ANÁLISE DE DADOS *CATEGÓRICOS* EM CIÊNCIA POLÍTICA

Uso de testes estatísticos em tabelas de contingência com fontes secundárias de dados

EMERSON
URIZZI
CERVI

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA POLÍTICA/**UFPR**

ANÁLISE DE DADOS CATEGÓRICOS EM CIÊNCIA POLÍTICA

Uso de testes estatísticos em tabelas de contingência com fontes secundárias de dados

Autor: Emerson Urizzi Cervi
Composição gráfica: Luciano B. Lemos
ISBN (e-book): 978-85-915195-0-7

CERVI, Emerson U.
Análise de Dados Categóricos em Ciência Política: Uso de testes estatísticos em tabelas de contingência com fontes secundárias de dados / Emerson U. Cervi
Independente. Curitiba: Pós-graduação em Comunicação e Pós-graduação em Ciência Política – Universidade Federal do Paraná (UFPR), 2014. 98 p.

Edição 1. E-book Versão PDF.

1. Comunicação 2. Ciência Política 3. Ciências Sociais

ISBN 978-85-915195-0-7

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA POLÍTICA – UFPR

SUMÁRIO

1. INTRODUÇÃO

Vivemos tempos de crescente disponibilidade de dados primários para produção de análises científicas no campo da política. O desenvolvimento das tecnologias digitais, aliado à tendência de transparência dos órgãos públicos junto com as práticas de prestação pública de contas tem feito com que mais e mais informações sejam facilmente acessadas para posteriores análises. Dadas as limitações de tempo de energia dos pesquisadores, podemos dizer que a disponibilidade de informações tende ao infinito. Então, por que fazer um curso sobre análises de dados categóricos a partir de fontes secundárias em tempos de *big data*? A resposta a essa pergunta deve ser dividida em duas partes.

Primeiro, porque a maior quantidade de dados disponíveis em fontes primárias não significa necessariamente melhor qualidade de informações. Muitas fontes de *big data* disponibilizam apenas relatórios com resultados sumarizados e não as bases de dados. Não raras vezes o pesquisador tem que copilar informações de diferentes relatórios para formar seu próprio banco de dados. O pior acontece quando não é possível a compilação de informações de diferentes fontes e a pesquisa "estaciona", ou melhor, "encalha" no imenso lamaçal de *terabytes* de informações disponíveis, mas inúteis.

Segundo, e principalmente, a utilização de dados secundários em análises nos permite cumprir uma das funções da pesquisa científica que tende a ser desconsiderada cada vez mais: a possibilidade de replicar dados para testar os resultados e os "achados" de outros pesquisadores. Uma das funções menos exercitadas da pesquisa científica é justamente a de testar resultados obtidos em trabalhos anteriores com novos dados ou utilizando outras ferramentas analíticas. Em tempos de *big data* isso é explicado pelo fato de sempre termos novas informações disponíveis para as atuais pesquisas. Parece que estamos sempre recomeçando a partir de novas informações, o que dificulta a realização de trabalhos que façam comparações com dados e resultados de pesquisas já desenvolvidas. Seja para atualizar as conclusões anteriores, seja para questionar a validade de conclusões para a realidade atual.

O objetivo deste *e-book* é contribuir para a difusão de técnicas de pesquisa empírica aplicada à área da ciência política que, apesar de bastante simples, permitem uma notável diferença na qualidade das análises e conclusões a que chegam os cientistas políticos. Este e-book nasceu de um minicurso ministrado por mim a alunos de pós-graduação em Ciência Política da Universidade de Campinas no segundo semestre de 2013, a convite do professor Bruno Wilhelm Speck. Poucas mudanças foram feitas após o curso. De mais significativo houve apenas o acréscimo de um teste que não foi discutido naquela ocasião. De saída, agradeço imensamente não apenas o convite para ministrar o curso mas, também, a leitura cuidadosa do professor Bruno W. Speck

após o curso e o apontamento de erros no texto inicial. Também agradeço a doutoranda em ciência política, Michele Goulart Massuchin, pela leitura e revisão textual. Ambos foram generosos e contribuíram muito para o resultado final do trabalho. Claro que nem todos os problemas foram ou serão resolvidos aqui. Os que ainda permanecem são de minha exclusiva responsabilidade.

Fazer pesquisas comparando resultados encontrados por outros pesquisadores acrescenta dimensão temporal, pois permite usar resultados do passado cujas fontes primárias não estão disponível, e dimensão espacial ao garantir comparabilidade entre áreas geográficas que não tenham disponibilidade de dados primários. Nem sempre trabalhar com um grande volume de dados é garantia de aumento na qualidade dos resultados. Não raras vezes, dado o recente desenvolvimento tecnológico, uma pesquisa em tempos de *big data* apresenta conclusões restritivas no tempo e no espaço, o que pode reduzir o escopo de suas conclusões.

Recorrer a fontes secundárias de informações, tais como sumarização de dados em frequências de uma variável simples ou em tabelas de contingência, pode elevar consideravelmente a qualidade de uma pesquisa científica ao produzir coeficientes que possam ser comparados entre informações de 50 anos atrás, quando os bancos de dados primários não estavam disponíveis, com os resultados atuais. Ou pode ser a base para uma pesquisa atual que queira testar conclusões apresentadas no passado. Por exemplo, em um livro sobre o coronelismo e as eleições municipais brasileiras entre final do século XIX e início do século XX, *Eul Soo Pang* apresenta uma lista com os números de eleitores nos que chama de municípios mais coronelistas da Bahia (Ver quadro 1) em cinco momentos distintos do tempo.

A tabela publicada por *Eul Soo Pang* na página 239 de seu livro "Coronelismo e Oligarquias" é um exemplo de fonte secundária de informações, pois os dados já foram organizados de determinada maneira. A descrição do número de eleitores por município em cinco momentos do tempo só nos permite identificar um crescimento no número total de eleitores entre 1905 e 1934 de mais de 100% no período, passando de 73,4 mil no início até 153,3 mil na última medição. No entanto, apenas com essas informações não é possível saber se houve um crescimento proporcional equivalente entre todos os municípios analisados ou se há diferenças significativas no crescimento apresentado por eles. Além disso, se completássemos as informações de eleitores a partir desse levantamento, teríamos uma ampliação no período analisado. A proposta aqui é usar dados sumarizados como base informacional para a produção de coeficientes estatísticos que permitam análises, comparações e conclusões.

NÚMERO DE ELEITORES NOS DEZ MUNICÍPIOS MAIS CORONELISTAS, 1905-34

Município	1905	1908	1910	1912	1934
Andaraí	608	686	686	999	939
Barreiras	715	727	1255	1260	1410
Carinhanha	514	660	660	813	301
Lençóis	695	850	848	848	644
Maracás	384	442	1.053	1.071	1.353
Mucugê	626	1.185	1.185	1.185	233
Pilão Arcado	235	435	509	480	627
Remanso	377	786	878	1.007	1.000
Rio Preto	369	558	558	609	78
Sento Sé	492	492	492	702	396
Bahia (total):	73.441	91.174	99.935	108.463	153.376

Fonte: Ministério da Agricultura, Indústria e Comércio, Diretoria do Serviço de Estatísticas, Estatística eleitoral da República dos Estados Unidos do Brasil, pp; 17-29, Guedes, Anuário... 1934, pp. 45-48.

NOTA: Esses números representam os eleitores que realmente votaram nessas eleições. O número verdadeiro de eleitores registrados devia ser maior.

QUADRO 1. EXEMPLO DE DADOS A PARTIR DE FONTE SECUNDÁRIA

Extraído de: PANG, Eul-Soo. Coronelismo e Oligarquias 1889-1943. Rio de Janeiro: Civilização Brasileira, 1979 (p. 239)

Como os dados são sumarizados em tabelas de frequências de categorias (p.ex. número de eleitores em 10 municípios da Bahia em 1905, onde cada município será uma categoria com determinada quantidade de eleitores) ou em cruzamento de categorias de variáveis independentes (var.1 = número de eleitores por município, var.2 = ano da eleição), estamos sempre falando em variáveis categóricas. O objetivo principal deste curso é apresentar ferramentas estatísticas quantitativas que servem para medir diferenças de qualidades em uma variável ou na relação entre duas ou mais variáveis associadas. Como estamos falando de tabelas de contingência com casos limitados ao número de linhas e número de colunas, podemos realizar todos os testes apenas com o auxílio de uma calculadora simples, dispensando o uso de softwares e pacotes estatísticos complexos.

É importante que o leitor tenha domínio de alguns conceitos que serão fundamentais para a compreensão da mecânica dos cálculos. Os principais que aparecerão são: "variável categórica", um tipo de variável que mede a distribuição das quantidades de qualidades das características

estudadas e cujos valores que dão sentido às qualidades são arbitrados pelo pesquisador. Há duas variáveis categóricas principais, as nominais, que apenas separam grupos distintos e as ordinais, que separam e ordenam por alguma ordem de grandeza. Serão apresentados testes para variáveis nominais e outros testes para variáveis ordinais. O segundo conceito importante está ligado ao "número de categorias" de uma variável. Isso porque para variáveis com o menor número possível de categorias, duas, há um tipo de teste que não se aplica às variáveis com três ou mais categorias. Outra ideia que o leitor já deve ter em mente para conseguir aproveitar melhor o curso é a de "transformações em escalas de variáveis". Ela considera ser possível transformar uma escala de variável desde que seja para reduzir o número de categorias da mesma. Não é possível fazer transformações aumentando o número de categorias. Isso significa que qualquer variável pode se transformar em uma categórica com apenas duas categorias, mas o contrário não é verdadeiro. Todos os testes para duas variáveis que serão apresentados daqui para frente partem da verificação da consistência nas relações entre "pares de categorias". Por isso o conceito de pares de categorias é importante. Em uma tabela de cruzamentos um par entre duas categorias de variáveis independentes pode assumir duas posições: é consistente, quando apresenta ou não as características medidas nas duas variáveis, ou é inconsistente quando só apresenta a característica de uma das variáveis testadas. Espera-se encontrar mais pares inconsistentes em variações independentes e os pares consistentes devem crescer quando as duas variáveis apresentam algum tipo de associação. O último conceito importante aqui é o de "independência estatística", fundamentado na teoria das hipóteses. Não aprofundaremos esse tema adiante por considerar que o leitor está familiarizado com os princípios dos testes de hipóteses (H_0 : hipótese nula e H_1 : hipótese alternativa). A meta comum em todos os testes estatísticos que serão apresentados aqui é verificar a chance de erro ao se rejeitar a hipótese nula e, sendo o erro aceitável, assumir a hipótese alternativa que é a de que existe alguma associação entre as variáveis testadas.

Aqui, os termos correlação e associação serão usados como sinônimos para representar um tipo de relação entre duas variáveis a partir do comportamento dos pares de categorias analisadas. Todos os dados utilizados nos exemplos são relativos ao processo eleitoral de 2012 e foram extraídos do repositório de dados eleitorais do portal TSE (http://www.tse.jus.br/eleicoes/repositorio-de-dados-eleitorais) como fonte primária de informações. As tabelas completas de distribuições de frequências e de contingência que dão origem aos dados usados em todos os testes de hipóteses são apresentadas no Anexo II. Optamos por considerar apenas os grandes partidos em número de prefeitos eleitos para os testes, pois a manutenção de todas as siglas nas tabelas de contingência quebraria o pressuposto de pelo menos cinco casos em cada frequência esperada da distribuição (esse pressuposto será apresentado em detalhes no momento adequado). Para definir quais foram os partidos grandes em 2012

padronizamos os números de eleitos pelo Z-score e selecionamos apenas aqueles que ficaram com Z-score acima de zero, ou seja, com número de eleitos acima da média de todos os partidos. Isso resultou em uma lista com 10 partidos, cf. consta na tabela 1 a seguir. Em seguida os grandes partidos foram agrupados por posição no espectro ideológico (esquerda, centro e direita), segundo o que mais aparece na literatura da área.

TABELA 1. DISTRIBUIÇÃO DOS PRINCIPAIS PARTIDOS POR POSIÇÃO IDEOLÓGICA E Z-SCORE

POSIÇÃO	PARTIDO	ELEITOS	Z-SCORE
centro	PMDB	1024	3,04
centro	PSDB	694	1,81
esquerda	PT	628	1,56
direita	PSD	495	1,06
direita	PP	464	0,94
esquerda	PSB	440	0,85
esquerda	PDT	308	0,36
direita	PTB	294	0,31
direita	DEM	276	0,24
direita	PR	272	0,22

Entre os dez partidos considerados aqui temos três de esquerda (PT, PSB e PDT), dois de centro (PMDB e PSDB) e cinco de direita (PDS, PP, PTB, DEM e PR). A essas informações são acrescentadas características como tamanho do município e localização no País (segundo IBGE), escolaridade e sexo do candidato (segundo autodeclaração ao TSE) para serem usadas nos testes de independência a partir daqui. Diferente do que normalmente acontece em manuais de técnicas, começaremos e terminaremos com os cálculos mais fáceis e intuitivos, não seguiremos uma ordem linear de crescimento da complexidade dos cálculos. Os primeiros, mais simples, são os testes de Risco Relativo (RR) e de qui-quadrado (χ^2) para uma variável ou para duas variáveis; em seguida passamos para os testes de associação para duas variáveis dicotômicas Q-yule (Q_{xy}), para seguirmos até os testes Q_{xy} para três variáveis. Esse teste pode ser aplicado para um número maior de variáveis testadas, mas limitaremos a abordagem aqui às relações entre três variáveis. Depois, apresentaremos um teste para medir associação entre duas variáveis categóricas ordinais, o Coeficiente Gama. Por último serão apresentadas técnicas bastante simples para análise do comportamento de pares de características em variáveis categóricas nominais; as

chamadas análises de Resíduos Brutos (R_b) e Resíduos Padronizados (R_p). Os cálculos e análises de resíduos são tão elementares quanto o teste de Risco Relativo, portanto, se você chegar ao meio do curso, não vale a pena desistir. No final, tudo ficará mais fácil. Ao final de cada teste são sugeridos alguns exercícios complementares para fixação dos cálculos.

Antes de tratar dos cálculos e interpretar os coeficientes, é bom discutir minimamente os conceitos que norteiam as técnicas específicas para análises de dados categóricos e os conceitos de variáveis qualitativas. É por esse ponto que começaremos.

2. ANÁLISE DE DADOS CATEGÓRICOS: DEFINIÇÕES BÁSICAS

A análise de dados categóricos permite que informações qualitativas a respeito dos eventos pesquisados sejam tratadas e analisadas a partir de técnicas quantitativas. Aqui, dado qualitativo e variável categórica são usados como sinônimos e a definição dada a eles é de que um dado qualitativo é uma representação atribuída a quantidades de manifestações de determinada qualidade. Então, chamamos de "variável categórica" a característica medida em determinado objeto de estudo que apresenta duas ou mais variações em quantidades distintas. O dado qualitativo classifica, assim, um fenômeno quase que imponderável a partir de premissas ontológicas e semânticas. Por exemplo: partido de direita e partido de esquerda é uma construção semântica, pois eles não existem na realidade. São construídos em função de comportamentos distintos que seus agentes políticos adotam em relação aos mesmos temas. Com a classificação é possível instrumentalizar o reconhecimento do evento, analisar seu comportamento e suas relações com outros eventos. Nesse sentido, trata-se de uma qualificação normativa que dá caráter objetivo à análise. Estar em determinado partido pode ser entendido como uma qualidade. Ter sido eleito em uma eleição é outra qualidade que se opõe à qualidade de ter sido derrotado. A análise qualitativa com técnicas quantitativas é considerada uma alternativa à pesquisa qualitativa, que se ocupa dos mesmos eventos, porém, com menor restritividade técnica e maior possibilidade de intervenção da subjetividade do pesquisador.

Antes de começar a medir características qualitativas é preciso ter em mente a distinção entre **objeto** e **atributo**. A informação qualitativa é uma estratégia de mensuração de atributos, ou seja, a mensuração não se dá sobre o objeto em si, a coisa, mas sobre uma ou algumas de suas características e predicados, aqui chamados de atributos. O que pretendemos aqui é analisar atributos dos objetos e não os objetos em si. Existem dois tipos de medidas que servem para identificar esses atributos: as **fundamentais** e as **derivadas**. As medidas fundamentais são aquelas em que a mensuração é feita diretamente sobre o objeto. Ex.: quando se usa uma balança para medir o peso das pessoas. Estamos medindo o atributo diretamente no objeto. No segundo caso, nas medidas derivadas, é feita uma projeção a partir de uma medida indireta e não diretamente sobre o objeto. Ex.: a opinião sobre preconceito medida a partir das respostas em um *survey*. Cada respondente emite sua opinião sobre comportamentos preconceituosos, mas o pesquisador não tem como ter certeza se o respondente adota o comportamento que ele diz preferir ou não. A opinião é uma medida derivada a respeito de comportamento sobre determinado tema. As técnicas de análise de dados categóricos em ciência política normalmente preocupam-se com a análise de atributos dos objetos a partir de medidas derivadas.

Antes de começar qualquer análise de dados categóricos é preciso processar os dados para ajustá-los às medidas necessárias para os objetivos pretendidos. Essa análise deve procurar estabelecer pelo menos uma das quatro finalidades a seguir (Günther, 2003):

a) relações de similaridade entre as categorias, quando se pretende verificar se a ocorrência de determinada qualidade é similar ou não à ocorrência de outra qualidade;

b) uma razão de ocorrência a partir da contagem entre duas variáveis, quando o objetivo é verificar quanto há de ocorrência de uma qualidade em comparação ao total das ocorrências observadas;

c) distribuição hierárquica das posições em escalas ordinais, quando o pesquisador não apenas identifica as diferenças das qualidades, mas também as elenca em função das diferenças de quantidades de determinada característica entre as categorias.

d) apenas uma correlação entre os valores encontrados para as variáveis, para quando o objetivo do pesquisador é estabelecer o grau de associação entre ocorrências de qualidades em variáveis distintas ou independentes.

Uma das formas mais comuns de análise de dados categóricos é a partir da redução dos valores de diferentes variáveis com a criação de uma escala de medida. Uma escala de medida nos permite organizar os valores de uma variável, viabilizando sua análise. Sendo assim, toda escala envolve a identificação de determinadas premissas de relação entre as qualidades analisadas e a representação dessas qualidades. Um ponto importante é saber que as escalas de valores categóricos atribuem rótulos (normalmente números) às características analisadas e, portanto, esses rótulos sempre são arbitrários, definidos pelo pesquisador. Por exemplo, um pesquisador pode dar os seguintes rótulos para as qualidades Homem, Mulher da variável Sexo (1=homem, 2=mulher). Porém, outro pesquisador pode dar os seguintes rótulos (1=mulher, 2=homem). Não há erro ou acerto aqui, pois os rótulos são arbitrários.

Uma vez rotuladas, precisamos começar a distinguir as variáveis em função de suas características internas, ou, em função do tipo de escala. Existem quatro tipos de escalas, mas aqui trataremos de apenas duas delas, aquelas que são identificadas como escalas categóricas:

a) Escala Nominal: Permite a medição de atributos apenas a partir do estabelecimento de relações de equivalência, ou seja, de igualdade (=) ou de diferença (≠), independente de quais sejam seus códigos numéricos. Essa escala não tem sentido de direção ou valor nulo. É a forma mais primária de medição e, portanto, é a que nos oferece o menor número de informações sobre o objeto pesquisado. Ex.: Sexo (homem ou mulher); Ideologia partidária (esquerda, centro ou direita). Ser homem não é melhor nem pior do que ser mulher. É apenas diferente. Escalas nominais indicam diferenças ou semelhanças entre as categorias de determinada característica.

b)Escala ordinal: Apresenta um volume de informações superior à anterior, pois além de distinguir os grupos pela presença de determinada característica, também hierarquiza essas

diferenças. A escala ordinal permite medir os atributos que conseguem se distinguir em termos de grau ou de intensidade, indo além das simples relações de igualdade/diferença. Aqui é possível identificar uma categoria que é maior que (>) outra ou menor que (<) outra. Sendo assim, apresenta direção e sentido. Permite o estabelecimento de uma hierarquia entre atributos e sentido de orientação da escala. Ex.: classificação dos partidos em três grupos a partir das votações obtidas em i) Alta votação, ii) Média votação, iii) Baixa votação. Partidos do primeiro grupo não são apenas diferentes dos dois outros. Eles também apresentam um volume maior da característica (votos) que o distingue dos grupos. E os partidos do grupo três apresentam menor quantidade da qualidade analisada do que os outros dois grupos. O quadro a seguir resume as principais características de cada uma das escalas categóricas de representação das variáveis.

QUADRO 2. PRINCIPAIS CARACTERÍSTICAS DOS TIPOS DE ESCALAS CATEGÓRICAS

Tipo	Características	Exemplos	Funções Formais
Nominal	Apenas para identificar pessoas, objetos ou categorias.	Cor de cabelo, estado civil, nome, marca de carro.	Igualdade ou diferença. "=" "≠"
Ordinal	Respostas podem ser ordenadas em uma dimensão própria.	Ordem de preferência, de chegada, status social, escala de Likert.	Além de igualdade ou diferença, mostra superioridade ou inferioridade. ">" "<"
Fonte: adaptado de Günther, 2003			

Todos os testes estatísticos apresentados aqui foram concebidos para analisar um desses dois tipos de variáveis: nominais ou ordinais. Se a variável é nominal ou ordinal importa para definir que tipo de teste usar. Outra característica importante para a definição do teste estatístico é o número de categorias de cada variável. Dentro do conjunto de variáveis categóricas, podemos distinguir dois grandes conjuntos em função no número de categorias:

a) variáveis dicotômicas ou binárias: são aquelas que apresentam apenas duas categorias. É a menor extensão possível de variação, pois espera-se pelo menos duas características para que haja o mínimo de variação. Por exemplo: Sexo (Homem ou Mulher); Resultado da eleição (Eleito ou não eleito).

b) variáveis politômicas: são as que apresentam três ou mais categorias. Elas podem assumir o número necessário de categorias para diferenciar as qualidades presentes na variável. Por exemplo: Região do País (norte, nordeste, centro-oeste, sudeste e sul), ideologia do partido (direita, centro e esquerda); Tamanho do município (micro, pequeno, médio e grande).

Ao reunirmos as duas características (tipo de escala e número de categorias) encontramos quatro formas possíveis de organização de variáveis categóricas:

- Dicotômica nominal: Sexo (homem, mulher).
- Dicotômica ordinal: Volume de votos (Alto, baixo).
- Politômica nominal: Região do país (norte, nordeste, centro-oeste, sudeste e sul).
- Politômica ordinal: Tamanho do município (micro, pequeno, médio e grande).

Uma das características mais importantes das variáveis categóricas diz respeito à possibilidade de redução do número de categorias existentes até o limite da dicotomização. Podemos ter uma variável contínua, tal como número de votos obtidos pelo partido, e transformá-la em uma variável categórica politômica ordinal agrupando os partidos em três grupos de votação: partidos com votação alta, partidos com votação média e partidos com votação baixa. Depois, podemos ainda reduzir o número de categorias para apenas duas: partidos com alta votação x todos os demais. No entanto, essa característica transitiva das variáveis se dá em uma única direção. Da direção do maior número de categorias para o menor. Não é possível transformar uma variável dicotômica em politômica e dessa para uma contínua.

Um tipo específico de variável categórica politômica muito usado nas pesquisas em ciência política é a chamada Escala de Likert, apresentada por Rensis Likert em 1932, que permite transformar variáveis contínuas em categóricas ordinais preservando i) as manifestações de qualidades, ii) reconhecimento de oposição entre contrários, iii) estabelecimento de gradientes e iv) a identificação de uma posição intermediária. Em outras palavras, a escala de Likert tem cinco pontos, com um ponto médio para manifestações intermediárias, de indiferença ou nula. O exemplo mais comum de Likert na ciência política é a escala de avaliação de governo, onde as alternativas são: (1) péssimo, (2) ruim, (3) regular, (4) bom, (5) ótimo. A oposição se dá entre as posições 1,2 e 4,5. Gradiente está nas diferenças possíveis dentro da mesma direção, entre 1+2 e entre 4+5. O ponto médio ou neutro é o 3. Com isso podemos distinguir avaliações positivas das negativas, podemos estabelecer diferentes níveis de positividade e negatividade, além de identificar o ponto médio ou neutro da distribuição.

Alguns dos testes que serão estudados a seguir têm o objetivo de identificar a existência de distribuições estatisticamente significativas em favor de determinado ponto de uma escala categórica. Permite verificar se a concentração de casos nas categorias positivas da avaliação de governo, por exemplo, é suficientemente alta para sustentar a afirmação de que os entrevistados avaliam o governo de maneira mais positiva do que negativa, por exemplo. Outros testes permitem verificar a força da relação entre categorias de duas variáveis independentes, chamadas de X e Y. Vale lembrar que as relações entre duas variáveis categóricas podem ser feitas por gráficos ou tabelas. Vários autores recomendam que a análise de dados qualitativos dê-se a

partir de representações visuais, como gráficos, em lugar de tabelas, pois o que se busca é a redução de dimensionalidades. No entanto, aqui não usaremos representações gráficas, mas sim as tabulares. Isso porque pretendemos encontrar um valor numérico, chamado de coeficiente, que seja capaz de reduzir toda a complexidade da relação entre as categorias de duas variáveis. No próximo tópico são apresentados os cálculos para identificação dos coeficientes mais simples, para identificação de relações entre categorias de uma mesma variável ou para teste de independência das distribuições de valores entre duas variáveis.

2.1. EXERCÍCIOS

Responda as perguntas:

2.1.a. É possível transformar uma variável categórica nominal em uma variável categórica ordinal? Por quê?

2.1.b. A partir dos dados a seguir, transforme a variável contínua notas dos alunos em uma variável dicotômica que separa em dois grupos de igual tamanho os alunos com menores notas dos alunos com maiores notas.

Notas dos alunos:

62	58	76	32	45	72	70	46	39	94	70	73	79	91	73	88	77	32	74	36

3.COEFICIENTES BÁSICOS

3.1. APLICAÇÃO DO RISCO RELATIVO NAS CIÊNCIAS SOCIAIS (RR)

Os testes de risco relativo são originários da área da saúde e usados para identificar diferentes probabilidades de ocorrência de determinado comportamento em alguns indivíduos quando comparado a outros, em função de características específicas. Nas ciências sociais ele equivale ao teste de razão de chance. Esse conceito pode ser adaptado para testar estatisticamente as diferenças de probabilidades para ocorrência de determinado fenômeno (chamado de risco) em função de características presentes na população pesquisada. Por exemplo, pensando em estudos eleitorais, qual o risco (probabilidade) de encontrar um eleitor sem candidato em períodos pré-eleitorais caso ele apresente baixo interesse por política. Com isso, é possível estabelecer uma escala de baixo ou quase nulo até risco máximo para os casos de alta probabilidade. Conceituamos "risco" em ciências sociais como a probabilidade que um indivíduo ou grupo de indivíduos têm de apresentar determinado comportamento/opinião em função de suas características atuais. Portanto, "risco" é um conceito probabilístico e não determinístico. Não se pode falar em risco quando se sabe que não há probabilidade alguma de ocorrência do fenômeno. Por exemplo, qual o risco de brasileiros menores de 16 anos votarem? Nenhum, pois eles são proibidos. Não há probabilidade nessa ocorrência.

O risco da ocorrência de determinado evento varia de probabilidade teórica zero, quando não há chance de ocorrência, até a probabilidade teórica de 1 (um), quando todos os indivíduos apresentarão ocorrência do fenômeno. Quanto mais próximo de zero, menor o risco e, portanto, menos diferenciação a característica estudada é capaz de apresentar nos integrantes da população pesquisada. A essa característica medida dá-se o nome de "**fator de risco**". Quanto maior a presença do fator de risco, maior a probabilidade de encontrarmos determinado fenômeno, ou seja, características que apresentam uma associação empírica e significativamente estatística com determinado efeito/fenômeno. A forma mais comum de quantificar o risco de determinada ocorrência a partir da presença de uma característica é através do cálculo de Risco Relativo (RR).

Risco Relativo é medido pelo risco de uma característica relacionar-se com outra em determinado indivíduo ou unidade de análise. Ele mede a potência da associação entre características distintas. Trata-se da relação entre o cociente do risco de apresentar determinada característica daqueles que estão expostos ou possuem o fator de risco possível de ser identificado, comparado com os que não apresentam esse fator. A fórmula pode ser representada por:

$$RR = \frac{Incidência\ da\ característica\ nos\ que\ possuem\ o\ fator\ de\ risco}{Incidência\ da\ característica\ entre\ os\ que\ não\ possuem\ o\ fator\ de\ risco}$$

Por exemplo, imagine que queremos analisar o risco relativo do eleitor que não possui preferência por partidos políticos de não ter candidato a prefeito antes do início da campanha eleitoral. Para facilitar o cálculo, os dados devem ser dispostos em uma tabela quádrupla, organizada de forma que apresente a presença ou ausência das características, como a tabela a seguir:

QUADRO 3.1. EXEMPLO DE DISTRIBUIÇÃO DE DADOS PARA CÁLCULO DO RR

Fator	Comportamento/opinião		TOTAL
	Sim	Não	
Sim	a	b	a + b
Não	c	d	c + d
TOTAL	a + c	b + d	a + b + c + d

Onde,

a = é a presença do fator (característica) e do comportamento ou opinião estudada, também chamado de verdadeiro positivo.

b = presença do fator (característica), sem a presença do comportamento ou opinião estudada, ou, falso positivo.

c = ausência do fator (característica), com a presença do comportamento ou opinião estudada, falso negativo.

d = ausência do fator (característica) e do comportamento ou opinião estudada, verdadeiro negativo.

a + b = total dos que apresentam o fator (característica).

c + d = total dos que não apresentam o fator (característica).

a + c = total dos que apresentam o comportamento/opinião estudado.

b + d = total dos que não apresenta o comportamento/opinião estudado.

a + b + c +d = total de indivíduos dos quais se tem informações.

No nosso exemplo, o fator de risco (característica) é declarar ter preferência por partido político em pesquisa de opinião pública e o comportamento estudado é ter candidato a prefeito antes do início da campanha eleitoral. O que se quer testar aqui é se existe um risco maior de um eleitor apresentar preferência por um candidato entre aqueles que dizem ter preferência por algum partido político ou não. A hipótese é que deve existir uma relação entre as duas variáveis e, nesse caso, um risco maior de um eleitor com preferência partidária declarar ter candidato a

prefeito antes mesmo do início da campanha do que um eleitor sem preferência partidária. Substituindo no quadro acima temos a seguinte tabela:

TABELA 3.1. DISTRIBUIÇÃO DA PREFERÊNCIA PARTIDÁRIA E TER CANDIDATO A PREFEITO

Preferência partidária	Possui candidato a prefeito		TOTAL
	Sim	Não	
Sim	114	110	224
Não	56	146	202
TOTAL	170	256	426

A fórmula para o cálculo da incidência do risco é probabilística e divide a proporção parcial dos que têm a característica medida pela presença da característica e do comportamento, dividido pela proporção dos que não têm à característica medida pelos que não têm a característica nem o comportamento esperado. Em termos matemáticos a fórmula é a seguinte:

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

No caso do exemplo acima, a aplicação é:

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{\frac{114}{224}}{\frac{56}{202}} = \mathbf{1{,}85}$$

A leitura dos resultados é a seguinte: existe 1,85 vezes mais chance de um eleitor com preferência partidária possuir candidato a prefeito antes do início da campanha eleitoral do que um eleitor sem preferência partidária. RR menores que 1,5 não são considerados não-práticos, ou seja, devem ser tido como fator que não interfere no comportamento estudado. Portanto, o risco acima, de 1,85 é relevante do ponto de vista prático. Ou seja, campanhas eleitorais são mais importantes para a decisão de voto dos eleitores sem preferência partidária do que os que já possuem simpatia por algum partido – independente de ter ou não candidato à eleição. Se tivéssemos os mesmos dados para eleições em outros momentos do tempo ou para eleitores de diferentes distritos eleitorais, poderíamos comparar os RRs obtidos em cada cálculo para saber se o risco relativo de encontrar eleitores com preferência partidária que dispensam as campanhas eleitorais para escolher seus candidatos está aumentando ou diminuindo.

Apesar da vantagem da simplicidade do cálculo, o RR apresenta algumas limitações. A primeira delas é que ele só se aplica para medir o risco relativo entre variáveis categóricas dicotômicas,

pois a ocorrência de determinada característica sempre é relativa à sua não ocorrência na variável. Outra limitação é o fato de que seu cálculo não leva em consideração os pares de ocorrências, mas sim a proporção de ocorrência do fator de risco em comparação com a proporção da não ocorrência desse fator. Porém, o maior problema desse indicador é a arbitrariedade na determinação do limite sobre o qual consideramos se o fator de risco é alto ou baixo. Um coeficiente muito usado para testes de independência de variáveis categóricas é o χ^2, por apresentar uma série de vantagens em relação ao RR. O χ^2 é um coeficiente que pode ser usado para identificar diferenças estatisticamente significativas nas distribuições de casos entre as categorias de uma única variável e nas tabelas de contingência para variáveis nominais. Por esses motivos, no próximo tópico trataremos em detalhes do χ^2.

3.2. TESTE QUI-QUADRADO (χ^2)

O teste de χ^2 foi proposto pelo estatístico Karl Pearson em 1900 e por isso é conhecido por χ^2 de Pearson. Serve para comprovar se existem diferenças estatisticamente significativas entre duas distribuições quaisquer ou entre casos de uma mesma distribuição. É usado em variáveis categóricas e também parte da hipótese nula de que não existem diferenças significativas entre as distribuições comparadas. O método de análise compara os resultados observados com os resultados esperados para verificar se há ou não diferenças significativa entre as distribuições. Se houver, então podemos rejeitar a hipótese nula e considerar a existência de alguma relação entre as variáveis. Ele depende apenas dos graus de liberdade como parâmetro externo para o cálculo.

O número de Graus de Liberdade (GL) em uma distribuição é calculado multiplicando o número de linhas - 1 pelo número de colunas -1. Assim, uma tabela quádrupla possui 1 grau de liberdade, pois tem 2 linhas e 2 colunas. Subtraindo um linha e uma coluna temos 1 x 1 = 1. Já uma tabela com 4 linhas e 5 colunas apresenta 12 graus de liberdade ((4-1) x (5-1) = 12).

3.2.1. APLICAÇÃO DO χ^2 PARA ANÁLISE DE UMA ÚNICA VARIÁVEL

Quando aplicamos o χ^2 para verificar se a distribuição de casos em uma única variável segue a normalidade, dizemos que estamos aplicando o χ^2 para testar as diferenças entre valores observados e valores teóricos. Digamos que se queira verificar como se deu a distribuição das mulhe-

res eleitas em 2012 para as prefeituras pelos principais partidos brasileiros. Nosso objetivo é verificar se as eleitas distribuíram-se de maneira igualitária entre todos os partidos ou se houve concentração de prefeitas eleitas em um ou alguns partidos. Para isso aplicamos um testes de χ^2 para comparação com uma distribuição teórica. A fórmula é a que segue:

$$\chi^2 = \sum \frac{(FO - FE)^2}{FE}$$

Onde:

Fo : Frequência observada

Fe : Frequência esperada

Como estamos verificando a distribuição de uma única variável, a frequência esperada nesse caso é a diferença entre a frequência observada e a média dos valores. A média dos valores apresentados na tabela abaixo é de 56,3 eleitas por partido. Então, a Fe para o PMDB é : 122 - 56,3 = +65,7 prefeitas. Ao passo que o partido que elegeu o menor número de prefeitas, PDT, apresentará a diferença entre Fo e Fe negativa : 24 - 56,3 = -32,3. Para encontrar o coeficiente da distribuição é preciso somar o resultado dos quadrados das diferenças de todas as categorias dividido pelo valor esperado. Todas as etapas do cálculo são descritas a seguir:

TABELA 3.2. NÚMERO DE PREFEITAS ELEITAS PELOS GRANDES PARTIDOS EM 2012

PARTIDO	NÚMERO DE ELEITAS				
	Fo	Fe	Fo - Fe	(Fo - Fe)²	(Fo - Fe)²/Fe
PMDB	122	56,3	65,7	4316,49	80,38
PSDB	93	56,3	36,7	1346,89	25,08
PT	70	56,3	13,7	187,69	3,49
PSD	57	56,3	0,7	0,49	0,00
PSB	51	56,3	-5,3	28,09	0,52
PP	47	56,3	-9,3	86,49	1,61
PR	37	56,3	-19,3	372,49	6,93
PTB	34	56,3	-22,3	497,29	9,26
DEM	28	56,3	-28,3	800,89	14,91
PDT	24	56,3	-32,3	1043,29	19,42
TOTAL	563				**161,64**

Assim, descobrimos que o coeficiente χ^2 para a distribuição das eleitas por partidos para as prefeituras em 2012 é de 161,64. Agora, precisamos saber se esse coeficiente é estatisticamente significativo ou não. Pelas proporções das diferenças e pela magnitude do coeficiente pode-se esperar que sim, mas para termos certeza precisamos comparar o coeficiente com o limite crítico na tabela de valores padronizados para χ^2 que encontra-se no Anexo II. Nessa tabela aparecem os limites críticos levando-se em conta os graus de liberdade da tabela e o intervalo de confiança. No caso do intervalo de confiança, adotaremos o mais usado internacionalmente, 95%, que significa 0,050 na tabela de valores padronizados. Seus valores estão na sexta coluna (0,050) do Anexo II. O segundo fator a se considerar é o número de graus de liberdade da distribuição. Considerando que se trata de uma única variável e que temos 10 partidos políticos na tabela, temos que GL = 10 -1 = 9 graus de liberdade. Buscando o valor do limite crítico na tabela do Anexo II para IC de 0,050 e 9 GL encontramos 16,919. Isso significa que qualquer coeficiente acima desse limite deve ser considerado significativo estatisticamente, com a consequente rejeição da hipótese nula. Como o nosso coeficiente foi de 161,64, muito acima do limite crítico, podemos dizer que as mulheres eleitas para as prefeituras em 2012 não se distribuíram proporcionalmente entre os partidos.

Para identificarmos quais partidos elegeram proporcionalmente mais mulheres basta olhar para os resíduos. A coluna de resíduos, que é a subtração do valor observado pelo esperado, mostra onde estão as maiores concentrações de casos. A concentração positiva, ou seja, mais do que o esperado, está em PMDB, PSDB e PT, enquanto a concentração negativa, indicando menos casos do que o esperado, fica em PDT, DEM, PTB e PR. Esses foram os partidos pelos quais menos mulheres foram eleitas.

É possível que o pesquisador queira fazer comparações entre duas variáveis e não apenas verificar a distribuição de casos em uma única. Por exemplo, para além de saber como foi a distribuição das mulheres eleitas por partidos, pode ser que o cientista político queira saber se a distribuição de eleitos e eleitas por partidos em 2012 apresentou independência ou se a variável sexo está relacionada com a variável partido no que diz respeito ao número de prefeitos eleitos pelos grandes partidos em 2012. Nesse caso, nós aplicamos o teste para comparações entre variáveis independentes.

3.2.2 APLICAÇÃO DO χ^2 PARA COMPARAR DISTRIBUIÇÕES INDEPENDENTES

Esse teste é usado para comparar se diferentes distribuições observadas em dois grupos independentes são estatisticamente significativos. A hipótese nula é a de que não existem diferenças significativas entre os dois grupos ou que as diferenças observadas são frutos do acaso ou ainda que as duas amostras procedem da mesma população. No nosso exemplo, significaria dizer que não há diferença na proporção de homens e mulheres eleitas por partidos em 2012.

O que queremos aqui é rejeitar a hipótese de independência, ou seja, saber se as duas variáveis categóricas estão ou não relacionadas. A hipótese nula afirma que elas são independentes, quer dizer, não apresentam nenhuma relação entre si. Antes de aplicar o teste é importante destacar que o coeficiente só permite aceitar ou rejeitar a hipótese nula e que no caso de rejeitá-la, não é possível saber em que medida as duas variáveis estão relacionadas.

A maneira mais simples de analisar a relação existente entre duas variáveis categóricas nominais através do χ^2 é partindo de uma tabela de contingência. Nela, a análise é realizada a partir da verificação da distribuição das ocorrências para identificar o padrão de comportamento. Se a distribuição não for aleatória, indicará uma relação entre as duas variáveis. Portanto, o χ^2 é um teste baseado no cálculo do total de desvios entre as ocorrências observadas e esperadas, segundo os graus de liberdade da tabela de contingência. A partir disso ele examina se um padrão da distribuição apresenta probabilidade suficiente de ocorrência para considerá-la não-aleatória. A fórmula principal é a mesma que já foi apresentada:

$$x^2 = \sum \frac{(FO-FE)^2}{FE}$$

A diferença é que nesse caso, como estamos trabalhando com mais de uma variável, a Frequência esperada não equivale à média delas. Ela é calculada a partir da seguinte fórmula:

$$Fe = \frac{(Mc\ x\ Ml)}{N}$$

Onde:
Mc : Marginal da coluna
Ml : Marginal da linha
N : Número total de casos.

Então, calculamos as Fe para cada casa, depois calculamos o χ^2 para cada uma das categorias e somamos os valores. O χ^2 da tabela de contingência será a soma dos coeficientes parciais. Aplicando o cálculo ao exemplo da comparação entre partido político e sexo do prefeito em 2012 temos o seguinte:

TABELA 3.3. FREQUÊNCIAS PARA HOMENS E MULHERES ELEITOS POR PARTIDO EM 2012

PARTIDO	ELEITO			ELEITA			TOTAL
	Fo	Fe	(Fo-Fe)²/ Fe	Fo	Fe	(Fo-Fe)²/ Fe	
PMDB	902	906,22	0,020	122	117,78	0,152	1024
PSDB	601	614,18	0,283	93	79,82	2,176	694
PT	558	555,77	0,009	70	72,23	0,069	628
PSD	438	438,07	0,000	57	56,93	0,000	495
PP	417	410,63	0,099	47	53,37	0,760	464
PSB	389	389,39	0,000	51	50,61	0,003	440
PDT	284	272,58	0,479	24	35,42	3,685	308
PTB	260	260,19	0,000	34	33,81	0,001	294
DEM	248	244,26	0,057	28	31,74	0,442	276
PR	235	240,72	0,136	37	31,28	1,044	272
TOTAL	4332		$\sum eleitos = 1{,}083$	563		$\sum eleitas = 8{,}331$	4895

Para conhecermos o coeficiente final do teste para as variáveis independentes basta somar o coeficiente dos homens (1,083) ao coeficiente das mulheres (8,331), então, $\chi^2 = 1{,}083 + 8{,}331 = 9{,}414$. Ainda estamos trabalhando com 9 GL, pois temos : (2 colunas - 1) x (10 linhas - 1) = 9 graus de liberdade. Mantendo o intervalo de confiança de 95% temos que o limite crítico (anexo II) é de 16,919, portanto, nosso coeficiente ficou abaixo do limite crítico, indicando que as diferenças entre homens e mulheres eleitos por partido para prefeito do Brasil em 2012 não são estatisticamente significativas e nós não devemos rejeitar a hipótese nula nesse caso.

O leitor menos atento pode estar se perguntando como é possível que a distribuição das mulheres eleitas apresente um coeficiente tão alto e a distribuição de homens e mulheres fique abaixo do limite crítico? Cuidado, não podemos confundir a distribuição de casos em uma variável com a comparação entre duas variáveis independentes. No primeiro caso estamos testando apenas como as eleitas se distribuem entre os partidos e elas concentram-se em alguns deles. No segundo, estamos testando como homens e mulheres se distribuem entre os partidos. E a distribuição é próxima da independência, ou seja, partidos que elegem mais homens tendem a eleger

mais mulheres, também. Enquanto que partidos que elegem menos homens também elegem menos mulheres. Por isso o coeficiente final fica abaixo do limite crítico no segundo caso. Em resumo, não podemos confundir, muito menos comparar, coeficiente χ^2 para a distribuição de uma variável com o coeficiente para variáveis independentes, ainda que uma delas esteja presente nos dois testes.

Para quando não se dispõe da tabela de valores padronizados de χ^2, existe uma forma alternativa para encontrar o nível de significância do coeficiente para quando se está testando a partir de uma tabela quádrupla. É através do cálculo do valor de alfa (α) para χ^2, como apresentado no próximo tópico.

3.2.3 CÁLCULO DE α PARA ESTABELECER LIMITE CRÍTICO DE CONFIANÇA AO χ^2

Uma vez encontrado o coeficiente χ^2 também é possível obter o valor de $\boldsymbol{\alpha}$ (em testes de média e de regressão, também chamado de *p-value*) para a significância do coeficiente quando não há disponibilidade dos dados originais e se está trabalhando com tabela quádrupla de resultados, ou seja, são testadas duas variáveis dicotômicas. O objetivo aqui é o mesmo: testar o grau de segurança para extrapolar os resultados do teste de uma amostra para a população. Para tanto, usa-se a fórmula:

$$\boldsymbol{\alpha\chi^2} = \frac{((a.d - b.c).0{,}5)^2}{(a+b)x(d+c)x(b+d)x(a+c)}$$

Aplicando ao exemplo anterior, para o coeficiente das respostas a um questionário de opinião pública sobre a relação entre "ter candidato a prefeito antes do início da campanha" e "ter preferência partidária", o resultado é:

$$\boldsymbol{\alpha\chi^2} = \frac{((114\ x\ 56) - (110\ x\ 146))\ x\ 0{,}5)^2}{(114+110)\ x\ (56+146)\ x\ (110+56)\ x\ (114+146)} = 0{,}011$$

O valor de $\boldsymbol{\alpha\chi^2}$ para a distribuição dos casos nas variáveis do exemplo é de 0,011. Este valor será significativo se ficar abaixo de 0,050 nos casos de Intervalo de Confiança de 95%. No nosso caso, sendo consistente com os resultados anteriores, o alfa indica significância para o fato de ter candidato antes do início da campanha e ter preferência partidária.

Se o coeficiente ficar acima do limite crítico na tabela de valores padronizados do χ^2, o valor de alfa para sempre estará abaixo de 0,050. Portanto, essa última fórmula oferece apenas uma informação adicional ao teste, que pode ser útil quando se compara coeficientes de diferentes variáveis explicativas em relação a uma variável dependente para verificar qual delas apresenta maiores coeficientes e menores valores de alfa. Quanto menor o valor de alfa, ou seja, quanto mais próximo de zero, mais confiável é a estatística que indica associação entre as duas variáveis dicotômicas.

Existe outra aplicação para o χ^2 que não será demonstrada aqui, apenas citada, por ser pouco usada na ciência política. Trata-se o teste χ^2 para comparar distribuições de dados relacionados ou pareados. Essa finalidade é muito comum na área de saúde, mas pouco aplicada pelos politólogos. Ocorre quando o objetivo do teste é medir a diferença gerada por um fator externo em dois momentos distintos do tempo. Para isso são feitas medições sobre os mesmos indivíduos que foram previamente igualados. Por exemplo, a um grupo de 190 estudantes aplicado-se um exame tradicional e outro teste com formato inovador. No tradicional foram aprovados 120 alunos e no teste inovador foram aprovados 130. Sabe-se que 110 alunos foram aprovados nos dois. Queremos saber se os dois tipos de provas oferecem os mesmos resultados. A comparação de dados pareados leva em conta apenas as frequências cujas categorias não coincidem, abrindo mão das respostas iguais antes e depois. Nesse caso, interessam apenas os alunos que foram aprovados em um teste e não aprovados em outro. Segue a mesma lógica do teste t para amostras pareadas, porém, aqui as duas variáveis podem ser categóricas. Ex.: teste A x teste B e candidatos aprovados x candidatos reprovados. A aplicação de dados pareados é bastante comum na área da saúde, pois ele indica se o uso de determinado medicamento, por exemplo, tem os efeitos esperados ou se as diferenças após o seu uso são oscilações do acaso. Isso pode se dar em experimento para um novo medicamento para controle da pressão arterial. Então, mede-se a pressão arterial de um grupo de indivíduos hipertenso. Depois, metade deles (escolhida aleatoriamente) recebe o medicamento. A outra metade receberá placebo. Depois, mede-se novamente as pressões arteriais. Se o remédio fez efeito, haverá diferença no coeficiente χ^2 do grupo que tomou o medicamento, porém, no que tomou o placebo, não.

Os testes de χ^2 nos mostram os coeficientes de associação entre duas variáveis categóricas, permitindo-nos, em função dos graus de liberdade, rejeitar ou não a hipótese nula de independência entre as variáveis. Se quisermos aprofundar a análise e identificar a magnitude da associação, devemos fazer testes complementares, que permitem a produção de coeficientes específicos para força da associação. Um dos mais usados é o Cramer´s V, que será apresentado a seguir.

3.2.4. COEFICIENTE CRAMER´S V PARA MEDIR A FORÇA DA ASSOCIAÇÃO EM TESTE DE χ^2

Uma vez identificado o valor do coeficiente de χ^2, um erro muito comum é tirar conclusões sobre a magnitude da relação entre as duas variáveis apenas a partir desse coeficiente ou do seu nível de significância. Quando o coeficiente é alto, podemos dizer que as variáveis testadas estão associadas, ou seja, não são independentes. Se considerarmos ainda os graus de liberdade do teste, podemos identificar o limite crítico do Intervalo de Confiança e se ficarmos abaixo desse limite podemos dizer que a associação é forte o suficiente para ser extrapolada a toda a população - caso estejamos trabalhando com uma amostra. Isso porque a significância do teste depende no número de casos (graus de liberdade). Quanto maior for a amostra ou população testada, maiores as chances do resultado ser estatisticamente significativo. No entanto, nenhum dos coeficientes tratados até aqui para testes entre duas variáveis categóricas indica a magnitude ou força da associação. Para isso, existem coeficientes específicos que medem a força do efeito de uma variável sobre a outra, quando estamos usando o teste de χ^2, que são o coeficiente Phi e o coeficiente Cramer´s V. Nesse curso não trataremos do coeficiente Phi porque dedicamos um capítulo específico para outro teste indicado para a mesma situação que Phi, o Q-yule. A seguir apresentamos como calcular o coeficiente de magnitude do efeito Cramer´s V para testes de χ^2. A indicação é calcular o Cramer´s V apenas quando o coeficiente χ^2 for estatisticamente significativo, caso contrário, a magnitude do efeito será muito baixa ou nula.

Para identificar a magnitude do efeito (*effect size*) em testes de χ^2 em que se rejeita a hipótese nula usa-se o coeficiente Phi para os casos de tabelas quádruplas (2x2) ou o coeficiente Cramer´s V para tabelas maiores (LxC). A leitura dos resultados do coeficiente V é equivalente à de um coeficiente de correlação de Pearson. Ele indica qual a força da associação direta entre o conjunto das categorias das duas variáveis testadas. A fórmula é a seguinte:

$$v = \sqrt{\frac{\chi^2}{N.(k-1)}}$$

Onde:

χ^2 = coeficiente qui-quadrado.

N = número de casos

K = número de categorias de uma das variáveis testadas. Utiliza-se sempre o menor número de categorias, independente de estar nas linhas ou nas colunas.

A leitura do resultado do Cramer´s V é similar à de um coeficiente de correlação de Pearson. Sendo assim, o quadrado do seu valor (V^2) nos indica qual a proporção da variância da relação que é explicada pelo χ^2. Por exemplo, um V = 0,12, se elevado ao quadrado e multiplicado por 100 nos indicará qual o percentual de variância explicada, no caso, apenas 1,44% de variância explicada pelo χ^2, o que é um percentual bastante baixo em termos gerais. Quando a tabela testada é quádrupla (2x2) a fórmula do teste Cramer´s V iguala-se à do Phi, pois nesse caso o número de categorias menos um sempre será um. Assim, a fórmula é reduzida a raiz quadrada de χ^2 dividido pelo número de casos.

Para exemplificar o uso do coeficiente Cramer´s V vamos fazer o teste de χ^2 para a associação entre sexo do vereador eleito em 2012 nas eleições municipais brasileiras (homem ou mulher) e região do País (norte, nordeste, centro-oeste, sudeste e sul). A hipótese nula defende que não há associação entre as duas variáveis e que, portanto, homens e mulheres distribuem-se igualmente entre os eleitos nas cinco regiões do País. Nunca é demais lembrar que estamos testando a associação entre duas variáveis categóricas nominais.

TABELA 3.4. DISTRIBUIÇÃO DOS VEREADORES ELEITOS POR SEXO E REGIÃO DO PAÍS EM 2012

Região	Homem	Mulher	Total
norte	4042	702	4744
nordeste	15943	2931	18874
centro-oeste	4110	588	4698
sudeste	15446	1903	17349
sul	10082	1519	11601
Total	49623	7643	57266

Em 2012, segundo dados oficiais do Tribunal Superior Eleitoral (TSE) foram eleitos 57.266 vereadores para os legislativos municipais brasileiros. Desses, 49.623 eram homens e 7.643 mulheres. O teste χ^2 indica se podemos ou não rejeitar a hipótese nula de independência entre as variáveis. O resultado é de χ^2 = 174,63 e α = 0,000, portanto, um resultado altamente significativo, permitindo a rejeição da hipótese nula de independência entre as variáveis. Agora, para identificar a magnitude do efeito, calcularemos o Cramer´s V. Como a variável com menor número de categorias é sexo, com 2 categorias, desconsideraremos o fator (k-1) na fórmula, pois multiplicaríamos o número de casos por um, o que não faz o menor sentido.

$$v = \sqrt{\frac{\chi^2}{N.(k-1)}} = \sqrt{\frac{174{,}63}{57.266}} = 0{,}055$$

Ou seja, a magnitude do efeito da região sobre a proporção de homens ou mulheres eleitos é de apenas 5,5%, embora o χ^2 seja alto e significativo, percebe-se que o efeito não é tão forte como se poderia pensar inicialmente. Se elevarmos o Cramer´s V ao quadrado teremos a proporção de variação que é explicada pela associação. No caso, v^2 = 0,003. Se multiplicarmos o valor por 100 teremos que apenas 0,3% de variação é explicada.

O próximo coeficiente faz a transição para o coeficiente do tópico seguinte (Q_{xy}). Trata-se do cálculo para encontrar as diferenças entre valores de frequências observadas e frequências esperadas. É o chamado coeficiente Delta das diferenças.

3.3. COEFICIENTE DELTA (Δ) PARA DIFERENÇAS ENTRE F_O E F_E

Um coeficiente **Δ** é uma medida que mostra a existência ou não de valores "sobrando" em determinados pares de categorias. Portanto, ele só pode ser aplicado em comparações entre duas variáveis, chamadas aqui de X e Y. Para deduzir se existe ou não alguma relação entre as variáveis X e Y comparam-se as frequências observadas a uma tabela com uma distribuição teórica onde as distribuições são independentes – chamada de frequência esperada. A hipótese estatística inicial é de que não há relação entre X e Y, logo, se a frequência esperada for a mesma ou estiver muito próxima da frequência observada, não podemos rejeitar a hipótese nula.

A existência de diferenças significativas entre as frequências observadas e as esperadas nos permite rejeitar a hipótese nula, que defende a independência entre as variáveis e passamos a considerar que X e Y estão associadas de alguma forma. O passo seguinte é medir a força da relação ou dependência entre as duas variáveis. O termo técnico para descrever a inexistência de relação entre duas variáveis é “independência estatística”. Portanto, temos independência estatística quando X e Y são estatisticamente independentes, o que ocorre quando as probabilidades das células esperadas igualam os produtos das probabilidades marginais relevantes. Nesse caso, também é indicado que as probabilidades de ocorrência de uma categoria em uma variável são as mesmas que as demais, independente da categoria da outra variável com a qual está ligada, ou seja, a frequência de casos na categoria da segunda variável não faz diferença

para a primeira. Em outras palavras, a relação não tem efeito estatístico. O que é diferente de dizer que não tem efeito algum.

O coeficiente **Δ** serve para indicar a existência de diferenças entre pares de frequências de uma tabela de contingência. Normalmente é utilizado em tabelas quádruplas, para variáveis dicotômicas. No entanto, nada impede que também seja utilizado nos cruzamentos entre variáveis com mais de duas categorias. O objetivo aqui é identificar se determinada característica conjunta de X e Y ocorre mais ou menos vezes do que seria o esperado. Se isso ocorrer, não podemos considerar as variáveis independentes para esse par de categorias. Para tanto, são usadas as probabilidades observada e esperada nas comparações. Então, **Δ** pode ser representado pela seguinte fórmula:

$$\Delta = \text{Prob. Observada} - \text{Prob. Esperada}$$

Onde a probabilidade observada é o valor da frequência de casos para determinado par de categorias e a probabilidade esperada é dada pela multiplicação das marginais da tabela, dividido pelo número total de casos. Aqui, a probabilidade esperada é encontrada da mesma forma que a Frequência esperada do teste anterior.

Digamos que o pesquisador queira encontrar o **Δ** para a probabilidade de homens que foram eleitos prefeitos por partidos de direita em 2012. A hipótese é que os partidos de direita tenham eleito mais homens do que mulheres, proporcionalmente. O primeiro passo é montar uma tabela de contingência entre ideologia do partido e sexo do eleito em 2012, como a que segue:

TABELA 3.5. DISTRIBUIÇÃO DAS PROPORÇÕES DE ELEITOS POR IDEOLOGIA E SEXO EM 2012

Ideologia	Sexo		Total
	Homem	Mulher	
Esquerda	1231 (0,251)	145 (0,030)	1376 (0,281)
Centro	1503 (0,307)	215 (0,044)	1718 (0,351)
Direita	1598 (0,326)	203 (0,041)	1801 (0,368)
Total	4332 (0,885)	563 (0,115)	4895 (1,000)

Olhando as marginais da tabela, é possível perceber que de maneira geral os homens representam quase nove em cada dez eleitos (0,88), enquanto as mulheres ficam em uma proporção de apenas 0,12 do total de eleitos em 2012. Já as marginais das linhas mostram que partidos de direita foram responsáveis pela maior proporção de eleitos (0,37), seguidos de partidos de cen-

tro, com 0,35, e de esquerda, com 0,28. Se olharmos para o corpo da tabela, podemos encontrar os valores das participações proporcionais de cada par de categorias. No caso que interessa aqui, homens de partidos de direita apresentam proporção de 0,33 (1598 / 4895 = 0,33). A questão é saber se essa proporção equivale a uma distribuição independente para o par de categorias ou se, ao contrário, ela indica a existência de algum grau de associação. O primeiro passo é encontrar a probabilidade esperada para prefeitos de partidos de direita. Como já estamos trabalhando com as proporções, basta multiplicar a marginal da linha "partidos de direita" pela marginal da coluna "homem" : 0,885 x 0,368 = 0,325. Então, a proporção esperada de homens eleitos por partido de direita é de 0,325. Para conhecer o **Δ** aplica-se a fórmula:

$$\Delta pd = \text{Prob. Observada} - \text{Prob. Esperada} = 0{,}326 - 0{,}325 = 0{,}001$$

O resultado do Δpd (delta para prefeitos de direita) é de 0,01, portanto, muito próximo de zero, ficando praticamente idêntica ao que seria a distribuição independente. Não podemos dizer que partidos de direita elegeram mais homens do que seria esperado em uma distribuição independente. A tabela abaixo mostra os valores Δ para todas as categorias do exemplo.

TABELA 3.6. VALORES DE Δ PARA TODOS OS PARES DE CATEGORIAS

Ideologia	Homem	Mulher
Esquerda	0,003	-0,003
Centro	-0,004	0,004
Direita	0,001	-0,001

A tabela acima mostra que as variações das proporções de homens e mulheres eleitos por ideologia partidária são muito pequenas, girando em torno do valor teórico. Todas ficam abaixo de 1% de diferença. Portanto, não podemos dizer que existam diferenças significativas na eleição de homens ou mulheres entre os partidos agregados nas três posições ideológicas acima. Todos elegeram praticamente as mesmas proporções de mulheres, portanto, não podemos identificar nenhuma associação ou tendência em determinado partido eleger mais homens do que mulheres proporcionalmente aos demais. O sinal do Δ indica a direção. Se as diferenças proporcionais fossem maiores, poderíamos dizer que o centro elegeu proporcionalmente menos homens que mulheres, enquanto a esquerda e a direita elegeram mais homens que mulheres. Na verdade esse resultado já tinha sido apontado no teste anterior para as duas variáveis, quando o coefici-

ente χ^2 ficou abaixo do limite crítico, já indicou a inexistência de dependência entre as categorias das variáveis "ideologia do partido" e "sexo do eleito".

O fato de não encontrarmos dependência entre as variáveis aqui não é um problema tão grande, pois um Δ diferente de zero também não significaria muita coisa. Isso porque esse é um coeficiente muito rústico. Seu forte não é a precisão. Ele apresenta dois problemas para a interpretação estatística: 1) é sensível ao tamanho da amostra. Se dobrássemos o N no exemplo anterior o valor de cada Delta também seria o dobro. Isso impossibilita a comparação de coeficientes Delta em amostras com N diferentes, e; 2) O coeficiente Delta não possui um limite superior. Do lado inferior o limite é zero, mas não é possível saber até quanto se pode chegar ao outro limite, tanto com sinal positivo, quanto negativo. Isso impossibilita estabelecer magnitudes comparativas quando não se tem limite superior.

No próximo tópico estudaremos coeficientes mais precisos e que permitem análises comparativas, além de serem específicas para cada tipo de variável. Ou seja, vamos melhorar a qualidade das "ferramentas estatísticas" para análise de dados categóricos a partir de fontes secundárias. Por outro lado, os cálculos começam a ficar um pouco mais trabalhosos, como veremos a seguir.

3.4 EXERCÍCIOS

Responda as questões a seguir:

3.4.a. A partir dos resultados apresentados abaixo de uma pesquisa de opinião pública conduzida nos EUA pela CNN, construa uma tabela quádrupla apenas com os valores válidos para (Branco/Não-branco e Opinião Favorável/Opinião desfavorável) para Hillary Clinton e calcule o risco relativo a partir dos percentuais de um eleitor branco apresentar opinião favorável para H. Clinton. Interprete o resultado.

CNN/ORC International Poll -- September 23 to 25, 2011.					
Question 4D					
We´d like to get your overall opinion of some people in the news. As I read each name, please say if you have a favorable or unfavorable opinion of thes people -- or if you have never heard of them.					
D. Hillary Clinton					
	Total	*Men*	*Women*	*White*	*Non-white*
Favorable Opinion	*69%*	*67%*	*71%*	*64%*	*82%*
Unfavorable Opinion	*26%*	*28%*	*24%*	*31%*	*15%*
Heard of, no opinion	*4%*	*5%*	*4%*	*4%*	*2%*
Never heard of	*	*	*1%*	*	*1%*
No opinion	*	*	*	*	*
Sampling error	*+/-3.0*	*+/-4.5*	*+/-4.5*	*+/-3.5*	*+/-6.5*

3.4.b. A partir do quadro 1, da introdução, calcule o valor de $\chi 2$ para as distribuições do número de eleitores nos municípios em cada um dos períodos analisados. Calcule um $\chi 2$ para cada ano individualmente. Não é necessário calcular a variação conjunta, pois já sabemos que o número de eleitores aumentou no período. O que queremos saber é como era a heterogeneidade do número de eleitores entre os municípios em cada período. Depois, comparando os coeficientes é possível identificar se essa heterogeneidade cresceu ou manteve-se estável ao longo do tempo. Para isso é necessário interpretar os resultados.

3.4.c. Nos dois quadros abaixo são apresentadas as seguintes informações em variáveis binárias para as eleições legislativas brasileiras de 2010: número de candidatos eleitos (El) e de derrotados (N-el) e sexo do candidato, Homem (H) ou Mulher (M). A partir dessas informações faça o seguinte:

i) calcule o $\chi 2$ para cada um dos cruzamentos e interprete os resultados. Já sabemos que as diferenças são significativas, dada a distribuição dos totais. Interprete comparativamente as disputas entre homens e mulheres nas esferas estaduais e federal;

ii) encontre o valor do limite crítico na tabela do Anexo II para cada um dos cruzamentos;

iii) calcule o α para o limite crítico de cada um dos casos (federal e estaduais).

iv) calcule o coeficiente de magnitude do efeito Cramer´s V para as duas tabelas.

Câmara de Deputados			
	El	N-el	Total
H	468	3486	3954
M	45	888	933
Total	513	4374	4887

Assembleias Estaduais			
	El	N-el	Total
H	921	9043	9964
M	138	2502	2640
Total	1059	11545	12604

3.4.d. A partir da tabela de distribuição do número de candidaturas de dois partidos brasileiros (PT e PSDB) às prefeituras municipais entre 2004 e 2012, calcule o coeficiente **Δ** para os pares de casos e interprete os coeficientes das três eleições para os dois partidos. Os números representam as candidaturas a prefeito que cada um dos partidos participou, seja com candidato a prefeito ou indicando o candidato a vice-prefeito. E, no caso das candidaturas a prefeito, estão somadas as campanhas em que os partidos não fizeram coligações com as campanhas coligadas a outros partidos. Com isso temos a participação total de cada um dos partidos nas disputas para prefeituras em três eleições distintas. Não deixe de considerar as limitações do **Δ** na hora de interpretar os resultados. Apresente possíveis explicações para as mudanças no coeficiente ao longo do tempo.

Ano	PSDB	PT	Total
2004	4379	3923	8302
2008	4870	4939	9809
2012	4246	4617	8863
Total	13495	13479	26974

4. PRINCIPAIS COEFICIENTES DE ASSOCIAÇÃO POR TIPO DE VARIÁVEIS

4.1. ASSOCIAÇÃO ENTRE VARIÁVEIS BINÁRIAS (Q-YULE)

Como já apresentado anteriormente, uma variável binária ou dicotômica é aquela que possui apenas duas possibilidades de categorias, que representam a presença ou a ausência de determinada característica. Normalmente a representação numérica das categorias é feita por 0 = ausência da característica e 1 = presença da característica. Pode ser aplicado à variável Sexo, quando se quer testar determinada característica das mulheres, então: 1 = mulher e 0 = homem. Ou quando se quer dividir o total de eleitores em dois grupos, sendo: 1 = eleitores que votaram no candidato K na última eleição ou 0 = eleitores que não votaram no candidato K na última eleição. Até aqui identificamos duas variáveis dicotômicas: sexo do eleitor e voto em determinado candidato. Digamos que nosso objetivo seja saber se o candidato K teve mais votos entre as mulheres quando comparado aos demais concorrentes. Nesse caso, precisaríamos cruzar as duas informações e teríamos quatro condições possíveis: a)é mulher e não votou em K; b) é mulher e votou em K; c) não é mulher e não votou em K e d) não é mulher e votou em K. Como existem quatro possibilidades em um cruzamento de duas variáveis dicotômicas, elas sempre são organizadas em tabelas quádruplas (2x2). Um teste estatístico para medir a existência ou não de relação entre duas variáveis dicotômicas e no caso de existir relação, a força e a direção da mesma, é o Q_{xy}, como veremos a seguir.

O teste de independência Q_{xy} serve para identificar se:

i) duas variáveis dicotômicas estão relacionadas entre si,

ii) de quanto é a intensidade da relação e,

iii) se os resultados podem ser usados em generalizações.

Como é aplicado em tabelas quádruplas (com duas variáveis dicotômicas) e qualquer variável pode ser dicotomizada, trata-se de um coeficiente bastante útil e que pode ser obtido com a aplicação de fórmulas simples, dispensando o uso de programas de computador. Uma variável pode ser dicotomizada quando se decide separar em dois grupos as categorias internas dela. Por exemplo, pode-se ter uma variável categórica na forma de Escala de Likert para avaliação de governo: Muito Boa, Boa, Regular, Ruim e Péssima. A dicotomização se dá quando o pesquisador divide os resultados entre Avaliação Positiva e as demais. Então, teríamos: 1 = (Muito Boa + Boa) e 0 = (Regular + Ruim + Péssimo). A dicotomização também pode ser a partir de uma variável escalar discreta, como idade em anos completos. Nesse caso, a opção pode ser usar o

valor da mediana para dividir em dois grupos de igual tamanho. Então, se quiséssemos testar o efeito entre os mais velhos, teríamos: 0 = grupo dos mais novos, até a mediana e 1 = grupo dos mais velhos, a partir da mediana.

Também é possível dicotomizar distribuições de frequências a partir de dados secundários, como, por exemplo, usando informações de uma tabela de distribuição das intenções de voto a seis candidatos em uma eleição qualquer. Nesse caso, separam-se as frequências de respondentes que dizem votar em um candidato (representado pela letra K) e essa será a característica analisada (1) de todos os demais, cuja soma será a frequência para o código zero. Ao final teremos apenas dois resultados possíveis: vota no candidato K ou não vota no candidato K.

O importante aqui é entender que qualquer variável pode ser dicotomizada através de processos defensáveis estatisticamente. Quando se tem duas variáveis dicotômicas, tais como votar ou não no candidato A e idade dos respondentes (jovem e não-jovem) é possível aplicar os cálculos do coeficiente de Q_{xy} para identificar se as duas variáveis são independentes ou não. Se não forem, significa que há alguma associação entre as características medidas. Então, o coeficiente também nos fornece a informação sobre o grau de associação entre elas, a direção, se é a mesma ou se está em direções opostas e, por fim, se os resultados dos testes em uma amostra são consistentes o suficiente para permitir a extrapolação para toda a população.

O mais comum quando se agregam variáveis escalares, proporcionais, ordinais ou de intervalo é considerar X e Y o conjunto de valores Altos ou a Presença da característica a ser medida e não-X e não-Y os valores Baixos ou a Ausência da característica a ser medida. Essa convenção é importante em função do sinal do coeficiente de associação no resultado do teste. Uma inversão das posições significaria inverter um sinal de relação na mesma direção (positivo) por relação em direções opostas (negativo). As tabelas quádruplas são compostas por quatro células de frequências, quatro células com frequências marginais e uma célula de total, chamada de N. Cada uma das células de frequências recebe uma letra como nome, sendo, A, B, C e D, como no quadro a seguir:

QUADRO 4.1. DISTRIBUIÇÃO QUÁDRUPLA PARA CÁLCULO DO Q-YULE

	Não-Y	**Y**	**Total**
X	**A**	**B**	Marginal X
Não-X	**C**	**D**	Marginal Não-X
Total	Marginal Não-Y	Marginal Y	Total de Casos (N)

Devem fazer parte das células de frequências apenas os casos válidos, o que sempre precisa ser explicitado aos leitores dos resultados. As variáveis analisadas são chamadas de X e Y. As categorias de grupamento dicotômico das variáveis são chamadas, por consequência, de X e não-X; Y e não-Y. Em um exemplo de pesquisa sobre intenção de voto relacionada a sexo dos eleitores para saber se determinado candidato (K) recebe votos de mulheres, os respondentes que dizem votar no candidato K compõem as casas da linha X e aqueles que dizem votar em qualquer outro candidato fazem parte da linha Não-X. Já as eleitoras são Y e os eleitores são não-Y. As somas dos casos nas linhas (horizontais) e nas colunas (verticais) formam o que se chama de Marginais. A somatória das marginais leva ao número total de casos analisados, representado pela letra N. Assim, teremos ao final uma tabela quádrupla que relaciona eleitores e não eleitores do candidato K com o fato de ser ou não ser mulher. O resultado apresentará se o candidato K tem uma concentração maior de votos entre as mulheres ou não.

Como todos os demais testes estatísticos probabilísticos, o Q_{xy} parte da hipótese inicial (H_0) de independência entre as variáveis. O que queremos identificar é se existe uma chance estatística forte o suficiente para garantir baixas possibilidades de erro caso a hipótese nula (H_0) seja rejeitada para podermos afirmar que há alguma relação entre as duas variáveis. No caso do exemplo, afirmar que o candidato K tem mais votos entre mulheres do que entre homens seria uma hipótese inicial de trabalho. Partiríamos do princípio de que não há diferença de sexo entre os eleitores do candidato K, ou seja, as duas variáveis são independentes, como nos diz a H_0. Nosso objetivo é realizar os testes para verificar se temos condições suficientes para afirmar que há uma associação entre as duas variáveis - ser mulher e votar em K. Nesse caso, rejeitaríamos H_0 e assumiríamos que podemos dizer que há uma probabilidade alta de que as duas variáveis estejam associadas, quer dizer, assumimos H_1. No próximo tópico veremos como fazer isso para duas variáveis dicotômicas.

4.1.1. TESTE DE INDEPENDÊNCIA Q DE YULE (Q_{XY})

As análises de independência visam identificar se duas variáveis apresentam alguma associação estatística. Se não, diz-se que a associação é nula, ou seja, as variáveis são independentes. Se sim, a associação pode ter diferentes intensidades: fraca, média, forte. Aqui, o teste de independência visa identificar a inexistência de relação entre duas variáveis. Portanto, relembrando, a hipótese inicial é de independência. Se houver alguma relação ou associação entre as variáveis, então, nega-se a hipótese de independência e mede-se o grau de relação entre elas.

Nas tabelas quádruplas cada casa representa a frequência encontrada para um par de características (par Não-Y,X; par não-Y,Não-X; par Y/X; par Y/Não-X). Se as variáveis forem independentes, a proporção de casos em cada par em relação ao total será a mesma, portanto, impedindo qualquer afirmação de associação entre as variáveis. Já, se houver uma distorção razoável entre a frequência relativa de casos em um ou alguns pares em relação aos demais, podemos negar a independência e medir o grau de associação entre as categorias das variáveis. Então, o coeficiente Q_{xy} nos fornece duas informações importantes: i) sobre a magnitude da relação, medida pelo tamanho do coeficiente. Quanto mais próximo de ±1 mais forte será a associação, e ii) a respeito da direção da relação. Se o sinal do coeficiente for positivo, então as duas categorias estão associadas e variam na mesma direção. Se o sinal for negativo, existe associação, mas as variações são em direções opostas. O quadro a seguir representa os sinais predominantes nas associações Positivas e Negativas entre duas variáveis dicotômicas.

Positiva

	Não Y	Y
X	-	+
Não X	+	-

Negativa

	Não Y	Y
X	+	-
Não X	-	+

QUADRO 4.2. RELAÇÃO DOS SINAIS NAS TABELAS QUÁDRUPLAS

No quadro acima a associação positiva indica uma concentração de casos com a característica da variável X e com a característica da variável Y, indicando que as presenças das características em X e Y "caminham na mesma direção". Já na associação negativa, a presença da característica na variável Y apresenta maior concentração de frequências na casa da ausência da característica na variável X, nesse caso, elas "caminham em direções opostas". Atenção para a diferença no uso dos termos “tende a ser” e “a maioria é”. Nas análises probabilísticas deve-se fazer, sempre, a primeira afirmação ao invés da segunda.

O coeficiente Q_{xy} apresenta as características desejadas em um coeficiente de associação que pretenda medir a força e a direção da relação. Ele é insensível ao tamanho da amostra, portanto, seus resultados não oscilam em função do N da tabela quádrupla e ele apresenta limites superior e inferior pré-estabelecidos. Dessas duas características saem os seguintes postulados para o coeficiente de associação Q_{xy}:

a) O coeficiente deve ser igual a zero quando X e Y forem independentes, e;

b) O coeficiente deve ser de no máximo + 1,00 para associação positiva e – 1,00 para associação negativa.

A partir desses postulados, Davis (1976) organiza as possíveis distribuições de valores de Q_{xy} por grau de intensidade e forma adequada de interpretação, como segue no quadro abaixo:

QUADRO 4.3. INTERVALOS DE VALORES PARA COEFICIENTE QXY

Valor de Q_{xy}	Leitura adequada
+0,70 ou mais	Associação positiva muito forte
+0,50 a +0,69	Associação positiva forte
+0,30 a +0,49	Associação positiva moderada
+0,10 a +0,29	Associação positiva baixa
+0,01 a +0,09	Associação positiva desprezível
0,00	Nenhuma associação
-0,01 a -0,09	Associação negativa desprezível
-0,10 a -0,29	Associação negativa baixa
-0,30 a -0,49	Associação negativa moderada
-0,50 a -0,69	Associação negativa forte
-0,70 ou mais	Associação negativa muito forte

Fonte: Davis, 1976.

O estatístico inglês G. Udny Yule apresentou uma proposta de coeficiente de correlação no início do século XX, respeitando as regras acima para aplicação aos resultados de uma tabela quádrupla. A primeira publicação do coeficiente foi em 1911 e Udny Yule o batizou de Q_{xy} em homenagem ao estatístico pioneiro Quételet (1796-1874). Com o tempo, o coeficiente passou a ser chamado de Q de Yule. Sua fórmula é a seguinte:

$$Q_{xy} = \frac{(BxC) - (AxD)}{(BxC) + (AxD)}$$

Trata-se da divisão entre os produtos cruzados de uma tabela quádrupla. Vamos aplicar a fórmula a um exemplo a partir dos resultados das eleições para prefeito no Brasil em 2012. Considere as seguintes variáveis dicotômicas: i) posição ideológica do partido do prefeito eleito em 2012; e ii) nível de escolaridade formal do prefeito eleito. Vamos testar se existe independência entre as seguintes características: prefeitos eleitos por partidos de esquerda e escolaridade superior. Nossa hipótese de pesquisa é que políticos de esquerda tendem a concentrar o maior nú-

mero de políticos com escolaridade superior. Se encontrarmos independência entre as características, significa que os eleitos com escolaridade superior distribuem-se igualmente entre partidos de esquerda e os demais. Se encontrarmos um coeficiente positivo, significa que há uma associação entre estar em partido de esquerda e ser prefeito com escolaridade superior. Se, por outro lado, o coeficiente for negativo, a associação indicará que entre os prefeitos eleitos por partidos de esquerda há menos com escolaridade superior do que esperaríamos encontrar caso as variáveis fossem independentes. As duas variáveis são apresentadas na tabela quádrupla a seguir:

TABELA 4.1. DISTRIBUIÇÃO POR ESCOLARIDADE E IDEOLOGIA DO PARTIDO DO PREFEITO ELEITO EM2012

Variáveis	Outros (Não-Y)	Esc. Superior (Y)	TOTAL
Esquerda (X)	567 (A)	879 (B)	1446 (X)
Outros (não-X)	1929 (C)	2142 (D)	4071 (não-X)
TOTAL	2496 (não-Y)	3021 (Y)	5517 (N)

Aplicando a fórmula do Q_{xy} teríamos que:

$$Q_{xy} = \frac{(B\ x\ C) - (A\ x\ D)}{(B\ x\ C) + (A\ x\ D)} = \frac{(879\ x\ 1929) - (567\ x\ 2142)}{(879\ x\ 1929) + (567\ x\ 2142)} = \frac{481077}{2910115} = \mathbf{0,165}$$

Resposta: há uma associação da ordem de +0,165, ou +16,5%, entre ser prefeito em partido de esquerda e ter escolaridade superior. Como o coeficiente é positivo, isso indica que os prefeitos de partidos de esquerda concentram-se acima da distribuição normal para o grupo dos prefeitos com escolaridade superior. Sendo mais precisos, ao olharmos o quadro de interpretação da magnitude do coeficiente podemos dizer que existe uma associação positiva baixa, pois fica entre +0,10 e +0,29 a associação entre ser prefeito de esquerda e ter escolaridade superior.

Uma das principais características do Q_{xy} é que por ser o resultado de produtos cruzados, em qualquer tabela quádrupla, quando os produtos dos pares consistentes e inconsistentes estão relacionados, o Q_{xy} cresce. Além disso, o coeficiente tem limite superior em +1,00 e inferior em -

1,00. No entanto, são necessários alguns cuidados suplementares. Por se tratar da divisão de produtos cruzados, quando uma das células for zero, o valor de Q_{xy} também será nulo, embora o cálculo matemático gere como resultado -1,00. Isso não significa necessariamente a existência de relação perfeita negativa. Outro cuidado a se tomar na aplicação do Q_{xy} é com a heterogeneidade da distribuição dos casos na tabela quádrupla. Uma distribuição muito heterogênea não é indicada para o coeficiente, pois ela já apontaria uma concentração de casos em determinada casa, linha ou coluna. A sugestão é que o calculo será realizado sempre que a distribuição dos casos na tabela ficar abaixo de uma relação 70:30 em pelo menos uma das variáveis. Ou seja, não mais de 70% dos casos em uma categoria e não menos de 30% em outra. No exemplo acima temos que os prefeitos eleitos por partidos de esquerda representam 26,2% dos casos, enquanto os eleitos por outros partidos, 73,8% dos casos. Portanto, precisamos olhar a distribuição na outra variável para verificar a viabilidade da consideração do coeficiente. No caso, são 45,2% do total para prefeitos com outros níveis de escolaridade e 54,8% para prefeitos com escolaridade superior. Nesse caso, a distribuição mais homogênea da variável escolaridade "salvou" a concentração de casos na categoria outros partidos políticos, validando o coeficiente de associação encontrado.

4.1.2. CÁLCULOS ADICIONAIS: PARES CONSISTENTES X PARES INCONSISTENTES E TAMANHO DA AMOSTRA

A interpretação do resultado do coeficiente de associação parte do princípio de que o significado interno do Q_{xy} está ligado à probabilidade de um par de casos diferir em ambos os itens, ou seja, em um ser de partido de esquerda e escolaridade superior e em outro não ser de partido de esquerda e não ter escolaridade superior, para ficarmos no exemplo acima. Um par [B, C] é chamado de consistente, pois em uma casa ele indica possuir a característica medida nas duas variáveis (X e Y) e na outra apresenta a ausência da característica nas duas variáveis (não-X e não-Y). Já um par [A, D] é chamado de inconsistente, pois em uma variável apresenta a característica analisada e em outra não (X e não-Y) e vice-versa.

A fórmula para encontrar a probabilidade de pares consistentes é a seguinte:

$$P_c = \frac{2\ x\ (B\ x\ C)}{N^2}$$

A fórmula para encontrar a probabilidade de pares inconsistentes é a seguinte:

$$P_i = \frac{2\ x\ (A\ x\ D)}{N^2}$$

Aplicando as fórmulas ao exemplo acima temos que:

$$P_c = \frac{2\ x\ (B\ x\ C)}{N^2} = \frac{2\ x\ (879\ x\ 1929)}{5517^2} = \frac{3391182}{30437289} = \mathbf{0,111}$$

$$P_i = \frac{2\ x\ (A\ x\ D)}{N^2} = \frac{2\ x\ (567\ x\ 2142)}{5517^2} = \frac{2429028}{30437289} = \mathbf{0,079}$$

Então, temos que a probabilidade de encontrar pares consistentes é de 0,111 e de inconsistentes de 0,079, portanto, temos proporcionalmente a participação de um pouco mais de pares consistentes em relação aos pares inconsistentes.

Outro elemento importante a se considerar quando estamos analisando a força preditiva de um Q_{xy} para associação entre duas variáveis é o tamanho da amostra, ou, a forma como as frequências se distribuem nas casas da tabela quádrupla. A recomendação é que existam pelo menos cinco casos esperados se houvesse independência entre as variáveis em cada casa de uma tabela quádrupla. Para saber se a distribuição mínima das frequências esperadas é respeitada sem precisar encontrar o valor para todas as casas basta multiplicar duas marginais e dividir por N e você terá o Menor Valor Esperado (MVE) para aquela tabela. Para não correr nenhum risco, opte por usar as menores marginais das linhas e das colunas. O resultado será o menor valor esperado para uma célula da tabela quádrupla. Portanto, se ficar acima de cinco, todas as demais apresentarão valor esperado superior ao limite mínimo. O cálculo é o seguinte:

$$MVE = \frac{MenorMarginalLinha\ x\ MenorMarginalColuna}{N}$$

Aplicando a fórmula ao nosso exemplo, teríamos que:

$$MVE = \frac{1446\ x\ 2496}{5517} = 654{,}19$$

Com o resultado de 654,19 para "Menor Valor Esperado" não precisamos temer, pois estamos respeitando o limite mínimo de casos em cada casa para a realização do teste Q_{xy}. Vamos em frente. Agora, atenção, se a sua tabela quádrupla apresentar um N < 20 o mais provável é que o MVE fique abaixo de cinco e, nesse caso, não devemos calcular o Q_{xy}. Na verdade essa distribuição só não terá MVE abaixo de cinco se houver uma distribuição totalmente homogênea em cada uma das casas, com frequência = 5 em cada uma das 4 casas (5x4 = 20). Porém, aqui também não faz sentido aplicar o Q_{xy} , pois já sabemos que as variáveis são independentes e o Q_{xy} será igual a zero.

Muitas vezes os cientistas políticos dispõem apenas de dados amostrais para fazer os testes estatísticos, mas seu objetivo é apresentar resultados que sejam válidos para toda a população. No caso da aplicação de inferências usando o Q_{xy} é preciso levar em conta um intervalo de confiança para os valores antes de afirmar se a associação encontrada na tabela amostral pode ser extrapolada para toda a população. Evidente que para isso estamos considerando que se trata de uma amostra probabilística. A forma de calcular o intervalo de confiança para inferências a partir do Q_{xy} será apresentada no próximo tópico.

4.1.3. INTERVALO DE CONFIANÇA PARA O TESTE DE CORRELAÇÃO Q DE YULE

Até aqui utilizamos o teste Q de Yule para indicar a correlação entre duas variáveis, considerando que o número de casos na tabela indica a totalidade, ou seja, o universo estudado. Porém, o coeficiente também pode ser usado em amostras, o que nos permite passar da estatística descritiva à inferencial, extrapolando os resultados de uma amostra para o universo de casos. Para que isso aconteça, nunca devemos nos esquecer que antes de qualquer coisa é preciso que a amostra da qual saíram as informações seja probabilística. Como teoria da amostragem não é nosso objeto aqui, parto do princípio de que se trata de uma amostra probabilística e que você sabe o que isso significa.

Para podermos inferir resultados de uma amostra ao universo usamos o conceito de intervalo de confiança (IC), pois ele permite dizer que dentro de determinado intervalo de valores amos-

trais encontra-se há uma probabilidade considerada alta o suficiente de que encontraríamos o valor da população na maioria das vezes das vezes que extraíssemos amostras dessa população. Portanto, nosso objetivo é encontrar um valor que seja o limite superior e outro para limite inferior do intervalo de confiança, ou seja, mínimo e máximo que indicam o intervalo dentro do qual é muito provável encontrar o parâmetro para a população. O Intervalo de Confiança mais usado é de 95%, que equivale a dizer que se tirássemos 100 amostras probabilísticas de determinada população, em 95 delas as estatísticas amostrais ficariam dentro do Intervalo de Confiança. Não devemos nos esquecer que a afirmação anterior implica em considerar que em cinco das 100 amostras o valor da estatística estará fora do intervalo de confiança. O valor padronizado do intervalo de confiança de 95% é de z = 1,96. Esse é o número padrão que entra na fórmula para estabelecer o intervalo de confiança de 95% para resultados amostrais. Com base nele, podemos usar a fórmula abaixo para calcular o limite superior e o inferior do intervalo de valores dentro do qual se deve encontrar o valor da correlação para toda população a partir dos dados obtidos na amostra. Se o resultado indicar que o limite passa pelo valor zero (superior positivo e inferior negativo), então, não podemos dizer que há uma diferença estatística forte o suficiente que permita extrapolar o coeficiente da amostra para a população. Se o intervalo não passar por zero, podemos fazer as inferências estatísticas. A fórmula para encontrar os limites superior e inferior é a seguinte:

Limite superior = Q_{xy} +

$$1{,}96\ x\ \sqrt{\frac{\left(1-\ {Q_{xy}}^2\right)^2\ x\ \frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}{4}}$$

Limite inferior = Q_{xy} -

Se estivéssemos trabalhando com uma amostra probabilística no exemplo da relação entre prefeito de esquerda e escolaridade superior, aplicaríamos a fórmula como segue:

Limite Superior = 0,165 +

$$1{,}96\ x\ \sqrt{\frac{(1-\ 0{,}165^2)^2\ x\ \frac{1}{567}+\frac{1}{879}+\frac{1}{1929}+\frac{1}{2142}}{4}}\ =\ 1{,}96\ x\ 0{,}030 = 0{,}059$$

Limite Inferior = 0,165 –

Assim, os limites seriam:

Limite Superior = 0,165 + 0,059 = +**0,224**.

Limite Inferior = 0,165 – 0,059 = +**0,106**.

Esses resultados mostram que se os dados fizessem parte de uma amostra probabilística, o valor do coeficiente de correlação Q_{xy} para a população estaria entre +0,224 e +0,106. Como o intervalo não inclui o zero, isso indica que as chances do coeficiente de associação para a população ser nulo está dentro do intervalo de confiança que assumimos no início, portanto, podermos considerar os valores da amostra para inferências à população. De outra forma, se o intervalo passasse por zero, ou seja, o valor do limite superior fosse positivo e do inferior negativo, não poderíamos seguir com as inferências, pois os resultados estatísticos não são fortes o suficiente para permitir a extrapolação do coeficiente.

Até aqui utilizamos o coeficiente Q_{xy} para identificar possíveis relações entre duas variáveis dicotômicas dispostas em uma tabela quádrupla. Para tanto, aplicamos os seguintes conceitos: i) o que estamos testando é se existe independência entre as categorias de duas variáveis dicotômicas. Se as variáveis forem correlacionadas, identificamos o nível de associação e a direção da mesma - se positiva ou negativa, e; ii) que podemos testar a relação entre as variáveis em uma população a partir de uma amostra probabilística, desde que os valores dos limites de intervalo de confiança não passem por zero - nesse caso poderemos fazer inferências estatísticas.
Além disso, há outra aplicação para o Q_{xy} que é quando o pesquisador insere uma terceira variável dicotômica para testar o quanto essa variável intervém na relação anterior, quando eram consideradas apenas as distribuições de X e Y. Essa terceira variável, chamada de interveniente ou de variável teste (T) pode interferir de quatro maneiras diferentes na relação anterior, tornando mais rica e detalhada a explicação sobre os fenômenos políticos estudados. Aprenderemos a aplicar o teste Q_{xy} para três variáveis (representado por $Q_{xy:t}$)no próximo tópico.

4.1.4. COEFICIENTE Q_{XY} PARA TRÊS VARIÁVEIS ($Q_{XY:T}$)

Uma vez realizado o teste de associação para duas variáveis dicotômicas, podemos identificar a existência ou não de relação entre as variáveis, assim como a direção dessa relação, caso ela

exista. Não raras vezes cientistas políticos pensam em incorporar uma terceira variável à relação para saber se ela é capaz de interferir de alguma forma na relação encontrada anteriormente. Quando a associação é apenas entre duas variáveis, ela é chamada de ordem zero, pois não há nenhum tipo de controle sobre a relação entre as categorias delas. Se controlamos a relação original por uma terceira variável, essa associação é chamada de ordem um, pois há uma variável controlando a relação identificada entre as duas originais. Portanto, ao acrescentarmos uma terceira variável na medida de associação, temos como resultado uma associação de ordem um ou uma associação parcial, pois está levando em conta o efeito de controle de uma variável externa.

No tópico anterior e, por tradição, chamamos as variáveis diretamente inseridas no teste de associação de variável X e variável Y. Aqui, a terceira variável será chamada de T (variável teste). Então, a variável teste deve exercer algum efeito sobre a relação entre X e Y, ou seja, ela deve "controlar" de alguma forma a associação medida entre X e Y. Por isso ela também é chamada de variável de controle nos testes de associação. No tópico anterior as duas variáveis que utilizamos para exemplificar o teste $\mathbf{Q}_{xy}$ foram ideologia do partido (esquerda ou outras) por escolaridade do eleito (superior ou outras). Testamos se é possível afirmar a existência de uma associação entre prefeitos eleitos por partidos de esquerda e prefeitos com escolaridade superior. Digamos que agora seja inserida uma terceira variável no teste, ou variável de controle, para verificar como a relação anterior se comporta, dado o efeito da variável teste. No caso, a terceira variável será Sexo do eleito (homem ou mulher). Vale lembrar que a variável teste também deve ser dicotômica para o teste $Q_{xy:t}$. Nossa pergunta aqui passa a ser a seguinte: dada a relação já identificada entre ser de esquerda e ter escolaridade superior, podemos dizer que homens e mulheres de esquerda apresentam comportamentos distintos em relação à escolaridade dos prefeitos eleitos? Ou, em outras palavras, será que há diferenças significativas entre prefeitos de esquerda com nível de escolaridade superior em relação ao nível de escolaridade das prefeitas eleitas por partidos de esquerda? Essas variáveis serão utilizadas a seguir para a realização do teste $Q_{xy:t}$ para três variáveis.

Na prática, realizam-se dois testes Q_{xy} entre as variáveis X e Y, uma para os casos em que há presença da característica da variável teste e outra para os casos em que não há a característica da variável T. Antes do exemplo é preciso discutir quais são os efeitos possíveis da variável teste sobre a relação entre duas variáveis. Basicamente existem quatro possíveis efeitos: explicação, supressão, especificação ou sem efeito algum. Esses efeitos são constatados quando há alguma diferença entre o coeficiente de associação obtido antes da inserção da variável de controle e depois dela. Havendo alguma diferença, isso é sinal de que a variável de teste exerceu algum tipo de controle sobre os resultados anteriores. A diferença entre os tipos de efeito é:

a) Efeito de Explicação: o efeito de explicação da variável de controle acontece quando o coeficiente de associação de ordem zero é significativo, mas após a inserção da variável teste ele se aproxima de zero. Nesse caso dizemos que "T explica Y", pois antes de considerarmos T havia uma relação aparente entre X e Y. Agora, com a inserção de T a relação é anulada, indicando que a relação anterior só existia enquanto se desconsiderava a característica de controle. É o caso, por exemplo, de termos uma relação significativa entre ideologia partidária e escolaridade do eleito. Porém, quando inserida a variável teste Sexo, a relação anterior passar a ser próxima de nula. Ou seja, Sexo explica escolaridade do eleito para além da ideologia partidária.

b) Efeito de Supressão: Acontece quando o coeficiente de associação parcial, após inserção da variável teste, é mais forte que a associação de ordem zero. Nesse caso dizemos que T é variável supressiva, visto que ela estava suprimindo a verdadeira relação entre X e Y, que se torna aparente apenas quando há o controle por T. Seria o caso de, após inserida a variável Sexo na associação entre ideologia e nível de escolaridade, o coeficiente aumentar. Diz-se que o efeito é supressivo porque Sexo estava escondendo ou suprimindo a verdadeira relação entre X e Y.

c) Efeito de Especificação: esse efeito é diferente dos dois anteriores, quando o coeficiente crescia ou diminuía após a inserção da variável teste na associação. Se considerarmos que uma associação parcial é a combinação de duas associações anteriores, devemos considerar a possibilidade de efeitos distintos sobre cada uma das anteriores. As diferenças podem ser em termos de magnitude dos coeficientes, assim como até mesmo em sinais invertidos – com associação positiva em uma e negativa em outra. Quando isso acontece, dizemos que há um efeito de especificação, ou, "T especifica XY". Por exemplo, a associação de ordem zero entre ideologia e escolaridade do eleito pode ser alta e positiva. Porém, quando inserimos a variável teste Sexo, podemos encontrar que para homens a relação entre ideologia e escolaridade é significativa e positiva, enquanto que para mulheres ela é significativa e negativa. Nesse caso, sexo está especificando a relação entre ideologia e escolaridade.

d) Sem efeito: o quarto tipo de efeito possível é justamente a ausência de qualquer efeito de T sobre a relação XY. Ele é percebido quando o coeficiente de associação entre as duas variáveis é exatamente o mesmo que o obtido após a inserção da variável teste, ou seja, não houve nenhum efeito da terceira sobre a relação identificada entre as duas anteriores. A ausência de efeito da variável teste é importante para demonstrar que as duas variáveis X e Y estão realmente associadas, pois a inserção da terceira variável não alterou a relação percebida inicialmente. No caso dos exemplos apresentados aqui, equivale a dizer que o coeficiente de correlação entre ideologia e escolaridade do prefeito eleito é o mesmo do que o coeficiente da correlação entre ideologia e escolaridade após a inserção da variável teste Sexo.

Como estamos falando, em termos práticos, da repetição do teste Q_{xy} para a presença da característica da variável T e para a ausência da característica na variável T, o que temos é a junção

de duas tabelas quádruplas no teste com três variáveis. Ou seja, se no teste entre X e Y tínhamos uma tabela quádrupla (2x2), agora, no teste T, X e Y temos uma tabela óctupla (2x2x2), como o que está representado no quadro a seguir:

Quadro 4.4. Formato das distribuições para $Q_{xy:t}$

		não-Y	Y	TOTAL
T	X	AT	BT	TX
	não-X	CT	DT	$T\overline{x}$
	TOTAL	$T\overline{Y}$	TY	
não-T	X	$A\overline{T}$	$B\overline{T}$	$\overline{T}X$
	não-X	$C\overline{T}$	$D\overline{T}$	$\overline{T}\ \overline{X}$
	TOTAL	$\overline{T}\ \overline{Y}$	$\overline{T}Y$	

O quadro acima é montado com as casas A, B, C e D repetindo-se duas vezes cada uma, o que comprova que o teste de $Q_{xy:t}$ com três variáveis equivale a dois testes de Q_{xy} simultâneos. Não existe nenhum problema para se calcular um $Q_{xy:t}$ para a tabela quádrupla da parte de cima e outro para a quádrupla da parte de baixo. A letra com o traço acima indica a ausência da característica ($^{-}$ representa "não"). Na parte de cima, onde aparece a característica da variável teste a frequência é representada por T. Na outra, onde não há característica da variável teste, o símbolo é $\overline{T}$. O mesmo vale para as marginais de X e não-X ($\overline{x}$) e Y e não-Y ($\overline{Y}$). As marginais das linhas apresentam três tipos de pares de valores: com presença de T e X (TX), com presença de apenas uma delas ($T\overline{x}$ ou $\overline{T}X$) e sem nenhuma das características ($\overline{T}\,\overline{x}$). O mesmo vale para as marginais da variável Y.

Feitas as descrições básicas da tabela óctupla, podemos dizer que existem dois tipos de pares de XY naquela relação: pares ligados a T e pares diferentes de T. Isso porque poderíamos calcular tranquilamente dois coeficientes Q_{xy}, um para pares ligados a T e outro para não ligados a T. O princípio por trás do teste $Q_{xy:t}$ para três variáveis é o de que podemos construir coeficientes Q_{xy} para pares ligados a T e para pares que diferem de T a partir das localizações no quadro acima. O procedimento é bastante simples e divide-se em três partes: i) calcula-se o coeficiente Q_{xy} para pares ligados e o coeficiente Q_{xy} para pares diferentes; ii) calcula-se o Peso para pares ligados e o Peso para pares diferentes; iii) soma-se o produto do coeficiente Q_{xy} para pares ligados multiplicado pelo Peso para pares ligados com o produto do coeficiente Q_{xy} para pares diferentes multiplicado pelo Peso para pares diferentes. A fórmula final é:

$$Q_{xy:t} = \left(Q_{xy}\ ligado\ x\ P.ligados\right) + \ (Q_{xy}\ diferente\ x\ P.diferentes)$$

Onde:
Q_{xy} ligado : coeficiente Q_{xy} para pares ligados a T;
P.ligados : peso para pares ligados a T;
Q_{xy} diferente : coeficiente Q_{xy} para pares diferentes de T;
P.diferentes : peso para pares diferentes.

Assim, o primeiro passo para calcular o $Q_{xy:t}$ para três variáveis é encontrar os valores de Q_{xy} para pares ligados e para pares diferentes de T. As fórmulas, também intuitivas, são as seguintes:

$$Q_{xy}ligado = \frac{\left[(BT\ x\ CT) + \ \left(B\overline{\mathrm{T}}\ x\ C\overline{\mathrm{T}}\right)\right] - \ [\ (AT\ x\ DT) + \ \left(A\overline{\mathrm{T}}\ x\ D\overline{\mathrm{T}}\right)]}{\left[(BT\ x\ CT) + \ \left(B\overline{\mathrm{T}}\ x\ C\overline{\mathrm{T}}\right)\right] + \ [\ (AT\ x\ DT) - \ \left(A\overline{\mathrm{T}}\ x\ D\overline{\mathrm{T}}\right)]}$$

Essa fórmula nos indica qual é o Q parcial para X e Y, controlado por T ou qual é o Q de X e Y em pares ligados a T. Trata-se da melhor forma de predizer a relação entre X e Y quando consideramos apenas os pares ligados a T. Agora é preciso fazer a mesma coisa para os pares diferentes de T na tabela óctupla. A fórmula é:

$$Q_{xy}diferente = \frac{\left[\left(BT\ x\ C\overline{\mathrm{T}}\right) + \ \left(B\overline{\mathrm{T}}\ x\ CT\right)\right] - \ [\ \left(AT\ x\ D\overline{\mathrm{T}}\right) + \ \left(A\overline{\mathrm{T}}\ x\ DT\right)]}{\left[\left(BT\ x\ C\overline{\mathrm{T}}\right) + \ \left(B\overline{\mathrm{T}}\ x\ CT\right)\right] + \ [\ \left(AT\ x\ D\overline{\mathrm{T}}\right) + \ \left(A\overline{\mathrm{T}}\ x\ DT\right)]}$$

O resultado desse cálculo pode ser interpretado como o coeficiente Q entre X e Y quando T diferente ou o Q entre X e Y para pares diferentes de T. Trata-se da melhor forma de predizer a relação X e Y apenas para os casos em que os pares são diferentes em T. Agora que já encontramos os dois coeficientes, para pares ligados a T e para diferentes de T, o próximo passo é definir os pesos de cada um dos tipos de pares na fórmula final. Com os pesos nós substituímos uma média simples entre os dois coeficientes por uma média ponderada pelas diferenças proporcionais dos pares ligados e diferentes de T. Assim, tornamos o coeficiente final mais preciso. As fórmulas são as que seguem:

Peso 1 : proporção de pares ligados em T entre pares diferentes em X e Y.

$$P1 = \frac{(BT\ x\ CT) + \left(B\overline{T}\ x\ C\overline{T}\right) + (AT\ x\ DT) + \left(A\overline{T}\ x\ D\overline{T}\right)}{\left[\left(BT + B\overline{T}\right) x \left(CT + C\overline{T}\right)\right] + \left[\left(AT + A\overline{T}\right) x \left(DT + D\overline{T}\right)\right]}$$

Peso 2 : proporção de pares diferentes em T entre pares diferentes em X e Y.

$$P2 = \frac{\left(BT\ x\ C\overline{T}\right) + \left(B\overline{T}\ x\ CT\right) + \left(AT\ x\ D\overline{T}\right) + \left(A\overline{T}\ x\ DT\right)}{\left[\left(BT + B\overline{T}\right) x \left(CT + C\overline{T}\right)\right] + \left[\left(AT + A\overline{T}\right) x \left(DT + D\overline{T}\right)\right]}$$

Em uma primeira mirada a fórmula assusta, mas no fundo o princípio é simples. Trata-se das frequências de pares T ligados e diferentes nos numeradores do Peso 1 e do Peso 2, respectivamente. O denominador é o mesmo nas duas fórmulas e trata-se do número total de pares diferentes em X e Y. Vamos ver como calcular $Q_{xy:t}$ para três variáveis inserindo como variável teste o sexo dos eleitos em 2012 para prefeito do Brasil no cruzamento entre ser eleito por partido de esquerda e ter nível de escolaridade superior. O cruzamento segue abaixo:

TABELA 4.2. CRUZAMENTO ENTRE POSIÇÃO IDEOLÓGICA DE ESQUERDA POR ESCOLARIDADE SUPERIOR CONTROLADO POR SEXO DO ELEITO

Sexo	Posição ideológica	Escolaridade		TOTAL
		Outras	Superior	
Homem	Esquerda	544	750	1294
	Outras	1786	1803	3589
	Total	2330	2553	4883
Mulher	Esquerda	23	129	152
	Outras	143	339	482
	Total	166	468	634

Aplicando as fórmulas, começamos por encontrar os coeficientes Q para pares ligados a T:

$$Q_{xy}ligado = \frac{\left[(BT\ x\ CT) + \left(B\overline{T}\ x\ C\overline{T}\right)\right] - \left[(AT\ x\ DT) + \left(A\overline{T}\ x\ D\overline{T}\right)\right]}{\left[(BT\ x\ CT) + \left(B\overline{T}\ x\ C\overline{T}\right)\right] + \left[(AT\ x\ DT) - \left(A\overline{T}\ x\ D\overline{T}\right)\right]}$$

$$= \frac{[(750\ x\ 1786) + (129x\ 143)] - [(544\ x\ 1803) + (23\ x\ 339)]}{[(750\ x\ 1786) + (129\ x\ 143)] + [(544\ x\ 1803) - (23\ x\ 339)]} = \frac{369318}{2330982}$$

$$= \mathbf{0,158}$$

Aplicamos a fórmula para encontrar o coeficiente Q para pares diferentes de T

$$Q_{xy}diferente = \frac{\left[\left(BT\ x\ C\overline{T}\right) + \left(B\overline{T}\ x\ CT\right)\right] - \left[\left(AT\ x\ D\overline{T}\right) + \left(A\overline{T}\ x\ DT\right)\right]}{\left[\left(BT\ x\ C\overline{T}\right) + \left(B\overline{T}\ x\ CT\right)\right] - \left[\left(AT\ x\ D\overline{T}\right) + \left(A\overline{T}\ x\ DT\right)\right]} =$$
$$= \frac{[(750\ x\ 143) + (129\ x\ 1786)] - [(544\ x\ 339) + (23\ x\ 1803)]}{[(760\ x\ 143) + (129x\ 1786)] + [(544\ x\ 339) + (23\ x\ 1803)]} = \frac{79119}{596169}$$
$$= \mathbf{0,198}$$

O próximo passo é encontrar os pesos para cada grupo de pares. Começamos pelo peso dos pares ligados a T, chamado P1:

$$P1 = \frac{(BT\ x\ CT) + \left(B\overline{T}\ x\ C\overline{T}\right) + (AT\ x\ DT) + \left(A\overline{T}\ x\ D\overline{T}\right)}{\left[\left(BT + B\overline{T}\right)x\left(CT + C\overline{T}\right)\right] + \left[\left(AT + A\overline{T}\right)x\left(DT + D\overline{T}\right)\right]}$$
$$= \frac{(750\ x\ 1786) + (129\ x\ 143) + (544\ x\ 1803) + (23\ x\ 339)}{[(750 + 129)\ x\ (1786 + 143)] + [(544 + 23)\ x\ (1803 + 339)]} = \frac{2346576}{2910105}$$
$$= \mathbf{0,807}$$

O passo seguinte é encontrar o peso dos pares diferentes de T, chamado de P2

$$P2 = \frac{\left(BT\ x\ C\overline{T}\right) + \left(B\overline{T}\ x\ CT\right) + \left(AT\ x\ D\overline{T}\right) + \left(A\overline{T}\ x\ DT\right)}{\left[\left(BT + B\overline{T}\right)x\left(CT + C\overline{T}\right)\right] + \left[\left(AT + A\overline{T}\right)x\left(DT + D\overline{T}\right)\right]}$$
$$= \frac{(750\ x\ 143) + (129\ x\ 1786) + (544\ x\ 339) + (23\ x\ 1803)}{[(750 + 129)\ x\ (1786 + 143)] + [(544 + 23)\ x\ (1803 + 339)]} = \frac{563529}{2910105}$$
$$= \mathbf{0,193}$$

Agora já temos todos os fatores necessários para o cálculo do coeficiente Q para as três variáveis. A fórmula a ser aplicada é:

$$Q_{xy:t} = \left(Q_{xy}\ ligado\ x\ P.ligado\right) + \left(Q_{xy}\ diferente\ x\ P.diferente\right)$$
$$Q_{xy:t} = (0{,}198 \times 0{,}807) + (0{,}132 \times 0{,}193)$$
$$Q_{xy:t} = 0{,}159 + 0{,}025$$

$$Q_{xy:t} = \mathbf{0,185}$$

Resultado: o coeficiente de associação $Q_{xy:t}$ para prefeito eleito por partido de esquerda e escolaridade superior, controlada por Sexo do prefeito é de 0,185 ou 18,5%. Como o coeficiente é positivo, significa que há uma associação de 18,5% entre ser eleito por partido de esquerda e ter escolaridade superior, quando controlado por sexo do prefeito. Perceba que quando comparamos com o coeficiente do Q_{xy} sem a variável de controle há uma pequena queda. Quando consideramos a escolaridade por ideologia do partido apenas o coeficiente de associação foi de 16,5%, contra 18,5% de associação quando controlado pela variável Sexo. A existência de diferença e a magnitude dela é que indica o efeito da variável teste sobre a relação de ordem zero. A questão é saber se essa diferença é grande o suficiente para dizer que a variável teste interferiu de fato na associação de ordem zero.

Podemos entender o $Q_{xy:t}$ para três variáveis como uma média ponderada do Q_{xy} parcial e do Q_{xy} diferencial, sendo os pesos as proporções de pares ligados em T e diferentes de T. Um dos mais importantes princípios do teste de associação $Q_{xy:t}$ com três variáveis é que qualquer que seja o valor da associação na ordem zero, o valor parcial poderá assumir também qualquer valor entre os limites teóricos +1,00 e -1,00. Esse princípio nos permite estabelecer relações possíveis para a análise de três variáveis em um espaço bidimensional, onde o eixo Y representa o valor que Q_{xy} e o eixo X representa os valores de $Q_{xy:t}$, ambos podendo variar nos limites teóricos de +1,00 a -1,00. Em um artigo publicado em 1950 *Kendall e Lazarsfeld* demonstram as regiões em que esse gráfico de ordenadas pode ser dividido, chegando a cinco regiões significativas, como indicado no gráfico a seguir.

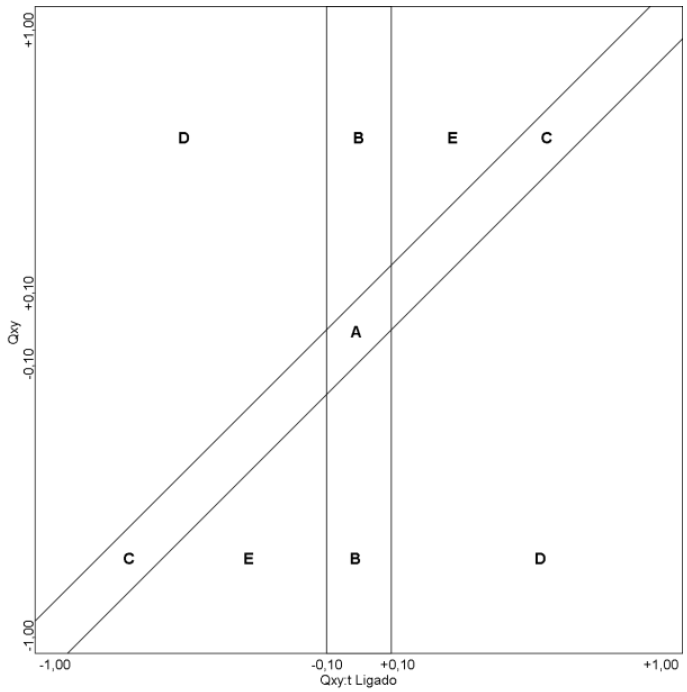


GRÁFICO 4.1. ÁREAS DE DISTRIBUIÇÕES DE POSIÇÕES DOS COEFICIENTES DE ORDEM ZERO E PARCIAL

A região A indica a área onde se situam as associações muito próximas a zero tanto na correlação de ordem zero, quanto na parcial. Ou seja, não há relação entre X e Y com ou sem o controle de T;

As regiões B indicam os resultados em que a relação de ordem zero é significativa, porém, a associação parcial fica próxima a zero, quer dizer, a variável de controle explica a relação entre X e Y, pois o coeficiente era alto (positivo ou negativo) antes do controle e caiu após a inserção da variável teste;

A região C, que segue a diagonal dos dois eixos indica os casos em que a correlação de ordem zero e a correlação parcial apresentam resultados muito próximos. Os coeficientes que caem nessa região apresentam uma correlação entre X e Y que independe do controle da variável T. Portanto, a variável teste não exerce efeito significativo sobre a associação de ordem zero.

Nas regiões D encontram-se os resultados em que o coeficiente parcial é maior que o coeficiente de ordem zero - com o mesmo sinal ou com sinal oposto. Isso é, quando a relação entre X e Y é controlada por T, ela aumenta na mesma direção ou aumenta mudando de sinal. Aqui há um efeito evidente da terceira variável sobre a relação de ordem zero.

A região E indica os casos em que a correlação de ordem zero é significativa e a correlação parcial, também. Porém, o coeficiente da correlação parcial é menor que o coeficiente de ordem zero. Nesses casos há uma oscilação entre poder explicar a relação e não poder explicá-la. Agora, conhecendo o gráfico de distribuição dos coeficientes de ordem zero e parcial, temos condições de verificar se o nosso coeficiente parcial (com variável de controle Sexo) exerce algum tipo de influência sobre o coeficiente de ordem zero (associação entre ideologia do partido e escolaridade do prefeito). O resultado está plotado no gráfico 4.2 a seguir:

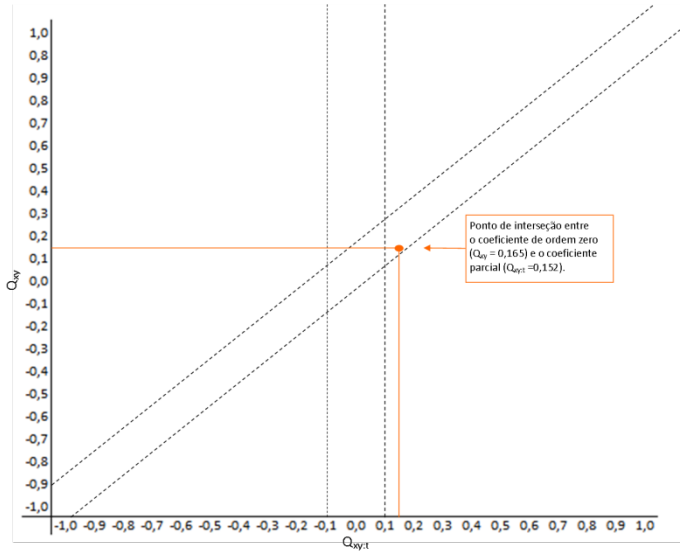


GRÁFICO 4.2. INTERSEÇÃO ENTRE Q_{XY} E $Q_{XY:T}$ PARA ASSOCIAÇÃO DE ORDEM ZERO E PARCIAL ANTERIOR

A imagem mostra que o ponto de interseção entre os dois coeficientes no gráfico de ordenadas X e Y fica na região C, indicando que a diferença entre o coeficiente de ordem zero e o coeficiente parcial é tão pequena que não podemos considerar a existência de qualquer efeito da variável de controle sobre a associação anterior. Em outras palavras, considerando os resultados para eleições municipais de 2012, a distribuição de prefeitos com escolaridade superior está relacionada com o fato de pertencer a partido de esquerda, porém, essa relação independente do sexo do candidato.

Até aqui apresentamos os testes para variáveis dicotômicas. O próximo tópico tratará da técnica adequada para medir a relação entre duas variáveis categóricas ordinais com pelo menos três categorias em uma delas. Trata-se do coeficiente de correlação Gama.

4.1.5. EXERCÍCIOS

A tabela a seguir foi extraída da pesquisa produzida pelo Latinobarômetro sobre a opinião pública brasileira em 2008. Fazem parte da tabela as respostas dadas às variáveis sobre avaliação geral da economia, avaliação do trabalho do então presidente Lula e a distribuição dos casos em dois grupos de idades distintas (baixa < 39 anos e alta > 39 anos). A partir dessas informações, reorganize as informações em tabelas quádruplas e responda as questões.

Como o sr/sra avalia a situação econômica do país? O sr./sra acha que é muito boa, boa, nem boa nem má, má ou muito má? * E falando em geral do atual governo, como o sr./sra. avalia o trabalho que o Presidente Lula está realizando? * Idade binária

			E falando em geral do atual governo, como o sr./sra. avalia o trabalho que o Presidente Lula está realizando?					
Idade			Muito bom	Bom	Nem bom, nem mal	Mal	Muito mal	Total
Idade Baixa	Como o sr/sra avalia a situação econômica do país?	Muito boa	5	6	3	1	1	16
		Boa	32	73	36	6	3	150
		Nem boa, nem má (regular)	35	170	172	18	12	407
		Má	5	32	39	12	6	94
		Muito má (péssima)	9	20	48	6	9	92
	Total		86	301	298	43	31	759
Idade Alta	Como o sr/sra avalia a situação econômica do país?	Muito boa	9	3	4	0	0	16
		Boa	30	69	31	3	4	137
		Nem boa, nem má (regular)	43	140	142	22	9	356
		Má	8	23	46	7	6	90
		Muito má (péssima)	7	21	46	8	11	93
	Total		97	256	269	40	30	692

4.1.5.a. Faça um teste de independência Q_{xy} para duas variáveis a partir das informações apresentadas acima. Para tanto, construa uma tabela quádrupla entre Avaliação Positiva x Avaliação

Negativa de Lula e Avaliação Positiva x Avaliação Negativa da economia. Para esse teste você deve desconsiderar as categorias das idades dos respondentes.

4.1.5.b. A partir da tabela quádrupla produzida no exercício anterior, aplique a fórmula para identificar a proporção de Pares Consistentes x Pares Inconsistentes.

4.1.5.c. Sabemos que os resultados das associações testadas aqui são obtidos a partir de uma pesquisa amostral, portanto, não podemos extrapolar os valores dos testes anteriores para a população antes de verificarmos se existem condições estatísticas suficientemente significativas para permitir a afirmação de que as variáveis estão relacionadas. Para tanto, identifique o intervalo de confiança para o teste Q_{xy} com os dados gerados para o exercício 4.1.5.a.

4.1.5.d. Insira no modelo a variável Teste (idade) produzindo uma tabela óctupla para: i) fazer o teste de Q_{xy} para três variáveis e ii) localizar o resultado no gráfico de dispersão. Depois interprete as diferenças da associação de ordem zero e a parcial.

4.2 - TESTE COM VARIÁVEIS ORDINAIS - COEFICIENTE GAMA (G)

Quando a tabela tem mais que duas colunas e duas linhas, deixa de ser tabela quádrupla e passa a ser classificada pelo número de linhas e colunas, chamada de "LxC", ou, linhas por colunas. Uma tabela com 3 linhas e 4 colunas é denominada de "3x4". Lembrando que cada linha representa o número de categorias válidas da variável (x) e cada coluna representa o número de categorias válidas da variável (y). Até aqui falamos de coeficientes que medem a força da relação entre duas variáveis dicotômicas, ou seja, em tabelas "2x2". A partir de agora abordaremos os coeficientes usados em tabelas com mais de duas linhas ou colunas, quer dizer, para aquelas variáveis que apresentam três ou mais categorias.

O número de categorias não basta para definir o tipo de teste, mas sim o nível de mensuração da variável. Como apresentado no início do texto, o Coeficiente Gama (G) é um teste indicado para medir o grau de relação entre duas variáveis categóricas ordinais ou pelo menos uma ordinal e outra nominal. As variáveis ordinais apresentam como característica um arranjo de categorias que é "transitivo", ou seja, ela nos dá uma noção de ordem entre as suas categorias. Há uma transição entre a categoria de menor valor, para a de maior valor. Ou entre a que representa os valores mais negativos para a que representa os valores mais positivos. Por exemplo, uma variável categórica é o nível de escolaridade das pessoas. Sabemos que uma pessoa que tem nível de escolaridade Fundamental passou menos tempo na escola do que alguém que tem nível Médio, que por sua vez tem menos escolaridade do que aqueles com nível Superior. Nesse caso, o nível Médio faz a transição entre o fundamental e o superior. Outro exemplo é tamanho dos municípios por categorias: micro, pequeno, médio e grande. Há uma transitividade de número de habitantes crescente da primeira para a última categoria dessa variável.

Se já vimos que qualquer variável pode ser dicotomizada para se verificar o nível de associação pelo teste Q_{xy}, então, por que optar por um teste de associação entre variáveis com mais de duas categorias? A resposta é: quando agregamos categorias para dicotomizar as variáveis, corremos o risco de esconder pares consistentes em conjuntos de categorias com predomínio de pares inconsistentes. Isso diminui a precisão dos resultados de testes como o Q_{xy}. Quando usamos testes com variáveis cujas categorias não foram agregadas, a precisão aumenta.

O cálculo para o coeficiente Gama segue o mesmo princípio do Q_{xy}. Ele compara o volume de pares consistentes com o de pares inconsistentes, dividindo pelo total de pares diferentes em X e Y. Só que o Gama faz isso para todas as categorias que aparecem nas variáveis, aumentando o detalhamento das relações entre as categorias das variáveis. Como o coeficiente Gama aplica o mesmo princípio para todos os pares de valores, a fórmula do seu cálculo não é tão simples (produtos cruzados) como a do Q_{xy}.

Vejamos como calcular a partir do exemplo a seguir entre as variáveis i) tamanho do município em número de habitantes em 2012 segundo IBGE e ii) escolaridade do prefeito eleito (lê e escreve, fundamental, médio e superior), de acordo com o declarado ao TSE. As perguntas que movem esse teste poderiam ser: será que existe alguma relação entre escolaridade do eleito e tamanho do município? Em municípios maiores devemos encontrar uma concentração de prefeitos com maior escolaridade? O objetivo é identificar se existe ou não associação entre escolaridade do eleito e tamanho do município. A hipótese nula é que não existe. A hipótese alternativa é que as duas variáveis estão associadas. Como elas são ordinais, deve-se calcular o Coeficiente Gama. Se o resultado por positivo, teremos uma associação positiva entre nível de escolaridade e tamanho do município; se for negativo, a associação será inversa - prefeitos com menor escolaridade tendem a se concentrar em maiores municípios (não é isso que esperamos encontrar). Se o coeficiente for zero ou próximo dele, a associação será muito baixa.

TABELA 4.3. CRUZAMENTO ENTRE ESCOLARIDADE DO PREFEITO ELEITO E TAMANHO DO MUNICÍPIO

Escolaridade	Tamanho do Município (número de habitantes)						Total
	até 5 mil	de 5 a 10 mil	de 10 a 20 mil	de 20 a 50 mil	de 50 a 100 mil	de 1000 a 500 mil	
Superior	552	637	734	675	237	172	3007
Médio	454	432	449	275	60	24	1694
Fundamental	245	215	169	97	18	10	754
Lê e escreve	7	12	19	7	1	1	47
Total	1258	1296	1371	1054	316	207	5502

Um detalhe importante para o teste Gama quando as duas variáveis são ordinais é que as categorias devem ser organizadas na tabela de forma que a linha superior e a coluna da direita correspondam aos extremos positivos das categorias das duas variáveis, conforme o esquema a seguir, onde a categoria que representa o valor mais alto de X(++) está na primeira linha da tabela e a categoria que representa o valor mais alto de Y(++) está na coluna da direita.

QUADRO 4.4. EXEMPLO DE ORGANIZAÇÃO DAS CATEGORIAS NA TABELA DE CONTINGÊNCIA PARA TESTE G

Var X	Var Y			Total
	Y--	Y-+	Y++	
X++				
X -+				
X --				
Total				

Como no nosso exemplo a variável X é nominal, precisamos tomar cuidado com a organização das categorias em ordem crescente apenas nas colunas da tabela, começando da escolaridade mais baixa à esquerda e passando para escolaridade mais alta à direita.
A fórmula para calcular o coeficiente Gama para a relação entre as categorias é:

$$G = \frac{PC - PI}{PC + PI}$$

Onde:
PC: Produto cruzado total para pares consistentes;
PI: Produto cruzado total para pares inconsistentes.

O cálculo do PC e do PI é um pouco trabalhoso, mas nada complicado. Para encontrar o Produto cruzado total para os pares consistentes (PC) faça o seguinte:

a) Comece multiplicando a frequência da célula superior direita com cada uma das frequências situadas abaixo e à direita dela. No nosso exemplo seria 172. Esse valor será multiplicado 15 vezes, pois há 15 casas à esquerda e abaixo dela.

b) em seguida repita o procedimento para a célula que encontra-se imediatamente à esquerda da anterior (no exemplo, 237) multiplicando-a por todas as frequências situadas abaixo e à esquerda dela. São 12 casas.

c) Repita o mesmo procedimento até chegar à penúltima coluna, pois não havendo nenhum valor à esquerda e abaixo da última coluna à esquerda, ela deve ser desconsiderada. A soma de todas essas multiplicações é o PC (no exemplo, PC = 3.790.399).

QUADRO 4.5. PROCEDIMENTOS PRÁTICOS PARA O CÁLCULO DO PC

	até 5 mil	de 5 a 10 mil	de 10 a 20 mil	de 20 a 50 mil	de 50 a 100 mil	de 1000 a 500 mil
Superior	552	637	734	675	237	172
Médio	454	432	449	275	60	24
Fundamental	245	215	169	97	18	10
Lê e escreve	7	12	19	7	1	1

78088	74304	77228	47300	10320	
42140	36980	29068	16684	3096	
1204	2064	3268	1204	172	
107598	102384	106413	65175		
58065	50955	40053	22989		
1659	2844	4503	1659		
306450	291600	303075			
165375	145125	114075			
4725	8100	12825			
333236	317088				
179830	157810				
5138	8808				
289198					
156065					
4459			**SOMA**	**3.790.399**	

Para o cálculo do PI, produto cruzado total para os pares inconsistentes, os passos são parecidos, porém, no sentido inverso:

a) Multiplique a frequência do canto superior esquerdo (552) com todas as frequências abaixo e à direita dela.

b) Repita o mesmo procedimento para a primeira frequência da segunda coluna à esquerda (637) e assim sucessivamente.

c) repita o procedimento até a penúltima coluna, pois não havendo frequência à direita e abaixo da última coluna, ela deve ser desconsiderada. A soma de todas essas multiplicações é o PI (PI = 2.154.445).

A seguir estão os cálculos para encontrar o PC e o PI do exemplo acima:

QUADRO 4.6. PROCEDIMENTOS PRÁTICOS PARA CÁLCULO DO PI

	até 5 mil	de 5 a 10 mil	de 10 a 20 mil	de 20 a 50 mil	de 50 a 100 mil	de 1000 a 500 mil
Superior	552	637	734	675	237	172
Médio	454	432	449	275	60	24
Fundamental	245	215	169	97	18	10
Lê e escreve	7	12	19	7	1	1

238464	247848	151800	33120	13248
118680	93288	53544	9936	5520
6624	10488	3864	552	552
	286013	175175	38220	15288
	107653	61789	11466	6370
	12103	4459	637	637
		201850	44040	17616
		71198	13212	7340
		5138	734	734
			40500	16200
			12150	6750
			675	675
				5688
				2370
SOMA	**2.154.445**			237

Antes mesmo de aplicarmos a fórmula, podemos perceber que o montante obtido em pares consistentes (PC) é bem superior ao montante de pares inconsistentes (PI), o que nos permite antecipar que há alguma associação entre as variáveis, pois os pares consistentes são em número distinto dos inconsistentes. Além disso, podemos dizer que a associação é positiva, pois os consistentes são em maior número que os inconsistentes. Aplicando a fórmula, temos que o coeficiente Gama é:

$$G = \frac{PC - PI}{PC + PI} = \frac{3.790.399 - 2.154.445}{3.790.399 + 2.154.445} = \frac{1.635.954}{5.944.844} = \mathbf{0,275}$$

O resultado é um coeficiente de associação Gama de +0,275 (+27,5%) entre escolaridade do prefeito eleito e tamanho do município, portanto, conforme aumenta o tamanho do município, maior tende a ser a escolaridade do prefeito eleito. O nível de detalhamento do coeficiente depende do número de categorias nas variáveis. Se, para reduzirmos o trabalho, agregarmos duas categorias em uma para tornar o teste mais simples, o resultado será uma redução na precisão do coeficiente. Essa redução depende do grau de "ocultação" de pares consistentes que acontecerá quando as categorias forem agregadas. Como tem mais de duas categorias, as variáveis ordinais usadas no cálculo do coeficiente Gama permitem resultados mais refinados do que ao usarmos um coeficiente Q-Yule para variáveis dicotômicas. Vejamos esse efeito na prática, refazendo o cálculo do exemplo anterior, porém, agora com as variáveis agregadas em Escolaridade (lê, escreve+fundamental / Médio+superior) e tamanho de município em até 10 mil, de 10 a 50 mil e acima de 50 mil habitantes. Os resultados são os que seguem:

TABELA 4.4. RELAÇÃO ENTRE ESCOLARIDADE E TAMANHO DO MUNICÍPIO AGREGADOS

Escolaridade	Tamanho do Município (número de habitantes)			Total
	Até 10 mil	De 10 a 50 mil	Acima 50 mil	
Médio/Superior	2075	2133	493	4701
Lê e Escreve/ Fundamental	479	292	30	801
Total	2554	2425	523	5502

Nesse caso: PC = 1.401.810 e PI = 732.140. Aplicando a fórmula do coeficiente G temos:

$$G = \frac{PC - PI}{PC + PI} = \frac{1.401.810 - 732.140}{1.401.810 - 732.140} = \frac{669.670}{2.133.950} = \mathbf{0,313}$$

Com as categorias agregadas, ainda que com os mesmos dados, obtivemos um coeficiente de associação Gama de 31,3%, superior ao anterior, que era de 27,5%. A explicação para a diferença é que as categorias agregadas "escondem" diferenças sutis entre os prefeitos com ensino fundamental e que sabem ler e escrever que deixaram de ser consideradas aqui. Então, a medida ficou mais rústica, o que aumenta a possibilidade de termos um coeficiente maior do que quando consideramos todas as sutilezas das diferentes categorias. No limite, se mantivermos o mesmo exemplo acima, mas dicotomizando a variável tamanho do município poderemos usar a

fórmula do Q-yule, porém, essa será uma medida ainda mais rústica, pois estará "escondendo" variações reais entre pares de casos das categorias agregadas. Vejamos como seria no caso da correlação entre escolaridade baixa X Município com até 10 mil habitantes. A tabela de contingência nesse caso seria:

TABELA 4.5. VARIÁVEIS DICOTÔMICAS PARA Q_{XY} ENTRE ESCOLARIDADE DO PREFEITO E TAMANHO DO MUNICÍPIO

Escolaridade	Tamanho do Município (número de habitantes)		Total
	Até 10 mil (não-y)	Acima de 10 mil (y)	
Alta (médio+superior) (x)	2075	2626	4701
Baixa (outras) (não-X)	479	322	801
Total	2554	2948	5502

O cálculo para encontrar o coeficiente Q_{xy} é:

$$Q = \frac{(BxC) - (AxD)}{(BxC) + (AxD)} = \frac{(2.626x479) - (2.075x322)}{(2.626x479) + (2.075x322)} = \frac{589.704}{1.926.004} = \mathbf{0,306}$$

O resultado é de um coeficiente de +30,6% de associação entre a variável dicotômica "tem nível de escolaridade superior" e a variável eleito em município com "tamanho acima de 10 mil habitantes". Como podemos perceber, a diferença entre a associação entre variáveis dicotomizadas (Q_{xy}) e as variáveis agregadas (em três categorias de municípios e duas de escolaridade) foi menor que no exemplo anterior, quando todas as categorias originais foram calculadas. Isso significa que na transferência de seis categorias de tamanho de município para três e de quatro categorias de escolaridade para duas a redução na qualidade das informações foi maior que na dicotomização das mesmas.

Quando devemos optar por Q_{xy} ou Gama? Não há uma fórmula para definir qual o melhor coeficiente. Sob o ponto de vista da facilidade do cálculo, Q_{xy} é mais simples e, portanto, preferível. No entanto, o Gama nos permite identificar sutilezas nas variações entre as categorias que podem ficar "escondidas" quando transformamos a variável em dicotômica. O fato é que se a ta-

bela de contingência apresentar uma distribuição dos casos muito distinta entre as categorias vizinhas, a agregação das mesmas em variável dicotômica pode "esconder" importantes diferenças que deixarão de existir no cálculo do coeficiente. Nesses casos, apesar da maior dificuldade para o cálculo, recomenda-se o uso do Gama.

Agora, se o seu objetivo for verificar em detalhes as diferenças de frequências por pares de ocorrências, então, precisa fazer uma análise de células na tabela de contingência. Ou quando se está trabalhando com duas variáveis nominais a ordenação da direção da relação não faz nenhum sentido analítico. Nesses casos, o ideal é analisar as distribuições nas células. Para isso são usados os cálculos de resíduos e resíduos padronizados, respectivamente. É o que veremos no próximo tópico.

4.2.3. EXERCÍCIO

4.2.3.a. O quadro a seguir faz parte do relatório da Pesquisa CNI-Ibope de Avaliação de Governo (Set. 2013). O instituto ouviu 2.002 eleitores entre 14 e 17 de setembro de 2013. Os percentuais da amostra por grau de instrução são (28% até 4 do fundamental, 21% de 5 a 8 do fundamental, 36% de ensino médio e 15% de ensino superior). O quadro selecionado apresenta os percentuais de respostas para avaliação do governo Dilma e grau de instrução (desconsidere as variáveis sexo e idade). Construa uma tabela de contingência com as frequências das duas variáveis ordinais (Avaliação do governo Dilma e Grau de Instrução do Respondente) desconsiderando as "não respostas" (não sabe/não respondeu), calcule o Coeficiente Gama para a associação entre as duas variáveis e interprete os resultados.

3.1 Segmentação por sexo, idade e grau de instrução - % respostas												
	TOTAL	Sexo		Idade					Grau de instrução			
		Masc.	Fem.	16 a 24	25 a 34	35 a 44	45 a 54	55 e mais	Até 4ª do fund.	5ª a 8ª do fund.	Ensino médio	Superior
Avaliação do governo Dilma												
Ótimo	6	6	6	5	5	6	9	7	8	8	5	4
Bom	31	31	30	28	29	30	32	33	35	31	28	28
Regular	29	38	40	38	42	41	26	37	33	40	42	41
Ruim	11	12	11	15	10	11	10	11	11	9	11	16
Péssimo	11	12	11	13	12	11	11	10	10	11	14	9
Não sabe/Não respondeu	1	1	1	1	1	1	1	2	2	1	1	1

4.3 - TESTE DE ASSOCIAÇÃO ENTRE VARIÁVEIS NOMINAIS (RESÍDUOS BRUTOS E RESÍDUOS PADRONIZADOS)

4.3.1. TABELAS DE CONTINGÊNCIA

Como vimos no tópico anterior, o coeficiente Gama considera as direções das categorias ordinais para o estabelecimento das somas de valores de pares consistentes e de pares inconsistentes para identificação da existência ou não de associação entre as variáveis. Porém, quando estamos analisando as relações entre categorias de variáveis nominais, não existe uma organização ordinal e transitiva entre as categorias. Logo, levar em conta a posição da categoria nas linhas ou nas colunas em relação a seus "vizinhos" não faz sentido nesse caso. Se uma das variáveis nominais tiver mais que duas categorias, não podemos usar o Q_{xy}, pois sua fórmula só leva em conta os valores de tabelas quádruplas. Então, sobra-nos o teste χ^2 para encontrar associações entre variáveis categóricas nominais. No entanto, quando usamos o teste χ^2 temos apenas um coeficiente que indica a existência de relação entre pelo menos duas das categorias das variáveis testadas. Mas, pode ser que queiramos uma indicação mais precisa sobre quais daquelas categorias apresentam relações mais fortes, ou seja, quais contribuem para a rejeição da hipótese de independência entre as variáveis. Para conhecer o peso das relações entre cada par de categorias, usa-se a análise de resíduos em tabelas de contingência.

Uma tabela de contingência é uma tabela que sumariza as frequências de ocorrências para cada par de categorias das variáveis X e Y. O conceito de independência entre as variáveis parte do princípio que a distribuição observada das frequências nas casas da tabela de contingência é muito próxima da distribuição esperada das frequências. Se houver diferenças entre a distribuição esperada e a observada podemos começar a pensar em rejeitar a hipótese nula de independência e pensar na existência de alguma associação entre as categorias das duas variáveis testadas. Um exemplo de tabela de contingência para duas variáveis nominais é a que segue, entre ideologia do partido do eleito por região do país.

TABELA 4.6. DISTRIBUIÇÃO DO NÚMERO DE ELEITOS POR IDEOLOGIA PARTIDÁRIA E REGIÃO DO PAÍS

Ideologia	Região					Total
	Norte	Nordeste	Sudeste	Sul	Centro-oeste	
Esquerda	86	545	357	303	85	1376
Centro	159	400	568	414	177	1718
Direita	156	593	485	400	167	1801
Total	401	1538	1410	1117	429	4895

Aplicando a fórmula do χ^2 teríamos um coeficiente de χ^2= 110,55 para as distribuições de frequências das duas variáveis. Considerando que temos 8 graus de liberdade (3-1 x 5-1 = 8), se olharmos na tabela padronizada do χ^2 no anexo II perceberemos que o limite crítico para o graus de liberdade e intervalo de confiança de 95% é de 15,50, portanto o valor χ^2 = 110,55 fica muito acima do limite crítico. Isso nos permite rejeitar a hipótese nula e aceitar a possibilidade de que as duas variáveis não são independentes. Nesse caso, podemos considerar que a relação entre as duas variáveis contingenciadas não é aleatória, pois o χ^2 aponta para uma possibilidade muito abaixo do limite crítico para a aceitação da aleatoriedade. No entanto, temos que parar as análises por aqui. O coeficiente não nos permite especular sobre porque essa associação ocorre, por exemplo. Outra questão que fica sem resposta é se existe associação entre todos os pares de categorias das variáveis ou em apenas parte deles. Para complementar as informações fornecidas pelo χ^2 é possível analisar uma tabela de resíduos contingenciados, também chamados de resíduos brutos.

4.3.2. CÁLCULO DOS RESÍDUOS BRUTOS (R_B)

O leitor atento já percebeu que os resíduos brutos nada mais são do que a diferença entre a Frequência Observada (F_o) e a Frequência Esperada (F_e). Os resíduos ajudam a evitar erros comuns na interpretação dos valores observados, pois quando as marginais não têm os mesmos valores, as frequências totais podem ser enganosas. Por exemplo, na tabela acima podemos olhar para a linha dos partidos de direita e comparar com a coluna da região sul concluindo que a direita elegeu menos que o centro nessa região (400 da direita no sul contra 414 do centro no sul). No entanto, se considerarmos as diferenças das marginais das linhas, perceberemos que proporcionalmente ao total de prefeito de direita e aos prefeitos de centro, a participação dos partidos de

direita no sul foi maior do que a participação dos partidos de centro. Por definição, o resíduo bruto de uma casa é a diferença entre a F_o e F_e. Na linguagem matemática seria:

$$R_b = F_o - F_e$$

Já a Frequência esperada é calculada da seguinte forma:

$$F_e = \frac{M.Linha\ x\ M.Coluna}{N}$$

Onde:
M.Linha : marginal da linha;
M.Coluna : marginal da coluna;
N : total de casos na tabela.

Para o exemplo anterior, a Fe do par "partido de esquerda" e "região norte" seria:

$$F_e = \frac{M.Linha\ x\ M.Coluna}{N} = \frac{1376\ x\ 401}{4895} = 112{,}72$$

E o resíduo bruto para esse par de casos seria:

$$R_b = F_o - F_e = 86 - 112{,}72 = -26{,}72$$

A interpretação desse resultado é que na distribuição observada os partidos de esquerda tiveram quase 27 eleitos a menos na região norte do que se esperaria se as distribuições fossem independentes.

	Frequência Esperada					**Resíduos Brutos**				
	norte	nordeste	sul	sudeste	Centro-oeste	norte	nordeste	sul	sudeste	Centro-oeste
Esquerda	112,72	432,34	396,36	313,99	120,59	-26,72	112,66	-39,36	-10,99	-35,59
Centro	140,74	539,79	494,87	392,03	150,57	18,26	-139,79	73,13	21,97	26,43
Direita	147,54	565,87	518,78	410,97	157,84	8,46	27,13	-33,78	-10,97	9,16

QUADRO 4.6. FREQUÊNCIA ESPERADA E RESÍDUOS BRUTOS DOS VALORES DA TAB. 4.6.

Agora, tendo gerado os resíduos brutos já podemos identificar as diferenças nas concentrações dos valores. Olhando para a tabela como um todo, percebemos que os maiores resíduos brutos

estão na região nordeste, onde partidos de esquerda elegeram 112 prefeito a mais do que o esperado e partidos de direita elegeram 139 prefeitos a menos que o esperado. A leitura também pode ser feita nas linhas. Por exemplo, os partidos de esquerda elegeram mais que o esperado no nordeste e menos em todas as outras regiões. O contrário dos partidos de centro, que elegeram menos que o esperado no nordeste e mais em todas as demais regiões. Os partidos de direita elegeram menos no sul e sudeste e tiveram concentrações positivas nas outras três regiões. A mesma leitura pode ser feita a partir das colunas, comparando as ideologias partidárias dentro de cada região do País.

A vantagem da análise dos resíduos brutos é que os valores estão na unidade de análise, no caso, em número de prefeitos eleitos. Por outro lado, isso também é uma desvantagem, pois ao estar na mesma dimensão da unidade de análise, não permite comparações diretas sobre as diferenças de magnitudes. O problema dos resíduos brutos é serem pouco informativos, pois não apresentam variância constante. Em outras palavras, são não-padronizados e não permitem a verificação de pontos extremos (*outliers*) por não poderem ser comparados diretamente. Para resolver esse problema, costuma-se padronizar os resíduos. Com isso, torna-se possível verificar quem são as relações com casos extremos e quais são as maiores concentrações de casos, em relações adimensionais.

Até padronizarmos os resíduos não é possível saber quais são os resíduos com tamanho suficiente para serem considerados estatisticamente significativos e quais estão abaixo do limite crítico. O χ^2 da tabela de contingência mostrou que as variáveis estão associadas. Os resíduos brutos indicaram diferenças que em alguns pares chegam a ser dez vezes maiores que em outros pares. Se por um lado os partidos de centro apresentaram o maior resíduo bruto (-139,79), os partidos de direita no norte ficaram com o menor resíduo bruto (+8,46). Será que os dois podem ser considerados válidos no teste de associação entre as variáveis. Até que ponto os resíduos podem ser usados como indicadores de diferenças significativas? Para responder a essa pergunta é preciso padronizar os valores dos resíduos e, partir de um limite pré-estabelecido, indicar se os resíduos são significativos ou não. Para isso existem os chamados Resíduos Padronizados (R_p), que são padronizados em valores *Z-score* para permitir a identificação dos pares que estão acima do limite crítico e, portanto, apresentam "acúmulos" de frequências acima ou abaixo do que seria esperado se a distribuição dos casos entre as variáveis fosse independente. Porém, antes de entrarmos nas explicações específicas do Resíduo Padronizado é preciso um lembrete: só faz sentido calcular o resíduo padronizado quando o resultado do χ^2 de uma tabela de contingência é significativo. Se o resultado não for estatisticamente significativo, todos os valores de Resíduos Padronizados ficarão abaixo do limite crítico, ou seja, não serão significativos. Porém, se tivermos um χ^2 significativo em uma tabela de contingência devemos calcular os resí-

duos padronizados para identificar quantos e quais pares de casos estão acima do limite crítico, quer dizer, concentram mais casos do que o esperado se as variáveis fossem independentes.

4.3.3 CÁLCULO DOS RESÍDUOS PADRONIZADOS (R_P)

A análise de resíduos padronizados nada mais é do que a verificação dos valores que representam a relação biunívoca (nas duas direções) com probabilidade de chances de ocorrências. Ou seja, são os valores que sobram (para mais ou para menos) quando a distribuição entre o valor observado e o esperado não é aleatória.

Ao se estabelecer 95% de intervalo de confiança, essas chances de ocorrência são de ± 1,96, valor que serve de ponto de corte para o nível de significância de falta ou excesso de ocorrência entre as variáveis, o que permite distinguir os valores de pares casuais dos não-casuais.

Como o valor na tabela z-score para o intervalo de confiança de 95% é de 1,96, pode-se considerar que valores de resíduos padronizados acima de +1,96 ou abaixo de –1,96 apresentam excessos de casos significativos, sendo, portanto, responsáveis pelas relações não-aleatórias apontadas pelo coeficiente χ^2.

O cálculo dos resíduos padronizados é bastante simples e quase intuitivo. Os valores são padronizações, ou seja, transformações adimensionais dos resíduos brutos. Todo resíduo, seja ele bruto ou padronizado, serve para indicar as diferenças entre o valor observado e o valor esperado em uma distribuição de frequências. Um resíduo padronizado (R_P) é calculado a partir da padronização dos resíduos brutos para que passem a ter variância igual e apresentem-se de maneira adimensional, passando a ser um coeficiente. Por ser padronizado, o R_P apresenta variância constante, o que permite a comparação direta entre os valores. Se a análise é feita a partir de uma grande amostra (n > 120) e intervalo de confiança de 95% (z = 1,96), qualquer resíduo acima de 1,96 deve ser considerado estatisticamente significativo, ou seja, o resíduo encontrado naquela relação é maior do que supunha a hipótese de independência entre as variáveis. Podemos calcular os resíduos padronizados em tabelas de contingência de variáveis categóricas a partir da seguinte fórmula:

$$R_p = \frac{R_b}{\sqrt{F_e}}$$

Onde:

R_P : resíduo padronizado

R_b : resíduo bruto

F_e : frequência esperada

Usando o mesmo exemplo das distribuições de prefeitos eleitos por ideologia partidária e região do País, temos que para a primeira célula da tabela de contingência - L1, C1 (partido de esquerda, região norte) -, o seguinte cálculo de Resíduo Padronizado:

$$R_{p(l1,c1)} = \frac{R_b}{\sqrt{F_e}} = \frac{-26{,}72}{\sqrt{112{,}72}} = -2{,}51$$

Considerando o limite crítico de 1,96, podemos afirmar que os resíduos negativos de partidos de esquerda na região norte são estatisticamente significativos, ou seja, a esquerda elegeu menos prefeitos no norte de fato. A tabela a seguir mostra todos os resultados de Resíduos padronizados para as duas variáveis. Nela podemos perceber que nem todos os resíduos são significativos, portanto, há pares de categorias que devem ser consideradas independentes, ainda que apresentem resíduos brutos. Dos 15 pares de categorias, mais da metade, oito deles ficam abaixo do limite crítico de ±1,96, portanto, não-significativas estatisticamente. Os resíduos significativos e positivos são três: a esquerda no nordeste (5,418), o centro no sul (2,287) e o centro no centro-oeste (2,154). Os resíduos significativos negativos também são três: esquerda no norte (-2,517), centro no nordeste (-6,017) e esquerda no centro-oeste (-3,241). Há ainda um resíduo que fica muito próximo do limite crítico, a esquerda no sul (-1,977), sendo necessários outros testes ou comparações com indicadores complementares para definir se essa é uma variação significativa ou não.

TABELA 4.7. RESÍDUOS PADRONIZADOS PARA IDEOLOGIA DO PARTIDO DO PREFEITO ELEITO POR REGIÃO

Ideologia	Região				
	norte	nordeste	sul	sudeste	Centro-oeste
Esquerda	**-2,517**	**5,418**	-1,977	-0,620	**-3,241**
Centro	1,539	**-6,017**	**3,287**	1,109	**2,154**
Direita	0,697	1,140	-1,483	-0,541	0,729

Para facilitar a visualização dos resultados, os resíduos padronizados significativos positivos foram marcados em azul e os negativos em vermelho. Os demais ficaram abaixo do limite crítico. Com isso, podemos dizer que do ponto de vista das regiões do país, apenas o sudeste apresentou uma distribuição próxima do esperado, ou seja, com proporções de eleitos parecidas entre as três posições ideológicas. No sul e no centro-oeste predominaram os partidos de centro

e no nordeste os partidos de esquerda. Já no norte houve uma participação menor dos partidos de esquerda. Se olharmos para os resíduos padronizados a partir das linhas, perceberemos que apenas os partidos de direita distribuíram de maneira equitativa seus eleitos em todas as regiões do País.

Em resumo, os resíduos padronizados foram necessários para a identificação individualizada da concentração de valores em pares de casos - acima ou abaixo - do esperado. Até então o que tínhamos feito era encontrar um coeficiente que representasse o conjunto das relações entre todos os pares de casos, normalmente partindo do produto de pares consistentes x pares inconsistentes.

4.3.3.1 RESÍDUOS PADRONIZADOS PARA ANÁLISES TEMPORAIS

Uma das principais limitações das técnicas quantitativas de análises temporais (as chamadas séries temporais) é a necessidade de um número mínimo de observações no tempo muito alto. Normalmente acima de 120 pontos observados ao longo do tempo para permitir uma análise estatística consistente. Muitas vezes isso não é possível em análises eleitorais, pois a distância entre as medições é grande, bianual, quadrienal ou até mais. Portanto, precisaríamos de dois séculos ou mais com dados disponíveis para podermos usar as técnicas tradicionais nesse caso. A tabela de contingência organizada em ordem temporal e os Resíduos Padronizados podem substituir as técnicas de séries temporais com a vantagem de ser possível trabalhar com poucos pontos no tempo.

Quando comparados entre si, os resíduos padronizados em uma tabela de contingência mostram as diferenças relativas entre cada par de categorias. Se uma das variáveis for temporal, a transição de uma categoria para outra indica uma mudança no tempo. Assim, diferenças de resíduos apontam para maior ou menor concentração de casos em determinado momento do tempo. Mas, atenção, os resíduos não são capazes de indicar quanto da mudança no tempo seguinte (t_1) é consequência ou "memória" da quantidade da característica no tempo anterior (t_0). Para conhecer a proporção da característica que influencia o tempo seguinte apenas usando as técnicas de análise de séries temporais que decompõem os valores - nesse caso, dependendo do número mínimo de observações no tempo.

Usaremos para exemplificar o cálculo dos resíduos padronizados a distribuição das emendas parlamentares (em milhões de R$) no Congresso Nacional brasileiro entre 1996 e 1999 por tipo de proponente. Aqui, serão utilizados grupos de proponentes: o deputado “relator” do orçamento em cada ano; as emendas coletivas de “bancadas estaduais e regionais”, as emendas a-

presentadas pela "comissão orçamentária" e as emendas "individuais" apresentadas pelos deputados e senadores. A tabela 4.8 mostra que há um constante crescimento nos totais de emendas nos quatro anos analisados, passando de R$ 1,5 bilhão em 1996 para R$ 3,1 bilhões em 1999. Nossa questão aqui é saber se esse aumento foi proporcional a todas as categorias de origem das emendas ou não. A princípio, olhando os valores brutos (N) na tabela 4.8 podemos perceber que as emendas de relatores foram as que menos cresceram, enquanto as demais categorias praticamente dobraram os valores em quatro anos. Como as diferenças entre os totais não são proporcionais, ou seja, não há padronização, esses números podem ser enganosos quando usados em comparações diretas. Por isso faremos a análise de resíduos padronizados para cada ano. A tabela 4.8 apresenta todos os resultados para cada categoria de origem X ano. Na linha (N) está o valor, em milhões de Reais, do total de emendas de cada categoria de origem. A linha (R_b) indica o resíduo bruto para as categorias e a linha (R_p) indica os resíduos padronizados. Como exemplo, apresentaremos apenas os cálculos para o primeiro par de categorias: valor da emenda apresentada pelo relator em 1996. Os demais seguem os mesmos passos:

1º Passo (encontrar a Frequência Esperada)

$$F_e = \frac{Ml \; x \; Mc}{N} = \frac{2.138 \; x \; 1.557{,}96}{9.757{,}50} = \mathbf{341,75}$$

2º Passo (encontrar o Resíduo bruto)

$$R_b = F_o - F_e = 409{,}73 - 341{,}75 = \mathbf{67{,}98}$$

3º Passo (encontrar o Resíduo Padronizado)

$$R_p = \frac{Rb}{\sqrt{Fe}} = \frac{67.98}{\sqrt{341{,}75}} = \mathbf{+3{,}677}$$

Esse mesmo procedimento deve ser repetido para todas as demais relações entre as categorias das variáveis analisada aqui.

TAB. 4.8 – EMENDAS PARLAMENTARES POR TIPO DE PROPONENTE EM MILHÕES R$ (1996 A 1999)

Origem da emenda		1996		1997		1998		1999		Total
		Fo	Fe	Fo	Fe	Fo	Fe	Fo	Fe	
Relator	N	409,73	341,75	645,12	500,05	495,53	614,13	587,85	682,30	2.138,25
	Rb	67,98		145,06		-118,60		-94,44		
	Rp	**3,677**		**6,487**		**-4,785**		**-3,615**		
Bancadas (estadual e regional)	N	797,66	795,18	1.037,21	1163,53	1.522,99	1428,96	1.617,38	1587,56	4.975,25
	Rb	2,47		-126,32		94,02		29,81		
	Rp	0,087		**-3,703**		**2,487**		0,748		
Comissão	N	20,25	76,44	66,10	111,85	167,97	137,37	223,94	152,61	478,28
	Rb	-56,19		-45,74		30,60		71,32		
	Rp	**-6,426**		**-4,325**		**2,611**		**5,773**		
Individual (Deputados e Senadores)	N	330,28	344,54	531,14	504,14	613,11	619,15	681,16	687,87	2.155,71
	Rb	-14,26		27,00		-6,03		-6,70		
	Rp	-0,768		1,202		-0,242		-0,255		
Total		1.557,93		2.279,59		2.799,62		3.110,36		9.747,50

Fonte: Assessoria de orçamento e fiscalização orçamentária da Câmara dos deputados

Como já vimos, resultados acima de ± 1,96 para Resíduos Padronizados devem ser considerados casos significativos. Se positivo, significa que aquele par de categorias apresenta mais casos do que deveria caso as variáveis fossem independentes. Se negativo, ele concentra menos casos do que seria esperado. Na tabela acima os resíduos em negrito são os positivos significativos, os em vermelho são os negativos significativos e os sem nenhuma indicação, ficaram abaixo de ± 1,96. Agora podemos comparar as diferenças proporcionais entre as categorias de origem por ano, ou seja, ao longo do tempo - as variações estão padronizadas.

Duas categorias de origem apresentaram comportamentos distintos naquele mandato. As emendas de relator diminuíram significativamente de volume ao longo do tempo, passando de R_p +3,667 para - 3,615 entre 1996 e 1999. Já as emendas de comissão apresentaram resíduos crescentes ao longo do tempo, R_p de -6,426 para +5,773. O volume de emendas individuais não apresentou mudança significativa nos quatro anos, com todos os coeficientes de R_p ficando abaixo de ± 1,96. As emendas de bancada tiveram uma participação negativa em 1997 (- 3,703) e positiva em 1998 (+ 2,487). Nos outros anos as variações ficaram abaixo do limite crítico. Ou seja, com o auxílio dos resíduos padronizados podemos dizer que entre 1996 e 1999 os relatores do projeto de orçamento perderam capacidade de aprovar suas emendas, enquanto a comissão

orçamentária ganhou força. Isso é uma maneira de analisar as mudanças ao longo do tempo. Já as emendas de bancadas e individuais mantiveram os mesmos volumes de emendas apresentadas nos quatro anos analisados. A comparação proporcional direta entre os anos seria impossível sem a padronização dos valores.

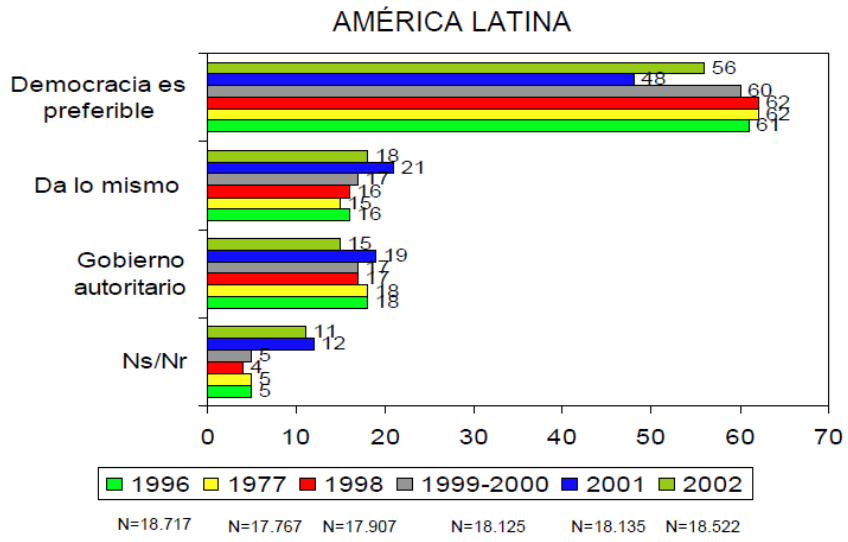
O objetivo desse curso foi apresentar algumas técnicas simples para cálculo de coeficientes específicos para variáveis categóricas e para dados secundários, a partir de tabelas de frequência ou de contingência. Essas ferramentas são úteis para o pesquisador que pretende, ou precisa, trabalhar com informações extraídas de relatórios ou publicações sobre os quais não é possível acessar o banco de dados primário. Recomendo que você procure sempre reunir o máximo possível dessas técnicas em um procedimento de análise. Dependendo do tipo de variável faça um teste de χ^2 ou coeficiente Gama agregado a uma análise de Resíduos Padronizados. Com isso, você poderá tirar conclusões não apenas a respeito da relação entre as variáveis, mas também para os pares de relações entre as categorias. Quando as variáveis forem dicotômicas ou for viável do ponto de vista lógico dicotomizá-las sem que com isso haja uma redução grande da precisão dos resultados, faça testes de Q_{xy}, que são simples, rápidos e fornecem coeficientes bastante explicativos. Por serem aplicados em tabelas quádruplas, os testes de Q_{xy} dispensam a análise de resíduos individuais. E, o mais importante: o conjunto de testes apresentado aqui não dá conta de uma ínfima parte do conjunto de ferramentas estatísticas disponíveis para análise de dados categóricos. Se um dia você se defrontar com a necessidade de um tipo de ferramenta diferente para medir a relação entre pares de categorias que não foi apresentada aqui, tenha certeza que o mais provável é que ela exista e eu é que não a conheço.

4.3.4. EXERCÍCIOS

4.2.4.a. Usando as mesmas informações do exercício anterior (4.2.3.a), calcule os Resíduos Padronizados para as categorias de "avaliação do governo Dilma" e "escolaridade do eleitor". Interprete os resultados.

4.2.4.b. A partir do gráfico abaixo construa uma tabela de contingência ordinal para "opinião sobre democracia" e "tempo" para as médias das opiniões dos eleitores latino-americanos sobre o assunto. Desconsidere as "não respostas" na tabela de contingência. Em seguida calcule os Resíduos Padronizados e interprete como se comportou a opinião do latino-americano sobre a democracia entre os anos de 1996 e 2002.

DEMOCRACIA-AUTORITARISMO-INDIFERENCIA



P. ¿Con cuál de las siguientes frases está Ud. más de acuerdo? La democracia es preferible a cualquier otra forma de gobierno. En algunas circunstancias, un gobierno autoritario puede ser preferible a uno democrático. A la gente como uno, nos da lo mismo un régimen democrático que uno no democrático

Fuente: LATINOBARÓMETRO 1996 - 2002

REFERÊNCIAS SUGERIDAS SOBRE ANÁLISE DE DADOS CATEGÓRICOS

BABBIE, Earl (2005). **Métodos de Pesquisas de Survey**. Belo Horizonte – MG: Editora UFMG.

BAUER, M. W. & GASKELL, G. (2003). **Pesquisa Qualitativa Com Texto, Imagem e Som: um manual prático**. Petrópolis – RJ: Editora Vozes.

BENZÉCRI, J. P. (1992) ***Correspondence analysis handbook***. New York: Marcel Dekker.

DANTAS, Carlos (2004). **Probabilidade: um curso introdutório**. São Paulo: Edusp.

DAVIS, J. A. (1976). **Levantamento de Dados em Sociologia: uma análise estatística elementar**. Rio de Janeiro – RJ: Zahar Editores.

GÜNTHER, H. (2003) **Como Elaborar um Questionário** (Série Planejamento de pesquisa nas Ciências Sociais, nº 01) Brasília: DF. UNB.

KENDALL, Patrícia L. & LAZARSFELD, Paul F (1950). **Problem of Survey Analysis**. In MERTON, Robert & LAZARSFELD, Paul F. (orgs.). Continuities in Social Research. New York: Free Press.

MAHONEY, James & GOERTZ, Gary (2006). **A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research.**Political Analysis Review, nº 14. p. 227 a 249.

MINGOTI, Sueli Ap (2013). **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG.

PEREIRA, Júlio C. R. (2004) **Análise de Dados Qualitativos – estratégias metodológicas para as Ciências da Saúde, Humanas e Sociais**. São Paulo: EdUsp/Fapesp.

RAGIN, Charles C (1994). **Constructing Social Research: the unit and diversity of method.**Pine Forge Press: Thousand Oaks.

ROSENBERG, Morris (1971).**A Lógica da Análise do Levantamento de Dados**. São Paulo: Editora Cultrix/Editora da Universidade de São Paulo.

SELLTIZ, C.; WRIGHTSMAN, L. S.; COOK, S. W (1981). **Métodos de pesquisa nas relações sociais**. 2 ed. São Paulo: EPU.

ANEXO I

Dados

Tab. A. - Número de eleitos, Z-score e eleitos por sexo por partido

POSIÇÃO	PARTIDO	ELEITOS	Z-SCORE	HOMEM	MULHER
centro	PMDB	1024	3,04	902	122
centro	PSDB	694	1,81	601	93
esquerda	PT	628	1,56	558	70
direita	PSD	495	1,06	438	57
direita	PP	464	0,94	417	47
esquerda	PSB	440	0,85	389	51
esquerda	PDT	308	0,36	284	24
direita	PTB	294	0,31	260	34
direita	DEM	276	0,24	248	28
direita	PR	272	0,22	235	37

Tab. B - Grau de instrução dos eleitos por partido

POSIÇÃO	PARTIDO	Lê e escreve	Fundamental	Médio	Superior
centro	PMDB	14	159	313	538
centro	PSDB	3	83	199	409
esquerda	PT	2	66	160	400
direita	PSD	2	65	189	239
direita	PP	3	70	143	248
esquerda	PSB	5	43	128	264
esquerda	PDT	1	47	90	170
direita	PTB	2	51	89	152
direita	DEM	3	44	97	132
direita	PR	2	49	94	127

Tab. C - Tamanho do município dos eleitos por partido

POSIÇÃO	PARTIDO	Até 5000	5001 até 10000	10001 até 20000	20001 até 50000	50001 até 100000	100001 até 500000	Maior que 500000
centro	PMDB	273	255	236	166	55	37	2
centro	PSDB	184	148	157	127	41	34	3
esquerda	PT	100	154	145	135	50	39	5
direita	PSD	99	112	151	97	19	17	0
direita	PP	134	109	113	77	17	14	0
esquerda	PSB	80	99	113	100	29	17	2
esquerda	PDT	69	59	80	65	23	11	1
direita	PTB	92	64	61	53	17	7	0
direita	DEM	67	81	67	46	10	3	2
direita	PR	68	66	73	47	13	5	0

Tab. D - Número de eleitos por partido e região do País

POSIÇÃO	PARTIDO	Norte	Nordeste	Sudeste	Sul	Centro-Oeste
centro	PMDB	92	285	243	294	110
centro	PSDB	67	115	325	120	67
esquerda	PT	51	186	194	158	39
direita	PSD	68	209	67	90	61
direita	PP	23	102	107	210	22
esquerda	PSB	27	268	87	34	24
esquerda	PDT	8	91	76	111	22
direita	PTB	17	107	106	44	20
direita	DEM	12	80	113	37	34
direita	PR	36	95	92	19	30

Tab. E. Distribuição da Escolaridade dos eleitos por Tamanho do Município

	Até 5000	5001 até 10000	10001 até 20000	20001 até 50000	50001 até 100000	100001 até 500000	TOTAL
Superior	552	637	734	675	237	172	3007
Médio	454	432	449	275	60	24	1694
Fundamental	245	215	169	97	18	10	754
Lê e escreve	7	12	19	7	1	1	47
TOTAL	1258	1296	1371	1054	316	207	5502

SEXO	POS. IDEOL.	LÊ E ESCREVE	FUNDAMENTAL	MÉDIO	SUPERIOR	TOTAL
HOMEM	ESQUERDA	8	152	384	750	1294
	CENTRO	16	226	471	790	1503
	DIREITA	21	336	716	1013	2086
	Total	45	714	1571	2553	4883
MULHER	ESQUERDA	0	6	17	129	152
	CENTRO	1	16	41	157	215
	DIREITA	1	18	66	182	267
	Total	2	40	124	468	634

ANEXO II

Valores Padronizados para Distribuição do Qui-quadrado

GL	0.995	0.975	0.900	0.500	0.100	0.050	0.025	0.01	0.005	0.001
1	0.000	0.001	0.016	0.455	2.706	3.841	5.024	6.635	7.879	10.827
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	13.815
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	16.266
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	18.466
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	20.515
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	22.457
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	24.321
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	26.124
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	27.877
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	29.588
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757	31.264
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300	32.909
13	3.565	5.009	7.041	12.340	19.812	22.362	24.736	27.688	29.819	34.527
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319	36.124
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801	37.698
16	5.142	6.908	9.312	15.338	23.542	26.296	28.845	32.000	34.267	39.252
17	5.697	7.564	10.085	16.338	24.769	27.587	30.191	33.409	35.718	40.791
18	6.265	8.231	10.865	17.338	25.989	28.869	31.526	34.805	37.156	42.312
19	6.844	8.907	11.651	18.338	27.204	30.144	32.852	36.191	38.582	43.819
20	7.434	9.591	12.443	19.337	28.412	31.410	34.170	37.566	39.997	45.314
21	8.034	10.283	13.240	20.337	29.615	32.671	35.479	38.932	41.401	46.796
22	8.643	10.982	14.041	21.337	30.813	33.924	36.781	40.289	42.796	48.268
23	9.260	11.689	14.848	22.337	32.007	35.172	38.076	41.638	44.181	49.728
24	9.886	12.401	15.659	23.337	33.196	36.415	39.364	42.980	45.558	51.179
25	10.520	13.120	16.473	24.337	34.382	37.652	40.646	44.314	46.928	52.619
26	11.160	13.844	17.292	25.336	35.563	38.885	41.923	45.642	48.290	54.051
27	11.808	14.573	18.114	26.336	36.741	40.113	43.195	46.963	49.645	55.475
28	12.461	15.308	18.939	27.336	37.916	41.337	44.461	48.278	50.994	56.892
29	13.121	16.047	19.768	28.336	39.087	42.557	45.722	49.588	52.335	58.301
30	13.787	16.791	20.599	29.336	40.256	43.773	46.979	50.892	53.672	59.702
31	14.458	17.539	21.434	30.336	41.422	44.985	48.232	52.191	55.002	61.098

32	15.134	18.291	22.271	31.336	42.585	46.194	49.480	53.486	56.328	62.487
33	15.815	19.047	23.110	32.336	43.745	47.400	50.725	54.775	57.648	63.869
34	16.501	19.806	23.952	33.336	44.903	48.602	51.966	56.061	58.964	65.247
35	17.192	20.569	24.797	34.336	46.059	49.802	53.203	57.342	60.275	66.619
36	17.887	21.336	25.643	35.336	47.212	50.998	54.437	58.619	61.581	67.985
37	18.586	22.106	26.492	36.336	48.363	52.192	55.668	59.893	62.883	69.348
38	19.289	22.878	27.343	37.335	49.513	53.384	56.895	61.162	64.181	70.704
39	19.996	23.654	28.196	38.335	50.660	54.572	58.120	62.428	65.475	72.055
40	20.707	24.433	29.051	39.335	51.805	55.758	59.342	63.691	66.766	73.403
41	21.421	25.215	29.907	40.335	52.949	56.942	60.561	64.950	68.053	74.744
42	22.138	25.999	30.765	41.335	54.090	58.124	61.777	66.206	69.336	76.084
43	22.860	26.785	31.625	42.335	55.230	59.304	62.990	67.459	70.616	77.418
44	23.584	27.575	32.487	43.335	56.369	60.481	64.201	68.710	71.892	78.749
45	24.311	28.366	33.350	44.335	57.505	61.656	65.410	69.957	73.166	80.078
46	25.041	29.160	34.215	45.335	58.641	62.830	66.616	71.201	74.437	81.400
47	25.775	29.956	35.081	46.335	59.774	64.001	67.821	72.443	75.704	82.720
48	26.511	30.754	35.949	47.335	60.907	65.171	69.023	73.683	76.969	84.037
49	27.249	31.555	36.818	48.335	62.038	66.339	70.222	74.919	78.231	85.350
50	27.991	32.357	37.689	49.335	63.167	67.505	71.420	76.154	79.490	86.660
51	28.735	33.162	38.560	50.335	64.295	68.669	72.616	77.386	80.746	87.967
52	29.481	33.968	39.433	51.335	65.422	69.832	73.810	78.616	82.001	89.272
53	30.230	34.776	40.308	52.335	66.548	70.993	75.002	79.843	83.253	90.573
54	30.981	35.586	41.183	53.335	67.673	72.153	76.192	81.069	84.502	91.871
55	31.735	36.398	42.060	54.335	68.796	73.311	77.380	82.292	85.749	93.167
56	32.491	37.212	42.937	55.335	69.919	74.468	78.567	83.514	86.994	94.462
57	33.248	38.027	43.816	56.335	71.040	75.624	79.752	84.733	88.237	95.750
58	34.008	38.844	44.696	57.335	72.160	76.778	80.936	85.950	89.477	97.038
59	34.770	39.662	45.577	58.335	73.279	77.930	82.117	87.166	90.715	98.324
60	35.534	40.482	46.459	59.335	74.397	79.082	83.298	88.379	91.952	99.608

RESPOSTAS DOS EXERCÍCIOS

2. ANÁLISE DE DADOS CATEGÓRICOS: DEFINIÇÕES BÁSICAS

2.1.a. Não é possível. Uma variável nominal não apresenta entre suas diferentes categorias a característica da transitividade, ou seja, não há uma direção no crescimento ou queda dos valores relacionados entre as categorias que compõem uma variável nominal.

2.1.b. para calcular o ponto médio da distribuição, dividindo-a em duas partes iguais é preciso encontrar a mediana (Md=71)
Grupo 1 - menores notas - valores de 32 a 70 (código 0).
Grupo 2 - maiores notas - valores de 72 a 94 (código 1).

3.COEFICIENTES BÁSICOS

3.4.a. Antes de aplicar a fórmula, vamos fazer a tabela quádrupla apenas com os percentuais válidos para as duas categorias de cada variável.

	W.	**N-W**
Fav.	64	82
Unf.	31	15

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{\frac{64}{146}}{\frac{31}{46}} = \frac{0{,}438}{0{,}674} = 0{,}650$$

Interpretação: existe 0,65% de chance de encontrar um eleitor branco favorável a H. Clinton em relação aos eleitores não-brancos. Em outras palavras, temos uma chance maior de encontrar avaliação positiva do desempenho de H. Clinton no Senado entre os norte-americanos não-brancos do que entre os brancos. No entanto, como a proporção ficou abaixo de (1,5), não devemos considerar o resultado significativo.

3.4.b. O primeiro passo é encontrar a média de cada ano, ela será a Frequência esperada dos 10 municípios. Depois, calcula-se a diferença entre a frequência observada e a esperada, para em seguida dividir o quadrado da diferença pelo valor esperado. Repete-se o procedimento para cada uma das medições (anos), cf. tabelas abaixo.

Município	1905	Fo-Fe	$(fo-Fe)^2$	$(fo-fe)^2/fe$	1908	Fo-Fe	$(fo-Fe)^2$	$(fo-fe)^2/fe$
Andaraí	608	-21,09	444,79	0,71	686	-107,54	11564,85	14,57
Barreiras	715	85,91	7380,53	11,73	727	-66,54	4427,57	5,58
Cartinhanha	514	-115,09	13245,71	21,06	660	-133,54	17832,93	22,47
Lençóis	695	65,91	4344,13	6,91	850	56,46	3187,73	4,02
Maracás	384	-245,09	60069,11	95,49	442	-351,54	123580,37	155,73
Mucugê	626	-3,09	9,55	0,02	1185	391,46	153240,93	193,11
Pilão Arcado	235	-394,09	155306,93	246,88	435	-358,54	128550,93	162,00
Remanso	377	-252,09	63549,37	101,02	786	-7,54	56,85	0,07
Rio Preto	369	-260,09	67646,81	107,53	558	-235,54	55479,09	69,91
Sento Sé	492	-137,09	18793,67	29,87	492	-301,54	90926,37	114,58
				621,20				742,05
Total	5015,00				6821,00			
Média	629,09				793,55			

Município	1910	Fo-Fe	$(fo-Fe)^2$	$(fo-fe)^2/fe$	1912	Fo-Fe	$(fo-Fe)^2$	$(fo-fe)^2/fe$
Andaraí	686	-226,18	51157,39	56,08	999	9,37	87,80	0,09
Barreiras	1255	342,82	117525,55	128,84	1260	270,37	73099,94	73,87
Cartinhanha	660	-252,18	63594,75	69,72	813	-176,63	31198,16	31,53
Lençóis	848	-64,18	4119,07	4,52	848	-141,63	20059,06	20,27
Maracás	1053	140,82	19830,27	21,74	1071	81,37	6621,08	6,69
Mucugê	1185	272,82	74430,75	81,60	1185	195,37	38169,44	38,57
Pilão Arcado	509	-403,18	162554,11	178,20	480	-509,63	259722,74	262,44
Remanso	878	-34,18	1168,27	1,28	1007	17,37	301,72	0,30
Rio Preto	558	-354,18	125443,47	137,52	609	-380,63	144879,20	146,40
Sento Sé	492	-420,18	176551,23	193,55	702	-287,63	82731,02	83,60
				873,05				663,75
Total	8124,00				8974,00			
Média	912,18				989,64			

Município	1934	Fo-Fe	(fo-Fe)²	(fo-fe)²/fe
Andaraí	939	128,55	16525,10	20,39
Barreiras	1410	599,55	359460,20	443,53
Cartinhanha	301	-509,45	259539,30	320,24
Lençóis	644	-166,45	27705,60	34,19
Maracás	1353	542,55	294360,50	363,21
Mucugê	233	-577,45	333448,50	411,44
Pilão Arcado	627	-183,45	33653,90	41,52
Remanso	1000	189,55	35929,20	44,33
Rio Preto	78	-732,45	536483,00	661,96
Sento Sé	396	-414,45	171768,80	211,94
				2552,75
Total	6981,00			
média	810,45			

Resultado: Os valores encontrados para cada ano são sumarizados na tabela a seguir.

ano	χ**2**
1905	621,20
1908	742,05
1910	873,05
1912	663,75
1934	2552,75

Interpretação: Percebe-se que em todos os anos o χ**2** é bastante superior ao limite crítico de 9 GL., portanto, podemos dizer que há diferenças significativas no número de leitores entre os municípios em cada ano com dados disponíveis. Além disso, nos quatro primeiros anos, de 1905 a 1912 (república velha) a distribuições ficam parecidas, variando entre coeficiente de 621,20 e 873,05. Porém, em 1934 o valor sobe para 2.552,75, indicando uma heterogeneidade bem maior nesse ano do que nos anteriores.

3.4.c. Resultados

i) Cálculo do χ**2** para as duas tabelas quádruplas:

Câmara dos deputados:

a) $\chi^2 = \frac{(FO-FE)^2}{FE} = \frac{(468-451{,}06)^2}{415{,}06} = 6{,}75$

b) $\chi^2 = \frac{(FO-FE)^2}{FE} = \frac{(3.486-3.538{,}94)^2}{3.538{,}94} = 0{,}79$

c) $\chi^2 = \frac{(FO-FE)^2}{FE} = \frac{(45-97{,}94)^2}{97{,}94} = 28{,}62$

d) $\chi^2 = \frac{(FO-FE)^2}{FE} = \frac{(888-835{,}06)^2}{835{,}06} = 3{,}36$

$\chi^2 = \sum(6{,}75 + 0{,}79 + 28{,}62 + 3{,}36) = \mathbf{39{,}52}$

Assembleias legislativas:

a) $\chi^2 = \frac{(FO-FE)^2}{FE} = \frac{(921-837{,}18)^2}{837{,}18} = 8{,}38$

b) $\chi^2 = \frac{(FO-FE)^2}{FE} = \frac{(9.043-9.126{,}82)^2}{9.126{,}82} = 0{,}77$

c) $\chi^2 = \frac{(FO-FE)^2}{FE} = \frac{(138-221{,}82)^2}{221{,}82} = 31{,}67$

d) $\chi^2 = \frac{(FO-FE)^2}{FE} = \frac{(2.052-2.418{,}18)^2}{2.418{,}18} = 2{,}91$

$\chi^2 = \sum(8{,}38 + 0{,}77 + 31{,}67 + 2{,}91) = \mathbf{43{,}74}$

Interpretação: O coeficiente $\boldsymbol{\chi 2}$ para as relações entre sucesso eleitoral e sexo do candidato foram de 39,52 e 43,74 para Câmara de Deputados e Assembleias Legislativas estaduais em 2010, respectivamente. Isso significa que podemos rejeitar H_0 e aceitar que as variáveis não são independentes. Comparando os dois coeficientes, podemos dizer que as diferenças de desempenho por sexo dos candidatos são maiores nas Assembleias Legislativas do que na Câmara dos Deputados.

ii) Consultando na tabela de valores padronizados (anexo II), percebemos que o Limite Crítico para Intervalo de Confiança de 0,001, o mais exigente da tabela, e para 1 Grau de Liberdade é de 10,827. Com os coeficientes do teste ficaram muito acima desse limite, podemos considerar as

categorias das variáveis relacionadas, como já era esperado, pois no Brasil sabemos que homens são eleitos em maior proporção em relação aos candidatos apresentados do que as mulheres.

iii) Cálculo do coeficiente $\boldsymbol{\alpha}$ para o intervalo de confiança (forma alternativa para verificar a significância do coeficiente $\boldsymbol{\chi 2}$:

Para Câmara dos Deputados:

$$\boldsymbol{\alpha\chi^2} = \frac{((a.d - b.c).0{,}5)^2}{(a+b)x(d+c)x(b+d)x(a+c)}$$

$$= \frac{\big((468x888) - (3.486x45\big).0{,}5)^2}{(468+3.486)x(888+45)x(3.486+888)x(468+45)}$$

$$= \frac{113.669.448.201{,}00}{8.277.790.914.684{,}00} = \mathbf{0,014}$$

Para Assembleias Legislativas:

$$\boldsymbol{\alpha\chi^2} = \frac{((a.d - b.c).0{,}5)^2}{(a+b)x(d+c)x(b+d)x(a+c)}$$

$$= \frac{\big((921x2502) - (9.043x18\big).0{,}5)^2}{(921+9.043)x(2.502+138)x(9.043+2.502)x(921+138)}$$

$$= \frac{2.823.660.140.625{,}00}{321.608.518.228.800{,}00} = \mathbf{0,009}$$

Interpretação: os valores de $\boldsymbol{\alpha}$ foram respectivamente 0,014 e 0,009 para Câmara de Deputados e para Assembleias Legislativas, indicando - como já havia sido apresentado no exercício anterior - que podemos rejeitar a hipótese nula, aceitando que existe alguma relação entre as duas variáveis. Além disso, comparativamente, essa afirmação pode ser feita com mais confiança no caso das Assembleias Legislativas porque o $\boldsymbol{\alpha}$ delas é menor (0,009) que o da Câmara de Deputados (0,014). Portanto, a chance de errar é menor na primeira do que na segunda ao rejeitar a H_0. Não poderíamos rejeitar a hipótese nula para Intervalo de Confiança de 95% se os coeficientes $\boldsymbol{\alpha}$ tivessem ficado acima de 0,050.

iv) Cálculo do Cramer´s V para Câmara de Deputados e Assembleias Legislativas:

1) Câmara de Deputados:

$$v = \sqrt{\frac{\chi^2}{N.(k-1)}} = \sqrt{\frac{39{,}52}{4.887}} = 0{,}089$$

2) Assembleias Legislativas:

$$v = \sqrt{\frac{\chi^2}{N.(k-1)}} = \sqrt{\frac{43{,}74}{12.604}} = 0{,}058$$

Interpretação: A magnitude do efeito do sexo do candidato sobre o fato de ele ter sido ou não eleito é maior na Câmara dos Deputados do que no conjunto das Assembleias Legislativa estaduais brasileiras na eleição de 2010. Enquanto a diferença de sexo dos eleitos e não eleitos tem magnitude de efeito de 8,9% para a Câmara dos Deputados, nas assembleias legislativas essa magnitude é de 5,8%. O resultado Pode parecer contraditório com os coeficientes de $\boldsymbol{\chi 2}$ gerado no exercício anterior (ii), mas não é. O coeficiente $\boldsymbol{\chi 2}$ das assembleias legislativas é maior e estatisticamente mais significativo do que o da Câmara de Deputados. Porém, a magnitude do efeito é maior na Câmara de Deputados do que nas Assembleias. Isso é explicado pelas diferenças de N dos dois grupos. Como há três vezes mais candidatos a deputado estadual do que federal, o coeficiente $\boldsymbol{\chi 2}$ é "diluído" pelo tamanho das amostras. Como os coeficientes dos dois grupos ficaram muito próximos, o que tem menor tamanho acabará apresentando maior magnitude de efeito.

3.4.d. Resultado para o cálculo do coeficiente Δ:

Encontrar as probabilidades observadas e esperadas na tabela original. A probabilidade observada é o valor da casa dividido pelo total da tabela. A probabilidade esperada é o total da linha multiplicado pelo total da coluna e o resultado dividido pelo total da tabela. Os resultados seguem abaixo.

Ano	PSDB	PT	Total
2004	4379	3923	8302
Prob. Obs.	0,162	0,145	
Prob. Esp.	0,154	0,154	
2008	4870	4939	9809
Prob. Obs.	0,181	0,183	
Prob. Esp.	0,182	0,182	
2012	4246	4617	8863
Prob. Obs.	0,157	0,171	
Prob. Esp.	0,164	0,164	
Total	13495	13479	26974

$$\Delta = \text{Prob. Observada} - \text{Prob. Esperada}$$

	PSDB	PT
2004	0,008	-0,008
2008	-0,001	0,001
2012	-0,007	0,007

Interpretação: Durante as primeiras três eleições para prefeituras municipais do século XXI as proporções de candidaturas apresentadas pelos partidos PSDB e PT inverteram-se. Na primeira disputa, em 2004, o PSDB apresentou um coeficiente Δ positivo de 0,008 e o PT um Δ negativo de -0,008. No final do período, em 2012, os coeficientes praticamente inverteram-se. Significa que ao considerarmos apenas os dois partidos, o PT passou a participar de mais candidaturas para executivos municipais do que o PSDB, proporcionalmente. Vale lembrar que os coeficientes foram muito próximos porque os números de casos em cada par manteve-se bastante parecido nas três disputas.

4.1. ASSOCIAÇÃO ENTRE VARIÁVEIS BINÁRIAS (Q-YULE)

4.1.5.a. A tabela quádrupla considerando apenas os valores válidos (Muito Bom + Bom x Mal + Muito Mal) para as variáveis (avaliação de Lula e avaliação da economia) é a seguinte:

		Avaliação Lula		
		NEGATIVA (não-Y)	POSITIVA (Y)	TOTAL
Avaliação Economia	POSITIVA (X)	18	227	245
	NEGATIVA (não-X)	65	125	190
TOTAL		83	352	435

A partir da tabela quádrupla acima é possível calcular o Q_{xy} aplicando a fórmula:

$$Q_{xy} = \frac{(B\ x\ C) - (A\ x\ D)}{(B\ x\ C) + (A\ x\ D)} = \frac{(227\ x\ 65) - (18\ x\ 125)}{(227\ x\ 65) + (18\ x\ 125)} = \frac{12.505}{33.255} = \mathbf{0,376}$$

Intepretação: A associação entre avaliação de Lula e avaliação da Economia apresenta um coeficiente Q de Yule de +0,376, ou, 37,6% de associação entre as categorias das duas variáveis.

4.1.5.b. Para calcular a proporção de pares consistentes, aplique a fórmula abaixo para a tabela quádrupla usada anteriormente.

$$P_c = \frac{2\ x\ (B\ x\ C)}{N^2} = \frac{2\ x\ (227\ x\ 65)}{435^2} = \mathbf{0,155}$$

Agora, para pares inconsistentes a fórmula é:

$$P_i = \frac{2\ x\ (A\ x\ D)}{N^2} = \frac{2\ x\ (18\ x\ 125)}{435^2} = \mathbf{0,023}$$

Interpretação: Enquanto a proporção de pares consistentes fica em 0,155 (ou 15,5% do total), os pares inconsistentes são 0,023 (2,3% do total), o que significa que há uma proporção de cerca de sete vezes mais pares consistentes nessa associação do que pares inconsistentes.

4.1.5.c. Para identificar os limites (superior e inferior) do intervalo de confiança, basta aplicar a seguinte fórmula:

$$IC_{Qxy} = 1{,}96\ x\ \sqrt{\frac{\left(1-{Q_{xy}}^2\right)^2\ x\ \frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}{4}} = 1{,}96\ x\ \sqrt{\frac{(1-0{,}376^2)^2\ x\ \frac{1}{18}+\frac{1}{227}+\frac{1}{65}+\frac{1}{125}}{4}} = 1{,}96\ x\ \sqrt{\frac{0{,}737\ x\ 0{,}083}{4}} =$$

$1{,}96\ x\ 0{,}123 = \mathbf{0,243}$

Limite Superior = $Q_{xy} + IC_{Qxy} = 0{,}376 + 0{,}243 = \mathbf{0,619}$
Limite Inferior = $Q_{xy} - IC_{Qxy} = 0{,}376 - 0{,}243 = \mathbf{0,133}$

Interpretação: Se quisermos extrapolar para toda a população os resultados obtidos a partir dessa amostra devemos dizer que o coeficiente de associação entre avaliação de Lula e avaliação da Economia tem uma chance estatisticamente significativa para IC de 95% de se encontrar entre 0,619 e 0,133, ou seja, uma associação entre 61,9% e 13,3%. Percebe-se que o intervalo é grande, porém, como ele não passa por zero (apresentando os dois valores positivos), podemos afirmar que as chances são superiores ao limite estatístico crítico de que as duas variáveis estejam associadas na população.

4.1.5.d. Para rodarmos o teste Q_{xy} para três variáveis, precisamos construir uma tabela óctupla, incluindo a variável idade nas distribuições, como a que segue:

		Negativo (N-Y)	Positivo (Y)	Total
Id Alta (T)	Positiva (X)	7	111	118
	Negativa (N-X)	32	59	91
Total		39	170	209
Id Baixa (N-T)	Positiva (X)	11	116	127
	Negativa (N-X)	33	66	99
Total		44	182	226

i) Para encontrar o coeficiente de correlação parcial, para relação entre avaliação de Lula por avaliação da economia controlada por idade, temos que seguir os seguintes passos:

1 - Encontrar os valores para pares ligados e para pares diferentes:

$$Q_{xy}ligado = \frac{\left[(BT\ x\ CT) + \left(B\overline{\mathrm{T}}\ x\ C\overline{\mathrm{T}}\right)\right] - \left[\ (AT\ x\ DT) + \left(A\overline{\mathrm{T}}\ x\ D\overline{\mathrm{T}}\right)\right]}{\left[(BT\ x\ CT) + \left(B\overline{\mathrm{T}}\ x\ C\overline{\mathrm{T}}\right)\right] + \left[\ (AT\ x\ DT) - \left(A\overline{\mathrm{T}}\ x\ D\overline{\mathrm{T}}\right)\right]}$$

$$= \frac{[(111\ x\ 32) + \ (116x\ 33)] - \ [\ (7\ x\ 59) + \ (11\ x\ 66)]}{[(111\ x\ 32) + \ (116x\ 33)] + \ [\ (7\ x\ 59) - \ (11\ x\ 66)]} = \frac{6589}{7415} = \mathbf{0,888}$$

$$Q_{xy}diferente = \frac{\left[\left(BT\ x\ C\overline{\mathrm{T}}\right) + \ \left(B\overline{\mathrm{T}}\ x\ CT\right)\right] - \ \left[\ \left(AT\ x\ D\overline{\mathrm{T}}\right) + \ \left(A\overline{\mathrm{T}}\ x\ DT\right)\right]}{\left[\left(BT\ x\ C\overline{\mathrm{T}}\right) + \ \left(B\overline{\mathrm{T}}\ x\ CT\right)\right] + \ \left[\left(AT\ x\ D\overline{\mathrm{T}}\right) + \ \left(A\overline{\mathrm{T}}\ x\ DT\right)\right]}$$

$$= \frac{[(111\ x\ 33) + \ (116x\ 32)] - \ [\ (7\ x\ 66) + \ (11\ x\ 59)]}{[(111\ x\ 33) + \ (116x\ 32)] + \ [\ (7\ x\ 66) + \ (11\ x\ 59)]} = \frac{6264}{8482} = \mathbf{0,738}$$

2 - Encontrar o valores dos pesos para proporção de pares ligados em T (peso 1) entre pares diferentes em X e Y (peso 2).

$$P1(pares\ ligados) = \frac{(BT\ x\ CT) + \left(B\overline{\mathrm{T}}\ x\ C\overline{\mathrm{T}}\right) + \ (AT\ x\ DT) + \ \left(A\overline{\mathrm{T}}\ x\ D\overline{\mathrm{T}}\right)}{\left[\left(BT + \ B\overline{\mathrm{T}}\right)x\left(CT + \ C\overline{\mathrm{T}}\right)\right] + \left[\left(AT + A\overline{\mathrm{T}}\right)x\left(DT + \ D\overline{\mathrm{T}}\right)\right]}$$

$$= \frac{(111x\ 32) + (116\ x\ 33) + \ (7\ x\ 59) + \ (11\ x\ 66)}{[(111 + \ 116)\ x\ (32 + \ 33)] + [(7 + 11)\ x\ (59 + \ 66)]} = \frac{5319}{17005} = 0{,}312$$

$$P2\ (pares\ diferentes) = \frac{\left(BT\ x\ C\overline{\mathrm{T}}\right) + \left(B\overline{\mathrm{T}}\ x\ CT\right) + \ \left(AT\ x\ D\overline{\mathrm{T}}\right) + \ \left(A\overline{\mathrm{T}}\ x\ DT\right)}{\left[\left(BT + \ B\overline{\mathrm{T}}\right)x\left(CT + \ C\overline{\mathrm{T}}\right)\right] + \left[\left(AT + A\overline{\mathrm{T}}\right)x\left(DT + \ D\overline{\mathrm{T}}\right)\right]}$$

$$= \frac{(111\ x\ 33) + (116\ x\ 32) + \ (7\ x\ 66) + \ (11\ x\ 59)}{[(111 + \ 116)\ x\ (32 + \ 33)] + [(7 + 11)\ x\ (59 + \ 66)]} = \ \frac{8486}{17005} = 0{,}499$$

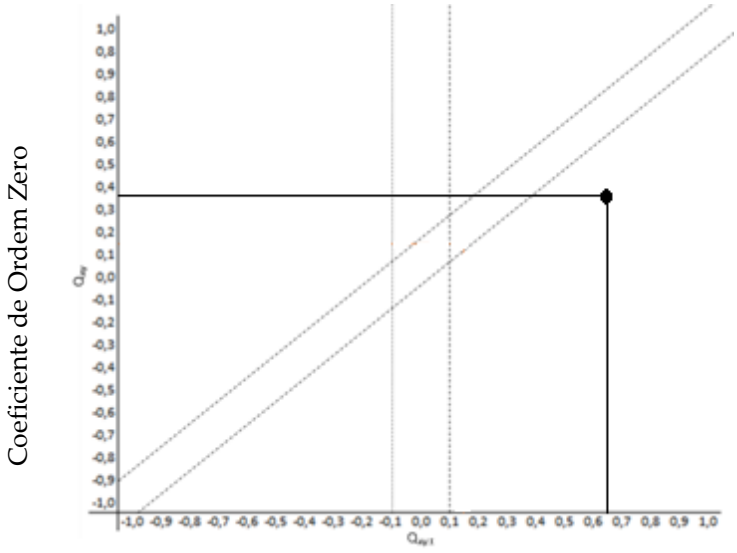
3. Calcular o Coeficiente parcial $Q_{xy:t}$:

$$Q_{xy:t} = \left(Q_{xy}\ ligado\ x\ P.ligados\right) + \ \left(Q_{xy}\ diferente\ x\ P.diferentes\right)$$

$$= (0{,}888\ x\ 0{,}312) + \ (0{,}738\ x\ 0{,}499) = \mathbf{0,645}$$

Interpretação: O coeficiente de associação parcial entre avaliação de Lula e avaliação da economia, controlados por idade dos respondentes, é de 0,645 ou 64,5% de associação parcial.

ii) Para interpretar o efeito do controle sobre a relação de entre as duas variáveis iniciais, marcamos o ponto onde os dois coeficientes encontram-se no gráfico de distribuição a seguir:
Coeficiente de ordem zero $Q_{xy} = 0{,}376$
Coeficiente parcial $Q_{xy:t} = 0{,}645$



Interpretação: Considerando as definições das relações a partir do gráfico 4.1, os resultados acima ficam na região D da distribuição dos coeficientes de ordem zero e parcial. Isso significa que ao controlarmos a relação inicial houve um aumento do coeficiente de associação, que passou de +0,376 para +0,645. Isso significa que a variável de controle interviu sobre a associação original, ou, em outras palavras, a associação de ordem zero era baixa porque desconsiderava o efeito da variável Teste. Portanto, há um efeito significativo da idade do respondente sobre a relação entre avaliação do governo Lula e avaliação da Economia.

4.2 - TESTE COM VARIÁVEIS ORDINAIS - COEFICIENTE GAMA (G)

4.2.3.a. O primeiro passo para resolver o exercício é encontrar os valores absolutos de casos para cada par de categorias. Para tanto usamos os percentuais de classe de escolaridade da amostra:

[Até 4ª Série 28% = 561 respondente; da 5ª a 8ª série 21% = 420; Ensino Médio 36% = 721 e Ensino Superior15% = 300]

Com essas informações podemos construir a tabela de contingência a partir da marginal de escolaridade. Ao final, ficará assim (O N=1975 é menor que o inicial porque desconsideramos as não respostas). Para encontrar o coeficiente Gama é preciso, antes, encontrar as proporções de Pares Consistentes (PC) e Pares Inconsistentes (PI).

	Até 4ª série	**Da 5ª a 8ª série**	**Ensino Médio**	**Ensino Superior**	**Total**
Ótimo	45	34	36	12	127
Bom	196	130	202	84	612
Regular	185	168	303	123	779
Ruim	62	38	79	48	227
Péssimo	56	46	101	27	230
Total	544	416	721	294	1975

Cálculo dos Pares consistentes e inconsistentes:

	até 4ª série	**de 5ª a 8ª série**	**ensino médio**	**superior**	**TOTAL**
Ótimo	45	34	36	12	127
Bom	196	130	202	84	612
Regular	185	168	303	123	779
Ruim	62	38	79	48	227
Péssimo	56	46	101	27	230
TOTAL	544	416	721	294	1975
	2356,20	1562,40	2422,56		
	2221,560	2016,00	3633,84		
	740,520	453,60	951,72		
	673,20	554,40	1211,28		
				Pares Consistentes	**18.797,28**
		5843,37	9060,37	3769,92	
		7539,84	13590,56	5520,24	
		1696,46	3559,43	2154,24	
		2073,45	4530,18	1211,76	
				Pares Inconsistentes	**60.549,85**

$$G = \frac{PC - PI}{PC + PI} = \frac{18.797{,}28 - 60.549{,}85}{18.797{,}28 + 60.549{,}85} = \frac{-41.752{,}57}{79.347{,}13} = \mathbf{-0{,}526}$$

Interpretação: O coeficiente Gama para as variáveis categóricas "avaliação do governo Dilma" e "Escolaridade" mostrou-se alto, com -0,526 de coeficiente. Isso representa uma associação de -52,6% entre as duas variáveis. Como o sinal é negativo, indica que conforme aumenta a escolaridade dos respondentes há uma tendência de redução da avaliação positiva do governo Dilma.

4.3 - ASSOCIAÇÃO ENTRE CATEGORIAS NOMINAIS (RESÍDUOS BRUTOS E RESÍDUOS PADRONIZADOS)

4.2.4.a. Usando os mesmos valores de frequências da tabela de contingência entre a avaliação do governo Dilma e a escolaridade do eleitor, devemos encontrar respectivamente a Frequência Esperada (F_e) para cada par; os Resíduos Brutos (R_b) e, depois, os Resíduos Padronizados (R_p) conforme a sequência das fórmulas abaixo:

Aval.	**1 - Frequência Esperada** (F_e) $F_e = \frac{M.Linha\ x\ M.Coluna}{N}$				**2 - Resíduos Brutos** (R_b) R_b =Fo - Fe				**3 - Resíduos Padronizados** (R_p) $R_p = \frac{R_b}{\sqrt{F_e}}$			
	até 4ª série	**de 5ª a 8ª**	**ensino médio**	**Supe-rior**	**até 4ª série**	**de 5ª a 8ª**	**ensino médio**	**supe-rior**	**até 4ª série**	**de 5ª a 8ª**	**ensino médio**	**superior**
ótimo	35	27	46	19	10	7	-10	-7	1,697	1,349	-1,492	-1,575
bom	169	129	224	91	28	1	-22	-7	**2,125**	0,111	-1,451	-0,751
regular	215	164	284	116	-29	4	18	7	**-2,013**	0,313	1,094	0,654
ruim	62	48	83	34	-1	-10	-3	14	-0,099	-1,440	-0,384	**2,450**
péssimo	63	48	84	34	-7	-2	17	-7	-0,921	-0,326	1,842	-1,242

Interpretação: Os resíduos padronizados mostram que embora as duas variáveis (avaliação do governo Dilma e Escolaridade dos respondentes) estejam associadas, segundo o que mostrou o Coeficiente Gama anteriormente, poucos pares apresentam resíduos padronizados superiores a (±1,96), podendo ser considerados significativos. Entre os eleitores com até a 4ª série há uma

concentração de resíduos padronizados na avaliação "bom" e uma ausência de casos na avaliação "regular". Entre os eleitores com escolaridade superior há uma concentração significativa de casos na avaliação "ruim". Todos os demais resíduos ficam abaixo do limite crítico de ±1,96.

4.2.4.b. Para encontrar os Resíduos Padronizados das opiniões sobre democracia entre 1996 e 2002 é necessário, antes, chegar aos valores das frequências dos pares na tabela de contingência, cf. segue abaixo:

	1996	1997	1998	1999/2000	2001	2002	TOTAL
Democracia es preferible	10482	8528	10744	11238	11244	11298	63534
Da lo mismo	3369	3731	3044	2900	2720	2964	18728
Gobierno autoritário	2808	3376	3044	3081	3264	3334	18907
TOTAL	16658	15635	16833	17219	17228	17596	101169

A partir das frequências observadas e dos totais das marginais acima podemos calcular as frequências esperadas, resíduos brutos e resíduos padronizados. Como já fizemos a mesma operação no exercício anterior, serão apresentados apenas os resultados aqui:

		1996	**1997**	**1998**	**1999/2000**	**2001**	**2002**	**TOTAL**
	Fo	10482	8528	10744	11238	11244	11298	63534
	Fe	10461	9819	10571	10813	10819	11050	
	Rb	20	-1290	173	424	424	248	
Democracia es preferible	**Rp**	0,199	**-13,024**	1,687	**4,078**	**4,081**	**2,362**	
	Fo	3369	3731	3044	2900	2720	2964	18728
	Fe	3084	2894	3116	3188	3189	3257	
	Rb	285	837	-72	-288	-469	-294	
Da lo mismo	**Rp**	**5,139**	**15,554**	-1,286	**-5,093**	**-8,305**	**-5,148**	
	Fo	2808	3376	3044	3081	3264	3334	18907
	Fe	3113	2922	3146	3218	3220	3288	
	Rb	-306	454	-102	-137	45	46	
Gobierno autoritário	**Rp**	**-5,477**	**8,395**	-1,811	**-2,410**	0,786	0,794	
TOTAL		16658	15635	16833	17219	17228	17596	101169

Interpretação: A partir dos resíduos padronizados das seis pesquisas realizadas entre 1996 e 2002 para opiniões sobre a importância da democracia na América Latina é possível perceber o crescimento da preferência pela democracia ao longo do tempo. Na primeira metade do período apresentava R_P abaixo do limite crítico ou negativo para nos últimos três anos apresentar R_P positivos e significativos. Já o ponto médio, "*da no mismo*" teve significativa queda ao longo do período, passando de R_P positivo e significativo para negativo e significativo. Ao passo que os resíduos das respostas para preferência pelo autoritarismo oscilaram ao longo do tempo, com dois pontos de resíduos negativos significativos, um ponto de resíduo positivo e significativo e três pontos sem significância estatística. Os resultados mostram que a tendência temporal de crescimento da preferência pela democracia deve-se à transferência dos que dizem não ter preferência. Já o número dos que preferem o regime autoritário mantém-se estável ao longo do tempo.