.2	0+0	13+5	1+0	0+0	14+5
2019.2-2020.1	0+0	1+0	5+8	0+0	6+8
2020.1-2020.2	0+0	6+1	4+2	0+0	10+3
2020.2-2021.1	0+0	2+0	19+4	0+0	21+4
2021.1-2021.2	0+0	0+0	11+4	0+0	11+4
2021.2-2022.1	0+0	0+0	20+3	11+1	31+3
2022.1-2022.2	0+0	0+0	12+3	8+0	20+3
2022.2-2023.1	0+0	1+0	7+4	13+2	20+6
2023.1-2023.2	1+1	3+0	9+1	0+0	13+2
2023.2-2024.1	0+0	16+0	9+0	0+0	25+0
2024.1-2024.2	10+0	0+0	13+0	0+0	23+0

Tabela 2: Dados relativos a abandono de alunos no curso de Engenharia de Telecomunicações do semestre 2018.1 para 2018.2 até do semestre 2024.1 para 2024.2 segundo modalidade de entrada. O abandono computado consta como x+y, com x tendo acontecido no semestre de ingresso e y no semestre seguinte.

Nível de confiança	Margem de erro
68.3%, $\sigma_{\bar{x}}$	$16.5385 \pm 1.987(\pm 12.01\%)$
90%, $1.645\sigma_{\bar{x}}$	$16.5385 \pm 3.269(\pm 19.76\%)$
95%, $1.960\sigma_{\bar{x}}$	$16.5385 \pm 3.894(\pm 23.55\%)$
99%, $2.576\sigma_{\bar{x}}$	$16.5385 \pm 5.118(\pm 30.95\%)$
99.9%, $3.291\sigma_{\bar{x}}$	$16.5385 \pm 6.539(\pm 39.54\%)$
99.99%, $3.891\sigma_{\bar{x}}$	$16.5385 \pm 7.731(\pm 46.75\%)$
99.999%, $4.417\sigma_{\bar{x}}$	$16.5385 \pm 8.777(\pm 53.07\%)$
99.9999%, $4.892\sigma_{\bar{x}}$	$16.5385 \pm 9.72(\pm 58.77\%)$

Tabela 3: $\mu \pm \sigma$ abandono até sem. 1 $\rightarrow 16.5385 \pm 7.1642$; $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \frac{7.1642}{\sqrt{13}} = 1.9869854203132$.

4 Conclusões

Em considerando-se margem de erro, não têm-se o interesse em se dar garantias de baixa margem, afinal não se deseja garantir uma margem exata contendo todas as possibilidades de abandono. Neste caso em específico a margem de 68.3%, a qual seria equivalente a garantia acumulada de 84.13% - desconsiderando-se o erro para mais -, é suficiente para o que deseja-se calcular.

Então, com garantia de 84.13%, considerando-se a evasão na primeira fase - alunos

Nível de confiança	Margem de erro
68.3%, $\sigma_{\bar{x}}$	$20.9231 \pm 2.155(\pm 10.30\%)$
90%, $1.645\sigma_{\bar{x}}$	$20.9231 \pm 3.545(\pm 16.94\%)$
95%, $1.960\sigma_{\bar{x}}$	$20.9231 \pm 4.224(\pm 20.19\%)$
99%, $2.576\sigma_{\bar{x}}$	$20.9231 \pm 5.552(\pm 26.53\%)$
99.9%, $3.291\sigma_{\bar{x}}$	$20.9231 \pm 7.092(\pm 33.90\%)$
99.99%, $3.891\sigma_{\bar{x}}$	$20.9231 \pm 8.386(\pm 40.08\%)$
99.999%, $4.417\sigma_{\bar{x}}$	$20.9231 \pm 9.519(\pm 45.50\%)$
99.9999%, $4.892\sigma_{\bar{x}}$	$20.9231 \pm 10.543(\pm 50.39\%)$

Tabela 4: $\mu \pm \sigma$ abandono até sem. 2 $\rightarrow 20.9231 \pm 7.7704$; $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \frac{7.7704}{\sqrt{13}} = 2.1551137436492$.

que não chegam a segunda fase -, têm-se que 18 alunos evadem considerando-se o número de vagas médio de 36 (50% de abandono). Note-se que o valor médio de evasão é de 16 alunos do primeiro para o segundo semestre.

Com as mesmas garantias, considerando-se os mesmo alunos, mas com evasão até a segunda fase - alunos que não chegam a terceira fase - têm-se que 23 alunos evadem (63.89% de abandono), com o valor médio de evasão sendo de 21 alunos.

Dadas as margens e realizadas considerações estatísticas, têm-se que aumentar-se para 54 vagas (18+36) parece razoável em se considerando dados históricos de abandono.

Referências

- [1] Alan Agresti. «A Survey of Exact Inference for Contingency Tables». Em: *Statistical Science* 7.1 (1992), pp. 131–153. DOI: <https://doi.org/10.1214/ss/1177011454>.
- [2] G. E. P. Box e D. A. Pierce. «Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models». Em: *Journal of the American Statistical Association* (1970), pp. 1509–1526. DOI: <https://doi.org/10.1080/2F01621459.1970.10481180>.
- [3] Aaron Clauset. *A brief primer on probability distributions*. 2011. URL: https://web.archive.org/web/20120120154739/http://tuvalu.santafe.edu/~aaronc/courses/7000/csci7000-001_2011_L0.pdf.
- [4] Karl Pearson F.R.S. «On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling». Em: *Philosophical Magazine* 50.5 (1900), pp. 157–175. DOI: <https://doi.org/10.1080/14786440009463897>.

- [5] R. A. Fisher. «On the interpretation of χ^2 from contingency tables, and the calculation of P». Em: *Journal of the Royal Statistical Society* 85.2 (1922), pp. 87–94. DOI: <https://doi.org/10.2307/2340521>. URL: <https://zenodo.org/records/1449484>.
- [6] R. A. Fisher. *Statistical Methods for Research Workers*. Oliver e Boyd, 1954. ISBN: 0-05-002170-2.
- [7] Gopal K Kanji. *100 Statistical Tests*. 3^a ed. SAGE, 2006. ISBN: 1446222500, 9781446222508.
- [8] F.R.S. Karl Pearson. *Mathematical contributions to the theory of evolution*. Dulau e Co, 1904. URL: <https://archive.org/details/cu31924003064833>.
- [9] Gary King. *Unifying Political Methodology : The Likelihood Theory of Statistical Inference*. New York: Cambridge University Press, 1989. ISBN: 0-521-36697-6. URL: <https://books.google.com/books?id=clig0wrD7XoC&pg=PA84>.
- [10] Nancy D. Lewis, Nigel Da Costa Lewis e N. D. Lewis. *100 Statistical Tests in R: What to Choose, how to Easily Calculate, with Over 300 Illustrations and Examples*. Heather Hills Press, 2013. ISBN: 1484052994, 9781484052990.
- [11] G. M. Ljung e G. E. P. Box. «On a measure of lack of fit in time series models». Em: *Biometrika* 2 (1978), pp. 297–303. DOI: <https://doi.org/10.1093/biomet/2F65.2.297>. URL: <https://apps.dtic.mil/sti/pdfs/ADA049397.pdf>.
- [12] Orestis Loukas e Ho Ryun Chung. *Entropy-based Characterization of Modeling Constraints*. 2022. DOI: <https://doi.org/10.48550/arXiv.2206.14105>. URL: <https://arxiv.org/pdf/2206.14105>.
- [13] Orestis Loukas e Ho Ryun Chung. *Total Empiricism: Learning from Data*. 2023. URL: <https://arxiv.org/abs/2311.08315>.
- [14] Sociology 3112, Department of Sociology, The University of Utah. *The Chi-Square Test for Independence*. URL: <https://soc.utah.edu/sociology3112/chi-square.php>.
- [15] Stephanie. *Statistics How To*. URL: <https://www.statisticshowto.com/iid-statistics/>.
- [16] F. Yates. «Contingency Tables Involving Small Numbers and the χ^2 Test». Em: *Journal of the Royal Statistical Society* 1.2 (1934). DOI: <https://doi.org/10.2307/2983604>.