unesp

UNIVERSIDADE ESTADUAL PAULISTA

Análise de Dados Categóricos e Aplicações

Jôira Conceição dos Santos Netto

INSTITUTO DE GEOCIÊNCIAS E CIÊNCIAS EXATAS

RIO CLARO

Universidade Estadual Paulista "Júlio de Mesquita Filho"
Instituto de Geociências e Ciências Exatas
Câmpus de Rio Claro

Análise de Dados Categóricos e Aplicações

Jôira Conceição dos Santos Netto

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Matemática, junto ao Programa de Pós-Graduação em Matemática, mestrado profissional, do Instituto de Geociências e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de Rio Claro.

Orientadora
Profa. Dra. Selene Maria Coelho Loibel

Rio Claro
2019

N476a

Netto, Jôira Conceição dos Santos
Análise de dados categóricos e aplicações / Jôira Conceição dos Santos Netto. -- Rio Claro, 2019
83 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Instituto de Geociências e Ciências Exatas, Rio Claro
Orientadora: Selene Maria Coelho Loibel

1. Dados Categóricos. 2. Testes de Hipótese Qui-Quadrado. 3. Modelo de Regressão Logística. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Geociências e Ciências Exatas, Rio Claro. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

Jôira Conceição dos Santos Netto

Análise de Dados Categóricos e Aplicações

Dissertação APROVADA como requisito parcial para a obtenção do grau de Mestre no Curso de Pós-Graduação em Matemática, mestrado profissional, do Instituto de Geociências e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", pela seguinte banca examinadora:

Profa. Dra. Selene Maria Coelho Loibel
Orientadora

Prof. Dr. José Silvio Govone
Departamento de Estatística, Matemática Aplicada e Computação - UNESP, Rio Claro

---

Prof. Dr. João Carlos Vieira Sampaio
Departamento de Matemática - UFSCar, São Carlos.

Rio Claro, 12 de setembro de 2019

Dedico essa dissertação a minha família:
André (Esposo), Leonardo (Filho) e Maria Eduarda (Filha)

AGRADECIMENTOS

Quero agradecer a Deus, por sempre estar me guiando e iluminando meu caminho. Agradeço à minha orientadora Selene Maria Coelho Loibel, por toda dedicação e paciência. Quero agradecer ao meu esposo André, pela paciência, amor e ajuda, para que eu pudesse estudar e levar esse mestrado adiante. Também quero agradecer aos meus filhos Leonardo e Maria Eduarda por todo amor e paciência com a Mamãe. Agradeço aos meus pais, Oliver Marcos e Marlú, que sempre apoiaram meus estudos. Agradeço minha amiga Rosângela, que me acompanhou em todas as viagens de Sorocaba a Rio Claro, batalhando também para concluir o mestrado. E agradecer também, a todos os amigos que de alguma forma colaboraram para que eu conseguisse dar andamento aos estudos. Agradeço a todos professores que ministraram as disciplinas desse mestrado.

Não importa o que aconteça,
continue a nadar.
WALTERS, GRAHAM; Procurando Nemo, 2003

Resumo

Esta dissertação tem como foco a análise de dados categóricos, uma parte integrante da Análise Multivariada que interpreta a informação que está contida em dados discretos provenientes de contagens de eventos, possuindo características definidas pela combinação das categorias de duas ou mais variáveis. A análise de dados categóricos é de grande importância dentro da Estatística pois tem aplicabilidade em variadas áreas do conhecimento. Os dados utilizados, foram coletados através de um questionário aplicado aos alunos de cinco Escolas Técnicas Estaduais (Etec) que finalizaram os cursos técnicos em 2018 e 2019. A pesquisa teve como objetivo obter dados locais e analisar se os alunos pretendem trabalhar ou continuar estudando na mesma área do curso que estão concluindo, se os alunos estão satisfeitos com os cursos que estão fazendo, se pretendem voltar para Etec e fazer outro curso complementar, entre outros questionamentos. Devido à natureza dos dados obtidos, as técnicas de análise de dados categóricos são adequadas e devem ser aplicadas para modelar e fazer inferências sobre os aspectos de interesse. Esta análise pode levar a resultados que serão de grande utilidade para essas Etecs.

Palavras-chave: Dados Categóricos, Testes de Hipótese Qui-Quadrado, Modelo de Regressão Logística.

Abstract

This dissertation focuses on the Categorical Data Analysis, an integral part of the Multivariate Analysis, which interprets embedded information in discrete data resulting from event counts, having characteristics defined by combinations of categories from two or more variables. The categorical data analysis is of considerable importance within Statistics since it has a wide applicability in several areas of knowledge. The data set used was collected through a questionnaire applied to students from five Public Technical Schools (Etec) that finished the technical courses in 2018 and 2019. The research aims to gather local data and analyze whether students intend to work or continue studying in the same field of the technical course they are completing, whether students are satisfied with the courses they are attending, whether they want to go back to Etec and take another complementary course, among other questions. Due to the nature of the data obtained, categorized data analysis techniques are adequate and should be applied to model and make inferences about the aspects of interest. This analysis can be leaded to outcomes that will be very useful to these Etecs.

Keywords: Categorial data, Chi-Square Hypothesis tests, Logistic Regression Models.

Lista de Figuras

Lista de Tabelas

Sumário

Introdução

A importância da análise de dados categóricos dentro da Estatística está relacionada com o fato de haver grande aplicabilidade em variadas áreas, tais como a medicina, a biologia, a psicologia, a economia, as ciências políticas e as ciências da educação, entre outras.

De acordo com [5] a análise de dados categóricos é uma parte da análise multivariada, que interpreta a informação que está contida em dados discretos resultantes de contagens de eventos ou de unidades (pessoas, lugares, objetos etc.) possuindo certas características ou atributos definidos pela combinação das categorias de duas ou mais variáveis de interesse ou apenas categorias de uma variável. A análise de dados discretos univariados, gerados, por exemplo, dos modelos binomial, multinomial, hipergeométrico ou Poisson são casos particulares dos métodos multivariados abordados nesse trabalho. Detalhes sobre os modelos univariados podem ser vistos no Apêndice B.

Os dados coletados experimentalmente, sob critérios estatísticos, são considerados como uma amostra de uma população real. Os objetivos da análise de dados categóricos residem na realização de inferências relativas a questões que se prendem com relações estruturais que possam existir entre as variáveis estudadas. Estes objetivos inferenciais pressupõem geralmente a adoção de um modelo probabilístico consistente com o processo de amostragem e os propósitos analíticos. As conclusões extraídas são então condicionadas à validade de tais suposições sobre esses modelos.

A escolha dos modelos depende do delineamento amostral e dos objetivos da análise. As questões de interesse em análise de dados categóricos muitas vezes podem ser respondidas testando-se hipóteses de associação.

Além desses testes que, em geral, são utilizados para verificar a independência entre duas variáveis, é possível modelar a associação entre um conjunto de variáveis explicativas e uma variável resposta dicotômica (modelo de regressão logística binomial) ou uma variável politômica (modelo de regressão logística multinomial). No capítulo 1 desta dissertação é apresentada a definição de dados categóricos e um conjunto de dados para exemplificar a classificação dessas variáveis. Também são apresentadas algumas medidas de associação entre variáveis. No capítulo 2 são descritos os modelos probabilísticos associados a três estratégias de amostragem e os respectivos testes de hipóteses com aplicações extraídas da literatura. No capítulo 3, os modelos de regressão logística binomial e multinomial são apresentados incluindo o método de estimação dos seus parâmetros, a interpretação dos resultados e avaliação da qualidade do modelo ajustado. No capítulo 4 os métodos apresentados nos capítulos anteriores são aplicados para um conjunto de dados reais obtido com a aplicação de um questionário com os alunos da ETEC Fernando Prestes, ETEC Rubens de Faria e Souza, ETEC Prof. Elias

Miguel Júnior, ETEC Armando Pannunzio e ETEC Piedade.
A verificação da correlação, dependência ou associações entre variáveis é de grande importância pelo fato de nos dar um melhor entendimento sobre o tema. Muitas vezes é preciso avaliar o grau de associação entre duas ou mais variáveis para descobrir o quanto uma variável interfere no resultado de outra, assim podemos encontrar uma possível solução de problemas.

1 Dados Categóricos e Medidas de Associação

1.1 Dados Categóricos

As categorias consideradas em uma análise podem ser respostas sobre variáveis qualitativas e/ou quantitativas discretas. No caso de variáveis qualitativas, se existir ou não uma relação de ordem entre as categorias, as variáveis serão denominadas ordinais ou nominais. No caso de variáveis quantitativas, seus valores podem ser discretizados ou agrupados num pequeno número de intervalos de variação. As variáveis categóricas, podem ser classificadas como dicotômicas ou binárias, quando apresentam duas categorias, ou politômicas, quando apresentam três ou mais categorias.

Os dados são apresentados em tabelas de frequência, denominadas tabelas de contingência, nas quais são descritos em uma matriz cujas linhas identificam cada evento ocorrido e cujas colunas indicam a categoria de interesse de cada variável de estudo. As tabelas de contingência podem ser bidimensionais (por exemplo do tipo 2×2 ou 3×3), tridimensionais (por exemplo do tipo $5 \times 2 \times 3$ ou $7 \times 2 \times 2$) e tetradimensionais (por exemplo do tipo $2 \times 2 \times 2 \times 2$).

As variáveis de interesse são classificadas em variáveis resposta ou explicativa. De acordo com [6], são denominadas variáveis resposta aquelas descrevendo a livre resposta de cada unidade amostral e que, por isso, estão sujeitas a modelos probabilísticos que estejam de acordo com o esquema de obtenção dos dados. Já aquelas consideradas fixas, seja pelo delineamento amostral ou pela ação casual atribuída a elas no contexto dos dados, são comumente denominadas variáveis explicativas (ou ainda fatores, covariáveis, dentre outros). A classificação de uma variável como variável resposta ou explicativa depende do esquema de amostragem e dos objetivos da análise.

Conforme [5], as tabelas de contingência são definidas unicamente por variáveis respostas ou por uma mistura de ambos os tipos. Com isso a tabela de contingência pode ser escrita num formato bidimensional. As subpopulações são listadas pela combinação dos níveis das variáveis explicativas e as categorias de resposta pela combinação dos níveis das variáveis respostas. Na Tabela 1.1 apresenta-se uma tabela de contingência genérica, com s linhas e r colunas.

A última coluna da tabela refere-se aos totais de cada linha, isto é,

$$N_i = n_{i+} = \sum_{j=1}^{r} n_{ij}, \quad i = 1, ..., s \tag{1.1}$$

Tabela 1.1: Forma bidimensional de uma tabela de contingência genérica ($s \times r$)

	Categorias de resposta						
Subpopulação	1	2	...	j	...	r	Total
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1r}	n_{1+}
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2r}	n_{2+}
...	...	...	...	...	...	....	...
i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ir}	n_{i+}
....	...	...	...	...	...	....	...
s	n_{s1}	n_{s2}	...	n_{sj}	...	n_{sr}	n_{s+}
Total	n_{+1}	n_{+2}	...	n_{+j}	...	n_{+r}	n_{++}

Fonte: Paulino e Singer (2006), adaptada pela autora

A última linha da tabela refere-se aos totais de cada coluna, isto é,

$$N_j = n_{+j} = \sum_{i=1}^{s} n_{ij}, \quad j = 1, ..., r \tag{1.2}$$

com n_{ij} representando a frequência absoluta observada associada à célula correspondente à $i-$ésima linha e $j-$ésima coluna. O símbolo "+" em qualquer quantidade indexada representa uma soma sobre o(s) índice(s).

Para exemplificar a classificação das variáveis categóricas consideramos um estudo realizado por [3]. Foram confrontadas nesta pesquisa a afirmação piagetiana de que o ensino da matemática deve basear-se no desenvolvimento das estruturas mentais da criança e a realidade do ensino dessa matéria na 1.ª série do primeiro grau. Estudou-se a relação existente entre a noção de conservação (I-Não conservação, II-Conservação parcial e III-Conservação, quando a criança domina a quantidade extensiva, é capaz de relacionar o todo com suas partes, por ex.: X é metade de Y ou X é duas vezes Y) e o grau de desempenho em matemática (Desempenho Bom: $76\% - 100\%$ de acertos , Desempenho Mediano: $51\% - 75\%$ de acertos e Desempenho Fraco: $0 - 50\%$ de acertos). Constituíram a amostra 47 alunos da 1.ª série do 1.° grau (17 do sexo masculino e 30 do feminino), nível sócioeconômico médio-inferior para baixo-superior e idade de 6 anos e meio a 11 anos, sem escolarização anterior. A avaliação do desempenho relativo ao domínio da noção de conservação foi feita através do teste de conservação de quantidades descontínuas e a do desempenho em matemática através da observação sistemática e de uma prova. Apresenta-se na Tabela 1.2, os dados obtidos na pesquisa na forma de tabela de contingência.

Tem-se neste caso que a variável resposta é a porcentagem de acertos, que é uma variável quantitativa (porcentagens) para a qual foi feito o agrupamento em 3 faixas, ou seja é politômica, pois possui três categorias de resposta: faixas de porcentagens de acerto 0% a 50%, 51% a 75% e 76% a 100%.

A variável explicativa é a noção de conservação, que é uma variável qualitativa politômica, pois originalmente possui três categorias de resposta, os níveis I,II e III.

Fases de resolução dos problemas típicos de dados categóricos :

Tabela 1.2: Distribuição dos alunos segundo a noção de conservação e % de acertos na prova

	Porcentagens de acertos			
Noção de conservação	$0 - 50\%$	$51\% - 75\%$	$76\% - 100\%$	Totais
I	14	5	2	21
II	1	5	6	12
III	3	1	10	14
Totais	18	11	18	47

Fonte: Rosamilha e Faria (1983), [3]

i) Definição das questões de interesse.

ii) Especificação do delineamento amostral.

iii) Escolha de um modelo probabilístico que seja adequado.

iv) Tradução das questões de interesse em termos dos parâmetros do modelo probabilístico adotado, ou seja, especificação de modelos estruturais.

v) Ajuste dos modelos especificados através de uma metodologia estatística (como por exemplo, metodologia de máxima verossimilhança ou metodologia de mínimos quadrados generalizados).

vi) Comparação do(s) modelo(s) ajustado(s) com outros modelos alternativos.

vii) Conversão das conclusões em termos das questões originais.

1.2 Medidas de associação em tabelas 2×2

Nesta dissertação, como em vários livros que abordam dados categóricos, a variável resposta é denotada por Y e as variáveis explicativas por X.

Na Tabela 1.3, temos a representação de uma tabela de contingência 2×2, considerando a notação utilizada por [6]. As categorias da variável X estão nas linhas das tabelas de contingência e as da resposta Y nas colunas.

Tabela 1.3: Representação de uma tabela de contingência 2×2

Categorias da variável X	Categorias da variável Y		Totais
	$j = 1$	$j = 2$	
$i = 1$	n_{11}	n_{12}	n_{1+}
$i = 2$	n_{21}	n_{22}	n_{2+}
Totais	n_{+1}	n_{+2}	$n_{++} = n$

Fonte: Giolo (2017), [6]

As frequências denotadas por n_{ij} $(i, j = 1, 2)$ correspondem aos totais de indivíduos observados simultaneamente na $i-ésima$ categoria da variável X e $j-ésima$ categoria da variável Y. As frequências denotadas por n_{i+} $(i = 1, 2)$ correspondem às somas das frequências n_{ij} na $i-ésima$ linha e são denominadas totais marginais-linha. Analogamente, as frequências denotadas por n_{+j} $(j = 1, 2)$ correspondem às somas das frequências n_{ij} na $j-ésima$ coluna, sendo denominadas totais marginais-coluna. O total amostral denotado por n_{++}, ou simplesmente n, corresponde à soma das frequências n_{ij}, para $i, j = 1, 2$.

A notação $p_{ij} = P(X = i, Y = j)$ é utilizada para denotar a probabilidade de um elemento amostral apresentar a categoria i de X e a categoria j de Y, para $i, j = 1, 2$. Tais probabilidades são denominadas probabilidades conjuntas. Por outro lado, probabilidades condicionais, tais como a probabilidade de um indivíduo apresentar a categoria j de Y, dado que pertence à categoria i de X, isto é, $P(Y = j|X = i)$, são denotadas por $p_{(i)j}$. As notações p_{+j} e p_{i+} são utilizadas para designar, respectivamente, as probabilidades marginais-coluna e marginais-linha, sendo $p_{+j} = P(Y = j)$ a probabilidade de um indivíduo apresentar a $j - sima$ categoria de Y (independente da categoria de X a que pertence) e $p_{i+} = P(X = i)$ a probabilidade de um indivíduo apresentar a $i-sima$ categoria de X (independente da categoria de Y a que pertence).

Estabelecida a existência de associação em uma tabela de contingência 2×2, pode haver o interesse em escrever a intensidade dessa associação. Algumas medidas úteis para essa finalidade são apresentadas a seguir. As medidas de associação tratadas aqui estão no contexto de medicina (epidemiologia) mas podem ser usadas em outros contextos.

1.2.1 Risco relativo

Para estudos de coorte e clínicos aleatorizados em que se tem duas amostras independentes tamanhos fixos n_{1+} e n_{2+} (associadas às categorias da variável X), a intensidade da associação usualmente é descrita por meio de uma medida denominada risco relativo (RR).

O RR é definido como a razão entre a probabilidade de resposta positiva entre os indivíduos expostos a um fator de interesse e esta mesma probabilidade entre os não expostos a esse fator, ou seja,

$$RR = \frac{P(D|E)}{P(D|E^C)} = \frac{p_{(1)1}}{p_{(2)1}} \geq 0, \tag{1.3}$$

em que $P(D|E)$ denota a probabilidade de resposta positiva entre os indivíduos expostos e $P(D|E^C)$ a probabilidade de resposta positiva entre os indivíduos não expostos.

Se $RR = 1$, tem-se que o risco de resposta positiva não difere entre os indivíduos expostos e não expostos. Se $RR > 1$, os indivíduos expostos têm risco maior de apresentar resposta positiva do que os não expostos. Consequentemente, $RR < 1$, tem-se os indivíduos não expostos com risco maior de apresentar resposta positiva.

Um estimador proposto para esta medida é dado por

$$\widehat{RR} = \frac{\widehat{p}_{(1)1}}{\widehat{p}_{(2)1}}, \tag{1.4}$$

em que $\widehat{p}_{(i)1} = \frac{n_{i1}}{n_{i+}}$, $i = 1, 2$. Desse modo, uma estimativa para o RR corresponde ao valor numérico assumido por este estimador.

Para a obtenção de um intervalo de confiança para o RR considera-se o logaritmo natural de RR, isto é, $f = \ln(RR) = \ln(p_{(1)1}) - \ln(p_{(2)1})$. Isso porque a distribuição amostral de $\widehat{RR}$ não é normal, mas para amostras grandes a de $\widehat{f} = \ln(\widehat{RR}) = \ln(\widehat{p}_{(1)1}) - \ln(\widehat{p}_{(2)1})$ se aproxima da normal com média $f = \ln(RR)$ e variância $V(\widehat{f})$ dada por

$$V(\widehat{f}) = \frac{(1 - p_{(1)1})}{(n_{1+})(p_{(1)1})} + \frac{(1 - p_{(2)1})}{(n_{2+})(p_{(2)1})}. \tag{1.5}$$

Assim, um intervalo de confiança (IC) para o RR, ao nível $100(1-\alpha)\%$ de confiança, com $z_{\alpha/2}$ denotando o $100(1 - \alpha/2)$ percentil da distribuição normal padrão, pode ser obtido por

$$IC(RR) = \exp\left(\widehat{f} \pm z_{\alpha/2}\sqrt{V(\widehat{f})}\right). \tag{1.6}$$

Estimativas para $\widehat{f}$ e $V(\widehat{f})$ podem ser obtidas substituindo-se $p_{(1)1}$ e $p_{(2)1}$ pelos respectivos valores numéricos assumidos por $\widehat{p}_{(1)1}$ e $\widehat{p}_{(2)1}$. Caso o valor 1 pertença ao $IC(RR)$, haverá evidências ao nível $100(1-\alpha)\%$ de confiança de que o risco de resposta positiva não difere entre os indivíduos expostos e não expostos ao fator de interesse.

1.2.2 Diferença entre proporções ou risco atribuível

Outra medida útil para a comparação entre os indivíduos expostos e os não expostos ao fator de interesse é a diferença entre as proporções $p_{(1)1}$ e $p_{(2)1}$, também conhecida entre os epidemiologistas por risco atribuível.

Um estimador proposto para esta diferença é dado por

$$\widehat{d} = \widehat{p}_{(1)1} - \widehat{p}_{(2)1}. \tag{1.7}$$

Ainda, um intervalo de confiança para d, a um nível $100(1 - \alpha)\%$ de confiança, pode ser obtido como segue [2]

$$IC(d) = (\widehat{d} - z_{\alpha/2}\sqrt{V(\widehat{d})}; \widehat{d} + z_{\alpha/2}\sqrt{V(\widehat{d})}), \tag{1.8}$$

em que $z_{\alpha/2}$ denota o $100(1 - \alpha/2)$ percentil da distribuição normal padrão e $V(\widehat{d})$ a variância não viesada de $\widehat{d}$, isto é,

$$V(\widehat{d}) = \frac{p_{(1)1}(1 - p_{(1)1})}{(n_{1+} - 1)} + \frac{p_{(2)1}(1 - p_{(2)1})}{(n_{2+} - 1)}. \tag{1.9}$$

Para obtenção de uma estimativa para $V(\widehat{d})$ é usual que $p_{(1)1}$ e $p_{(2)1}$ sejam substituídos, respectivamente, pelos valores numéricos assumidos por $\widehat{p}_{(1)1}$ e $\widehat{p}_{(2)1}$. Caso o valor 0 (zero) pertença ao $IC(d)$, haverá evidências ao nível $100(1 - \alpha)\%$ de confiança de que $p_{(1)1}$ não difere de $p_{(2)1}$.

1.2.3 Razão de chances nos estudos de coorte

Nos estudos de coorte, indivíduos expostos e não expostos a um fator de interesse são acompanhados ao longo do tempo a fim de se observar quantos deles desenvolvem a doença. Embora nesses estudos seja possível calcular o risco relativo, outra medida denominada razão de chances (*odds ratio)* ou razão de produtos cruzados (*cross product ratio*) também pode ser obtida.

Para entender essa medida, é importante compreender a diferença entre chance (*odds)* e probabilidade. A chance de ocorrência de um evento de interesse (ou chance de sucesso) é definida por

$$\text{chance} = \frac{\text{probabilidade do evento ocorrer}}{\text{probabilidade do evento não ocorrer}}. \tag{1.10}$$

Para enfatizar a diferença entre probabilidade e chance, [2] observa que a probabilidade de ocorrência de um evento (sucesso) pode ser expressa em função de sua chance de ocorrência, ou seja,

$$\text{probabilidade de sucesso} = \frac{\text{chance}}{(\text{chance} + 1)}$$

No contexto desses estudos, a razão de chances (denotada por OR) fica definida como a razão entre a chance de ocorrência da doença entre os expostos e a chance de ocorrência da doença entre os não expostos. Nos estudos de coorte, indivíduos expostos e não expostos a um fator de interesse são acompanhados ao longo do tempo a fim de se observar quantos deles desenvolvem a doença.

$$\begin{aligned} OR_{coorte} &= \frac{\text{chance de doença entre os expostos}}{\text{chance de doença entre os não expostos}} \\ &= \frac{p_{(1)1}/(1-p_{(1)1})}{p_{(2)1}/(1-p_{(2)1})} = \frac{p_{(1)1}}{p_{(1)2}}\frac{p_{(2)2}}{p_{(2)1}}. \end{aligned} \tag{1.11}$$

Similar ao RR, a OR também não assume valores negativos. Quando $OR = 1$, não existe associação entre as variáveis (fator e doença). Já se $OR > 1$, a chance de doença entre os indivíduos expostos é maior do que entre os não-expostos. O contrário ocorre se $OR < 1$.

Uma estimativa proposta para a OR é dado por

$$\widehat{OR} = \frac{\widehat{p}_{(1)1}}{\widehat{p}_{(1)2}}\frac{\widehat{p}_{(2)2}}{\widehat{p}_{(2)1}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{1.12}$$

Para a obtenção de um intervalo de confiança para a OR, considera-se também o logaritmo natural de OR, isto é, $f = \ln(OR)$, tendo em vista a distribuição amostral de $\widehat{f} = \ln(\widehat{OR})$ ser aproximadamente normal com média f e variância assintótica estimada por

$$\widehat{V}(\widehat{f}) = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right). \tag{1.13}$$

Assim o intervalo de confiança (IC) para a OR, ao nível de $100(1-\alpha)\%$ de confiança, pode ser obtido por

$$IC(OR) = \exp\left(\widehat{f} \pm z_{\alpha/2}\sqrt{\widehat{V}(\widehat{f})}\right), \tag{1.14}$$

em que $z_{\alpha/2}$ denota o $100(1-\alpha/2)$ percentil da distribuição normal padrão.

Caso o valor 1 pertença ao intervalo, haverá evidências de chances não diferentes de ocorrência da doença entre os indivíduos expostos e não expostos.

1.2.4 Razão de chances nos estudos caso-controle

Nos estudos caso-controle, indivíduos doentes e não doentes são inicialmente selecionados para então se investigar quantos deles estiveram expostos ao fator de interesse. No contexto desses estudos, a razão de chances fica definida como a razão entre a chance de exposição entre os casos e a chance de exposição entre os controles, ou seja,

$$\begin{aligned} OR_{coorte} &= \frac{\text{chance de exposição entre os casos}}{\text{chance de exposição entre os controles}} \\ &= \frac{p_{(1)1}/(1-p_{(1)1})}{p_{(2)1}/(1-p_{(2)1})} = \frac{p_{(1)1}}{p_{(1)2}}\frac{p_{(2)2}}{p_{(2)1}}. \end{aligned} \tag{1.15}$$

Se $OR = 1$, não existe associação entre as variáveis. Já se $OR > 1$, a chance de exposição ao fator de interesse entre os indivíduos doentes é maior do que entre os não doentes. O contrário ocorre se $OR < 1$.

Uma estimativa proposta para essa medida é dada por

$$\widehat{OR} = \frac{\widehat{p}_{(1)1}}{\widehat{p}_{(1)2}}\frac{\widehat{p}_{(2)2}}{\widehat{p}_{(2)1}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{1.16}$$

Um intervalo de confiança (IC) para a OR pode ser obtido de forma análoga ao apresentado (expressão 1.14). Caso o valor 1 pertença ao respectivo intervalo, então haverá evidências de que a chance de exposição ao fator de interesse não difere entre os casos e controles.

1.2.5 Razão de chances nos estudos transversais

A partir das definições de chance e de razão de chances apresentadas, nota-se não ser apropriado defini-las nos estudos transversais, pois não se tem nenhum dos totais marginais fixos no delineamento amostral. Contudo, condicional aos totais marginais-linha, considera-se que uma estimativa para essa medida nesses estudos que consiste do valor numérico obtido por

$$\widehat{OR} = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{1.17}$$

Embora alguns autores, dentre eles Castro-Costa e Ferri (2008) *apud* [6], mencionem que utilizar a OR nos estudos transversais não seja necessariamente errado, é necessário

ter cautela quanto à sua interpretação. Nesses estudos, a OR deveria ser vista apenas como uma medida que auxilia a investigar possíveis associações entre variáveis, a fim de embasar novas pesquisas.

Outra medida usual nesses estudos é denominada razão de prevalências, que também está condicionada aos totais marginais-linha. Estimativa para essa medida consiste do valor numérico obtido por

$$\widehat{RP} = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}. \tag{1.18}$$

1.2.6 Relação entre risco relativo e razão de chances

Da definição de risco relativo discutida anteriormente tem-se que

$$RR = \frac{p_{(1)1}}{p_{(2)1}} = \frac{P(D|E)}{P(D|E^C)} = \frac{P(D)P(E|D)/[P(D)P(E|D) + P(D^C)P(E|D^C)]}{P(D)P(E^C|D)/[P(D)P(E^C|D) + P(D^C)P(E^C|D^C)]}.$$

Sendo D e D^C doença e não doença e, E e E^C exposição e não exposição ao fator de interesse, respectivamente. Relembrando $P(D^C) = 1 - P(D)$, segue que

$$RR = \frac{P(E|D)\{P(E^C|D^C) + P(D)[P(E^C|D) - P(E^C|D^C)]\}}{P(E^C|D)\{P(E|D^C) + P(D)[P(E|D) - P(E|D^C)]\}}$$

Sob a suposição de doença rara, $P(D) \to 0$. Logo,

$$RR \approx \frac{P(E|D)P(E^C|D^C)}{P(E^C|D)P(E|D^C)} = \frac{p_{1(1)}p_{2(2)}}{p_{2(1)}p_{1(2)}} = OR \tag{1.19}$$

como consequência desse resultado, tem-se que a razão de chances obtida em um estudo de caso-controle conduzido com o objetivo de avaliar a associação entre um fator de interesse e uma doença rara, pode, nesses casos, ser interpretada como aproximação do risco relativo.

2 Amostragem e Modelos associados

2.1 Estratégias

Nesta seção são descritos os modelos probabilísticos utilizados dependendo do delineamento amostral e dos objetivos da análise. Apresentam-se a seguir três delineamentos ou estratégias de amostragem usuais e os modelos associados, segundo [5] e [6].

2.1.1 Estratégia I: Modelo Produto de distribuições de Poisson

Em alguns experimentos é conveniente pré-fixar a duração de realização do experimento (T). Por exemplo, em uma pesquisa de opinião é possível entrevistar tantas pessoas quanto possível em um determinado período de tempo. Para a definição de um modelo apropriado para estudos desse tipo se faz necessário considerar algumas suposições:

i) O número de ocorrências do evento de interesse no intervalo de tempo especificado é independente do número de ocorrências em qualquer outro intervalo de tempo disjunto;

ii) A probabilidade de haver k ocorrências do evento de interesse no intervalo de tempo $(s, s+t]$ depende somente de t e não de s;

iii) A probabilidade condicional de ocorrer 2 ou mais vezes em $(0, t]$, dado que ocorreu 1 ou mais vezes em $(0, t]$, tende a zero quando $t \to 0$.

Definimos $N_{ij}, i, j = 1, 2$ como sendo o número de ocorrências do par (i, j) associado respectivamente ao par de variáveis (X, Y).

Com tais suposições é possível assumir, para cada $N_{ij}, i, j = 1, 2$ uma distribuição de Poisson com parâmetro $\mu_{ij} = T\lambda_{ij}$, sendo λ_{ij} a taxa média de ocorrência por unidade de tempo e T a duração do experimento. Nesse caso temos que $N = \sum_{i,j} N_{ij}$ é uma variável aleatória com distribuição de Poisson com parâmetro $\mu = \sum_{i,j} \mu_{ij}$.

Se for assumido que $N_{ij}, i, j = 1, 2$ são variáveis aleatórias independentes, segue que o modelo probabilístico associado ao estudo é o produto de distribuições de Poisson independentes com função de probabilidade dada por:

$$\begin{aligned} P(\mathbf{N} = \mathbf{n}) &= P(N_{11} = n_{11}, N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22}) = \\ &= \prod_{i=1}^{2}\prod_{j=1}^{2} \frac{e^{-\mu_{ij}}\mu_{ij}^{n_{ij}}}{n_{ij}!} \end{aligned} \tag{2.1}$$

para $n_{ij} \in \mathbb{N}_0$ e $\mu_{ij} \in \mathbb{R}^+, i, j = 1, 2$.

Sob esse modelo a questão que se pretende averiguar pode ser definida nos seguintes termos: a proporção de ocorrências de N_{11} é a mesma que a proporção de ocorrencias de N_{12}, ou seja não há associação entre as variáveis X e Y. Assim, em termos das médias pode-se expressar a hipótese de interesse como:

$$\frac{\mu_{11}}{\mu_{+1}} = \frac{\mu_{12}}{\mu_{+2}} \left(= \frac{\mu_{1+}}{\mu_{++}} \right) \tag{2.2}$$

sendo $\mu_{+j} = \sum_i \mu_{ij}, \mu_{i+} = \sum_j \mu_{ij}$ e $\mu_{++} = \sum_{i,j} \mu_{ij}$.

O modelo Produto de distribuições de Poisson segundo [5] baseia-se na suposição de que as frequências em cada célula são geradas por processos poissonianos independentes. Estes processos visam descrever a ocorrência aleatória de um dado evento, geralmente raro, classificado segundo certas características, em algumas unidades de exposição. Este modelo é adequado quando o interesse se centra na comparação das diversas taxas de ocorrência baseadas em medidas de exposição eventualmente diferentes de um evento surgindo aleatória e independentemente.

A aplicação deste delineamento e das suposições inerentes no cenário genérico da Tabela 1.1 , de dimensão $s \times r$, conduz ao modelo probabilístico para o vetor das frequências $\mathbf{n} = (\mathbf{n}_1', ..., \mathbf{n}_s')'$, com $\mathbf{n}_i = (n_{i1}, ..., n_{ir})', i = 1, ..., s$ dado por:

$$P(\mathbf{N} = \mathbf{n}) = \prod_{i=1}^{s} \prod_{j=1}^{r} \frac{e^{-\mu_{ij}} \mu_{ij}^{ij}}{n_{ij}!} \tag{2.3}$$

em que $\mu = (\mu_1', ..., \mu_s')'$, $\mu_i = (\mu_{i1}, ..., \mu_{ir})'$, $\mu_{ij} > 0$ para $i = 1, ..., s$, $j = 1, ..., r$.

2.1.2 Estratégia II: Modelo Multinomial

Um delineamento, em geral mais frequente, consiste em fixar antecipadamente o número total de elementos amostrais (n) e selecioná-los de modo aleatório de uma população de interesse. São registradas as frequências n_{ij} de elementos que apresentam simultaneamente as categorias (i, j) associadas, respectivamente, ao par de variáveis (X, Y).

Sob este delineamento amostral, nota-se que ambas as variáveis são consideradas respostas. Contudo, dependendo dos objetivos do estudo, uma delas é usualmente considerada como variável explicativa (ou fator). Isso equivale a um processo de amostragem aleatória simples em que é selecionada uma amostra aleatória de tamanho n de uma população suficientemente grande.

Supondo válidas as condições e considerando o caso de tabelas de contingência (2×2):

i) Ocorrência de somente uma das $k = 4$ possibilidades de resposta mutuamente exclusivas $(i = 1, j = 1), (i = 1, j = 2), (i = 2, j = 1)$ e $(i = 2, j = 2)$ para cada elemento da amostra;

ii) Independência entre as respostas dos n elementos;

iii) Para os n elementos da amostra o mesmo vetor $\mathbf{p} = (p_{11}, p_{12}, p_{21}, p_{22})'$ de probabilidades de ocorrência das k possibilidades de resposta.

Segue que o vetor aleatório $\mathbf{N} = (N_{11}, N_{12}, N_{21}, N_{22})'$ com N_{ij} denotando as frequências de elementos que apresentam simultaneamente as categorias (i, j) de $(X, Y), i, j =$

$1, 2$ segue a distribuição multinomial com parâmetros n e $\mathbf{p}$, com função de probabilidade dada por:

$$\begin{aligned} P(\mathbf{N} = \mathbf{n}) &= P(N_{11} = n_{11}, N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22}) = \\ &= n! \prod_{i=1}^{2} \prod_{j=1}^{2} \frac{p_{ij}^{n_{ij}}}{n_{ij}!} \end{aligned} \tag{2.4}$$

em que $n_{ij} \geq 0, \sum_{i,j=1}^{2} n_{ij} = n$ e $\sum_{i,j=1}^{2} p_{ij} = 1$.

Condicional a n, tem-se que as variáveis N_{ij} que compõem o vetor $\mathbf{N}$ não são independentes e que os elementos da matriz de variância e covariância de $\mathbf{N}$ são dadas por:

$$Cov(N_{ij}, N_{i^*j^*}) = \begin{cases} np_{ij}(1 - p_{ij}) & \text{para } i = j = i^* = j^* \\ -np_{ij}p_{i^*j^*} & \text{para } i \neq i^* \text{ e/ou } \ j \neq j^* \end{cases} \tag{2.5}$$

Considere um estudo composto de n ensaios independentes no qual cada ensaio resulta em um número r finito de categorias de respostas com probabilidades $p_1, p_2, ..., p_r$, tal que $p_j \geqslant 0, j = 1, ..., r$, e $\sum_{j=1}^{r} p_j = 1$. Denotando por N_j o número de vezes que $j - ésima$ resposta ocorre nos n ensaios, segue que a distribuição associada ao vetor aleatório $\mathbf{N} = (N_1, ..., N_r)'$ é a multinomial com parâmetros n e $\mathbf{p} = (p_1, p_2, ..., p_r)$, denotada aqui por $\mathbf{N} \sim Multi(n, \mathbf{p})$, cuja função de probabilidade é dada por:

$$\begin{aligned} P(\mathbf{N} = \mathbf{n}) &= P(N_1 = n_1, N_2 = n_2, ..., N_r = n_r) = \\ &= \frac{n!}{n_1!n_2!...n_r!} {p_1}^{n_1} {p_2}^{n_2} ... {p_r}^{n_r} = n! \prod_{j=1}^{r} \frac{(p_j)^{n_j}}{(n_j)!} \end{aligned} \tag{2.6}$$

em que $n_j \geq 0$ e $\sum_{j=1}^{r} n_j = n$.

Como em cada ensaio ocorre somente uma das j categorias possíveis de resposta $(j = 1, ..., r)$, defina:

$$X_j = \begin{cases} 1 & \text{se a } \ j - \text{ésima resposta ocorreu no } i - \text{ésimo ensaio} \quad (i = 1, ..., n) \\ 0 & \text{caso contrário} \end{cases}$$

de modo que $N_j = \sum_{i=1}^{n} X_j, j = 1, ..., r$. Assim, para $j = j'$, segue que

$$\begin{aligned} Cov(X_j, X_{j'}) = Var(X_j) &= E[X_j X_j] - E[X_j]E[X_j] = \\ &= E\left[X_j^2\right] - E[X_j]E[X_j] = \\ &= E[X_j] - E[X_j]E[X_j] = \\ &= p_j - (p_j)^2 = p_j(1 - p_j) \end{aligned} \tag{2.7}$$

com $X_j^2 = X_j$ devido ao fato de X_j assumir somente os valores 1 ou 0. Por analogia, para $j \neq j'$, segue que

$$Cov(X_j, X_{j'}) = E[X_j X_{j'}] - E[X_j]E[X_{j'}] = -E[X_j]E[X_{j'}] = -p_j p_{j'} \tag{2.8}$$

com $X_j X_{j'} = 0$ devido ao fato X_j e $X_{j'}$ não poderem ser ambos iguais a 1 ao mesmo tempo. Logo, segue que

$$Cov(N_j, N_{j'}) = nCov(X_j, X_{j'}) = \begin{cases} np_j(1-p_j) & \text{para } j = j' \\ -np_j p_{j'} & \text{para } j \neq j' \end{cases} \tag{2.9}$$

2.1.3 Estratégia III: Modelo Produto de distribuições Multinomiais

O experimento pode também ser planejado fixando antecipadamente o número N_j de elementos de cada categoria. Um motivo para esse tipo de planejamento é evitar que na amostra surjam poucos elementos de alguma(s) categoria(s). Note que enquanto que na estratégia II somente o total geral da tabela é fixo, aqui os totais marginais das colunas ou linhas também são fixos. A variável fixa serve apenas para indicar as subpopulações de onde são tomadas as observações da outra variável. Então uma amostra de tamanho fixo n_1 é extraída de uma subpopulação e outra amostra de tamanho n_2 de outra subpopulação, independente da primeira. Se a variável resposta Y apresentar três ou mais categorias ($r > 2$ finito), com N_{ij} denotando o número de vezes que a $j-$ésima categoria ocorre na $i-$ésima amostra ($j = 1, ..., r$ e $i = 1, 2$), o modelo probabilístico associado é o produto de distribuições multinomiais. A justificativa para isso é que, condicional aos totais n_{i+}, tem-se duas distribuições multinomiais independentes, uma associada a cada amostra (isto é, a cada linha da tabela de contingência $2 \times r$). Assim obtem-se a distribuição do vetor $(N_{11}, N_{12}, ..., N_{1r}, N_{21}, N_{22}, ..., N_{2r})'$. O modelo produto de multinomiais independentes é descrito pela função de probabilidade:

$$\begin{aligned} &P(N_{11} = n_{11}, N_{12} = n_{12}, ..., N_{1r} = n_{1r}, N_{21} = n_{21}, N_{22} = n_{22}, ..., N_{2r} = n_{2r}) = \\ &= \prod_{i=1}^{2} \left[(n_{i+})! \prod_{j=1}^{r} \frac{p_{(i)j}^{n_{ij}}}{n_{ij}!} \right] \end{aligned} \tag{2.10}$$

com $p_{(i)j} = P(Y = j | X = i)$, em que para $i = 1, 2$ tem-se $\sum_{j=1}^{r} p_{(i)j} = 1$.

2.2 Testes de Hipóteses

A seção a seguir, exibe a metodologia dos testes usuais em tabelas de contigência. A escolha do teste depende do delineamento amostral, do modelo associado e do tipo de variáveis. A seguir apresenta-se a síntese dos delineamentos amostrais, modelos e

estatísticas de teste em tabelas de contigência, considerando as s categorias de X nas linhas e as r categorias de Y nas colunas.

Delineamento: Totais aleatórios
Modelo associado: Produto de Poisson

Variáveis X	Y	*Hipótese nula*	*Estatística de teste*
Nominal Nominal Ordinal	Nominal Ordinal Ordinal	H_0: multiplicatividade	Q_P, Q_L ou $Q_N \sim \chi^2_{(s-1)(r-1)}$

Delineamento: Total n fixo e demais aleatórios
Modelo associado: Multinomial

Variáveis X	Y	*Hipótese nula*	*Estatística de teste*
Nominal	Nominal	H_0: independência	Q_p, Q_L ou $Q_n \sim \chi^2_{(s-1)(r-1)}$
Nominal	Ordinal	H_0: independência	Q_p, Q_L ou $Q_n \sim \chi^2_{(s-1)(r-1)}$
Ordinal	Ordinal	H_0: ausência de tendência linear	$Q_{CS} \sim \chi^2_{(1)}$

Delineamento: Marginais-linha fixos (n_{i+} fixos)
Modelo associado:
a) se $s = 2$ e $r > 2$ (Produto de multinomiais)

Variáveis X	Y	*Hipótese nula*	*Estatística de teste*
Nominal	Ordinal	H_0: escores médios não diferem	$Q_S \sim \chi^2_{(1)}$

b) se $s > 2$ e $r > 2$ (Produto de multinomiais)

Variáveis X	Y	*Hipótese nula*	*Estatística de teste*
Nominal	Ordinal	H_0: escores médios não diferem	$Q_S \sim \chi^2_{(s-1)}$
Ordinal	Ordinal	H_0: ausência de tendência linear	$Q_{CS} \sim \chi^2_{(1)}$ ou $Q_S \sim \chi^2_{(s-1)}$

Delineamento: Marginais-coluna fixos (n_{+j} fixos)
Modelo associado: se $s > 2$ e $r \geq 2$ (Produto de multinomiais)

Variáveis X	Y	*Hipótese nula*	*Estatística de teste*
Ordinal	Nominal	H_0: escores médios não diferem	$Q_S \sim \chi^2_{(r-1)}$
Ordinal	Ordinal	H_0: ausência de tendência linear	$Q_{CS} \sim \chi^2_{(1)}$

Nos casos de variáveis Y e X nominais e totais fixos ou aleatórios, pode-se fazer o uso da estatística qui-quadrado de Pearson

$$Q_p = \sum_{i=1}^{s}\sum_{j=1}^{r} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{2.11}$$

em que $e_{ij} = \frac{(n_{i+})(n_{+j})}{n}$, $i = 1, ..., s$ e $j = 1, ..., r$. Quando todas as células apresentem valores esperados maiores do que 5, Q_p segue a distribuição aproximada qui-quadrado com $(s-1)(r-1)$ graus de liberdade.

Nos casos de variáveis Y ordinal, X nominal e totais marginais n_{i+} fixos, a estatística Q_p pode ser usada, para avaliar a associação global. Mas se o interesse for levar em conta a natureza ordinal da variável resposta, pode-se usar a estatística escore médio, Q_S.

Então de acordo com a literatura, atribui-se escores para as categorias da variável resposta e, então, definir um escore médio para cada subpopulação (linha) da tabela de contigência, tal que

$$\overline{F}_i = \sum_{j=1}^{r} a_j p_{(i)j} \qquad i = 1, .., s \tag{2.12}$$

os quais podem ser estimados a partir dos estimadores

$$\overline{f}_i = \sum_{j=1}^{r} a_j \widehat{p}_{(i)j} = \sum_{j=1}^{r} a_j \left(\frac{N_{ij}}{n_{i+}}\right), \qquad i = 1, ..., s \tag{2.13}$$

Estimativas dos escores médios correspondem, desse modo, aos valores assumidos por tais estimadores, isto é, $\sum_{j=1}^{r} a_j \left(\frac{N_{ij}}{n_{i+}}\right), i = 1, ..., s$. Nota-se que os valores de tais estimativas estarão sempre no intervalo que tem como limites o menor e o maior valor assumidos para os escores.

Sob a hipótese nula de não associação entre as variáveis Y e X, formulada em termos de escores médios por $H_0 : \overline{F}_1 = ... = \overline{F}_s$, segue que a esperança e a variância de $\overline{f}_1$ são dadas, respectivamente por

$$E(\overline{f}_1) = \sum_{j=1}^{r} a_j \left[\frac{E(N_{1j})}{n_{1+}}\right] = \sum_{j=1}^{r} \frac{a_j}{n_{1+}} \left(\frac{n_{1+}n_{+j}}{n}\right) = \sum_{j=1}^{r} a_j \left(\frac{n_{+j}}{n}\right) = \mu_a \tag{2.14}$$

e

$$V(\overline{f}_1) = \left(\frac{n - n_{1+}}{(n_{1+})(n-1)}\right) \sum_{j=1}^{r} (a_j - \mu_a)^2 \left(\frac{n_{+j}}{n}\right). \tag{2.15}$$

De acordo com Giolo (2017), em decorrência do teorema do limite central, temos que $\overline{f}_1$ converge para a distribuição normal. Logo, a estística Q_S, denominada estatística escore médio e expressa por

$$Q_S = \frac{\left[\overline{f}_1 - E(\overline{f}_1)\right]^2}{V(\overline{f}_1)} = \frac{\left(\overline{f}_1 - \mu_a\right)^2}{V(\overline{f}_1)} = \frac{(n-1)}{(n - n_{1+})} \frac{(n_{1+})\left(\overline{f}_1 - \mu_a\right)^2}{v_a}, \tag{2.16}$$

com $\upsilon_a = \sum_{j=1}^{r}(a_j - \mu_a)^2 \left(\frac{n_{+j}}{n}\right)$, segue distribuição aproximada qui-quadrado com $(s-1)$ graus de liberdade. Valores de Q_S aos quais se associam probabilidades pequenas (usualmente $\leq 0,05$) conduzem a rejeição de H_0.

Na situação em que as variáveis Y e X ordinais e total n ou totais n_{i+} fixos , asumimos escores para as duas variáveis, e a estátistica de teste apropriada,é a estatística de correlação, Q_{CS}. Se houver dificuldade em assumir escores para as categorias Y e X, a estatistica Q_p pde ser utilizada para avaliar a asssociação global entre as variáveis. Naqueles delineamentos amostrais em que os totais marginais-linha n_{i+} são fixos, pode-se, alternativamente, de acordo com [6], considerar a variável X como nominal e, então, utilizar a estatistica Q_S apresentada anteriormente.

Então, consideramos escores para ambas, é possível assumir o escore médio

$$\overline{F} = \sum_{i=1}^{s}\sum_{j=1}^{r} c_i a_j p_{ij} \tag{2.17}$$

bem como estimá-lo por meio do estimador

$$\overline{f} = \sum_{i=1}^{s}\sum_{j=1}^{r} c_i a_j \widehat{p}_{ij} = \sum_{i=1}^{s}\sum_{j=1}^{r} \frac{c_i a_j n_{ij}}{n} \tag{2.18}$$

Sob H_0, esperança e variâcia de $\overline{f}$ são dadas, respectivamente, por

$$\begin{aligned} E(\overline{f}) &= \sum_{i=1}^{s}\sum_{j=1}^{r} \frac{c_i a_j}{n} E(N_{ij}) = \sum_{i=1}^{s}\sum_{j=1}^{r} \frac{c_i a_j}{n} \frac{(n_{i+})(n_{+j})}{n} \\ &= \sum_{i=1}^{s} c_i \left(\frac{n_{i+}}{n}\right) \sum_{j=1}^{r} a_j \left(\frac{n_{+j}}{n}\right) = \mu_c \mu_a \end{aligned} \tag{2.19}$$

e

$$V(\overline{f}) = \sum_{i=1}^{s} (c_i - \mu_c)^2 \left(\frac{n_{i+}}{n}\right) \sum_{j=1}^{r} \frac{(a_j - \mu_a)^2 (n_{+j}/n)}{n-1} \tag{2.20}$$

Para amostras grandes, tem-se, em decorrência do Teorema central do limite, que $\overline{f}$ segue distribuição aproximadamente normal. Logo,

$$\begin{aligned} Q_{CS} &= \frac{\left[\ \overline{f}\ -\ E(\ \overline{f}\ )\right]^2}{V(\ \overline{f}\ )} = \frac{(n-1)\left[\sum_{i=1}^{s}\sum_{j=1}^{r}(c_i - \mu_c)(a_j - \mu_a) n_{ij}\right]^2}{\left[\sum_{i=1}^{s}(c_i - \mu_c)^2 n_{i+}\right]\left[\sum_{j=1}^{r}(a_j - \mu_a)^2 n_{+j}\right]} = \\ &= (n-1)(r_{ac})^2 \end{aligned} \tag{2.21}$$

sendo r_{ac} o coeficiente de correlação de Pearson, se $Z = \frac{\overline{f} - E(\overline{f})}{\sqrt{V(\overline{f})}} \sim N(0,1)$, sabe-se que $Z^2 \sim X^2_{(1)}$ (Stokes *et al.*, 2000; *apud* [6]).

Arrumando a equação (2.22), temos que o r_{ac} é calculado da seguinte forma:

$$r_{ac} = \frac{\left[\sum_{i=1}^{2}\sum_{j=1}^{2}(c_i - \mu_c)(a_j - \mu_a)n_{ij}\right]}{\sqrt{\left[\sum_{i=1}^{2}(c_i - \mu_c)^2 n_{i+}\right]\left[\sum_{j=1}^{2}(a_j - \mu_a)^2 n_{+j}\right]}} \tag{2.22}$$

Por envolver o coeficiente de correlação de Pearson, Q_{CS} é denominada estatástica de correlação. Ainda, como tal coeficiente mede a intensidade de associação linear entre duas variáveis, é possível, nesses casos, expressar a hipótese nula como ausência de tendência linear ($H_0 : r_{ac} = 0$).

2.3 Exemplos retirados da literatura

2.3.1 Exemplo da Estratégia I

Exemplo 2.1: *Problema do melanoma*

Num estudo epidemiológico envolvendo a população norte-americana de homens brancos entre $1969 - 1971$ contou-se o número de novos casos de melanoma (tipo de tumor cancerígeno na pele) classificados pela região norte ou sul do país e pela faixa etária (em 6 níveis) do indivíduo atingido. A Tabela 2.1 registra os resultados obtidos assim como uma estimativa do tamanho de cada uma das populações expostas ao risco (valores entre parênteses), dados obtidos em Gail, 1978 *apud* [5]. Uma das questões de interesse neste problema é saber se a razão das taxas de incidência dessa doença por unidade de exposição entre as duas regiões (i.e., o risco relativo) varia com o grupo etário.

Tabela 2.1: Número de novos casos de melanoma (Populações expostas)

X_1: Faixa	X_2: Região				
Etária	Norte	População em risco	Sul	População em risco	Totais
$1 : < 35$	61	(2880262)	64	(1074226)	$n_{1+} = 125$
$2 : 35 - 44$	76	(564535)	75	(220407)	$n_{2+} = 151$
$3 : 45 - 54$	98	(592983)	68	(198119)	$n_{3+} = 166$
$4 : 55 - 64$	104	(450740)	63	(134084)	$n_{4+} = 167$
$5 : 65 - 74$	63	(270908)	45	(70708)	$n_{5+} = 108$
$6 : \geqslant 75$	80	(161850)	27	(34233)	$n_{6+} = 107$
Totais	$n_{+1} = 482$		$n_{+2} = 342$		$n_{++} = 824$

Fonte: Paulino e Singer (2006)

Foi feito o cálculo dos totais marginais-linha ($n_{i+}, i = 1, 2, ..., 6$) que correspondem às somas das frequências n_{ij} na $i-ésima$ linha e dos totais marginais-coluna ($n_{+j}, j = 1, 2$) que correpondem às somas das frequências n_{ij} na $j-ésima$ coluna. Também foi calculado o total amostral (n_{++} ou n) que corresponde a soma das frequências n_{ij},

para $i = 1, 2, ..., 6$ e $j = 1, 2$.
Observa-se que nesse conjunto de dados, temos duas variáveis definidoras de estratos e subpopulações: X_1 : Faixa etária, que é uma variável politômica ordinal ($i = 1, 2, ..., 6$) e X_2 :Região do país, que é uma variável dicotômica nominal ($j = 1, 2$). A resposta é obtida com a introdução de uma variável binária adicional que indica a ocorrência ou não de caso de melanoma.
Como para este exemplo tem-se um estudo epidemiológico envolvendo a população norte-americana de homens brancos entre $1969 - 1971$, este é um estudo com tempo de duração $T = 3$ anos (ou 36 meses) e temos totais marginais e amostral aleatórios. Com isso, segundo Giolo (2017), um modelo possível para esse estudo é o produto de Poisson, em que $N_{ij} \sim Poisson(\mu_{ij} = T\lambda_{ij})$, $i, j = 1, 2$. Testar a ausência de associação entre X_1 e X_2 significa testar, se para as categorias $j = 1, 2$ de X_2 (em termos das médias μ_{ij}) as proporções de respostas dos indivíduos nas categorias $i = 1, ..., 6$ de X_1 não diferem, ou seja,

$$H_0 : \begin{cases} \frac{\mu_{11}}{\mu_{+1}} = \frac{\mu_{12}}{\mu_{+2}} \left(= \frac{\mu_{1+}}{\mu}\right) \\ \frac{\mu_{21}}{\mu_{+1}} = \frac{\mu_{22}}{\mu_{+2}} \left(= \frac{\mu_{2+}}{\mu}\right) \\ \vdots \qquad \vdots \qquad \vdots \\ \frac{\mu_{61}}{\mu_{+1}} = \frac{\mu_{62}}{\mu_{+2}} \left(= \frac{\mu_{6+}}{\mu}\right) \end{cases}$$

H_1 : Há diferença(s) entre as proporções

Sendo $\mu_{i+} = \sum_{j=1}^{2} \mu_{ij}$; $\mu_{+j} = \sum_{i=1}^{6} \mu_{ij}$ e $\mu_{++} = \sum_{i=1}^{6}\sum_{j=1}^{2} \mu_{ij}$.

Como a hipótese H_0, pode também ser expressa por

$$H_0 : \mu_{ij} = \frac{(\mu_{i+})(\mu_{+j})}{\mu}, i = 1, ..., 6 \text{ e } j = 1, 2 \tag{2.23}$$

o que evidência uma forma multiplicativa nas médias, ela é denominada hipótese de multiplicatividade.

Sob H_0, tem-se que $E(N_{ij}) = \mu_{ij} = \frac{(\mu_{i+})(\mu_{+j})}{\mu}$. Assim os valores das frequências esperadas são obtidos por

$$e_{ij} = \frac{(n_{i+})(n_{+j})}{n}, i = 1, 2, ..., 6 \text{ e } j = 1, 2 \tag{2.24}$$

visto que $\widehat{\mu}_{i+} = N_{i+}$, $\widehat{\mu}_{+j} = N_{+j}$ e $\widehat{\mu} = N$.

Com isso, as frequências esperadas, considerando os dados da Tabela 2.1 são:

$e_{11} = \frac{(n_{1+})(n_{+1})}{n} = \frac{125 \cdot 482}{824} = 73,1189$
$e_{12} = \frac{(n_{1+})(n_{+2})}{n} = \frac{125 \cdot 342}{824} = 51,8811$

e assim por diante, calculando para todas as células temos os resultados apresentados na Tabela 2.2.

Tabela 2.2: Frequências esperadas (e_{ij})

Faixa Etária	Região	
	Norte	Sul
< 35	$73,1189$	$51,8811$
$35-44$	$88,3277$	$62,6723$
$45-54$	$97,1019$	$68,8981$
$55-64$	$97,6869$	$69,3131$
$65-74$	$63,1748$	$44,8252$
$\geqslant 75$	$62,5898$	$44,4101$

Os cálculos da estatística do teste, definida por (2.11) podem ser vistos na Tabela 2.3. Maiores detalhes sobre os testes qui-quadrado podem ser vistos no Apêndice C.

Tabela 2.3: Cálculo de Q_p

n_{ij}	e_{ij}	$\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$
61	$73,1189$	$2,008615$
76	$88,3277$	$1,72055$
98	$97,1019$	$0,008307$
104	$97,6869$	$0,40799$
63	$63,1748$	$0,000484$
80	$62,5898$	$4,82883$
64	$51,8811$	$2,830852$
75	$62,6723$	$2,42487$
68	$68,8981$	$0,011707$
63	$69,3131$	$0,575003$
45	$44,8252$	$0,000682$
27	$44,4101$	$6,825285$
		$Q_p = 21,65723$

Obtém-se, então o valor da estatística do teste $Q_p = \sum_{i=1}^{6}\sum_{j=1}^{2}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = 21,6573$

Temos que a região crítica para esse teste, ao nível de 5%, é dada por $RC = \{Q_p : Q_p > 11,0705\}$, então a hipótese H_0 é rejeitada e podemos afirmar que a distribuição do número de novos casos de melanoma por idade, não é a mesma nas duas regiões.

2.3.2 Exemplo da Estratégia II

Exemplo 2.2 *Problema do uso de tabaco por adolescentes*

Os dados dispostos na Tabela 2.4 são de um estudo realizado com adolescentes com o objetivo de investigar a existência de associação entre o uso de tabaco e a consciência do risco em usá-lo. Observa-se uma tendência crescente de não uso de tabaco à medida que a consciência do risco em usá-lo aumenta. Como nesse exemplo apenas o total

amostral n é fixo, segue que o modelo associado ao estudo descrito é o multinomial. Desse modo, as hipóteses de interesse são estabelecidas como

$$\begin{cases} H_0 : p_{ij} = (p_{i+})(p_{+j}) \text{ para } i = 1, 2, 3 \text{ e } j = 1, 2 \\ H_1 : p_{ij} \neq (p_{i+})(p_{+j}) \text{ para pelo menos um par } (i, j). \end{cases}$$

Tabela 2.4: Estudo sobre o uso de tabaco por adolescentes

	Uso de Tabaco		
Consciência do risco	Não	Sim	Totais
Mínima	70	33	103
Moderada	202	40	242
Substancial	218	11	229
Totais	490	84	574

Fonte: Bauman et al. (1989), *apud* Giolo (2017).

Como a variável *uso de tabaco* é dicotômica nominal e a variável *consciência do risco* é politômica ordinal, é possível considerar escores para ambas. É usual assumir os valores 0 e 1 para as categorias de uma variável dicotômica. Assim, se forem assumidos os escores $c = (c_1, c_2, c_3) = (1, 2, 3)$ para as categorias "*mínima*", "*moderada*" e "*substancial*"da variável consciência do risco de uso do tabaco, e os escores $a = (a_1, a_2) = (0, 1)$ para as categorias "*não usa*" e "*usa*" tabaco, respectivamente, é possível definir o escore médio de acordo com a equação (2.17).

Logo, $\overline{F} = \sum_{i=1}^{3} \sum_{j=1}^{2} c_i a_j p_{ij}$, e calculando o estimador através da equação (2.18) temos que $\overline{f} = 0,254355$. Sob H_0, a esperança e a variância de $\overline{f}$ são dadas, respectivamente, pelas equações 2.19 e 2.20, assim temos que $E(\overline{f}) = 0,324807$ e $V(\overline{f}) = 0,000116$.

Em decorrência do Teorema central do limite, que $\overline{f}$ segue distribuição aproximadamente normal e considerando os dados dispostos na Tabela 2.4, tem-se $Q_{CS} = 42,94 (p < 0,0001)$, o que evidência associação entre o uso de tabaco e a consciência do risco em usá-lo. Ainda, calculando r_{ac} através da equação 2.22, tem-se que $r_{ac} = -0,274$, assim é possível concluir que o uso do tabaco diminui à medida que a consciência aumenta.

2.3.3 Exemplo da Estratégia III

Exemplo 2.3 *Problema do uso de fio dental*

Num estudo odontopediátrico envolvendo uma escola da Rede Pública do Município de São Paulo foram selecionadas 60 crianças de ambos os sexos, divididas igualmente pelas faixas etárias de 5-8 anos e de 9-12 anos, e cada uma delas foi classificada segundo as seguintes variáveis:

i) **Frequência de uso do fio dental**, nas categorias "nunca usa ou usa raramente"e "usa regular ou frequentemente", denotadas abreviadamente por "insuficiente"e "boa";

ii) **Capacidade motora,** traduzida na maior ou menor habilidade no uso do fio dental, com as categorias "não consegue usar ou consegue usar embora incorretamente nos dentes anteriores (incisivos e caninos)"e "consegue usar corretamente pelo menos nos dentes anteriores", descritas abreviadamente por "inábil"e "razoável".

Este estudo visava verificar se a faixa etária influenciava a frequência e habilidade do uso de fio dental para ambos os sexos. As frequências observadas estão resumidas na Tabela 2.5.

Tabela 2.5: Frequências observadas de 120 crianças (Escola municipal de São Paulo)

	Faixa		Habilidade	
Sexo	etária	Frequência	Inábil	Razoável
M	$5-8$	insuficiente	19	5
M	$5-8$	boa	4	2
M	$9-12$	insuficiente	5	8
M	$9-12$	boa	0	17
F	$5-8$	insuficiente	11	6
F	$5-8$	boa	7	6
F	$9-12$	insuficiente	2	5
F	$9-12$	boa	1	22

Fonte: Paulino e Singer (2006)

Similiar ao que foi feito nas Estratégia I e na Etratégia II, as estatísticas Q_p e Q_{CS}, podem ser usadas para testar tais hipóteses. Para este caso, será utilizado a estatística Q_{CS}, pois, além do grau de evidência para uma associação, pode ser descrito a força (ou intensidade) dessa associação. A Tabela 2.5 é subdividida em quatro tabelas. Como as variáveis Faixa Etária e Frequência são dicotômicas, qualitativas e ordinais, é possível considerar escores para ambas. Assim, assumimos o escores $c = (c_1, c_2) = (1, 2)$ para as categorias $5-8$ (Faixa I) e $9-12$ (Faixa II) da variável Faixa Etária, e os escores $a = (a_1, a_2) = (1, 2)$ para as categorias Insulficiente e Boa.

Tabela 2.6: Frequência com Faixa Etária (Sexo Masculino)

Sexo Masculino			
	Faixa Etária		
Frequência	Faixa I	Faixa II	Total
Insuficiente	24	13	37
Boa	6	17	23
Total	30	30	60

O escore médio e o seu estimador, são calculados de acordo com as equações 2.17 e 2.18. Substituindo os valores da Tabela 2.6 na equação 2.18, temos que $\overline{f} = 2,1666667$.

Calculando $E(\overline{f})$ e $V(\overline{f})$ com as equações (2.19) e (2.20), obtem-se $E(\overline{f}) = 2,075$ e $V(\overline{f}) = 0,001001654$.

Em decorrência do Teorema central do limite, que $\overline{f}$ segue distribuição aproximadamente normal e considerando os dados dispostos na Tabela 2.6, tem-se $Q_{CS} = 8,38895$.

Temos que a região crítica para esse teste, ao nível de 5%, é dada por $RC = \{Q_{cs} : Q_{cs} > 3,841\}$ e a hipótese H_o é rejeitada. Podemos verificar isso também através do $p - value$, que é dado por $P(Q > Q_{cs}) = P(Q > 8,39) = 0,0038$, logo $p < 0,05 = \alpha$.

Com isso podemos dizer que existe uma relação linear entra as Faixa Etária e a Frequência.

Ainda, como $r_{ac} = 0,377$, é possível concluir que com o aumento da idade das crianças do sexo masculino há um aumento na frequência com que ela usa o fio dental.

Analogamente, foram realizados os testes considerando o restante dos dados da Tabela 2.5, subdivididos em três tabelas.

Tabela 2.7: Habilidade com Faixa Etária (Sexo Masculino)

Sexo Masculino		
	Faixa Etária	
Habilidade	Faixa I	Faixa II
Inábil	23	5
Razoável	7	25

Considerando os dados dispostos na Tabela 2.7, tem-se $Q_{CS} = 21,3348$ $P(Q > Q_{cs}) = P(Q > 21,3348) = 0,0000038567 \Rightarrow p < 0,05 = \alpha$, o que evidencia uma relação linear entra as Faixa Etária e a Habilidade. Ainda, como $r_{ac} = 0,6013$, é possível concluir que com o aumento da idade das crianças do sexo masculino há um aumento na habilidade no uso do fio dental.

Tabela 2.8: Frequência com Faixa Etária (Sexo Feminino)

Sexo Feminino		
	Faixa Etária	
Frequência	Faixa I	Faixa II
Insuficiente	17	7
Boa	13	23

Considerando os dados dispostos na Tabela 2.8, tem-se $Q_{CS} = 6,8287$ $P(Q > Q_{cs}) = P(Q > 6,8287) = 0,009 \Rightarrow p < 0,05 = \alpha$, o que evidencia uma relação linear entra as Faixa Etária e a Frequência. Ainda, como $r_{ac} = 0,3402$, é possível concluir que com o aumento da idade das crianças do sexo feminino há um aumento na Frequência no uso do fio dental.

Tabela 2.9: Habilidade com Faixa Etária (Sexo Feminino)

Sexo Feminino		
	Faixa Etária	
Habilidade	Faixa I	Faixa II
Inábil	18	3
Razoável	12	27

Considerando os dados dispostos na Tabela 2.9, tem-se $Q_{CS} = 16,0209$ $P(Q > Q_{cs}) = P(Q > 16,0209) = 0,00005673 \Rightarrow p < 0,05 = \alpha$, o que evidencia uma relação linear entra as Faixa Etária e a Habilidade. Ainda, como $r_{ac} = 0,5241$, é possível concluir que com o aumento da idade das crianças do sexo feminino há um aumento na habilidade no uso do fio dental.

3 Regressão Logística

3.1 Regressão logística dicotômica

Os modelos de regressão binomial são utilizados para modelar a associação entre um conjunto de variáveis explicativas (X) e uma variável resposta (Y) binária ou dicotômica. As variáveis explicativas são, em geral, um misto de variáveis categóricas e contínuas.

Considerando que a variável resposta assume os valores 0 e 1, $P(Y = 1|x) = E(Y|x)$, visto que $E(Y|x) = 0P(Y = 0|x) + 1P(Y = 1|x) = P(Y = 1|x)$. Aqui a relação entre X e $E(Y|x)$ não é linear e sim sigmoidal e $E(Y|x) \in [0, 1]$, o que motivou o uso da distribuição logística para modelar $E(Y|x)$ que tem função distribuição acumulada expressa por

$$F(X) = \frac{1}{1+e^{-x}} = \left[1+e^{-x}\right]^{-1} = \left[1+\frac{1}{e^x}\right]^{-1} = \left[\frac{e^x+1}{e^x}\right]^{-1} = \frac{e^x}{1+e^{-x}} \tag{3.1}$$

Considerando um conjunto de variáveis explicativas $\mathbf{X} = (X_1, ..., X_P)$ e denotando $E(Y|x) = P(Y = 1|\mathbf{x}) = p(\mathbf{x})$ segue que o modelo de regressão logística é expresso por:

$$p(\mathbf{x}) = \frac{\exp(\beta'\mathbf{x})}{1+\exp(\beta'\mathbf{x})} = \frac{\exp(\beta_0 + \sum_{k=1}^{p}\beta_k x_k)}{1+\exp(\beta_0 + \sum_{k=1}^{p}\beta_k x_k)} \tag{3.2}$$

em que $\mathbf{x} = (1, x_1, ..., x_P)$ denota o vetor com constante 1 e os valores observados das variáveis explicativas $\mathbf{X}$, β_0 é uma constante e $\beta_k (k = 1, ..., p)$ são os p parâmetros de regressão.

O modelo (3.2) fornece, portanto, a probabilidade de um indíviduo com valores observados $\mathbf{x}$ apresentar a resposta de interesse. Consequentemente,

$$1 - p(\mathbf{x}) = 1 - \frac{\exp(\beta_0 + \sum_{k=1}^{p} \beta_k x_k)}{1 + \exp(\beta_0 + \sum_{k=1}^{p} \beta_k x_k)} =$$
$$= \frac{1 + \exp(\beta_0 + \sum_{k=1}^{p} \beta_k x_k) - \exp(\beta_0 + \sum_{k=1}^{p} \beta_k x_k)}{1 + \exp(\beta_0 + \sum_{k=1}^{p} \beta_k x_k)} =$$
$$= \frac{1}{1 + \exp(\beta_0 + \sum_{k=1}^{p} \beta_k x_k)} \tag{3.3}$$

fornece a probabilidade deste indíviduo não apresentar a referida resposta.
A transformação a seguir fornece um modelo linear e é denominado logito.

$$\ln\left[\frac{p(\mathbf{x})}{1 - p(\mathbf{x})}\right] = \beta_0 + \sum_{k=1}^{p} \beta_k x_k = \beta'\mathbf{x} \tag{3.4}$$

então a chance segundo a definição dada em (1.10) é

$$chance = \frac{p(\mathbf{x})}{1 - p(\mathbf{x})} = \exp(\beta'\mathbf{x})$$

3.1.1 Estimação dos parâmetros

A estimação do vetor de parâmetros β em regressão logística é realizada, em geral, pelo método da máxima verossimilhança. Sendo assim, é necessário obter a função de verossimilhança, que expressa a probabilidade dos dados observados como função dos parâmetros desconhecidos.
Por definição, a função de verossimilhança para um conjunto de n observações independentes ($l = 1, ..., n$) é expressa pelo produto de suas "contribuições"individuais, isto é,

$$L(\beta) = \prod_{l=1}^{n} P(Y = y_l|\mathbf{x}_l) = \prod_{l=1}^{n} [p(\mathbf{x}_l)]^{y_l} [1 - p(\mathbf{x}_l)]^{1-y_l} \tag{3.5}$$

sendo $y_l = \begin{cases} 1 & \text{se o } l - \textit{ésimo} \text{ indíviduo apresentou resposta de interesse} \\ 0 & \text{caso o contrário.} \end{cases}$

Os estimadores dos parâmetros que compõem o vetor β são os valores que maximizam a função de verossimilhança $L(\beta)$ ou o logaritmo dessa função $l(\beta) = \ln[L(\beta)]$, dada por

$$l(\beta) = \ln[L(\beta)] = \sum_{l=1}^{n} \{y_l \ln [p(\mathbf{x}_l)] + (1 - y_l) \ln [1 - p(\mathbf{x}_l)]\} \tag{3.6}$$

Diferenciando $l(\beta)$ com respeito a cada parâmetro $\beta_k = (k = 0, 1, ..., p)$, obtem-se o sistema com $p + 1$ equações de máxima verossimilhança.

$$\frac{\partial l(\beta)}{\partial \beta_0} = \frac{\partial \left[\sum_{l=1}^{n} y_l \ln\left[\frac{e^{\beta_0}}{1+e^{\beta_0}}\right] + \sum_{l=1}^{n}(1-y_l)\ln\left[\frac{1}{1+e^{\beta_0}}\right]\right]}{\partial \beta_0} =$$

$$= \sum_{l=1}^{n} y_l \left[\frac{1+e^{\beta_0}}{e^{\beta_0}} . \frac{e^{\beta_0}(1+e^{\beta_0}) - e^{\beta_0}e^{\beta_0}}{(1+e^{\beta_0})^2}\right] + \sum_{l=1}^{n}(1-y_l)\left[\frac{1+e^{\beta_0}}{1} . \frac{0.(1+e^{\beta_0}) - 1.e^{\beta_0}}{(1+e^{\beta_0})^2}\right] =$$

$$= \sum_{l=1}^{n} y_l \left[\frac{e^{\beta_0}(1+e^{\beta_0}-e^{\beta_0})}{e^{\beta_0}(1+e^{\beta_0})}\right] + \sum_{l=1}^{n}(1-y_l)\left[\frac{-e^{\beta_0}}{(1+e^{\beta_0})}\right] =$$

$$= \sum_{l=1}^{n} y_l \left[\frac{1}{1+e^{\beta_0}}\right] - \sum_{l=1}^{n}(1-y_l)\left[\frac{e^{\beta_0}}{1+e^{\beta_0}}\right] =$$

$$= \left[\frac{1}{1+e^{\beta_0}}\right] \sum_{l=1}^{n} y_l - \frac{e^{\beta_0}}{1+e^{\beta_0}}(n - \sum_{l=1}^{n} y_l) =$$

$$= \left[\frac{1}{1+e^{\beta_0}}\right] \left[\sum_{l=1}^{n} y_l - e^{\beta_0} n + e^{\beta_0} \sum_{l=1}^{n} y_l\right] =$$

$$= \left[\frac{1}{1+e^{\beta_0}}\right] \left[\sum_{l=1}^{n} y_l [1+e^{\beta_0}] - n e^{\beta_0}\right] =$$

$$= \sum_{l=1}^{n} y_l - n\frac{e^{\beta_0}}{1+e^{\beta_0}} = \sum_{l=1}^{n} y_l - \sum_{l=1}^{n} 1\frac{e^{\beta_0}}{1+e^{\beta_0}} =$$

$$= \sum_{l=1}^{n} y_l - \sum_{l=1}^{n} p(\mathbf{x}_l) = \sum_{l=1}^{n} [y_l - p(\mathbf{x}_l)]$$

e

$$\frac{\partial l(\beta)}{\partial \beta_1} = \frac{\partial \left[\sum_{l=1}^{n} y_l \ln\left[\frac{e^{\beta_1 x_1}}{1+e^{\beta_1 x_1}}\right] + \sum_{l=1}^{n}(1-y_l)\ln\left[\frac{1}{1+e^{\beta_1 x_1}}\right]\right]}{\partial \beta_1} =$$

$$= \sum_{l=1}^{n} y_l \left[\frac{1+e^{\beta_1 x_1}}{e^{\beta_1 x_1}} \frac{x_1 e^{\beta_1 x_1}(1+e^{\beta_1 x_1}) - x_1 e^{\beta_1 x_1} e^{\beta_1 x_1}}{(1+e^{\beta_1 x_1})}\right] +$$

$$+ \sum_{l=1}^{n}(1-y_l)\left[\frac{1+e^{\beta_1 x_1}}{1} . \frac{0.(1+e^{\beta_1 x_1}) - 1.x_1 e^{\beta_1 x_1}}{(1+e^{\beta_1 x_1})^2}\right] =$$

$$= \sum_{l=1}^{n} y_l \left[\frac{x_1 e^{\beta_1 x_1}(1+e^{\beta_1 x_1}-e^{\beta_1 x_1})}{e^{\beta_1 x_1}(1+e^{\beta_1 x_1})}\right] + \sum_{l=1}^{n}(1-y_l)\left[\frac{-x_1 e^{\beta_1 x_1}}{(1+e^{\beta_1 x_1})}\right] =$$

$$= \sum_{l=1}^{n} y_l \left[\frac{x_1}{1+e^{\beta_1 x_1}}\right] - \sum_{l=1}^{n}(1-y_l)\left[\frac{x_1 e^{\beta_1 x_1}}{(1+e^{\beta_1 x_1})}\right] =$$

$$
\begin{aligned}
&= \sum_{l=1}^{n} y_l x_1 [1 - p(\mathbf{x}_1)] - \sum_{l=1}^{n} (1 - y_l) [x_1 p(\mathbf{x}_1)] = \\
&= \sum_{l=1}^{n} x_1 [y_l - y_l p(\mathbf{x}_1)] - \sum_{l=1}^{n} [x_1 p(\mathbf{x}_1) - y_l x_1 p(\mathbf{x}_1)] = \\
&= \sum_{l=1}^{n} x_1 [y_l - y_l p(\mathbf{x}_1)] - \sum_{l=1}^{n} x_1 [p(\mathbf{x}_1) - y_l p(\mathbf{x}_1)] = \\
&= \sum_{l=1}^{n} x_1 [y_l - p(\mathbf{x}_1)]
\end{aligned}
$$

$$
Logo, \begin{cases} \sum_{l=1}^{n} [y_l - p(\mathbf{x}_l)] = 0 \\ \sum_{l=1}^{n} x_{l_1} [y_l - p(\mathbf{x}_{l_1})] = 0 \\ \vdots \\ \sum_{l=1}^{n} x_{lp} [y_l - p(\mathbf{x}_{l_p})] = 0 \end{cases} \tag{3.7}
$$

O sistema (3.7) não é linear em $\beta_k, k = 0, 1, ..., p$, e portanto é necessário o uso de métodos interativos para a sua solução. O método de Newton-Raphson, implementado via um algoritmo de mínimos quadrados reponderados interativamente (IRLS), é usualmente utilizado para essa finalidade (McCullagh; Nelder,1989; Pawitan,2001;*apud* [6]).

A solução do sistema produz o vetor com os estimadores de máxima verossimilhança para os parâmetros, denotado por $\widehat{\beta}$.

Para a estimação das variâncias e covariâncias do vetor $\widehat{\beta}$ utiliza-se o resultado denominado de Normalidade assintótica dos EMV.

$$
\text{Se } \widehat{\beta} \text{ é o } EMV(\beta) \text{ então } \widehat{\beta} \overset{a}{\sim} N[\beta, I_n^{-1}(\widehat{\beta})] \tag{3.8}
$$

sendo que $I_n(\beta)$ é a matriz de informação de Fisher definida por $I_n(\beta) = -\left[\frac{\partial^2 l(\beta)}{\partial \beta^2}\right]$.

Assim a estimação das variâncias-covariâncias do vetor $\widehat{\beta}$ pode ser feita a partir da matriz de derivadas parciais de segunda ordem de (3.6). Tais derivadas, para $k, k' = 0, 1, ..., p$, têm a seguinte forma geral

$$
\frac{\partial^2 \ln L(\beta)}{\partial \beta_k^2} = -\sum_{l=1}^{n} x_{lk}^2 p(\mathbf{x}_l)[1 - p(\mathbf{x}_l)] \tag{3.9}
$$

$$
\frac{\partial^2 \ln L(\beta)}{\partial \beta_k \partial \beta_{k'}} = -\sum_{l=1}^{n} x_{lk} x_{lk'} p(\mathbf{x}_l)[1 - p(\mathbf{x}_l)] \tag{3.10}
$$

O $k-ésimo$ elemento da diagonal da matriz, denotado por $\widehat{Var} = (\widehat{\beta}_k)$, corresponde à variância de $\widehat{\beta}_k$. Já o elemento na $k-ésima$ linha e $k'-ésima$ coluna, denotada por $Cov(\widehat{\beta}_k, \widehat{\beta}_{k'})$, corresponde à covariância entre $\widehat{\beta}_k$ e $\widehat{\beta}_{k'}$.

A matriz de informação é expressa por $I_n(\beta) = \mathbf{X'VX}$, em que $\mathbf{X}$ é uma matriz com n linhas e $(p+1)$ colunas contendo valores iguais 1 na primeira coluna e os valores das p variáveis explicativas nas demais colunas e $\mathbf{V}$ é uma matriz diagonal de n linhas e n colunas com elementos $p(x_l)[1-p(x_l)]$ na diagonal. Isto é,

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}$$

$$\mathbf{V} = \begin{bmatrix} p(\mathbf{x}_1)[1-p(\mathbf{x}_1)] & 0 & \dots & 0 \\ 0 & p(\mathbf{x}_2)[1-p(\mathbf{x}_2)] & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & p(\mathbf{x}_n)[1-p(\mathbf{x}_n)] \end{bmatrix}$$

3.1.2 Significância dos efeitos das variáveis

Obtidas as estimativas dos parâmetros $\beta_k (k = 0, 1, ..., p)$, faz-se necessário avaliar a adequação do modelo ajustado. O interesse está em acessar a significância dos efeitos das variáveis presentes no modelo. O princípio em regressão logística é o mesmo usado em regressão linear, ou seja, comparar os valores observados da variável respostas com os valores preditos pelos modelos com e sem a variável sob investigação.

Em regressão logística, a comparação pode ser feita por meio de testes como o da razão de verossimilhança (TVR), em que a função de verossimilhança do modelo sem as variáveis (L_S) é comparada com a função de verossimilhança do modelo com as variáveis (L_C). Formalmente, o teste é expresso pela estatística

$$TRV = -2\ln\left[\frac{L_S}{L_C}\right] = 2\ln(L_C) - 2\ln(L_S). \tag{3.11}$$

Nota-se que o logaritmo da razão das verossimilhanças é multiplicado por -2. Isso é feito para que se obtenha uma quantidade cuja distribuição seja conhecida (no caso a distribuição qui-quadrado) de modo que tal quantidade possa ser utilizada para a realização de teste de hipóteses. Em regressão logística, a estatística

$$D = -2\ln\left[\frac{\text{verossimilhança do modelo sob estudo}}{\text{verossimilhança do modelo saturado}}\right] \tag{3.12}$$

é denominada *deviance* e assim, a estatística TRV apresentada anteriormente, pode ser vista como a diferença entre duas *deviances*, a do modelo sem as variáveis explicativas e a do modelo com tais variáveis, isto é, o valor observado da variável resposta como sendo o valor predito pelo modelo saturado. Um modelo saturado é aquele que contém tantos parâmetros quantos dados existirem.

$$TRV = \left[-2\ln\left(\frac{\text{verossimilhança do modelo sem as variáveis}}{\text{verossimilhança do modelo saturado}}\right)\right] - \left[-2\ln\left(\frac{\text{verossimilhança do modelo com as variáveis}}{\text{verossimilhança do modelo saturado}}\right)\right] \tag{3.13}$$

de modo que $TRV = 2\ln(L_C) - 2\ln(L_S)$.

Sob a hipótese nula de que os p coeficientes associados às variáveis no modelo não diferem de zero, a estatística TRV segue distribuição qui-quadrado com p graus de liberdade. A rejeição da hipótese nula tem, a interpretação análoga àquela em regressão linear, ou seja, a de que pelo menos um dos p coeficientes difere de zero.

3.1.3 Análise de *deviance* e seleção de modelos

Uma tabela de análise de variância similar àquela obtida em regressão linear pode ser construída em regressão logística. Nesse caso, tal tabela é denominada análise de *deviance* (ANODEV), podendo ser vista como uma generalização da análise de variância. O objetivo da ANODEV é obter, a partir de uma sequência de modelos encaixados, os efeito de fatores, variáveis e suas interações.

Para uma sequência de modelos encaixados, tendo estes a mesma distribuição e função de ligação, utiliza-se a *deviance* como uma medida de discrepância do modelo,a fim de construir uma tabela contendo as diferenças de *deviance* como a apresentada na Tabela 3.1.

Tabela 3.1: *Deviance* e suas diferenças associadas a um estudo com resposta binária e duas variáveis explicativas categóricas X_1 e X_2 binárias

Modelo	*g.l.*	*Deviances*	*TRV*	*Dif.g.l.*
$Nulo$	gl_n	D_N		
X_1	gl_{n-1}	D_{X_1}	$D_N - D_{X_1}$	1
$X_2 \vert X_1$	gl_{n-2}	D_{X_1,X_2}	$D_{X_1} - D_{X_1,X_2}$	1
$X_1 * X_2 \vert X_1, X_2$	gl_{n-3}	D_{X_1,X_2,X_1*X_2}	$D_{X_1,X_2} - D_{X_1,X_2,X_1*X_2}$	1

Nota: gl_n =graus de liberdade do modelo nulo = n^o de subpopulações -1, dif. = diferença. Fonte: Giolo (2017)

A partir das *deviances* e de suas diferenças, pode-se usar o teste da razão de verossimilhanças, descrito anteriormente, para testar a significância da inclusão de determinadas variáveis, bem como de suas interações no modelo. Em outras palavras, pode-se avaliar o quanto da *deviance* associada ao modelo nulo é explicada pela inclusão de termos no modelo.

Uma observação importante sobre o TRV é que, na presença de variáveis explicativas com dados ausentes (do inglês *missing data*), sua utilização fica inviável. Isso porque o tamanho amostral nos modelos sequenciais ajustados dependerá das variáveis que o compõem e, desse modo, não seria apropriado fazer uso das diferenças de *deviances* entre esses modelos. Uma alternativa para testar a significância dos coeficientes na

presença de dados ausentes seria o teste de Wald (1943) *apud* [6], frequentemente utilizado para testar hipóteses relativas a um único parâmetro $\beta_k, k = 0, 1, ..., p$. Sob a hipótese nula $H_0 : \beta_k = 0$, a estatística para esse teste fica expressa por

$$W = \frac{\left(\widehat{\beta}_k\right)^2}{Var\left(\widehat{\beta}_k\right)} \tag{3.14}$$

que, sob H_0, segue a distribuição qui-quadrado com 1 grau de liberdade.

A comparação de modelos pode também ser realizada por meio de critérios que sumarizam o quão próximas as probabilidades preditas pelo modelo tendem a estar das probabilidades verdadeiras. Um desses critérios, o de informação de Akaike (AIC), é dado por

$$AIC = -2(\log \text{ verossimilhança } - \text{ número de parâmetros do modelo}) \tag{3.15}$$

O modelo que minimiza o AIC é considerado como sendo o que fornece as melhores probabilidades preditas.

3.1.4 Qualidade do modelo ajustado

Uma vez selecionado o modelo, o passo seguinte é avaliar o quão bem ele se ajusta aos dados, ou seja, o quão próximos os valores preditos por este modelo se encontram de seus correspondentes valores observados. As estatísticas de teste utilizadas para essa finalidade são, em geral, denominadas estatísticas de qualidade do ajuste, uma vez que comparam, de maneira apropriada, as diferenças entre os valores observados e preditos.

Duas estatísticas tradicionais de qualidade do ajuste são: a) a estatística qui-quadrado de Pearson, Q_p que é baseada nos resíduos de Pearson; e b) a estatística qui-quadrado da razão de verossimilhanças, Q_L, também conhecida por qui-quadrado *deviance* por se basear nos resíduos *deviance*.

Tais estatística são expressas, respectivamente, por

$$Q_p = \sum_{i=1}^{S}\sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \text{ e } Q_L = \sum_{i=1}^{S}\sum_{j=1}^{2} n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right), \tag{3.16}$$

em que e_{ij} são as quantidades preditas pelo modelo e definidas por

$$\begin{aligned} e_{ij} &= n_{i+}\widehat{p}(\mathbf{x}_i) \quad \text{para } j = 1 \\ e_{ij} &= n_{i+}(1 - \widehat{p}(\mathbf{x}_i)) \quad \text{para } j = 2. \end{aligned}$$

Sob a hipótese nula de que o modelo se ajusta bem aos dados, Q_p e Q_L seguem distribuição aproximada qui-quadrado com os graus de liberdade definidos pela diferença entre o número de subpopulações (linha da tabela de dados) e o número de parâmetros do modelo. Na prática, a aproximação para a distribuição qui-quadrado será razoável se: i) cada n_{i+} for > 10; ii) 80% das frequências preditas for ≥ 5; iii) todas as demais frequências preditas > 2; e iv) nenhuma frequência observada for zero.

3.1.5 Diagnóstico em regressão logística

As estatística Q_p e Q_L, descritas na Secção 3.14 e utilizadas para verificar a qualidade de ajuste do modelo de regressão logística, fornecem um único valor o qual resume a concordância entre os valores observados e preditos pelo modelo. A limitação dessas estatísticas é que este único valor é utilizado para resumir uma quantidade considerável de informação. Assim, é importante que outras medidas sejam examinadas a fim de se averiguar se o ajuste é válido sobre todo o conjunto de padrões (combinações das categorias) das variáveis explicativas ou fatores.

Com essa finalidade, Pregibon(1981) *apud* [6] estendeu os métodos de diagnóstico utilizados em regressão linear para a regressão logística e, para isso, fez uso dos componentes individuais das estatísticas qui-quadrado de Pearson (Q_p) e *deviance* (Q_L), uma vez que esses componentes são funções dos valores observados e preditos pelo modelo.

Assim, se em uma tabela de contigência $s \times 2$ existirem n_{i+} indivíduos em cada uma das s linhas, dos quais n_{i1} apresentam a resposta de interesse ($Y = 1$), define-se o $i - ésimo$ resíduo de Pearson por

$$c_i = \frac{n_{i1} - (n_{i+})\widehat{p}(\mathbf{x}_i)}{\sqrt{(n_{i+})\widehat{p}(\mathbf{x}_i)[1 - \widehat{p}(\mathbf{x}_i)]}}, i = 1, ..., s, \tag{3.17}$$

com $\widehat{p}(\mathbf{x}_i)$ a probabilidade $P(Y = 1|\mathbf{x}_i)$ predita pelo modelo para a $i - ésima$ linha (subpopulação). Tais resíduos são denominados resíduos de Pearson devido à soma deles ao quadrado ser igual a Q_p, isto é, $Q_p = \sum_{i=1}^{s} c_i^2$.

A inspeção dos resíduos c_i, $i = 1, ..., s$, auxilia a determinar quão bem o modelo se ajusta às subpopulações individuais. Resíduos excedendo os valores $\pm 2,5$ (ou $\pm 3,0$) indicam possível falta de ajuste.

Quanto ao $i - ésimo$ resíduo *deviance*, este é definido por

$$d_i = \pm \left[2n_{i1} \ln \left(\frac{n_{i1}}{e_{i1}} \right) + 2(n_{i+} - n_{i1}) \ln \left(\frac{n_{i+} - n_{i1}}{n_{i+} - e_{i1}} \right) \right]^{\frac{1}{2}} \tag{3.18}$$

para $i = 1, ..., s$, em que $e_{i1} = (n_{i+})\widehat{p}(\mathbf{x}_i)$. O sinal de d_i é definido a partir da diferença $(n_{i1} - e_{i1})$. Se esta é negativa, d_i será negativo. Caso contrário, é positivo. A soma dos resíduos *deviance* ao quadrado resulta na estatística Q_L, isto é, $\sum_{i=1}^{s}(d_i)^2$. A partir da inspeção dos resíduos *deviance* é possível observar a presença de resíduos não usuais (demasiadamente grandes), bem como a presença de valores atípicos (do inglês *outliers*) ou, ainda, padrões sistemáticos de variação indicando a escolha de um modelo possivelmente não muito adequado.

Nota-se que as estatísticas de diagnóstico apresentadas permitem que o analista identifique padrões de variáveis que estão com ajuste pobre. Após essa identificação, pode-se avaliar a importância que eles têm na análise. Essa avaliação é similar ao que é feito em regressão linear, em que os padrões com ajuste pobre são removidos a fim de se verificar o seu impacto nas estimativas dos parâmetros, bem como nas estatísticas Q_p e Q_L.

3.2 Regressão logística multinomial

O modelo proposto para análise de dados caracterizados por uma variável resposta Y politômica nominal com Y seguindo a distribuição multinomial é denominado

MLCR.

Considera-se Y com r categorias ($r > 2$) e denota-se por $p_j(\mathbf{x})$ a probabilidade de ocorrência da categoria j ($j = 1, ..., r$) para um dado vetor $\mathbf{x}$ de valores de p variáveis explicativas tal que $\sum_{j=1}^{r} p(\mathbf{x}) = 1$.

Os logitos no MLCR se baseiam em fixar uma categoria de referência, usualmente a última. Fixada a categoria r como referência, o modelo fica expresso em termos nos logitos por

$$\ln\left[\frac{p_j(\mathbf{x})}{p_r(\mathbf{x})}\right] = \ln\left[\frac{P(Y=j|\mathbf{x})}{P(Y=r|\mathbf{x})}\right] = \beta_{0j} + \beta_j'\mathbf{x}, \tag{3.19}$$

em que $j = 1, ..., r-1$ indexa os $r-1$ logitos.

O MCLR assume intercepto β_0 e o vetor β diferentes para cada logito, o que implica que os efeitos das covariáveis variam de acordo com a categoria de resposta que está sendo comparada com a categoria de referência.

Em termos das probabilidades de resposta, as equações que expressam o modelo são:

$$p_j(\mathbf{x}) = \frac{\exp(\beta_{0j} + \beta_j'\mathbf{x})}{1 + \sum_{j=1}^{r-1} \exp(\beta_{0j} + \beta_j'\mathbf{x})}, j = 1, ..., r-1, \tag{3.20}$$

e

$$p_r(\mathbf{x}) = \frac{1}{1 + \sum_{j=1}^{r-1} \exp(\beta_{0j} + \beta_j'\mathbf{x})}, \tag{3.21}$$

Tal que $\sum_{j=1}^{r} p_j(\mathbf{x}) = 1$.

A estimação dos parâmetros do modelo pode ser realizada por meio do método da máxima verossimilhança.

Para $i = 1, ..., n$ em que y_{ij} é dado por:

$$y_{ij} = \begin{cases} 1 & \text{se a resposta do indivíduo } i \text{ está na categoria } j,\ j = 1, ..., r \\ 0 & \text{caso contrário} \end{cases}$$

Com $\sum_{j=1}^{r} p_j(\mathbf{x}) = 1$ tem-se o logaritmo da função de verossimilhança dado por

$$l = \ln \prod_{i=1}^{n} \left\{ \prod_{j=1}^{r} [p_j(\mathbf{x}_i)]^{y_{ij}} \right\} = \ln \prod_{i=1}^{n} \left\{ \prod_{j=1}^{r-1} [p_j(\mathbf{x}_i)]^{y_{ij}} \, [p_r(\mathbf{x}_i)]^{y_{ir}} \right\}.$$

Como $y_{ir} = 1 - \sum_{y=1}^{r-1} y_{ij}$, segue que

$$
\begin{aligned}
l &= \ln \prod_{i=1}^{n} \left\{ \prod_{j=1}^{r} \left[p_j(\mathbf{x}_i) \right]^{y_{ij}} \left[p_r(\mathbf{x}_i) \right]^{1-\sum_{y=1}^{r-1} y_{ij}} \right\} = \\
&= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{r-1} y_{ij} \left(\beta_{0j} + \beta_j' \mathbf{x}_i \right) - \ln \left[1 + \sum_{j=1}^{r-1} \exp \left(\beta_{0j} + \beta_j' \mathbf{x}_i \right) \right] \right\}.
\end{aligned}
$$

A maximização de l para obtenção dos estimadores de máxima verossimilhança dos parâmetros é realizada com método de Newton-Raphson.

Os estimadores seguem a distribuição assintótica normal, com suas variâncias assintóticas dadas pelos elementos da diagonal da inversa da matriz de informação. Maximização de l para obtenção dos estimadores de máxima verossimilhança dos parâmetros é realizada com auxílio do método de Newton-Raphson. Os estimadores seguem distribuição assintótica normal, com seus respectivos erros-padrão assintóticos correspondendo à raiz quadrada dos elementos da diagonal da inversa da matriz de informação.

Tendo em vista que a maximização de l deve satisfazer simultaneamente os $r-1$ logitos que especificam o MLCR, vale mencionar que o tamanho amostral necessita ser grande o suficiente para que não haja problemas quanto à estimação dos parâmetros. Uma abordagem alternativa para o ajuste do MLCR é considerar modelos de regressão logística binária separados para os $r-1$ logitos. Contudo, sobre essa abordagem, os erros-padrão dos estivadores tendem a ser maiores do que quando $r-1$ logitos são considerados simultaneamente (Agresti, 2002).

No que diz respeito à seleção de covariáveis e a verificação da qualidade do ajuste do MLCR, são frequentemente utilizados procedimentos similares ao discutidos para os modelos de regressão de dicotômica.

4 Aplicações para dados das Etecs

O Centro Estadual de Educação Tecnológica Paula Souza (CEETEPS) é uma autarquia do governo do estado de São Paulo, vinculada à Secretaria de Desenvolvimento Econômico, Ciência, Tecnologia e Inovação do Estado de São Paulo, que administra as 220 Escolas Técnicas (ETECs) e as 66 Faculdades de Tecnologia (FATECs) do estado. Foi criado pelo governador Abreu Sodré em 1969. O CEETEPS possui mais de 290 mil estudantes matriculados em cursos técnicos e superiores.

O órgão nasceu com intuito de gerar os primeiros cursos superiores de tecnologia, porém, ao longo do tempo, o Centro realizou projetos de Educação Tecnológica para o ensino médio também. A intenção é expandir o ensino profissional da área de tecnologia para diferentes regiões do Estado de São Paulo.

História: A instituição foi idealizada em 1963 e começou suas atividades em 1969, na gestão do governador Roberto Costa de Abreu Sodré (1967-1971). Em 1970, adotou o nome de Centro Estadual de Educação Tecnológica de São Paulo (CEET). Os primeiros cursos superiores instalados foram - Construção Civil (Construção de Obras Hidráulicas, Construção de Edifícios e Movimento de Terra e Pavimentação), e Construção Mecânica (Desenhista Projetista e Oficinas). O centro só foi firmado como órgão mantenedor, depois que a Faculdade de Tecnologia de São Paulo e a Faculdade de Tecnologia de Sorocaba começaram a administrar os dois cursos. Entre 1981 e 1982, o órgão foi incorporado com mais doze unidades de ensino técnico, conhecidas como Escolas Técnicas Estaduais, informalmente chamadas de Etecs. Já em 1994, mais 82 unidades foram ligadas diretamente à Secretaria da Tecnologia, Desenvolvimento Econômico e da Ciência.

O Centro Paula Souza conta com a participação de 220 Etecs, 66 Faculdades de Tecnologia estaduais, conhecidas como FATECS em 300 municípios de São Paulo.

Os cursos oferecidos pelas Escolas Técnicas Estaduais são: 119 para setores industriais, agropecuária e de serviços, sendo 3 disponíveis na modalidade semipresencial; 20 cursos técnicos, os quais podem ser feitos junto ao Ensino Médio; 2 cursos técnicos vinculados ao Ensino Médio na modalidade de Educação de Jovens e Adultos.

Desde 1996 o Centro Paula Souza preocupa-se em saber se os técnicos e tecnólogos que forma estão trabalhando, se estão com dificuldades no desempenho profissional e se obtiveram melhorias pessoais e profissionais. As respostas a essas indagações permitem perceber se o ensino oferecido contribuiu para integrar o egresso como cidadão e profissional aos setores em que atua e às necessidades do mercado. Auxiliam também a aprimorar o perfil do tecnólogo para estar sempre em sintonia com as exigências e mudanças do mercado de trabalho [8].

Com a dificuldade em acessar os dados obtidos nas pesquisas oficiais realizadas pelo Centro Paula Souza,

Para esse trabalho de dissertação de mestrado foi elaborado um questionário que foi aplicado aos alunos que estão finalizando os Cursos Técnicos em 2018 e meio de 2019 em cinco Escolas Técnicas Estaduais: ETEC Fernando Prestes, ETEC Rubens de Faria e Souza, ETEC Prof. Elias Miguel Júnior, ETEC Armando Pannunzio e ETEC Piedade. A pesquisa teve como objetivo obter dados locais e analisar se os alunos pretendem trabalhar ou continuar estudando na mesma área do curso que estão concluindo, se os alunos estão satisfeitos com os cursos que estão fazendo, se pretendem voltar para Etec e fazer outro curso complementar, entre outros questionamentos. O questionário completo pode ser visto no Apêndice D.

Devido à natureza dos dados obtidos, as técnicas de análise de dados categorizados são adequadas e devem ser aplicadas para modelar e fazer inferências sobre os aspectos de interesse. Esta análise pode levar a resultados que serão de grande utilidade para essas ETECs, pois em reuniões pedagógicas há sempre dúvidas sobre essas questões e faltam informações locais para embasar essas discussões.

4.1 Aplicação da Estratégia I: Modelo Produto de distribuições de Poisson

Já sabemos, que em alguns experimentos é conveniente pré-fixar a duração da realização do experimento, (ver cap. 2). Com isso fixamos o tempo de duração de um ano e meio (2018 e meio de 2019) no estudo feito com os alunos de duas das Etecs pesquisadas (ver Introdução).

A tabela 4.1 registra os resultados obtidos quando relacionamos duas escolas da pesquisa (ETEC Fernando Prestes e ETEC Rubens de Faria e Souza) com a variável *O Curso que fez, atendeu suas expectativas?* (sim, parcialmente ou não). Uma das questões de interesse neste problema é saber se a opinião dos alunos com relação ao curso que está fazendo (expectativas atendida completamente, parcialmente ou não), está associada com a escola que está matriculada.

Tabela 4.1: Escolas com O curso que fez, atendeu suas expectativas?

	O curso que fez, atendeu suas expectativas?			
Escolas	Sim	Parcialmente	Não	Total
ETEC Fernando Prestes	57	73	7	137
ETEC Rubens de Faria e Souza	18	24	6	48
Total	75	97	13	185

Observa-se que nesse conjunto de dados, temos duas variáveis definidoras de estratos e subpopulaçõe: X_1 : Escolas, que é uma variável dicotômica nominal ($i = 1, 2$) e X_2 : O curso que fez, atendeu suas expectativas?, que é uma variável politômica nominal ($j = 1, 2, 3$).

Como para este exemplo tem-se um estudo envolvendo os alunos de duas ETECs nos anos de 2018 e meio de 2019, este é um estudo com tempo de duração fixado em $T = 1,5$ anos (ou 18 meses) e temos totais marginais e amostral aleatórios. Com isso, segundo Giolo (2017), um modelo possível para esse estudo é o produto de Poisson, em

que $N_{ij} \sim Poisson(\mu_{ij} = T\lambda_{ij})$. Para testar a ausência de associação entre X_1 e X_2 significa testar, se para as categorias $j = 1, 2, 3$ de X_2 (em termos das médias μ_{ij}) as proporções de respostas dos indivíduos nas categorias $i = 1, 2$ de X_1 não diferem, ou seja,

$$H_0 : \begin{array}{l} \frac{\mu_{11}}{\mu_{+1}} = \frac{\mu_{12}}{\mu_{+2}} = \frac{\mu_{13}}{\mu_{+3}} \left(= \frac{\mu_{1+}}{\mu}\right) \\ \frac{\mu_{21}}{\mu_{+1}} = \frac{\mu_{22}}{\mu_{+2}} = \frac{\mu_{23}}{\mu_{+3}} \left(= \frac{\mu_{2+}}{\mu}\right) \end{array}$$

H_1 : Há diferença(s) entre as proporções

Sendo $\mu_{i+} = \sum_{j=1}^{3} \mu_{ij}$; $\mu_{+j} = \sum_{i=1}^{2} \mu_{ij}$ e $\mu_{++} = \sum_{i=1}^{2}\sum_{j=1}^{3} \mu_{ij}$.

As frequências esperadas, considerando os dados da Tabela 4.1 são:

$e_{11} = \frac{(n_{1+})(n_{+1})}{n} = \frac{137 \cdot 75}{185} = 55,5405$
$e_{12} = \frac{(n_{1+})(n_{+2})}{n} = \frac{137 \cdot 97}{185} = 71,8324$

e assim por diante, calculando para todas as células temos os resultados e os cálculos da estatística do teste, definida por (2.11) podem ser vistos na Tabela 4.2.

Tabela 4.2: Cálculo de Q_p

n_{ij}	e_{ij}	$\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$
57	55,5405	0,0384
73	71,8324	0,0190
7	9,6270	0,7169
18	19,4595	0,1095
24	25,1676	0,0542
6	3,3730	2,0460
		$Q_p = 2,9839$

Obtém-se, então o valor da estatística do teste $Q_p = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(n_{ij}-e_{ij})^2}{e_{ij}} = 2,9839$

Temos que a região crítica para esse teste, ao nível de 5%, é dada por $RC = \{Q_p : Q_p > 3,841\}$, então a hipótese H_0 não é rejeitada e podemos afirmar que não há associação entre a escola que o aluno estuda e sua opinião sobre se o curso que fez atendeu suas expectativas.

4.2 Aplicação da Estratégia II: Modelo Multinomial

De acordo com a estratégia II, na seção 2.1.2, fixa-se antecipadamente o número total de elementos amostrais (n) e selecioná-se de modo aleatório de uma população de interesse. São registradas as frequências n_{ij} de elementos que apresentam simultaneamente as categorias (i, j) associadas, respectivamente, ao par de variáveis (X, Y).

Para este exemplo temos um estudo feito com 200 alunos ($n = 200$), envolvendo cinco etecs: ETEC Fernando Prestes, ETEC Rubens de Faria e Souza, ETEC Prof. Elias Miguel Júnior, ETEC Armando Pannunzio e ETEC Piedade.

Foi realizado o cruzamento das variáveis "Idade"que é politômica ($s = 3$) ordinal com variável "Pretende trabalhar na área do seu curso?"que é politômica ($r = 3$) nominal.

Como nesse estudo apenas o total amostral n é fixo, segue que o modelo associado ao estudo descrito é o multinomial. Desse modo, as hipóteses de interesse são:

$$\begin{cases} H_0 : \text{A decisão por trabalhar na área do curso independe da idade do aluno} \\ H_1 : \text{A decisão por trabalhar na área do curso depende da idade do aluno} \end{cases}$$

Tabela 4.3: Pretende trabalhar na área do curso? com Idade

	Pretende trabalhar na área do seu curso?			
Idade	Sim	Não decidi	Não	Total
$16-17$	52	50	50	152
$18-19$	10	7	2	19
≥ 20	22	5	2	29
Total	84	62	54	200

De acordo com a seção 2.2 podemos observar que para este delineamento o total n fixo, com $s = 3$ e $r = 3$, o modelo associado é o Multinomial, com a variável X nominal e a variável Y ordinal. A hipótese nula (H_0) é de independência, então uma das estatísticas de teste que podemos utilizar é $Q_P \sim \chi^2_{(s-1)(r-1)}$, ou seja, $Q_P \sim \chi_{(4)}$.

Obtendo-se então o valor da estatística do teste $Q_p = \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = 20,7156$. Temos que a região crítica para este teste, ao nivel de 5% é dada por $RC = \{Qp : Qp > 9,5\}$, então $Qp = 20,7156 \in RC$ e rejeita-se H_0. Conclui-se que a decisão sobre trabalhar na área do curso depende da idade do aluno. Quanto maior a idade, maior a proporção de respostas "Sim".

4.3 Aplicação da Estratégia III: Modelo Produto de distribuições Multinomiais

De acordo com a estratégia III (ver seção 2.1.3), o experimento pode também ser planejado fixando antecipadamente o número N_i de elementos de cada categoria.

Com $n_{1+} = n_{2+} = 70$, foi feito o cruzamento de "*Período*", que é uma variável dicotômica nominal ($s = 2$) com "*Esse curso atendeu suas expectativas?*", que é uma variável politômica ordinal ($r = 3$).

Como nesse estudo foi fixado antecipadamente o número N_i de elementos de cada categoria, segue que o modelo associado ao estudo descrito é o Modelo Produto de distribuições Multinomiais. Desse modo, as hipóteses de interesse são:

$$\begin{cases} H_0 : \text{A opinião sobre o curso não está associada} \\ \qquad \text{com o período em que o aluno estuda} \\ H_1 : \text{A opinião sobre o curso está associada} \\ \qquad \text{com o período em que o aluno estuda.} \end{cases}$$

Tabela 4.4: Esse curso atendeu suas expectativas? com Período

	Esse curso atendeu suas expectativas?			
Período	Não	Parcialmente	Sim	Total
Integral	4	43	23	70
Outros	5	21	44	70
Total	9	64	67	140

De acordo com a seção 2.2 podemos observar que para este delineamento de marginais -linha fixos, com $s = 2$ e $r > 2$, o modelo associado é o produto de multinomiais, com a variável X nominal e a variável Y ordinal, na hipótese nula (H_0) os escores médios não diferem, então utilizamos a estatistica de teste $Q_s \sim \chi^2_{(1)}$, para que, a natureza ordinal da variável resposta ser levada em consideração.

Portanto, usando a alternativa sugerida por Giolo (2017), vamos atribuir escores $a = (a_1, a_2, a_3) = (0, 1, 2)$ para as categorias "*não*", "*parcialmente*"e "*sim*" da variável "*Esse curso atendeu suas expectativas?*", e então, definir um escore médio, de acordo com a equação (2.12) para cada subpopulação (linha) da tabela de contigência. Com os dados calcula-se as estimativas dos escores médios (equação 2.13) e temos que $\overline{f}_1 = 1,2714$ e $\overline{f}_2 = 1,5571$.

Sob a hipótese nula de não associação entre o período e a opinião sobre o curso, formulada em termos de escores médios por $H_0 : \overline{F}_1 = \overline{F}_2$, segue que a esperança e a variância de $\overline{f}_1$ são dadas, respectivamente por $E(\overline{f}_1) = 1,4143$ e $V(\overline{f}_1) = 0,0027$ (equações 2.14 e 2.15)

Logo, a estística Q_S, denominada estatística escore médio, é calculada utilizando-se a equação (2.16) e segue distribuição aproximada qui-quadrado com 1 grau de liberdade.

Temos, portanto que $Q_S = 7,6416$ (valor p=0,0057, g.l.=1). Sendo assim é possível concluir que os períodos *integral* e *outros* diferem, bem como que os alunos que estudam no o período *integral* estão menos satisfeitos em relação ao seu curso do que os que estudam em *outros* (períodos não integral), pois $\overline{f}_1 < \overline{f}_2$. Devido aos escores assumidos, vale observar que os valores de $\overline{f}_1$ e $\overline{f}_2$ estão restritos ao intervalo $[0, 2]$, bem como que valores mais próximos de zero indicam um contigente maior de indivíduos que dizem que o curso não atendeu suas expectativas e, valores mais próximos de dois, um contigente maior de indivíduos que dizem que o curso atendeu suas expectativas.

4.4 Aplicação do modelo de regressão logística multinomial

Uma aplicação dos modelos logitos categoria de referência (MLCR) foi feita com os dados obtidos no questionário aplicado aos alunos das ETECs e os cálculos foram feitos no software R [7](pacote VGAM). O objetivo foi avaliar a associação entre uma variável resposta e três variáveis explicativas produzidas nas questões aplicadas aos alunos:

Questão 9: Esse curso atendeu às suas expectativas? Variável resposta politômica nominal Y: Sim, parcialmente ou não.

Questão 5: Qual é o seu curso técnico? As respostas a esta questão foram agrupadas pelos eixos tecnológicos, considerando somente os que constam na amostra em número suficiente para a análise : Variável politômica nominal $(E_i, i = 2, ..., 5)$.

A seguir apresenta-se a descrição dos principais eixos tecnológicos de acordo com o site ETEC: [4]. A quantidade de alunos na amostra que estão matriculados nos cursos dos eixos E1, E6 e E7 é insuficiente para a análise e os dados relativos a esses alunos foram descartados.

E1 - PRODUÇÃO ALIMENTÍCIA - Alimentos

O TÉCNICO EM ALIMENTOS é o profissional que atua no processamento e conservação de matérias-primas, produtos e subprodutos da indústria alimentícia e de bebidas, realizando análises físico-químicas, microbiológicas e sensoriais. Auxilia no planejamento, na coordenação e controle de atividades do setor. Promove a sanitização das indústrias alimentícias e de bebidas. Controla e corrige desvios nos processos manuais e automatizados. Acompanha a compra e a manutenção de equipamentos. Participa do desenvolvimento de novos produtos e processos. Auxilia na implantação de sistema de garantia de qualidade e segurança em organizações da área de alimentos. Realiza trabalho em equipe, assumindo papéis de liderança e tomada de decisões. Busca atualização e ampliação dos seus conhecimentos em linguagens, capacidade de comunicação oral e escrita. Articula com iniciativa e capacidade de adaptação a novos ambientes e situações. Exerce atitude profissional, postura ética, com visão na sustentabilidade e responsabilidade social.

E2 - GESTÃO E NEGÓCIOS - Logística

O TÉCNICO EM LOGÍSTICA é o profissional que executa e colabora na gestão dos processos de planejamentos, operações e controles de programação da produção de bens e serviços, programação de manutenção de máquinas e de equipamentos, de compras, de recebimento, de armazenamento, de estoques, de movimentação, de expedição, transporte e distribuição de materiais e produtos, utilizando tecnologia de informação. Presta atendimento aos clientes. Implementa os procedimentos de controle de custos, qualidade, segurança e higiene do trabalho no sistema logístico.

E3 - INFRAESTRUTURA - Edificações

O TÉCNICO EM EDIFICAÇÕES é o profissional que desenvolve e executa projetos de edificações conforme normas técnicas de segurança, de acordo com legislação específica, conforme limites regulamentares e normativos ambientais. Planeja a execução, elabora orçamento e memorial descritivo de obras. Supervisiona a execução de diferentes etapas do processo construtivo. Presta assistência técnica no estudo e desenvolvimento de projetos, pesquisas e controle tecnológico de materiais na área da Construção Civil. Orienta e coordena a execução de serviços de manutenção de equipamentos e de instalações em edificações. Orienta na assistência técnica para compra, venda e utilização de produtos e equipamentos especializados.

E4 - CONTROLE E PROCESSOS INDUSTRIAIS - Mecânica

O TÉCNICO EM MECÂNICA é o profissional que elabora projetos mecânicos e sistemas automatizados. Planeja, aplica e controla procedimentos de instalação e de manutenção mecânica de máquinas e equipamentos. Desenvolve e controla processos de fabricação e montagem de conjuntos mecânicos. Aplica técnicas de medição e ensaios. Especifica materiais para construção mecânica. Elabora documentação, realiza

compras e vendas técnicas e cumpre normas e procedimentos de segurança no trabalho e de preservação ambiental.

E5 - CONTROLE E PROCESSOS INDUSTRIAIS - Eletrotécnica

O TÉCNICO EM ELETROTÉCNICA é o profissional que instala, opera e mantém elementos de geração, transmissão e distribuição de energia elétrica. Participa na elaboração e no desenvolvimento de projetos de instalações elétricas e de infraestrutura para sistemas de telecomunicações em edificações. Atua no planejamento e execução da instalação e manutenção de equipamentos e instalações elétricas. Aplica medidas para o uso eficiente da energia elétrica e de fontes energéticas alternativas. Participa no projeto e instala sistemas de acionamentos elétricos. Executa a instalação e manutenção de iluminação e sinalização de segurança.

E6 - CONTROLE E PROCESSOS INDUSTRIAIS - Mecatrônica

O TÉCNICO EM MECATRÔNICA é o profissional que atua no projeto, na execução e na instalação de máquinas e equipamentos automatizados e sistemas robotizados. Realiza manutenção, medições e testes dessas máquinas, equipamentos e sistemas, conforme especificações técnicas. Opera equipamentos, utiliza softwares específicos e linguagens de programação adequadas. Organiza local de trabalho. Coordena, equipes e oferece treinamento operacional. Realiza manutenções preditiva, preventiva e corretiva, em conformidade com as normas técnicas e higiene, segurança, qualidade e proteção ao meio ambiente. Programa e opera estas máquinas observando as normas de segurança.

E7 - INFORMAÇÃO E COMUNICAÇÃO - Informática para Internet

O TÉCNICO EM INFORMÁTICA PARA INTERNET é o profissional que desenvolve e realiza manutenções em websites, portais na Internet e Intranet. Utiliza ferramentas de desenvolvimento de projetos para construir soluções que auxiliam o processo de criação de interfaces e aplicativos empregados no comércio e marketing eletrônicos.

Questão 6: Em que período você estuda? Variável dicotômica nominal: Integral ou outro (manhã, tarde e noite)

Questão 8: Você já trabalha na área do seu curso? Variável dicotômica nominal: Sim ou não.

A Tabela 4.5 apresenta os dados em forma de tabela de contingência. Foram considerados somente os eixos tecnológicos $E_i, i = 2, ..., 5$ pois para os eixos E_1, E_6 e E_7 não há dados suficientes para o cruzamento com as outras variáveis explicativas.

Tabela 4.5: Estudo sobre os cursos das Etecs

Eixo	Período	Trabalha na área ?	Curso atendeu suas expectativas?		
			Sim	Parcialmente	Não
E_2	I	Não	4	2	0
E_2	I	Sim	0	0	1
E_2	O	Não	9	4	0
E_2	O	Sim	2	1	0
E_3	I	Não	3	14	0
E_3	I	Sim	0	0	1
E_3	O	Não	1	0	0
E_3	O	Sim	1	0	0
E_4	I	Não	1	6	1
E_4	I	Sim	1	0	0
E_4	O	Não	1	0	0
E_4	O	Sim	0	0	1
E_5	I	Não	4	23	6
E_5	I	Sim	1	2	0
E_5	O	Não	0	2	1
E_5	O	Sim	0	1	0

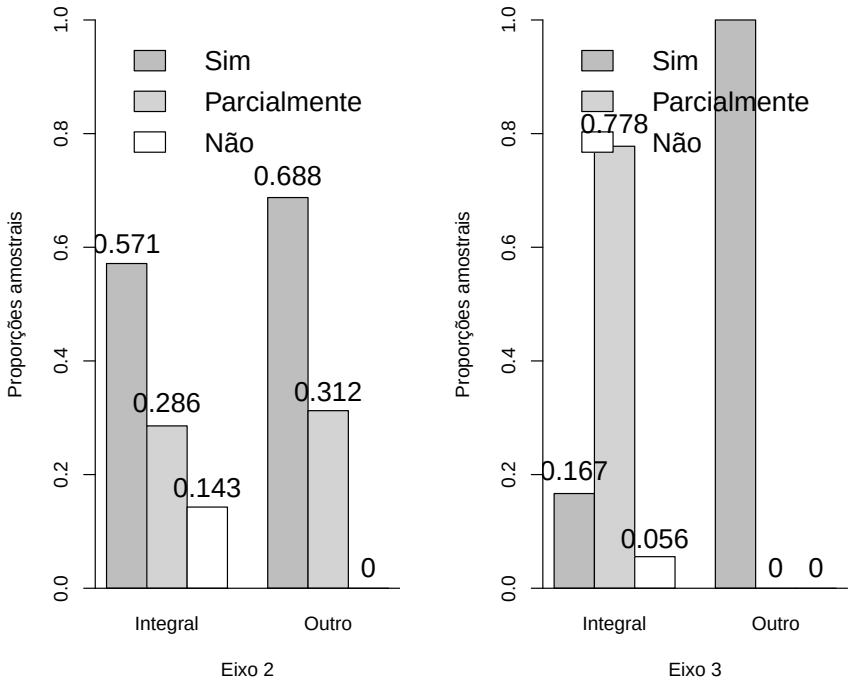

Figura 4.1: Gráficos de colunas dos dados Eixos 2 e 3

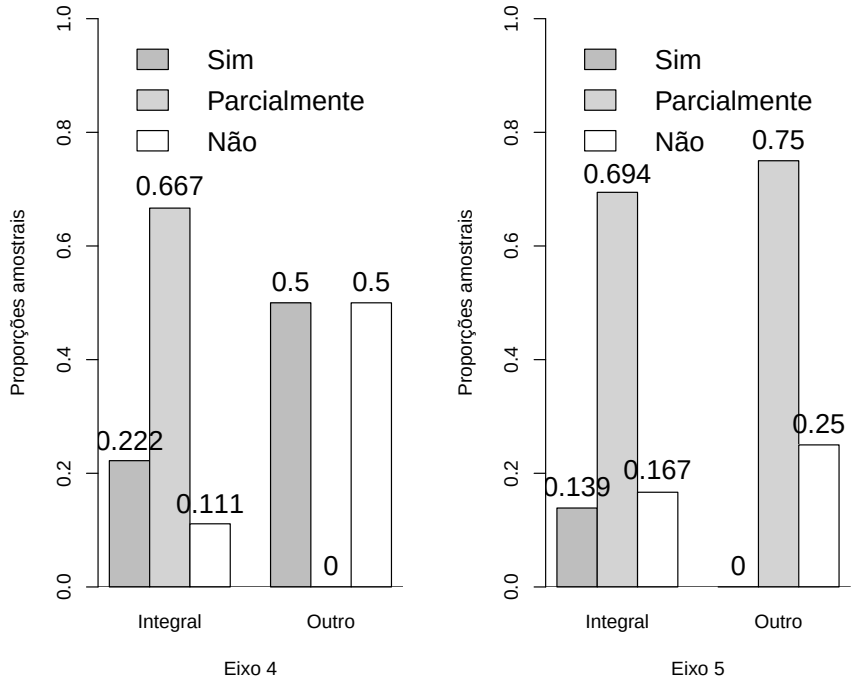

Figura 4.2: Gráficos de colunas dos dados Eixos 4 e 5

As variáveis explicativas foram consideradas nos modelos MLCR por meio de variáveis fictícias (do inglês *dummy*), como segue:

Eixos tecnológicos: $X_{11} = \begin{cases} 0 & \text{eixos } E_2, E_4, E_5 \\ 1 & \text{eixo } E_3 \end{cases}$

$X_{12} = \begin{cases} 0 & \text{eixos } E_2, E_3, E_5 \\ 1 & \text{eixo } E_4 \end{cases}$

$X_{13} = \begin{cases} 0 & \text{eixos } E_2, E_3, E_4 \\ 1 & \text{eixo } E_5 \end{cases}$

Período: $X_2 = \begin{cases} 0 & \text{Período:Integral} \\ 1 & \text{Período:Outro} \end{cases}$

Aluno trabalha na área do curso: $X_3 = \begin{cases} 0 & \text{Não} \\ 1 & \text{Sim} \end{cases}$

Foram ajustados 4 modelos MLCR considerando como categoria de referência a resposta $Y = 3$ (Não) e sendo $\mathbf{x} = (X_{11}, X_{12}, X_{13}, X_2, X_3)$ descritos a seguir:

MLCR0 - Modelo sem covariáveis = modelo nulo (Logito j, com $j = 1, 2.$)

$$\ln\left[\frac{p_j(\mathbf{x})}{p_3(\mathbf{x})}\right] = \ln\left[\frac{P(Y=j|\mathbf{x})}{P(Y=3|\mathbf{x})}\right] = \beta_0^j$$

MLCR1 - Modelo com a covariável Eixo tecnológico (Logito j, com $j = 1, 2$.)

$$\ln\left[\frac{p_j(\mathbf{x})}{p_3(\mathbf{x})}\right] = \ln\left[\frac{P(Y=j|\mathbf{x})}{P(Y=3|\mathbf{x})}\right] = \beta_0^j + \beta_{11}^j X_{11} + \beta_{12}^j X_{12} + \beta_{13}^j X_{13}$$

MLCR2 - Modelo com as covariáveis Eixo tecnológico e Período (Logito j, com $j = 1, 2$.)

$$\ln\left[\frac{p_j(\mathbf{x})}{p_3(\mathbf{x})}\right] = \ln\left[\frac{P(Y=j|\mathbf{x})}{P(Y=3|\mathbf{x})}\right] = \beta_0^j + \beta_{11}^j X_{11} + \beta_{12}^j X_{12} + \beta_{13}^j X_{13} + \beta_2^j X_2$$

MLCR3 - Modelo com as covariáveis Eixo tecnológico, Período e se o aluno trabalha na área do curso (Logito j, com $j = 1, 2$.)

$$\ln\left[\frac{p_j(\mathbf{x})}{p_3(\mathbf{x})}\right] = \ln\left[\frac{P(Y=j|\mathbf{x})}{P(Y=3|\mathbf{x})}\right] = \beta_0^j + \beta_{11}^j X_{11} + \beta_{12}^j X_{12} + \beta_{13}^j X_{13} + \beta_2^j X_2 + \beta_3^j X_3$$

Tabela 4.6: Resultados do ajuste dos modelos

Modelo	$G.L.$	$Dev.$	$ln\ L(.)$	TRV	$\neq G.L.$	Valor - p	AIC
MLCR0	30	$57,27$	$-40,17$	$-$	$-$	$-$	$84,35$
MLCR1	24	$35,94$	$-29,51$	$21,33$	4	$0,00027$	$75,02$
MLCR2	22	$33,98$	$-28,53$	$1,97$	2	$0,3739$	$77,05$
MLCR3	20	$31,04$	$-27,06$	$2,94$	2	$0,2299$	$78,11$

Na Tabela 4.6, $G.L.$ indica os graus de liberdade e $TRV-$ indica o valor da estatística do teste da razão de verossimilhança calculada de acordo com a equação (3.11), o "Valor - p" é calculado com a função *chi2cdf(TRV,G.L.)* do Matlab e o valor do critério de Akaike com a equação (3.15). Segundo os resultados apresentados na Tabela 4.6 temos que o modelo MLCR1 é o mais adequado uma vez que o critério de Akaike (AIC) resultou no menor valor para este modelo. Além disso o teste da razão de verossimilhança indica a rejeijão dos modelos com mais covariáveis MLCR2 e MLCR3. A seguir, na Tabela 4.7 estão as estimativas e os erros padrão dos parâmetros do modelo MLCR1.

Tabela 4.7: Estimativas dos parâmetros do modelo MLCR1

	Logito $j=1$		Logito $j=2$	
Parâmetro	Estimativa	Erro padrão	Estimativa	Erro padrão
β_0^j	2,7080	1,0328	1,9459	1,0690
β_{11}^j	-1,0986	1,5055	0,6931	1,4880
β_{12}^j	-2,3026	1,3784	-0,8473	1,3452
β_{13}^j	-3,0445	1,1872	-0,5596	1,1495

As estimativas das chances de ocorrência das respostas "sim"e "parcialmente"em relação à categoria de referência "não"para a questão "O curso atendeu suas expectativas?"por eixo tecnológico estão apresentadas na Tabela 4.8.

Tabela 4.8: Estimativas das chances por eixo tecnológico

Eixo tecnológico	$\frac{\widehat{p_1}}{\widehat{p_3}}$	$\frac{\widehat{p_2}}{\widehat{p_3}}$
E2: Logística	15,00	7,00
E3: Edificações	5,00	14,00
E4: Mecânica	1,500	3,00
E5: Eletrotécnica	0,7143	4,00

$\frac{\widehat{p_1}}{\widehat{p_3}}$: Estimativa da chance de um aluno responder "sim"em comparação com a resposta "não".

Vemos na Tabela 4.8 que nos eixos E2, E3 e E4 a chance de responder "sim"é maior que responder "não"sendo 15 vezes maior no eixo E2, 5 vezes maior no eixo E3 e 1,5

vezes maior no eixo E4. No eixo E5 ocorre o inverso, a chance de responder "não"é 1,4 vezes maior que responder "sim".

$\frac{\widehat{p_2}}{\widehat{p_3}}$:Estimativas da chance de um aluno responder "parcialmente"em comparação com a resposta "não".

Vemos na Tabela 4.8 que nos 4 eixos a chance de responder "parcialmente"é maior que responder "não"sendo 7 vezes maior no eixo E2, 14 vezes maior no eixo E3, 3 vezes maior no eixo E4 e 4 vezes maior no eixo E5.

5 Conclusão

A importância da análise de dados categóricos dentro da Estatística está relacionada com o fato de haver grande aplicabilidade em variadas áreas. Na literatura as aplicações mais comuns estão voltadas para a área de medicina em estudos clínicos, mas é possível utilizar essa mesma metodologia em diversas outras áreas como nas ciências da educação, entre outras. Os dados, em geral, são dispostos em tabelas de contingência para a análise, mas não são todas as tabelas de contingência que podem ser geradas pelas estratégias de amostragem e modelos apresentados nesta dissertação. O uso de delineamentos amostrais complexos, envolvendo estágios múltiplos de estratificação e/ou de agrupamento, em variados casos, especialmente em pesquisas de larga escala, coloca a necessidade de escolher modelos probabilísticos mais sofisticados. Os testes qui-quadrado discutidos apresentam limitações. Por exemplo, eles necessitam de amostras grandes para que se tenha uma aproximação apropriada para distribuição qui-quadrado. Além disso, eles indicam somente o grau de evidência para uma associação, não descrevendo, contudo, a força (ou a intensidade) dessa associação. Em algumas situações, o tamanho amostral não é suficientemente grande, podendo ocorrer diversos valores esperados menores do que 5 associados às células da tabela de contingência $s \times r$. Nesses casos, as estatísticas discutidas anteriormente não são recomendáveis, uma vez que aproximação para a distribuição qui-quadrado não é razoável. Uma alternativa é fazer o uso do teste exato de Fisher, sendo que, nesses casos, as probabilidades de interesse são calculadas a partir da distribuição hipergeométrica multivariada.

Os dados coletados para este trabalho de dissertação de mestrado são de uma amostra de uma população real (alunos das ETECs que estão próximos da formatura). Os objetivos da análise desses dados categóricos residem na realização de inferências relativas a questões que se prendem com relações estruturais que possam existir entre as variáveis estudadas. Estes objetivos inferenciais pressupõem geralmente a adoção de um modelo probabilístico consistente com o processo de amostragem e os propósitos analíticos. As conclusões extraídas são então condicionadas à validade de tais suposições sobre esses modelos.

Os modelos probabilísticos utilizados dependem do delineamento amostral (estratégia) e dos objetivos da análise. Na dissertação foram apresentadas três estratégias: Modelo Produto de distribuições de Poisson, Modelo Multinomial e Modelo Produto de distribuições Multinomiais. Identificar o delineamento amostral que produziu os dados foi importante para que fosse determinada a análise apropriada e feitas as inferências de interesse. Foram descritos nesse trabalho alguns delineamentos amostrais e os modelos probabilísticos assumidos para eles.

Primeiramente foi preestabelecido o tempo de duração da pesquisa de um ano e meio (2018 e metade de 2019) de duas ETECs de interesse, com o objetivo de saber

se existia uma associação entre a escola que o aluno estuda e a sua opinião sobre se o curso atendia suas expectativas. O delineamento amostral utilizado foi o que considera os totais aleatórios, logo um modelo que foi possível para esse estudo foi o Produto de Poisson. Conclui-se que não houve associação entre as variáveis de interesse.

Logo após, foi fixado antecipadamente o número total de elementos amostrais, para isso consideramos 200 alunos das cinco ETECs consideradas para coletar os dados, com o objetivo de saber se a decisão de trabalhar na área do curso depende ou não da idade do aluno. Usamos o delineamento amostral que considera o total n fixo, sendo que o modelo associado foi o Multinomial. Concluímos que existe associação entre as variáveis.

Também realizamos um estudo onde fixamos o número de elementos das marginais-linha. O objetivo era saber se a opinião sobre se o curso atendeu suas expectativas estava associada ou não com o período que o aluno estuda. Para isso usamos o delineamento que considera as marginais linha fixas e o modelo associado foi o produto de binomiais. Com isso concluímos que existe associação, ou seja, os alunos que estudam no período integral estão menos satisfeitos em relação ao seu curso do que os que estudam em outros (períodos não integrais).

Com o objetivo de modelar a associação entre um conjunto de variáveis explicativas e uma variável resposta politômica, utilizamos um dos modelos logísticos de regressão multinomial, no caso, o modelo logitos categoria de referência (MLCR). Como variável resposta, foi utilizada a Questão 9 (Esse curso atendeu às suas expectativas?), que é uma variável politômica nominal (sim, parcialmente ou não). Para as variáveis explicativas foram utilizadas a Questão 5 (Qual é o seu curso técnico?), a Questão 6 (Em que período você estuda?) e a Questão 8 (Você já trabalha na área do seu curso?). Concluiu-se que apenas o eixo tecnológico correspondente ao curso técnico em que o aluno está matriculado interfere na opinião sobre o curso.

A verificação da correlação, dependência ou associações entre variáveis é de grande importância pelo fato de nos dar um melhor entendimento sobre o tema. Muitas vezes é preciso avaliar o grau de associação entre duas ou mais variáveis para descobrir o quanto uma variável interfere no resultado de outra, assim podemos encontrar uma possível solução de problemas.

Para o futuro, pretende-se dar continuidade ao trabalho com análise de dados categóricos, continuando a coleta de dados com alunos e, possivelmente professores e funcionários, afim de obter a quantidade suficiente para novos cruzamentos de variáveis e estudar mais associações que sejam úteis para as ETECs.

Referências

[1] M. N. Magalh aes. *Noções de Probabilidade e Estatística*. Ed. da Universidade de São Paulo, SP, 7 edition, 2015.

[2] A. Agresti. *Categorical data analysis*. Johon Wiley e Sons, New Work, 2 edition, 2002.

[3] N. Rosamilha e A. R. Faria. Evolução da noção de conservação de quantidade e desempenho em matemática. *Psicol. cienc. prof*, 3:25–44, 1983.

[4] Centro Paula Souza (Cursos e Eixos Tecnológicos). Disponível em: < https://www.vestibulinhoetec.com.br/unidades-cursos/>. Acesso em: 20 jul. 2019.

[5] C. D. Paulino e J. M.Singer. *Análise de dados categorizados*. Ed. Blucher, SP, 1 edition, 2006.

[6] S. R. Giolo. *Introdução à análise de dados categóricos com aplicações*. Revista Brasileira de Psiquiatria, SP, 1 edition, 2018.

[7] A language and environment for statistical computing R. Disponível em: < http://www.R project.org/>.

[8] Sobre o Centro Paula Souza. Disponível em: < https://www.cps.sp.gov.br/sobre-o-centro-paula-souza/>. Acesso em: 24 jul 2019.

A Notação utilizada na dissertação

Para a notação matricial, os vetores e matrizes são representadas por símbolos grafados em negrito. Assim:

- $\mathbf{a} = (a_1, ..., a_n)'$ denota um vetor (coluna) de dimensão $n \times 1$ com elementos $a_1, ..., a_n$; o símbolo "$'$"no expoente denota a operação de transposição;
- $\mathbf{A}$ denota uma matriz de dimensão $s \times r$ com elementos (ordenados lexicograficamente) $a_{11}, .., a_{1r}, a_{21}, ..., a_{2r}, ..., a_{s1}, ..., a_{sr}$, isto é,

$$\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1r} \\ a_{21} & \cdots & a_{2r} \\ \vdots & \ddots & \vdots \\ a_{s1} & \cdots & a_{sr} \end{pmatrix}$$

- $\mathbf{1}_p$ denota um vetor de dimensão $p \times 1$ com todos elementos iguais a 1;
- $\mathbf{I}_n$ denota a matriz identidade de dimensão n e para evitar confusão com a matriz de informação de Fisher correspondente a uma amostra de tamanho n de uma distribuição indexada por um vetor de parâmetros β, esta será denotada $\mathbf{I}_n(\beta)$;
- $\mathbf{D}_a$ ou $diag(a_1, ..., a_n)$ denotam uma matriz diagonal com os elementos do vetor $\mathbf{a}$ dispostos ao longo da diagonal principal;
- $\exp(\mathbf{a})$ e $\ln(\mathbf{a})$ denotam, respetivamente vetores cujos elementos são $\exp(a_i)$ e $\ln(a_i)$, $i = 1, ..., n$.

B Modelos probabílisticos discretos

Funções de probabilidade $P(X = x)$, esperança $E(X)$ e variância $V(X)$ dos principais modelos probabílisticos discretos:

1. Modelo Bernoulli(p)

$$P(X = x) = p^x(1-p)^{1-x}, \tag{B.1}$$

$p \in [0,1], x = 0, 1$ com $E(X) = p$ e $V(X) = p(1-p)$

2. Modelo Binomial(n, p)

$$P(X = x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}, \tag{B.2}$$

$n > 0$, com $n \in \mathbb{Z}^*, p \in [0,1], x = 0, 1, ..., n$, com $E(X) = np$ e $V(X) = np(1-p)$

3. Modelo Multinomial $(n, p_1, ..., p_r)$

$$P(X = x) = \frac{n!}{x_1!...x_r!}{p_1}^{x_1}...{p_r}^{x_r}, \tag{B.3}$$

$n > 0$,com $n \in \mathbb{Z}^*, p_j \in [0,1], x_j \in \{0, 1, ..., n\}$, $\sum_{j=1}^{r} p_j = 1, \sum_{j=1}^{r} x_j = n$ com

$$E(X_j) = np_j \quad \text{e} \quad V(X_j) = np_j(1-p_j)$$

4. Modelo Hipergeométrico (n, m, r)

$$P(X = x) = \frac{\binom{m}{x}\binom{n-m}{r-x}}{\binom{n}{r}}, \tag{B.4}$$

$x = 0, 1, ..., \min(r, m), n > 0$, com $n \in \mathbb{Z}^*, r < n$, com

$$E(X_j) = \frac{rm}{n} \quad \text{e} \quad V(X_j) = \frac{rm(n-m)(n-r)}{n^2(n-1)}$$

5. Modelo Poisson(μ)

$$P(X = x) = \frac{e^{-\mu}\mu^x}{x!}, \tag{B.5}$$

$\mu > 0, x = 0, 1, 2, ...,$ com $\quad E(X) = V(X) = \mu$

C Testes Qui-quadrado

Teste de Independência e *Teste de Homogeneidade*, de acordo [1].

Apresentamos uma forma de testar a independência entre duas variáveis. Se dispomos da função de probabilidade conjunta de duas variáveis aleatórias, podemos verificar se, para todos os possíveis valores das variáveis, o produto das probabilidades marginais é igual a probabilidade conjunta.

Na situação mais comum em que não temos informação sobre a ocorrência conjunta das variáveis aleatórias, o procedimento usual é coletar uma amostra anotando a frequência conjunta da ocorrência dos valores das variáveis. Pode-se, então, utilizar um teste de hipóteses conhecido como *Teste de Independência*. Este teste será apresentado através do exemplo a seguir:

Exemplo: A tabela abaixo contém os resultados obtidos por estudantes do ensino médio, em um exame com questões das disciplinas de física e matemática. Deseja-se testar se existe associação entre as notas dessas duas disciplinas que, para efeito de apresentação na tabela e análise de comportamento, foram classificadas nas categorias alta, média e baixa.

Física/Matemática	Alta	Média	Baixa	Total
Alta	56	71	12	139
Média	47	163	38	248
Baixa	14	42	85	141
Total	117	276	135	528

Hipóteses:

$$\begin{cases} H_0\text{: As notas de física e matemática são independentes entre si;} \\ H_1\text{: Elas não são independentes.} \end{cases}$$

Construção da tabela de frequências esperados. Para casela (i, j) esse valor é:

$$e_{12} = \frac{\text{Total da linha } i \times \text{Total da coluna } j}{\text{Total geral}}$$

Note que os valores das frequências esperados são calculados sob a hipótese H_0 de independência e, por essa razão, utilizamos os totais de linha e coluna que representam as frequências marginais das variáveis. Por exemplo, para célula $(1, 2)$, temos:

$$e_{ij} = \frac{\text{Total da linha 1} \times \text{Total da coluna 2}}{\text{Total geral}} = \frac{139 \times 276}{528} = 72,66$$

A tabela completa de frequências esperados é

Física / Matemática	Alta	Média	Baixa
Alta	30,80	72,66	35,34
Média	54,95	129,64	63,41
Baixa	31,25	73,70	36,05

para medir a diferença entre os valores das frequências observadas e esperadas utiliza-se a estatística, dada por:

$$Q^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

r e s representando o número de linhas e de colunas, respectivamente. Para um número grande de observações, a distribuição de Q^2 se comporta como um modelo Qui-Quadrado com $(r-1)\times(s-1)$ graus de liberdade. A região crítica contém valores grandes de Q^2, isto é,

$$RC = \{w : w \geq q_c\}$$

com q_c sendo determinado pelo nível de significância do teste (α), ou seja,

$$\alpha = P(Q^2 \geq q_c | H_0 verdadeiro)$$

Para $\alpha = 0,01$ a tabela da Qui-quadrado com 4 graus de liberdade fornece $q_c = 13,28$. Obtemos assim,

$$RC = \{w : w \geq 13,28\}.$$

Cálculo do valor observado de Q^2:

$$q^2_{obs} = \frac{(56-30,80)^2}{30,80} + \frac{(71-72,66)^2}{72,66} + ... + \frac{(85-36,05)^2}{36,05} = 145,78$$

Conclui-se pela rejeição da hipótese nula, ou seja, as notas de física e matemática não são independentes (estão associados).

Na construção da tabela de frequências esperadas, caso alguma célula tenha valor menor que 5, será necessário agrupar categorias. Este procedimento visa garantir uma melhor aproximação para o uso do modelo Qui-Quadrado para Q^2.

Consideremos agora o chamado *Teste de Homogeneidade*. Esse teste consiste em verificar se uma variável aleatória se comporta de modo similar, ou homogêneo, em várias subpopulações. Apesar da mecânica de realização do teste semelhante a do Teste de Independência, uma distinção importante se refere à forma como as amostras são coletadas. No teste de homogeneidade, fixamos o tamanho da amostra em cada uma das subpopulações e, então selecionamos uma amostra de cada uma delas. Na tabela apresentada a seguir, as linhas representam as subpopulações e, as colunas, os diferentes valores ou categorias da variável.

Subpopulações	valores da variável			total de linha
1	o_{11}	o_{12}	$\cdots$	n_1
2	o_{21}	o_{22}	$\cdots$	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
total da coluna				Total Geral

Supondo homogeneidade entre as subpopulações, utilizamos para o cálculo da frequência esperada da célula (i, j) a seguinte expressão:

$$e_{ij} = n_i \times \frac{\text{Total da coluna } j}{\text{Total geral}}.$$

O total de linha n_i indica o tamanho da amostra da subpopulação i, ao passo que o quociente, total da coluna j dividido pelo total geral, representa a proporção de ocorrências do valor da variável correspondente à coluna j. Caso haja homogeneidade no comportamento da variável, esperamos que essa proporção seja a mesma, em todas as subpopulações.

Exemplo: Estamos interessados em saber se a preferência por certo tipo de filme se altera com estado civil. Selecionamos pessoas em que cada uma das subpopulações: solteiro, casado, divorciado e viúvo. Os resultados estão na tabela a seguir:

Estado Civil \ Filme	Policial	Comédia	Romance	tam. amostra
Solteiro	45	25	30	100
Casado	36	61	43	140
Divorciado	39	36	35	110
Viúvo	14	19	17	50
Total	134	141	125	400

Na tabela, a última coluna representa o tamanho da amostra selecionada em cada subpopulação. Observe que esses valores foram fixados antes de a coleta ser realizada. As hipóteses a serem testadas são:

$$\begin{cases} H_0\text{: A preferência por certo tipo de filme é igual para qualquer estado civil;} \\ H_1\text{: A preferência muda.} \end{cases}$$

A proporção dos indivíduos que preferem filmes policiais é de 134/400. Se a variável Filme for homogênea entre subpopulações de Estado Civil, devemos ter essa mesma preferência por Filmes policiais, para qualquer estado civil. Logo, o valor esperado de preferência pelo gênero Policial, na subpopulação dos solteiros, deve ser $100 \times 134/400$. Para as outras subpopulações, multiplicamos 134/400 pelos respectivos valores do tamanho da amostra, que são diferentes nesse exemplo. A tabela de frequências esperadas é apresentada a seguir:

Estado Civil \ Filme	Policial	Comédia	Romance	tam. amostra
Solteiro	$33,50$	$35,25$	$31,25$	100
Casado	$46,90$	$49,35$	$43,75$	140
Divorciado	$36,85$	$38,78$	$34,37$	110
Viúvo	$16,75$	$17,62$	$15,63$	50
Total	134	141	125	400

Quantifica-se a quantidade Q^2 da mesma forma como fizemos anteriormente, isto é, vamos quantifica-se a "distância"entre os valores observados ($o_{i,j}$) e aqueles esperados ($e_{i,j}$), se houvesse homogenidade. Assim,

$$Q^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Para um número grande de observações, a distribuição de Q^2 é Qui-Quadrado com $(r-1) \times (s-1)$ graus de liberdade (r número de linhas e s número de colunas). A região crítica contém valores grandes de Q^2, isto é,

$$RC = \{w : w \geq q_c\},$$

com q_c sendo determinado pelo nível de significância do teste, ou seja,

$$\alpha = P(Q^2 \geq q_c | H_0 \text{ verdadeiro}).$$

Para $\alpha = 0,05$ obtemos, da tabela da densidade Qui-quadrado com 6 graus de liberdade, $q_c = 12,59$. Portanto,

$$RC = \{w : w \geq 12,59\}.$$

Para o valor observado de Q^2 temos:

$$q^2_{obs} = \frac{(45 - 33,50)^2}{33,50} + \frac{(36 - 46,90)^2}{46,90} + ... + \frac{(17 - 15,63)^2}{15,63} = 13,29$$

Concluímos pela rejeição da hipótese nula, ou seja, preferência de filmes não é a mesma nas diferentes subpopulações definidas pelo estado civil.

D Questionário

Questionário para a dissertação

1. Qual seu sexo?
   (a) Feminino
   (b) Masculino
2. Qual sua idade?
3. Qual a cidade onde você mora?
4. Em que escola você estuda?
   (a) ETEC Fernando Prestes
   (b) ETEC Rubens de Faria e Souza
   (c) ETEC Prof. Elias Miguel Júnior (Votorantim)
   (d) ETEC Armando Pannunzio
   (e) ETEC Piedade
5. Qual seu curso técnico?
   (a) Administração
   (b) Agenciamento de Viagem
   (c) Alimentos
   (d) Automação Industrial
   (e) Contabilidade
   (f) Design de interiores
   (g) Edificações
   (h) Elétrica
   (i) Eletrotécnica
   (j) Eletroeletrônica
   (k) Enfermagem
   (l) Eventos

(m) Informática
(n) Informática para Internet
(o) Logística
(p) Marketing
(q) Mecânica
(r) Mecatrônica
(s) Nutrição
(t) Química
(u) Recursos Humanos
(v) Segurança do Trabalho
(w) Serviços Jurídicos

6. Período do Curso?

(a) Manhã
(b) Tarde
(c) Noite
(d) Integrado

7. Porque escolheu este curso?
8. Você já trabalha na área do seu curso?

(a) Sim
(b) Não

9. Este curso atendeu suas expectativas?

(a) Sim
(b) Não
(c) Parcialmente

10. Justifique?
11. Qual a renda familiar?

(a) Até 1 salário mínimo
(b) Mais de 1 até 2 salários mínimos
(c) Mais de 2 até 3 salários mínimos
(d) Mais de 3 até 4 salários mínimos
(e) Mais de 4 salários mínimos

12. Quantas pessoas vivem em sua casa?

(a) 1 (somente você)
(b) 2 pessoas
(c) 3 pessoas
(d) 4 pessoas
(e) 5 pessoas
(f) 6 pessoas
(g) Mais de 6 pessoas

13. Você trabalha?

(a) Sim
(b) Não

14. Se sim, em que trabalha? (Descrição: caso não trabalhe, escreva "não trabalho")

15. Você já faz estágio?

(a) Sim
(b) Não

16. Quando terminar o curso, você pretende trabalhar na área?

(a) Sim
(b) Não
(c) Ainda não decidi

17. Qual foi (ou ainda é) a profissão do seu pai? (Descrição: se não souber a profissão do pai, escrever "não sei")

18. Qual o grau de escolaridade do seu pai?

(a) Não sabe ler nem escrever.
(b) Ensino fundamental incompleto.
(c) Ensino fundamental completo.
(d) Ensino médio incompleto.
(e) Ensino médio completo.
(f) Ensino superior incompleto.
(g) Ensino superior completo.
(h) Pós-graduação.
(i) Não sei.

19. Qual foi (ou ainda é) a profissão da mãe? (Descrição: se não souber a profissão da mãe, escrever "não sei")

20. Qual o grau de escolaridade da sua mãe?

(a) Não sabe ler nem escrever.

(b) Ensino fundamental incompleto.

(c) Ensino fundamental completo

(d) Ensino médio incompleto.

(e) Ensino médio completo.

(f) Ensino superior incompleto.

(g) Ensino superior completo.

(h) Pós-graduação.

(i) Não sei.

21. Qual é a profissão que você sempre sonhou em seguir? (Descrição: responder apenas uma profissão. Caso não tenha decidido, escrever "nunca pensei nisso")

22. Quando terminar o curso você pretende?

(a) Fazer uma faculdade particular na área do seu curso

(b) Fazer uma faculdade particular em uma área diferente do meu curso

(c) Fazer uma faculdade pública na área do seu curso

(d) Fazer uma faculdade pública em uma área diferente do meu curso

(e) Não pretendo fazer faculdade

(f) Pretendo fazer outro curso técnico

(g) Não sei

23. Qual a sua escolaridade?

(a) Estou fazendo integrado com o Ensino médio na ETEC

(b) Não sou do curso do integrado, mas estou cursando o ensino médio da ETEC e fazendo o curso técnico da ETEC.

(c) Não sou do curso do integrado, mas estou cursando o ensino médio em outra escola e fazendo o curso técnico da ETEC.

(d) Já conclui o Ensino médio

(e) Ensino superior completo

(f) Cursando o ensino superior

(g) Já tenho pós-graduação.

24. Vc já fez outro Curso Técnico na ETEC?

(a) Não, este é o primeiro

(b) Sim, já possuo 1 curso completo.

(c) Sim, já possuo mais de 1 curso completo.

E Sobre a escolha dos escores

As estratégias utilizadas para análise de dados ordinais comumente requerem a escolha de escores para as categorias da variável resposta. Duas maneiras usuais de escolhas são:

i) Escores inteiros e: são definidos por $a_j = j$ ou $a_j = j - 1$, para $j = 1, ..., r$,sendo úteis quando as $r > 2$ categorias ordenadas da variável resposta são assumidas como sendo igualmente espaçadas, bem como quando as categorias correspondem as contagens inteiras.

ii) Escores padronizados (*standardized midranks*): esses escores são restritos a valores entre 0 e 1 sendo definidos por

$$a_j = \frac{2\left[\sum_{k=1}^{j} n_{+k}\right] - (n_{+j}) + 1}{2\,(n+1)}.$$

A diferença entre os escores padronizados e os escores inteiros é que os dados são utilizados para obtenção dos escores padronizados. Assim, o analista não se responsabiliza diretamente pela a escolha dos escores, o que não representa necessariamente uma vantagem, como observado a seguir.

Para muitos conjuntos de dados, a escolha dos escores apresenta pequeno efeito nos resultados. Escolhas diferentes de escores inteiros usualmente fornecem resultados similares. Contudo isso pode não acontecer quando os dados são desbalanceados, como quando algumas categorias apresentam muito mais observações do que outras. Com os escores padronizados, isso também ocorre, visto que aquelas categorias com poucas observações em relação ás demais apresentarão escores muito próximos. A consequência é que as distâncias entre as categorias da variável resposta podem ser consideradas muito mais próximas do que elas realmente são.

A escolha de escores não constitui, portanto, em uma tarefa simples. [2] recomenda que os dados sejam analisados considerando diversos conjuntos da escores a fim de se observar se conclusões importantes depende da escolhas feitas. Vale ressaltar, ainda, que interação com pesquisador é extremamente importante para o entendimento das distâncias entre as categorias e consequentemente escolha adequada dos escores.