EMERSON URIZZI CERVI

Manual de

MÉTODOS QUANTITATIVOS

PARA INICIANTES EM CIÊNCIA POLÍTICA

Vol. 1

CPOP
UFPR

Manual de Métodos Quantitativos para iniciantes em Ciência Política

Volume 1

EMERSON URIZZI CERVI

MANUAL DE MÉTODOS QUANTITATIVOS PARA INICIANTES EM CIÊNCIA POLÍTICA

VOLUME 1

CURITIBA
2017

Dados Internacionais de Catalogação na Publicação
Fundação Biblioteca Nacional

C419m Cervi, Emerson U.
Manual de métodos quantitativos para iniciantes em Ciência Política – Volume 1 / Emerson Urizzi Cervi - Curitiba: CPOP-UFPR, 2017. (1ª edição). 256 p.

ISBN 978-85-915195-4-5

1. Pesquisa – Métodos Quantitativos. 2. Ciência Política. I. Título.

CDD-320

Copyright @ 2017 do autor

Editoração
Grupo de Pesquisa em Comunicação Política e Opinião Pública - CPOP/UFPR

Edição do Autor
Emerson Urizzi Cervi

Planejamento Gráfico e Diagramação
Fernanda Cavassana de Carvalho

Grupo de Pesquisa em Comunicação Política e Opinião Pública – CPOP
Programa de Pós-graduação em Ciência Política – PPGCP-UFPR
Universidade Federal do Paraná – Campus Reitoria
Rua General Carneiro, nº46 – Ed. Dom Pedro I, 9º andar.
CEP: 80.060-000. Curitiba – Paraná – Brasil

www.cpop.ufpr.br | www.facebook.com/cpopufpr | nucleocpopufpr@gmail.com

Sumário

INTRODUÇÃO - PARA QUE SERVEM OS MÉTODOS QUANTITATIVOS? 9
I.i Só se quantifica o que se sabe que existe 10
I.ii Mãos à obra, quantitativistas! 19
I.iii Exercícios propostos 26
I. iv Bibliografia citada no capítulo 26

CAPÍTULO I - CONCEITOS FUNDAMENTAIS 27
1.1 Teorias, métodos e observações 29
1.2 Tipos de relações entre variáveis 31
1.3 Exercícios propostos 36
1.4 Bibliografia citada no capítulo 36

CAPÍTULO II - DEFINIÇÕES ESTATÍSTICAS 37
2.1 Tipos de variáveis 39
2.2 Validade e confiabilidade de medidas 46
2.3 Distribuição de frequências em variável categórica (*Sturges*) 48
2.4 Distribuição em três categorias (ABC) pelo Índice de *Nihans* 52
2.5 Exercícios propostos 58
2.6 Bibliografia citada no capítulo 58

CAPÍTULO III - DADOS, INDICADORES, ÍNDICES, ESCALAS E FORMAS DE REPRESENTAÇÃO 59
3.1 Matriz de dados 59
3.2 Índices e escalas 62
3.2.1 Tipos de Escalas 64
3.3 Criação de escalas por soma de valores e teste de Confiabilidade (α de *Cronbach*) 67
3.4 Principais indicadores estatísticos 76
3.5 Principais indicadores parlamentares e eleitorais 77
3.6 O Índice de Presença Eleitoral (*IPE*) 81
3.7 Indicadores educacionais mais comuns 88
3.8 Indicadores sócio-demográficos 89
3.9 Representações tabulares de dados 91
3.10 Representações gráficas de dados 92
3.11 Exercícios propostos 103
3.12 Bibliografia citada no capítulo 103

CAPÍTULO IV - MEDIDAS ESTATÍSTICAS 105
4.1 Medidas de tendência central 105
4.2 Medidas de dispersão, variabilidade ou assimetria 114
4.3 Coeficiente de variação de pearson (CV) 120
4.4 Medidas de posição 121
4.5 Medidas de assimetria e curtose 123
4.6 Exercícios propostos 126
4.7 Bibliografia citada no capítulo 126

CAPÍTULO V - AS AMOSTRAS 127
5.1 Conceito de amostragem 127
5.2 Tipos de amostras 131
5.3 Tamanho das amostras 134
5.4 Exercícios propostos 142
5.5 Bibliografia citada no capítulo 142

CAPÍTULO VI - *SURVEYS* PARA COLETA DE DADOS INDIVIDUAIS 143
6.1 Questionário 144
6.2 Erros não-amostrais 145
6.3 Principais tipos de *Surveys* 147
6.4 Estrutura do Questionário 151
6.5 Características Técnicas das Respostas 157
6.6 Organização do Questionário na Prática 159
6.7 O Pré-Teste 161
6.8 Exercícios propostos 162
6.9 Bibliografia citada no capítulo 162
Anexos do Capítulo VI 164

CAPÍTULO VII – TESTES DE HIPÓTESES DE DIFERENÇAS DE MÉDIAS 171
7.1 Conceito de probabilidade 171
7.2 Distribuição normal 173
7.3 Testes de Hipóteses Estatísticas (NHST) 177
7.4 O erro nos testes de hipóteses 178
7.5 Nível de significância no teste de hipóteses 180
7.6 Prova t de *Student* 180
7.7 Análise de Variância (ANOVA) 194
7.8 Prova de Qui-quadrado (x^2) 206
7.9 Exercícios propostos 215
7.10 Bibliografia citada no capítulo 215
Anexos do Capítulo VII 216

ADENDO I - GABARITO DE RESPOSTAS AOS EXERCÍCIOS PROPOSTOS 219

ADENDO II - PASSO A PASSO PARA USO DO RCOMMANDER 237

Apresentação

Um manual é um conjunto de instruções
que servem para o bom funcionamento
de uma ferramenta. É meio, não fim.

Este "Manual de Métodos Quantitativos para Iniciantes em Ciência Política" é resultado de mais de dez anos dando aulas de introdução aos métodos quantitativos para alunos de graduação em ciência política e em ciências sociais. Apesar disso, não está terminado. Nunca estará. Para atender ao objetivo de ser um texto didático a iniciantes em análises quantitativas, tive que fazer cortes e escolhas. Com certeza há mais coisas importantes fora do que dentro dele. Por isso, conto com a colaboração dos leitores para melhorias permanentes.

Como introdução às ferramentas aplicadas à pesquisa empírica em ciência política, este livro não pretende ser uma obra de estatística; não tem profundidade para tanto. E não é nada útil para quem já é iniciado nos conceitos e ferramentas básicas de análise quantitativa. Trata-se de manual introdutório para iniciantes que, a partir dele, devem alçar voos mais altos em suas pesquisas. Para não cansar o leitor, o livro está dividido em dois volumes. O primeiro, apresentado aqui, é composto por sete capítulos com apresentações conceituais e de testes estatísticos simples para serem usados por graduandos em seus trabalhos. O segundo conta com um capítulo para cada técnica de abordagem empírica dentre as mais usadas na ciência política.

Em toda a obra, a opção foi por usar apenas programas estatísticos com código-fonte aberto, os chamados *softwares* livres. Não foi possível incluir aqui cursos de introdução para todos os *softwares* usados no livro, coisa que o leitor terá que fazer de forma complementar. Ainda assim, ao final deste volume, há um apêndice explicativo de como usar o pacote estatístico *Rcmdr* nos exemplos do livro. O banco de dados usado em todos os exercícios práticos pode ser acessado a partir do link: *https://blogempublico.files.wordpress.com/2017/09/b_exerc_v1.xlsx*.

Para encerrar, preciso agradecer a todos os alunos que passaram pelas minhas aulas de métodos quantitativos, especialmente os que integram o grupo de pesquisa em Comunicação Política e Opinião Pública (CPOP). Sem eles, o conteúdo desse e-book não existiria. Também agradeço a participação da doutoranda Fernanda Cavassana na editoração do livro. Sem ela, o e-book não teria forma.

Curitiba, outubro de 2017.

Introdução

Para que servem os Métodos Quantitativos?[1]

Começando do início: se tem quantidade,
pode ser contado. Se não tem, não pode.

Esta introdução tem o objetivo de discutir, ainda que brevemente, o papel dos métodos quantitativos como ferramenta de pesquisa empírica das ciências sociais. O argumento central é que o quantitativismo tem seu lugar na ciência por se tratar de um conjunto de técnicas de pesquisa social e análise que, ao ser bem aplicado, permite relacionar descobertas sobre padrões de comportamento social com implicações nas teorias sociais já existentes. Para isso, é preciso evitar duas armadilhas principais. A primeira é o excesso de quantitativismo, que não raras vezes transforma a ferramenta estatística, que deveria servir como forma de acesso à realidade social, em finalidade última da pesquisa empírica. A segunda é o debate estéril entre quantitativistas e qualitativistas que procura, indefinidamente, identificar a melhor técnica de pesquisa entre as duas. Esse debate não produzirá nenhum resultado relevante para a ciência porque ele desconsidera solenemente que os métodos de pesquisa só existem em função de um objeto de análise, sem o qual perdem qualquer sentido. Assim, dependendo do objeto de análise e objetivos do estudo que se pretendem empreender, um dos métodos acima

1 Publicado originalmente como: CERVI, Emerson C. Métodos Quantitativos em Ciências Sociais: uma abordagem alternativa ao fetichismo dos números e ao debate com os qualitativistas. In BOURGUIGNON, Jussara A. (org.) Pesquisa Social: reflexões teóricas e metodológicas. Editora Toda Palavra: Ponta Grossa, 2009.

se mostrará mais adequado que o outro.

Este capítulo está dividido em duas partes principais. Na primeira, são apresentados os debates conceituais que definem o conceito e caracterizam os métodos quantitativos, em especial no que diz respeito às ciências sociais. Para isso, mostra-se a relação entre cada uma das etapas da pesquisa científica e faz-se uma breve comparação entre as características de cada uma das três principais estratégias de pesquisa: qualitativa, comparativa e quantitativa. A segunda parte é destinada a um tratamento mais aplicado do quantitativismo, descrevendo, ainda que sumariamente, as principais técnicas de coleta de dados e análises empíricas relacionadas a essa ferramenta disponível à comunidade científica para a produção de estudos sobre a realidade social e, por consequência, de novas representações dessa realidade.

I.i. Só se quantifica o que se sabe que existe

Normalmente, costuma-se estabelecer o espaço específico da pesquisa quantitativa como aquele em que se busca contar e medir características dos fatos sociais; enquanto que por outro lado, os métodos qualitativos serviriam para classificar as características do objeto de análise. O problema, ou melhor, a principal limitação desse tipo de distinção está na impossibilidade de aplicá-la na prática da pesquisa científica. Considerando o *status* atual da pesquisa científica das ciências sociais, baseada na síntese entre racionalismo e empirismo, torna-se impossível quantificar, contar ou medir qualquer coisa da qual não se tenha uma classificação ou definição compartilhada a respeito dessas características medidas. Da mesma forma, qualquer classificação só é válida quando permite a distinção entre diferentes quantidades. Portanto, torna-se impossível fazer pesquisa científica exclusivamente em termos quantitativos ou qualitativos, segundo os termos apresentados acima.

De maneira alternativa, prefiro ressaltar que os métodos quantitativos para pesquisa social incluem uma série de técnicas de pesquisa que têm como principal finalidade a medição de quantidades e quantificação de qualidades. Para definir se essa técnica deve ou não ser utilizada em pesquisas científicas específicas, é preciso, antes, delimitar

o objeto do trabalho. Em outras palavras, o objeto a ser analisado é que determina se as técnicas quantitativas são ou não as mais adequadas. A pesquisa quantitativa deve ser usada quando o que se quer medir no objeto pesquisado já é conhecido. Caso contrário, é recomendável o uso de técnicas qualitativas, que são mais adequadas para análises exploratórias dos objetos pesquisados. Assim, evita-se a discussão e polêmica histórica entre quantitativistas e qualitativistas. Os métodos quantitativos são apropriados para estudar características do objeto pesquisa que, sabe-se, existem. Ao passo que os métodos qualitativos devem ser usados quando a pesquisa busca entender características do objeto que não se sabe se existem, porém não há pesquisa social produtiva que utilize com exclusividade técnicas quantitativas ou qualitativas.

Sendo assim, o que determina o predomínio dos métodos a serem utilizados é o objetivo principal da pesquisa e o grau de conhecimento prévio que já exista a respeito do objeto em análise. Evidente que esta opção não significa que ambas as metodologias devam ter suas características confundidas. Pelo contrário. Por princípio, pode-se definir que a pesquisa quantitativa caracteriza-se como uma forma de explicação causal a partir de uma predição linear, enquanto a qualitativa busca um compreensão estruturalista a partir de uma predição formal. Como o objetivo deste capítulo é apresentar uma discussão sobre as características dos métodos quantitativos, passo a discutir o que considero fundamentalmente explicações causais e preditivas lineares (CONDE, 2005).

Dadas as características relatadas acima, a perspectiva mais comum das pesquisas quantitativas é empírica em busca da descrição mais objetiva possível de dados da realidade – daí a relação que vários autores fazem da pesquisa quantitativa com o positivismo de Auguste Comte e com o neopositivismo do Círculo de Viena (HERSCOVITZ, 2007). Para análise empírica da realidade, podem ser quantificados dados secundários sobre a realidade, quando eles já estão disponíveis, ou a partir de dados primários, que surgem da própria ação de pesquisa.

Invariavelmente, quantitativistas consideram que os fenômenos sociais podem ser explicados a partir da representação deles em números, usados em análises que permitem generalizações, indicação de relações de causalidade e, como consequência, para validação ou rejeição de teorias. Tudo isso é feito através da contrastação a partir da medição empírica de fenômenos sociais. Para tanto, as técnicas mais comuns

usadas pelos pesquisadores para coleta de informações da realidade são a realização de pesquisas por aplicação de questionários e a partir de fontes de dados primários estatísticos. Esse tipo de pesquisa tem como principais objetivos, portanto, a descrição das quantidades de características de determinada população, o estabelecimento de relações causais entre variáveis já conhecidas e a realização de inferências a partir de resultados obtidos em amostras representativas de populações mais amplas. Para tanto, torna-se indispensável para o início do trabalho quantitativista a utilização de alguns conjuntos de medidas, como as de tendência central e de dispersão; assim como as análises bivariadas e multivariadas entre diferentes características da população.

A partir do "instrumental" básico descrito acima, assim como todos os demais cientistas sociais, os pesquisadores quantitativistas produzem representações de fatos relevantes da vida social. Mas é preciso considerar que os cientistas não são os únicos a produzirem esse tipo de representação – nem os únicos a usarem os instrumentos quantitativos. Profissionais de outras áreas, por exemplo do jornalismo ou da produção literária, também estudam e produzem representações sobre a vida social. De diferentes maneiras, ou através de múltiplas técnicas, pesquisadores das mais diversas áreas reconstroem eventos históricos importantes, identificam padrões de comportamento ou mapeiam formas de organização do cotidiano social.

Ao contrário das teses correntes sobre metodologia de pesquisa científica, Ragin (1994) descarta as diferenciações puristas da pesquisa social em relação às demais representações da sociedade e entre diferentes ramos científicos. Para isso, não toma como base os argumentos comuns de que a pesquisa social tem uma matriz própria, utiliza uma linguagem especial ou utiliza uma metodologia específica. O autor defende que, assim como o cientista, o jornalista coleta, organiza, analisa e reporta dados sociais, portanto, ele faz ciência social, sem ser considerado um cientista social (RAGIN, 1994).

O que diferencia a pesquisa social do jornalismo é basicamente o público que cada uma dessas áreas de representação da sociedade pretende atingir e não, necessariamente, os instrumentos utilizados para a coleta de informações da realidade. Avançando nesse mesmo sentido, o que distancia os quantitativistas dos qualitativistas não passa, muitas vezes, das demandas mais imediatas dos públicos a que se destinam as pesquisas que utilizam essas técnicas. Voltando ao argumento de Ragin (1994), en-

quanto o jornalista consegue divulgar seu trabalho a grandes públicos, o cientista social escreve principalmente para outros cientistas sociais – o que não significa que os resultados de sua pesquisa não possam ser divulgados para o grande público. A publicização de resultados ou descobertas de pesquisas científicas pelos meios de comunicação apenas reforça o fato de que o jornalismo e a pesquisa social têm por objetivo criar e divulgar representações sociais, porém para públicos iniciais distintos – o que resultará em diferenças de linguagem para a publicização de dados e resultados.

Ao se dirigir ao grande público, o jornalista tem dificuldade em tratar do método utilizado para produzir determinada representação. Como a ciência social dirige-se principalmente aos próprios cientistas, ela pode apresentar e discutir também o método de abordagem dos dados sobre a realidade social. Sendo assim, o "público" do cientista social espera, ao mesmo tempo em que tem acesso às conclusões da pesquisa, localizar um fenômeno socialmente significativo – através dos dados - e identificar uma teoria social ligada direta ou indiretamente ao tema pesquisa. Nem sempre a comunidade científica como um todo tem condições de avaliar diferentes ferramentas metodológicas utilizadas na pesquisa social – principalmente porque discutir o método não é objetivo final da grande maioria dos pesquisadores. Falando sobre uma das metas do cientista social ao divulgar os resultados de seu trabalho ao público especializado, Ragin (1994) afirma que "quando sua audiência primária é formada por cientistas sociais e outros profissionais, eles enfatizam, entre outras coisas, aspectos técnicos de suas pesquisas e o posicionamento dela em uma literatura específica sobre a pesquisa – isto é, em relação ao trabalho de outros cientistas que têm pesquisado tópicos similares" (RAGIN, 1994 p. 22). Trata-se não apenas de falar sobre resultados, mas, ao mesmo tempo, defender uma forma específica de chegar até eles.

De maneira complementar a esta, Morris Rosenberg (1971) afirma que o valor de um instrumento de pesquisa reside naquilo que se pode apreender com sua utilização. "Todo cientista social busca chegar a generalizações teóricas significativas sobre a natureza da vida social" (ROSENBERG, 1971 p. 16). O ganho do método científico em relação às demais formas de representação social surge daí. Por se destinar principalmente a uma comunidade específica - a acadêmica -, cada pesquisa social individual tem pouco valor, mas ao agregar-se a um conjunto de trabalhos científicos que

compartilham os mesmos instrumentos metodológicos, é possível fornecer informações mais ricas e completas do que qualquer outra forma de representação. Nos termos apresentados por Imre Lakatos (1989), há uma distinção entre pesquisa isolada e que não é capaz de produzir grandes "achados" científicos e o que ele chama de "programa de pesquisa", que é formado por um núcleo central e um cinturão de proteção. Para que haja um diálogo entre pesquisas isoladas visando à formação de um programa de pesquisa é necessário, antes, uma linguagem comum – em especial do ponto de vista metodológico – entre os pesquisadores (LAKATOS, 1989).

Outro ponto relevante a se destacar na discussão sobre o uso de métodos quantitativos na pesquisa social é que, assim como as demais técnicas empíricas, ele também necessita de um diálogo entre as ideias e as evidências. O resultado da relação entre conceitos e evidências da realidade forma uma representação social da realidade. Essa representação será inovadora e permitirá um avanço real da ciência quanto mais qualidade houver no "programa de pesquisa" que lhe deu origem. Porém, como a realidade que os cientistas buscam representar é multifacetada e complexa, torna-se necessário fazer um recorte, selecionando apenas alguns aspectos dela como objeto de atenção da pesquisa. Em termos quantitativos, trata-se de medir e contar à exaustão alguns aspectos da realidade – o que não significa ser possível exaurir toda a realidade em um programa de pesquisa, muito menos em pesquisas isoladas. Na pesquisa quantitativa, as etapas para transição do mundo das ideias para o das "coisas" materiais é similar às das demais metodologias empíricas. Como registrado no quadro (i.i) abaixo, parte-se dos conceitos para fazer as análises da realidade selecionada a partir de evidências coletadas empiricamente. Como resultado, tem-se um novo conjunto de representações e imagens a respeito do objeto pesquisado.

Quadro i.i – Etapas da pesquisa científica empírica

Ideias/Conceitos ⟶	Evidências ⟶	Análises ⟶	Representações

As ideias e os conceitos que servem de premissa para a pesquisa empírica

são resultados de transmissões intergeracionais de pesquisadores. A partir delas, os cientistas buscam informações empíricas através da coleta de evidências da realidade. Como existe uma quase infinidade de informações possíveis de serem captadas da realidade, as técnicas de pesquisa empírica selecionadas ajudam a delimitar o campo de trabalho. Quando se opta por técnicas relativas aos métodos quantitativos, sabe-se que as evidências relacionam-se a características já conhecidas minimamente e que podem ser contadas ou medidas. Por fim, as análises das evidências selecionadas e tratadas a partir das "ferramentas" de pesquisa disponíveis permitem a criação de novas representações da realidade, ou seja, faz com que sejam incorporadas explicações a respeito do objeto de interesse delimitado pelo pesquisador.

Considerando a argumentação aqui, pode-se afirmar que a pesquisa social empírica é obrigada a selecionar, através do recorte, alguns aspectos de uma realidade (quase infinita) que deverão ser submetidos às análises científicas. O passo seguinte é o tratamento das evidências, que está relacionado às técnicas escolhidas ou disponíveis para a etapa analítica. Sendo assim, é possível relacionar diretamente a etapa dos objetivos a serem alcançados com determinada pesquisa e as técnicas de análise, visto que os resultados obtidos a partir da utilização de determinadas "ferramentas" são consequências, também, de escolhas feitas sobre quais aspectos da realidade pesquisar. Apesar da complexidade da vida social, há uma ordem que faz com que as pessoas identifiquem situações familiares. Nas palavras de Max Weber (2001, p. 124), "a ciência social que pretendemos exercitar é uma ciência da realidade. Procuramos entender na realidade o que está ao nosso redor, e na qual nos encontramos situados, aquilo que ela tem de específico". A identificação de regularidades na sociedade é a principal meta do cientista social que opta por trabalhar com métodos quantitativos. Existem outras metas secundárias que ajudam a alcançar essa principal. Como a sociedade é multifacetada, a pesquisa social também será, pois só assim apresenta condições de retratar múltiplos padrões sociais.

Dentre os principais objetivos da pesquisa científica empírica, podemos destacar a identificação de padrões gerais dos elementos integrantes da realidade e as relações que existem entre eles. Conhecer padrões permite identificar, entre outras coisas, casos atípicos - o que resulta na possibilidade de comparação. Como a pesquisa social

produtiva, aquela que se vincula a um programa de pesquisa orgânico do ponto de vista metodológico, é um diálogo entre pesquisadores, permite o surgimento de novas ideias que modificam a forma de representar a realidade, o que leva a um refinamento das teorias já existente. Esse refino teórico é outro objetivo importante da pesquisa social empírica.

As predições a partir da descrição de determinados aspectos da realidade, outra meta da pesquisa científica, estão baseadas nos testes de hipóteses em relação a um determinado conjunto de dados. Considerando que a descrição permite relacionar os comportamentos de duas variáveis da realidade, quando se altera uma delas é possível identificar efeitos possíveis e prováveis no comportamento da outra variável, mesmo antes de acontecer. Porém, as predições da ciência social têm limites de precisão, mesmo que sejam utilizadas as mais avançadas técnicas de medição e contagem dos aspectos dos fatos sociais. Explorar a diversidade da vida social, identificando padrões ou diferenças nas relações sociais e comparando-as, é o que permite criar uma representação da sociedade. Como em exemplo citado por Ragin (1994), saber porque alguns grupos de imigrantes para determinado país são bem-sucedidos socialmente é importante para responder porque outros grupos não conseguem ser. Ao tratar de grupos específicos, a pesquisa social possibilita visibilidade a eles. Garante voz a diferentes grupos sociais, a partir de seus conhecimentos e experiências.

Sendo assim, a pesquisa social empírica quantitativa permite - pelos resultados vinculados a programas de pesquisa - avançar em novas teorias. Para isso, é preciso que as ideias sejam elaboradas de maneira atualizada. O avanço tem qualidade indutiva e está baseada em novas evidências. É da interação entre ideias e evidência que surge uma descrição com bases teóricas da vida social. Para alcançar essas grandes metas existem três estratégias principais de pesquisa científica: qualitativa, comparativa e quantitativa. Dependendo do tipo de objeto a ser pesquisado e das metas a serem alcançadas, o pesquisador deve optar por uma delas como instrumento principal de análise, o que não significa que seja exclusivo. O quadro (i.ii) abaixo, adaptado de Ragin (1994), resume as principais características dos métodos, considerando a relação deles com o universo (parcela do mundo social) que se pretende estudar; com o número de variáveis (características dos integrantes do mundo social) a serem pes-

quisadas; e com os objetivos mais gerais da pesquisa (a que as análises pretendem remeter).

Quadro i.ii – Comparação entre métodos de pesquisa social empírica

Estratégias	Universo Pesquisado	Variáveis Analisadas	Objetivos
Método qualitativo	Pequeno	Muitas	Examinar detalhes
Método comparativo	Médio	Número moderado	Comparar características
Método quantitativo	Grande	Poucas	Identificar padrões

Fonte: Adaptado de Ragin (1994)

Cada uma das estratégias metodológicas apresentadas acima permite um tratamento específico da realidade, que está relacionado a determinados objetivos. O método qualitativo é usado em estudos de comunidades com o objetivo de examinar casos específicos, ou detalhes. Em alguns casos essa estratégia de pesquisa parece não ser científica por não permitir o estabelecimento de grandes regras. Mas é preciso entender que este não é o objetivo da estratégia qualitativa de análise da realidade. O que se espera nesse tipo de trabalho é o exame de casos específicos, pois na construção de grandes representações sociais perdem-se os detalhes. Com o método comparativo, o objetivo é estudar diversidades. Essa estratégia permite o exame de padrões de diferenças e similaridades em um número moderado de casos – maior que o da pesquisa qualitativa e menor que o da quantitativa. A diferença em relação à pesquisa qualitativa é que o estudo comparado permite identificar casos padrões em diferentes comunidades, enquanto o método qualitativo estuda as características dos casos no interior de uma comunidade. O método comparativo é usado em estudos de configurações, ou seja, a combinação específica de atributos comuns em um dado número de casos.

Por oposição, a estratégia que mais interessa neste capítulo, a quantitativa, é indicada quando o objetivo da pesquisa é definir padrões e identificar relações entre muitos casos. Se essa metodologia quantitativa for aplicada a um número reduzido de casos, a representação será distorcida por não haver um universo mínimo que garanta

a representatividade e, por consequência, a análise inferencial para o universo como um todo. Nesses casos, os métodos quantitativos não se aplicam. Como se vê, a meta dessa estratégia é identificar padrões gerais e relações entre eles – as chamadas covariações, correlações e, em alguns casos, relações deterministas – além de testar teorias e fazer predições. Em relação às estratégias anteriores, a quantitativa aplica-se quando se quer pesquisar um grande número de indivíduos, analisando muitas características deles. Como aqui é preciso conhecer pelo menos superficialmente as características que se pretende estudar de maneira antecipada, essa distinção a partir de Ragin (1994) reforça o posicionamento no início do capítulo de que a distinção entre as estratégias dá-se principalmente pelo nível de informação prévia em relação ao que se pretende estudar.

A pesquisa social quantitativa pode ser definida como uma das formas de representação de fenômenos significativos da sociedade a partir de um conjunto de técnicas que permite medir e contar os padrões e relações entre as características do objeto da pesquisa. Tem como particularidade o fato de que se conheça a existência prévia das características que se pretende estudar e os seus resultados destinam-se a um público específico, o dos cientistas. As diferenças entre as três estratégias de pesquisa dão-se no tamanho do universo analisado, no número de variáveis envolvidas em cada pesquisa e nos objetivos a que cada uma pode cumprir. Nesse sentido, nenhum cientista, inclusive quantitativistas, pode esquecer que antes do acesso à realidade existe a etapa das ideias. Assim como lembra Fábio Wanderlei Reis (1991), no artigo “O Tabelão e a Lupa”, toda representação científica da sociedade precisa de uma concepção teórica que a antecede. Esse marco teórico é o que vai fornecer as grandes questões para o trabalho empírico, além de oferecer a problemática teórica. Essa problemática deve ser o fio condutor da pesquisa, o que não significa que deva determinar o quadro final da representação social. As conclusões das pesquisas empíricas quantitativas servem para validar, questionar ou avançar as teorias – nesse caso, as demais estratégias têm a mesma finalidade. Todo aspecto estudado pela ciência social, independente da estratégia utilizada, deve ser reduzido a sua dimensão real e transformado em um elo de uma cadeia geral de ocorrências.

I.ii. Mãos à obra, quantitativistas!

Assim como toda pesquisa empírica, a operação lógica é mais importante que a operação matemática nos métodos quantitativos. A matemática, mais especificamente a estatística, é uma linguagem eficiente para a descrição das operações lógicas necessárias à análise dos dados da realidade (BABBIE, 2005). O que justifica a utilização de instrumentos estatísticos e matemáticos na pesquisa quantitativa é que eles permitem reduzir uma grande massa de informações a alguns indicadores que são capazes de representar as principais características do objeto analisado. A partir dessa linguagem, é possível medir quantidades de uma mesma característica ou fazer associações entre características distintas com distribuições de frequências conjuntas. Como se vê, uma das principais fontes de equívocos das críticas aos métodos quantitativos está na inversão da relação que se faz entre as ferramentas matemáticas e estatísticas e os objetivos mais amplos da pesquisa empírica quantitativa.

De fato, o que merece críticas são a mera produção de resultados matemáticos, por um lado, e indicadores de relações estatisticamente significativas entre variáveis desprovidas das devidas análises e contextualização por outro. Porém, é preciso reconhecer que esse tipo de pesquisa estéril, que não consegue relacionar os resultados obtidos a partir do uso de instrumentos científicos com a realidade que deve ser representada, não é exclusividade dos métodos quantitativos. A partir de qualquer estratégia metodológica, pode-se abstrair da realidade qualquer conclusão factível a partir de análises subjetivas, mesmo nos trabalhos empíricos destinados a análises de poucas características de pequenas comunidades. Para evitar os resultados estéreis em pesquisas quantitativas, apresento a seguir uma sumarização de finalidades que o pesquisador deve levar em consideração no uso de técnicas de pesquisa como forma para produzir análises e conclusões a respeito da realidade social a ser representada.

A primeira delas diz respeito às concepções para explicar possíveis causas. Através desse critério, o pesquisador usa os métodos quantitativos para medir os efeitos possíveis de determinadas variáveis, chamadas de independentes, em outras, denominadas de dependentes. Busca-se aqui identificar pela teoria probabilística, possíveis

correlações e até mesmo relação de causa e efeito entre dois ou mais fenômenos que podem ser identificados, contados e medidos no mundo social. Isso pode ser feito de duas formas: ou pela análise de resultados de experiência controlada; ou pelos estudos observacionais. Pelas próprias características do objeto de análise da pesquisa social, os estudos observacionais são muito mais comuns que os experimentos controlados nas pesquisas empíricas. O uso de instrumentos estatísticos tem a finalidade de tentar reproduzir, pelo menos em parte, as análises de resultados de experiências controlados em contextos de pesquisas observacionais. Busca-se identificar causas que afetam, aumentando ou diminuindo, valores de características de grandes populações pela correlação ou medição de causalidade entre variáveis. Não se devem desconsiderar as relevantes dificuldades em se passar do âmbito de uma experiência controlada para um estudo observacional, pois elas estão presentes em todos os momentos da pesquisa. No entanto, a questão central a se estudar, de acordo com este critério, é a relação de causa e efeito a partir de possíveis explicações palpáveis e não encontrar todas as respostas possíveis para determinado comportamento social. Trata-se de buscar uma explicação de efeito médio de uma ou mais causas para determinado comportamento de uma população.

Outra finalidade da pesquisa quantitativa deve ser prestar explicações multivariadas, demonstrando que elementos da realidade não têm relevância para a determinação de causas e efeitos dos fenômenos sociais. O pesquisador deve ter em mente que, em se tratando de relações sociais, torna-se impossível explicar comportamentos a partir de cadeias causais simples. Da mesma forma como existem causas indiretas, aditivas ou ocasionais por interação entre os elementos do mundo social, também há variáveis que só apresentam relação com determinado efeito quando agregadas a outras. Muitas vezes, há um grande número de efeitos implícitos no trajeto causal para explicação de determinados efeitos. A literatura está repleta de demonstrações e exemplos de que a aproximação individual sobre efeitos continua sendo o que mais aparece nas abordagens com uso de ferramentas estatísticas pelas ciências sociais, o que demonstra que ainda há muito que fazer nessa área para obter um resultado mais próximo do limite potencial que esses instrumentos podem oferecer às ciências sociais. Da mesma forma que se identificam as limitações dos modelos causais univariados,

percebe-se também uma presença de variáveis com efeitos causais comuns somando-se para explicar determinado fenômeno. Para a explicação da realidade, isso não traz nenhum ganho. Pode, no máximo, melhorar o desempenho dos resultados de alguns testes estatísticos utilizados como instrumentos para a análise do pesquisador. Nesse caso, há um sério risco de transformação dos dados estatísticos, que deveriam ser "meio" para se produzirem análises relevantes do mundo social, em "fim", o que pode ser chamado em "fetichismo estatístico". Quando a obtenção de um modelo mais adequado estatisticamente se sobrepõe à análise contextualizada dos fenômenos sociais que dão substância à estatística aplicada, temos um problema de finalidade do pesquisador e não da técnica utilizada para produzir pesquisa científica. Responsabilizar o instrumento quantitativo por possíveis desvios de finalidade da pesquisa é o mesmo que culpar o pincel e as tintas por um quadro de qualidade artística duvidosa.

Uma terceira finalidade é a seleção de casos a serem analisados que permitam a generalização dos resultados. Para isso, os métodos quantitativos costumam adotar grandes populações para maximizar a força dos resultados estatísticos de generalização e, com isso, incentivam uma compreensão mais ampla do espaço social já retratado nas teorias. Some-se a isso a heterogeneidade causal nas pesquisas empíricas sobre sociedade e tenha como resultados diferentes objetivos para as generalizações. As grandes populações podem servir para estimar o efeito médio de uma ou algumas variáveis sobre determinado fenômeno, assim como a exclusão de variáveis associadas à explicação total. Seguindo essa finalidade da pesquisa quantitativa, variáveis importantes apenas para um pequeno subconjunto de casos podem ser desconsideradas ou transformadas em termo de erro. Para evitar esse tipo de "desvio", os pesquisadores quantitativistas precisam se preocupar em selecionar casos a serem estudados, sem levar em conta o peso individual de cada um deles na variável dependente (nesse sentido, a pesquisa qualitativa é diametralmente oposta). A seleção de casos pelo seu valor individual na característica que se quer estudar resulta em viés estatístico e impossibilita qualquer descrição mais objetiva da realidade estudada. Para evitar isso, quantitativistas usam sempre que possível a seleção de casos aleatórios, pelo menos no que diz respeito às variáveis a serem pesquisadas.

Uma finalidade complementar à anterior, para evitar o extremismo do trata-

mento igualitário de variáveis com papel desigual nas explicações sobre fenômenos sociais, é a atenção especial a determinados casos e ajuste a determinadas observações. Aqui, como acima, o ponto de partida é que todas as observações são igualmente importantes, ou seja, não se faz nenhuma suposição a respeito das observações. Pesquisadores quantitativistas buscam estabelecer um padrão de comportamento para confrontá-lo em um teste padrão de hipóteses. Em função disso, uma única ou poucas observações são insuficientes para sustentar uma teoria e os resultados que se afastam da norma já constatada devem ser tratados como observações específicas (*outliers*), ou seja, com suspeição. É a partir dessa desconfiança inicial que o pesquisador define que casos ganham relevância para a explicação posteriormente à análise, mas nunca antes dela. Posteriormente, é possível examinar os efeitos de cada caso e verificar se os *outliers* têm mais poder explicativo do que os resultados estatísticos esperados.

Por fim, a última finalidade a ser apontada aqui é a que trata da definição de medida empírica relacionada a um conceito teórico. Como todas as demais pesquisas, as que utilizam os métodos quantitativos também partem de conceitos teóricos e precisam transformá-los em indicadores empíricos. Com isso, é possível que exista um conflito entre os modelos teóricos prescritivos e resultados dos modelos estatísticos, cujos resultados mostram-se incompatíveis com as ideias/conceitos apropriadas para dar início à pesquisa. Há duas maneiras para resolver essa incongruência, quando ocorre. A primeira é tratar como erro os fatores causais não sistemáticos que geraram viés nos resultados empíricos, o que garante a manutenção da capacidade explicativa modelo teórico anterior. A outra é considerar o erro como uma ocorrência verdadeira da realidade, porém, como a atenção principal direciona-se para o centro dos indicadores, esse fator de erro pode ser considerado como um novo indicador, até então não identificado pela ciência. Assim, pode ser agregado à teoria e conceitos anteriores, permitindo não apenas a preservação deles, mas também a atualização. Por estes motivos é que, na pesquisa quantitativa, a falha de um modelo teórico, quando usado para explicar casos particulares, não é um problema tão grande, visto que muitos fatores idiossincráticos podem influenciar casos particulares. Quando esses fatores não são importantes para uma teoria geral, que explique o conjunto de fenômenos de grandes populações, não deve ser considerado pelos métodos quantitativos. Assim como a exclusão ou o não

tratamento analítico das características específicas não altera ou enviesa as estimativas para as variáveis mais gerais do modelo.

Nos termos utilizados por Mahoney e Goertz (2006), a pesquisa quantitativa focaliza sua atenção menos nos erros de medida que podem ser derivados das definições e preocupa-se, principalmente, com a operacionalização e uso de indicadores. Para os pesquisadores, o erro de medida ocorre no nível dos indicadores e não dos conceitos, o que permite que as discussões metodológicas sobre esse tipo de erro fiquem limitadas às formas de modelagem dos indicadores, distanciando-se das origens e aplicações próprias do nível conceitual. Assim, as definições de finalidades principais da pesquisa quantitativa podem ser sumarizadas em:

a) identificação das causas dos fenômenos;

b) identificação de fatores que não explicam os fenômenos;

c) seleção de casos para análise que representem grandes populações;

d) tratamento igualitário de todos os casos *a priori*;

e) tratamento próprio das possíveis inconsistências entre resultados das medidas de variáveis e conceitos anteriores.

A partir delas, o pesquisador quantitativista pode contar com uma série de técnicas de coleta de informações e de análise dos elementos da realidade para desenvolver seus estudos. As técnicas mais usadas na pesquisa quantitativa para análises empíricas podem ser agrupadas em três grandes conjuntos de finalidades:

a) experimento de campo

b) pesquisa descritiva

c) pesquisa exploratória

O experimento de campo, como o próprio nome indica, é a técnica presente em estudos nos quais o pesquisador promove alguma alteração intencional no objeto de análise – de maneira direta ou indireta – para, a partir dessa mudança, medir os padrões de comportamento e de relação entre diferentes variáveis. A pesquisa descritiva, por sua vez, utiliza-se de uma série de informações coletadas em uma grande massa de variáveis e reduzidas a indicadores representativos da dinâmica social, para descrever como os fenômenos acontecem. O estudo de caso é um tipo específico de pesquisa descritiva que visa encontrar o maior número de explicações e as descrições

mais abrangentes a respeito de uma situação social específica. Trata-se de uma aproximação entre a estratégia comparativa, que procura estudar comunidades, e a estratégia quantitativa, que busca medir o maior número possível de características da população analisada. Já a pesquisa exploratória aplica-se aos objetos de estudo pouco descritos ou analisados pela literatura já existente. Trata-se, no caso dos métodos quantitativos, de aplicar instrumentos para medição e descrição inicial de determinados comportamentos sociais que até então receberam pouca atenção dos cientistas.

Para tornar as técnicas descritas acima funcionais à pesquisa científica, os pesquisadores utilizam várias técnicas de coleta de dados. As mais comuns são as entrevistas, na maioria das vezes com questionários dotados de grande número de perguntas e, quase sempre, com alternativas de respostas pré-estabelecidas. Isso se deve ao fato de que os questionários são aplicados a um grande número de respondentes ou corre-se o risco de perder a representatividade do todo. O exemplo mais comum desse tipo de técnica de coleta de informações são as chamadas pesquisas de opinião pública. Também podem ser realizadas consultas a bases de dados e relatórios primários, de onde são extraídos dados sobre as variáveis do objeto que se pretende estudar. Um exemplo dessa técnica é o recurso da busca de informações em arquivos público de dados sobre inflação do país ou a busca de informações no Tribunal Superior Eleitoral (TSE) sobre características socioculturais dos candidatos que disputam eleições no Brasil. Outro método de pesquisa quantitativa é a Análise de Conteúdo, que busca investigar em documentos públicos, tais como discursos, literatura, publicações oficiais, músicas e imprensa, entre outros, indicadores sobre o comportamento, demandas e preferências sociais de determinado grupo humano. Em todos os casos, independente da técnica de coleta utilizada, a análise quantitativa depende da aplicação de métodos estatísticos para identificação de frequências, correlações, associações e causalidades.

Por fim, a título de breve conclusão sobre o uso dos métodos quantitativos nas ciências sociais, é preciso ressaltar que o conjunto de técnicas apresentado aqui visa oferecer aos estudos científicos ferramentas para contar e medir atitudes, opiniões e características da população pesquisada que podem ser materializadas. Essas medições, quando aplicadas para relacionar um conceito teórico a determinadas ocorrências empíricas, são legítimas representações de regularidades sociais, porém, não exclusivas

dos cientistas. Outros profissionais também podem produzir representações sociais legítimas. A particularidade aqui está no fato de que os cientistas usam os métodos quantitativos para poder compartilhar as representações de regularidades sociais com outros cientistas, que por usa vez conhecem e utilizam as mesmas ferramentas de pesquisa.

Por mais que os pesquisadores usem adequadamente os métodos de pesquisa de que dispõem, ou seja, para relacionar as finalidades de um estudo específico às teorias e conceitos transmitidos por gerações anteriores de cientistas, toda análise empírica está sujeita a algumas críticas. A primeira delas é que o resultado pode ser considerado trivial, ou seja, apesar do uso de técnicas avançadas de pesquisa social, pode-se afirmar o que "todo mundo já sabia". Outra crítica comum nasce do fato de que os resultados de determinado estudo podem ser questionados pela citação de casos contraditórios. Além disso, os resultados da pesquisa social podem ser desconsiderados porque os pesquisadores podem ter perturbado a regularidade social por opção própria e com o objetivo de relatar as consequências desse ato.

De qualquer maneira, os métodos quantitativos podem servir como antídoto - ainda que parcial -, para os três conjuntos de críticas. Em relação à primeira, é possível que todos (inclusive cientistas ensaístas) já tivessem uma impressão sobre a forma como ocorrem os fenômenos sociais, porém, sem a medição estatística e o estabelecimento de relações causais significativas, não se pode ter precisão a respeito da dinâmica das relações sociais. A existência de casos contraditórios não reduz a validade da ciência, pelo contrário, a medição quantitativa parte do princípio de que nem toda a heterogeneidade social poderá ser tratada em um único estudo. E, se considerarmos que o objeto das ciências sociais – a sociedade - é vivo e se encontra em permanente transformação, nem mesmo com um grande conjunto de estudos científicos será capaz de dar conta de toda a diversidade. Em terceiro lugar, o uso de métodos quantitativos permite uma significativa redução da intervenção direta do pesquisador no seu objeto de estudo, a realidade social, por estar tratando de grandes populações através de um número significativo de variáveis e sobre características já conhecidas.

I.iii. Exercícios propostos

A partir do texto da introdução, responda as seguintes questões de forma objetiva:

a) Descreva as quatro etapas da pesquisa empírica apresentadas na introdução. Relacione cada uma delas em um desenho de pesquisa empírica específico.

b) Discorra sobre os objetivos de cada uma das estratégias: quantitativa, comparativa e qualitativa. Para que serve cada uma delas? Exemplifique.

c) O que caracteriza e o que há de específico nas técnicas de pesquisa de "experimento de campo", "pesquisa descritiva" e "pesquisa exploratória"? Dê exemplo para cada uma delas usando o mesmo objeto empírico.

I.iv Bibliografia citada no capítulo

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

CONDE, Maria Rosa B.; ROMÁN José A. **Investigar en comunicación**: guía práctica de métodos y técnicas de investigación social en comunicación. Madrid – España: Mc Graw Hill, 2005.

HERZCOVITZ, Heloiza G. Análise de conteúdo em jornalismo. In: LAGO, Claudia e BENETTI, Marcia. **Metodologia de Pesquisa em Jornalismo.** Petrópolis – RJ: Editora Vozes, 2007.

MAHONEY, James; GOERTZ, Gary. A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research. **Political Analysis Review**, n. 14, p. 227-249, 2006.

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

REIS, Fábio W. O Tabelão e a Lupa. **Revista Brasileira de Ciências Sociais**, n. 16, ano 6, p. 27-41, 1991.

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

LAKATOS, Imre. **La Metodología de los Programas de Investigación Científica.** Madrid – España: Editora Alianza Universidad, 1989.

WEBER, Max (2001). **Metodologia das Ciências Sociais**. Cortez Editora: São Paulo – SP.

Capítulo 1

Conceitos Fundamentais

A base do conhecimento científico
é formada por conceitos.
Sem compreender os fundamentos deles
não é possível fazer ciência.

Em sentido amplo, ciência significa conhecimento, enquanto em um sentido estrito, ela visa explicar fenômenos com evidências possíveis. Logo, ciência deve ser considerada como uma parte do conhecimento, que por sua vez integra a cultura. Isso significa que existem outras formas de conhecimento além das científicas. O que distingue a ciência dos demais conhecimentos é o método de explicação dos eventos, que nada tem a ver com o senso comum ou da religião, baseados na fé e, até mesmo, de parte da filosofia. Este manual trata dos principais aspectos de um dos métodos de pesquisa científica nas ciências humanas, que é a pesquisa empírica utilizando o método quantitativo, ou, simplesmente pesquisa quantitativa.

Método, do grego, significa "*caminho para...*". O método científico possui uma lógica própria para validar seus resultados. Ele é basicamente empírico-dedutivo, chegando a conclusões a partir da análise do mundo real. Existem três pares de conceitos distintos usados para esclarecer essa definição de conhecimento científico:

- Método x Técnica: as técnicas de pesquisa variam muito dentro do mesmo método nas diferentes áreas da ciência.

- Contexto de Validação x Contexto da Descoberta: o contexto da descoberta está ligado ao início do trabalho, mas ele precisa ser validado pela metodologia científica, que é a segunda etapa do processo, na qual acontece a justificação científica das descobertas iniciais. A validação acontece tanto no nível lógico (teórico), quanto no nível prático (com a realidade).

- Ciência como Produto x Ciência como Processo: a ciência como produto é um conjunto de afirmações linguísticas, de proposições. Já o processo de construção do conhecimento é muito variado. O produto é o resultado do processo científico, ou seja, são as teorias. O processo diz respeito aos procedimentos para se chegar ao produto e, nesse caso, existem muitas variações.

Há três grandes tipos de abordagens científicas: **as descritivas, as verificativas e as exploratórias**. A última vem antes dos estudos descritivos e busca conhecer um objeto de pesquisa novo ou ainda pouco explorado. Nos estudos descritivos, as perguntas principais são: *o que é? Como é?* Já nos estudos verificativos, a pergunta principal é: *por quê?*

Na primeira fase das interpretações científicas, após a análise das características e observações do objetivo estudado, o pesquisador usa um conjunto de conceitos e suas definições para descrever algo. Esta é a etapa da **taxonomia**, pois ainda não há uma teoria presente no processo. Apenas quando a proposição relaciona vários conceitos pode-se considerar a existência de uma teoria. As taxonomias podem ser avaliadas como mais ou menos ricas, porém não se pode considerá-las como verdadeiras ou falsas. Já as teorias, quando se sustentam, são verdadeiras, quando não, são falsas.

Quadro 1.1 – Taxonomia X Teoria

Taxonomias - Ricas ou Pobres (em função da utilidade para a teoria). **Teorias -** Verdadeiras ou Falsas (em função da capacidade de sustentação no mundo empírico).

Tanto as taxonomias quanto as teorias são tratadas a partir de supostos, pressupostos e orientações gerais, que podem ser verdadeiras ou falsas, úteis ou não úteis.

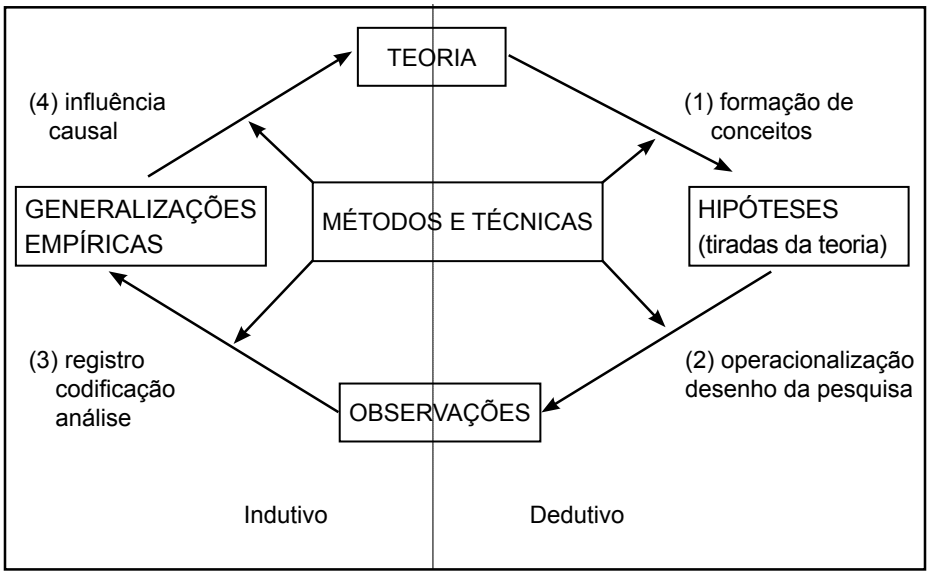
Muitas vezes existe uma teoria e taxonomia juntas quando o autor teoriza sobre uma taxonomia, misturando as duas esferas. Uma diferença entre teorias e pressupostos é que as primeiras são confrontadas com a realidade, enquanto as últimas não.

1.1 Teorias, métodos e observações

Toda teoria surge como resultado de pesquisas científicas. Já essas pesquisas **nascem de um problema científico**, que é um problema da realidade. Quando o pesquisador pergunta "*por quê?"*, ele transforma o problema social em um problema de pesquisa científica. Portanto, todo problema científico surge de uma pergunta. Os fatores responsáveis pelas origens do problema são muito variados. Eles estão ligados ao momento da descoberta. Podem ser individuais ou coletivos, porém o problema de pesquisa sempre é apresentado em forma de pergunta. Depois disso, são formuladas as **Hipóteses**, que podem ser definidas como tentativas de respostas a um problema científico. São conjecturas que tentam responder ao problema, explicá-lo através de uma teoria ou teorias diferentes. As hipóteses podem surgir do conhecimento prático, da observação de alguns fatos ou podem até mesmo serem deduzidas de uma teoria existente, pela analogia com outras ciências ou por acaso. Uma característica importante das hipóteses é que elas precisam ser passíveis de confrontação com a realidade. Se não puderem ser testadas, não serão aceitas pela ciência.

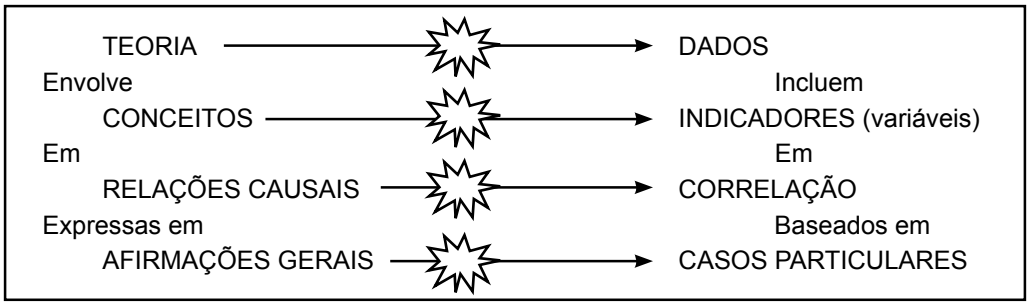
Para testar as hipóteses, são feitas **inferências**, que é o momento da pesquisa onde se considera inicialmente a validade das hipóteses que parecem lógicas a partir da dedução. Se as hipóteses parecem lógicas para as inferências, elas são testadas no mundo real. Os testes devem ser feitos mesmo que não haja uma teoria para explicá-los no primeiro momento. O ***feedback*** é a análise comparativa dos resultados dos testes em relação às hipóteses. Nesse momento, os resultados dos testes podem negar a hipótese inicial. Se isso acontecer e a hipótese inicial for descartada pelos resultados do *feedback,* é preciso formular novas hipóteses e recomeçar o processo. O ciclo de todo o processo está representado na figura a seguir (ROSENBERG, 1971).

Figura 1.1 – Ciclo da pesquisa científica proposto por Rosenberg



Nunca se pode afirmar a verdade de uma proposição. Pode-se, no máximo, afirmar que uma hipótese é falsa. Falseabilidade e refutabilidade são critérios para estabelecer o que é ciência ou não. Dentro do processo de produção de conhecimento científico, há uma constante tensão entre teoria e empiria, como representado no esquema a seguir:

Figura 1.2 – Representação das tensões entre teoria e empiria



A falta de sintonia entre teoria e empiria pode gerar o que se chama de falácia científica. Existem dois tipos principais de falácias científicas:

- Falácia ecológica – a partir de dados que dizem respeito a regiões (macro informações), tirar conclusões sobre indivíduos. **Exemplo:** negros e crimes em uma região

dos Estados Unidos. É um risco a passagem de afirmações gerais para casos particulares.

- Falácia individualista – a partir de dados individuais (Micro informações), tiram-se conclusões a respeito de regiões. **Exemplo:** uma menina tem melhores notas do que seu colega de turma no colégio. É um risco passar desse caso específico para afirmações gerais sobre o desempenho escolar de homens e mulheres.

1.2 Tipos de relações entre variáveis

Variáveis são características observáveis, com alguma definição prévia conceitual e que apresenta algum tipo de diferenciação das ocorrências, ou seja, deve variar e não ser constante. O objetivo básico de todo teste estatístico é verificar a existência, o tipo e a intensidade das variações intra e extra variáveis. Todos os testes de hipóteses científicas envolvem dois tipos principais de variáveis: **Variável independente** (x), que é o fator explicativo e a **Variável dependente** (y), que é o fator que se pretende explicar.

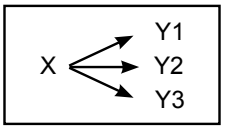
O primeiro esforço de uma pesquisa é fazer a relação com a teoria com a qual se pretende trabalhar. A teoria precisa ser explicitada para identificar a hipótese inicial.

Modelos de teorias:

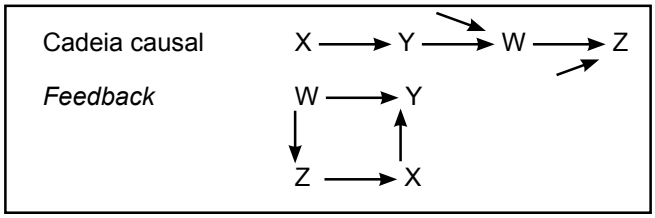
1) TIPOLOGIAS – quando a teoria estabelece tipos, ou seja, predominam as definições e classificações para distinguir situações diferentes da variável dependente.

2) INVENTÁRIO SIMPLES DE CAUSAS – explica uma ocorrência a partir de fatores externos. Esses fatores podem estar relacionados ou não.

3) INVENTÁRIO DE EFEITOS – tenta mostrar que uma variável independente tem efeitos sobre vários outros fatores.



4) CADEIAS SIMPLES OU MODELO COM *FEEDBACK* – considera que um fator influencia outro e este tem influência em um terceiro, etc.



Todo conceito é uma abstração que serve para identificar ocorrências do mundo real. A ciência precisa ter um mínimo de clareza sobre os seus conceitos. Quanto mais precisos forem os conceitos, melhor será para a compreensão. São critérios para definições claras de conceitos:

1) Deve-se procurar o aspecto essencial daquele termo;

2) A definição não deve ser tautológica (circular);

3) Não devem ser formuladas em termos negativos;

4) Não deve ser expresso em termos obscuros ou figurado.

É preciso entender que um conceito não é uma variável. Para ser estudado empiricamente, o conceito precisa ser transformado em variável e então ser testado na realidade.

Exemplos:

Status social (conceito) -------- medida de renda (variável)

Desempenho escolar (conceito) ------- notas nas provas (variável).

Para se conseguir uma boa variável é preciso estabelecer quais aspectos observáveis traduzem melhor os conceitos que devem ser medidos. Uma variável “é um conceito que pode ter vários valores e que se define de tal maneira que se pode conhecer, mediante observação, que valor tem em cada caso particular” ou “símbolos aos quais números ou valores são atribuídos para representar sua variação” (ROSE; SULLIVAN, 1996).

Para se transformar um conceito em variável operacional é preciso de definições operacionais que sejam exaustivas e mutuamente exclusivas. Uma variável precisa ter nome, definição verbal e um processo de classificação que permita identificar cada caso. Quando se trabalha com mais de uma variável, buscam-se explicações para as relações entre elas. Segundo Zetterberg (1963), as relações causais entre variáveis

podem ser de cinco tipos:

I

Reversível – se X, então Y; se Y, então X. **Exemplo:** intenção de voto e prestar atenção na campanha de determinado candidato.

Irreversível – se X, então Y; se Y, nenhuma conclusão sobre X. **Exemplo:** posição social e intenção de voto.

II

Determinista – se X, então sempre haverá Y. Essas relações são muito raras nas ciências sociais. **Exemplo:** burocratização leva à oligarquização.

Probabilística – se X, então provavelmente Y. **Exemplo:** dada certa idade, então as atitudes dos homens são mais conservadoras.

III

Sequencial – se X, então mais tarde ocorrerá Y. **Exemplo:** influência do grupo de relações sociais e a direção do voto.

Coexistente – se X, também Y. **Exemplo:** urbanização e industrialização.

IV

Suficiente – se X, então Y, independente de qualquer outra coisa. **Exemplo:** ter câncer é suficiente para morrer.

Contingente – se X, então Y apenas se houver Z. A relação contingente é mais comum nas ciências sociais. **Exemplo:** raça e direção do voto - depende da classe social.

V

Necessárias – se X e somente X, então Y. **Exemplo:** crise de legitimidade e outros fatores que levaram aos processos de democratização nos anos 80.

Substituíveis – se X, então Y, mas se Z, então também Y. **Exemplo:** se um avião cair há morte, mas se houver câncer também há morte.

Em relação aos tipos de relações entre as variáveis, elas podem ser:

Relações simétricas – que não envolvem causalidade. As relações simétricas podem ser indicadores alternativos de um mesmo conceito. **Exemplo:** aumento dos batimentos cardíacos, suor nas mãos e boca seca são indicadores alternativos de ansiedade.

Relações espúrias – equívoco na avaliação. **Exemplo:** relação entre cegonhas e nascimentos na Suécia. Quando as cegonhas voltam do inverno, ocorre um maior número de nascimentos, mas não há relação direta entre elas. Essas duas variáveis ocorrem juntas, mas uma não depende da outra.

Relação entre fatores interdependentes – quando dois fatores juntos, como as diversas dimensões da burocracia, não apresentam necessariamente uma relação de causalidade entre eles. **Exemplo:** presença de coração e pulmões no corpo humano.

Partes de um sistema comum – quando os dois fatores fazem parte de um estilo de vida. **Exemplo:** ser membro de um clube de campo e assistir a óperas.

Relações fortuitas e ocasionais – quando não há relação entre os dois fatores que acontecem no mesmo período. **Exemplo:** surgimento do rock e a corrida espacial.

Relações Recíprocas – Quando há reversibilidade entre as variáveis. Isso envolve provas separadas. **Exemplo:** Interação social e simpatia.

Relações assimétricas – Quando existe causalidade entre dois fatores.

Relação entre estímulo e resposta. **Exemplo:** assistir a um filme e adotar determinadas atitudes. Para provar essa relação é preciso comparar as atitudes finais com as anteriores ao filme.

Relação entre disposições e respostas – tendência a reagir de certa maneira em certas circunstâncias. São disposições. **Exemplo:** liberalismo e voto.

Relação entre propriedades e disposições – são as atitudes, idade, valores. **Exemplo:** idade e conservadorismo (ROSENBERG, 1971).

Uma variável independente é pré-condição para dado efeito – por exemplo, o desenvolvimento tecnológico e armamento nuclear. Uma pré-condição é necessária para ter armamento nuclear.

Relação entre meios e fins – entre o fim que se quer atingir e o comportamento adotado para isso. **Exemplo:** número de horas de estudo para notas altas.

Para afirmar que existe causalidade entre variáveis é preciso levar em conta alguns critérios teóricos. A simples observação da realidade não garante a afirmação de causalidade. Precisa haver nexo entre o fator independente e o dependente em relação à teoria que se justifica na realidade. Tem que eliminar a possibilidade das relações serem simétricas ou espúrias. Do ponto de vista empírico, é preciso demonstrar que há regularidade ou correlação entre dois fatores; sequência temporal, pois um fator só causa outro se ele vier antes e em contiguidade temporal e espacial. Já para estabelecer no nível empírico a relação de causalidade entre duas variáveis, é preciso observar diferentes categorias ou valores da variável causal e estabelecer a covariação, direção causal e ausência de relação espúria entre as variáveis. Uma causalidade entre dois fatores só é provada quando se considera a variação da variável causal historicamente. **Exemplo:** tomando os casos em que houve crise e os em que não houve crise para explicar a existência ou não de revoluções e não apenas nos casos em que houve revoluções.

Em suma, para inferir relação de causalidade entre dois fatores é preciso observar casos em diferentes categorias ou valores da variável causal; estabelecer a covariação e direção causal e mostrar que a relação não é espúria. Ainda que todas as condições acima tenham sido cumpridas, deve-se considerar a possibilidade de teorias alternativas para explicar a relação entre duas variáveis geradas porque as observações e suas relações estão ligadas a uma amostra enviesada, porque as observações se devem a um conjunto de pequenas causas, ou porque estão ligadas a um terceiro fator não considerado.

1.3 Exercícios propostos

Com base no capítulo 1, responda as questões a seguir de forma objetiva:

a) Qual a diferença entre taxonomia e teoria? Dê exemplos.

b) O que é um problema de pesquisa e o que é uma hipótese de trabalho? Por que um problema sempre é na forma interrogativa e uma hipótese sempre é uma afirmação provisória?

c) Explique o que são falácias científicas e quais os dois principais tipos de falácias apresentadas aqui?

d) Quais são os cinco tipos possíveis de relações causais entre variáveis apresentadas no capítulo? Dê exemplos.

e) O que diferencia, principalmente, as relações simétricas, espúrias e assimétricas entre variáveis?

I.4 Bibliografia citada no capítulo

ROSE, D.; SULLIVAN, O. **Introducing data analysis for social scientists.** Philadelphia, Pa: Open University Press, 1996.

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

ZETTERBERG, Hans L. **Social Theory and Social Practice**. New Jersey: The State University, 1963.

Capítulo II

Definições Estatísticas

O que se quer medir são as diferenças nas características visíveis dos indivíduos estudados e as que existem entre eles.

A nomenclatura principal da pesquisa quantitativa tem origem na estatística. A partir daqui, começaremos a discutir alguns dos principais termos utilizados na área[1].

Chama-se **População** o conjunto de elementos que queremos abranger em nosso estudo e cujas características que devemos identificar são passíveis de observação. Trata-se do conjunto de todos os elementos que possuem uma ou mais características que se pretende estudar. Isso significa que os indivíduos ou objetos de uma população devem apresentar pelo menos uma característica em comum. Essa população pode ser "finita/fechada", quando se obtém uma lista de todos os seus integrantes. Ou, pode ser "infinita/aberta", quando é impossível o estabelecimento exato de seus limites. Faz-se a delimitação da população quando se estabelece o campo de validade do conhecimento que se quer ter.

Se a pesquisa envolve grandes populações ou população infinita, torna-se necessária a realização de uma **Amostragem**, que é a seleção de uma parte da população para ser observada. Portanto, amostra é um subconjunto qualquer da população a

1 Todos os cálculos realizados na segunda parte deste capítulo foram feitos com calculadora simples, planilha de dados do LibreOfice e pacote estatístico com código aberto R-Studio.

que se quer extrair dados e informações. Uma amostra de centenas ou poucos milhares de casos pode representar uma população de milhões com grande precisão, desde que esta amostra seja extraída sob um rigoroso plano de amostragem, capaz de garantir a representatividade do todo. Portanto, amostra é qualquer subconjunto da população que deve representar as características de todo o conjunto.

Os procedimentos de retirada dos indivíduos que farão parte da amostra requerem certas técnicas e cuidados para garantir que a mesma seja representativa da população. A amostra não deve ser confundida com **espaço amostral**, visto que este último é o conjunto que contém todos os possíveis resultados de um experimento. Do espaço amostral é que sairá a amostra, portanto, podem sobrar resultados nele que não foram coletados na amostra, porém não podem faltar resultados possíveis de uma amostra no espaço amostral.

Estatística descritiva:

É parte da estatística que se preocupa com a coleta, descrição e apresentação de dados observados, sem tirar conclusões mais genéricas. É a parte da estatística que trabalha com valores totais das variáveis, por meio de uma série de fórmulas. Ela não se ocupa de valores amostrais.

Estatística indutiva ou inferencial:

É parte da estatística que, baseando-se em resultados obtidos a partir de uma amostra, procura inferir ou tirar conclusões para a população como um todo, mostrando o grau de precisão dos resultados e com que probabilidade pode-se confiar neles. A partir dela é possível chegar a conclusões a respeito das características da população, considerando os resultados obtidos nas estatísticas descritivas amostrais. No entanto, nunca haverá certeza absoluta da fidelidade de representação dos resultados nas inferências.

Parâmetros são funções de valores populacionais

Estimadores são funções de elementos amostrais.

Estimativa é o valor numérico obtido pelo estimador numa certa amostra.

Estimação é o processo que usa resultados extraídos da amostra para pro-

duzir inferências sobre a população da qual foi extraída. São dois tipos:

Estimação por ponto, quando a partir da amostra procura-se calcular uma estimativa de certo parâmetro populacional (apresenta grandes chances de erro).

Estimação por intervalo, quando procura construir um intervalo de variação com certa probabilidade de conter um verdadeiro parâmetro populacional em função das incertezas existentes.

Todos os conceitos apresentados até aqui são aplicados ao estudo da principal unidade de análise nos métodos quantitativos, que é a variável. Entende-se por **variável** cada característica própria ou adquirida pelos indivíduos que apresenta a possibilidade de variar. Então, uma variável é a característica que pode mudar entre os indivíduos analisados. A característica que não muda é chamada de **constante**. São **exemplos** de características próprias: “sexo”, “cor”, “localização geográfica do município”; e características adquiridas: “preferência partidária”, “opinião sobre temas públicos”, “número de moradores do município”. Sendo assim, o menor número de categorias em uma variável é dois. As variáveis com apenas duas categorias são chamadas de dicotômicas ou *dummies*. Por ser importante para a introdução aos métodos quantitativos, o próximo tópico abordará os diferentes tipos de variáveis em detalhes.

2.1 Tipos de variáveis

Toda variável nasce da necessidade de operacionalização de um conceito teórico. Portanto, uma variável não tem “vida independente” de uma conceituação teórica, ela sempre é o resultado de uma formulação conceitual anterior. As variáveis são utilizadas para representar características presentes em determinada população. Podemos classificar as variáveis segundo dois critérios básicos: de acordo com o número de valores que a variável pode ter ou de acordo com a sua origem.

Quadro 2.1 – Classificação das variáveis

De acordo com o número de valores
Dicotômicas
Classificatórias
Contínuas
De acordo com a origem
Naturais
Conceituais
Construídas

Exemplos:

Dicotômica/natural: sexo, nacionalidade (sim ou não)

Dicotômica/conceitual: ocupação (manual ou intelectual)

Dicotômica/construída: preferência partidária (republicana ou democrata).

Classificatória/natural: nacionalidade (por país de origem)

Classificatória/conceitual: tipo de moradia (depende dos conceitos usados)

Classificatória/construída: salário (quando se estabelece uma escala de número de salários)

Contínua/natural: idade, anos de escolaridade.

Contínua/conceitual: IDH, medidas de QI.

Contínua/construída: grau de informação sobre política

Uma variável de um tipo pode ser transformada em outro. **Exemplo**:

Idade – variável contínua – 18, 19, 20, 21, 22, 23, etc.

Idade – variável classificatória – 18 a 20, 21 a 23, etc.

Também é possível reunir duas variáveis em uma terceira. **Exemplo**:

Sexo e estado civil como variáveis independentes podem ser transformadas em outra variável com as categorias: homem casado, homem solteiro, mulher casada, mulher solteira.

Como pressuposto, para que a transformação de um conceito teórico em variável operacional seja correta é preciso considerar a existência de algumas características indispensáveis para a operacionalidade da mesma:

1º - para ser operacional, uma variável deve ser exaustiva.

2º - para ser operacional, uma variável precisa possuir categorias mutuamente exclusivas.

Uma vez respeitadas exaustividade e exclusividade de categorias na variável, ela precisa apresentar algum processo de classificação que permita identificar cada caso. Quando os possíveis resultados de uma variável são números numa certa escala, dizemos que esta variável é chamada de **quantitativa** e tem seus valores expressos em termos numéricos. Ela será quantitativa discreta quando contiver uma quantidade finita de valores e quantitativa contínua quando for medida com números fracionados. Se os possíveis resultados são atributos ou qualidades, a variável é chamada de **qualitativa ou de atributos**, já que entre uma qualidade e outra não é possível encontrar gradações. Existem dois grandes grupos de variáveis qualitativas:

Variável qualitativa nominal ou de atributos:

Quando os números são usados para nomear ou categorizar características observadas que não são diferenciadas por ordenação. Podem ser binomiais ou dicotômicas, quando a variável apresenta apenas duas categorias possíveis; e polinomiais ou politômicas, quando têm mais de duas categorias. O número atribuído à categoria da variável serve apenas como etiqueta, tendo a função de distingui-la das demais. Sendo assim, essa variável não é capaz de medir a realidade objetiva no sentido estrito da palavra, mas de classificar os fatos em determinadas categorias. Elas também são chamadas de nominais, pois sua função é dar nome aos elementos classificados. Os nomes das categorias em uma variável qualitativa nominal devem representar categorias exaustivas e excludentes. **Exemplos**:

Dicotômica: sexo do respondente – Homem ou Mulher.

Politômica: marcas de veículos, regiões de planejamento de um Estado.

Variável qualitativa ordinal ou de ordem:

Quando os números identificam, categorizam ou ordenam as unidades observadas segundo um processo de comparação. Por exemplo, nível socioeconômico toma-se a partir do rendimento entre as pessoas. Não serve para contagem propriamente dita, pois definem grupos. Além de dar nomes, elas atribuem uma hierarquia ao conjunto de categorias que representam a realidade. Os dados ordinais permitem estabelecer desigualdades, passando a significar posições graduadas de uma qualidade na série. É uma relação quantitativada. Além do "igual ou diferente" é possível falar em "maior que ou menor que". Forma-se um encadeamento lógico entre os níveis, por isso elas são chamadas de ordinais. É possível afirmar que toda variável ordinal é também uma nominal. **Exemplos**:

Escolaridade (por grau: 1º, 2º e 3º).

Desempenho do governo (ótimo, bom, regular, ruim e péssimo), etc.

Variável quantitativa discreta:

Quando os casos observados na variável podem assumir um número finito de valores em um intervalo finito. Os números passam a representar distinções reais dentro da variável, deixando de ser apenas etiquetas. Estabelece-se uma ordem entre os elementos da realidade, sendo possível medir as distâncias entre eles, o que não acontece com a variável ordinal. Quando quantitativa, seus números representam distintas quantidades de um mesmo atributo. Aqui, o que se mede é o mesmo atributo para todos e o que muda é a quantidade do atributo para cada um. **Exemplos**:

Número de projeto de leis no Congresso,

Número de estudantes em uma família,

Número de erros em um livro.

Variável quantitativa contínua:

Mantém todas as características da anterior, mas ocorre quando a variável pode assumir, teoricamente, quaisquer valores em certo intervalo finito, inclusive números quebrados. **Exemplos**:

Peso corporal de crianças,

Número de horas na Internet por dia,

Índice de inflação.

Tanto a variável quantitativa discreta quanto a quantitativa contínua podem apresentar-se em dois tipos de escalas distintas: escala intervalar ou de razão.

Escala Intervalar:

Ocorre quando o zero é arbitrário. Apresenta as mesmas características da escala ordinal, sendo possível quantificar a diferença entre dois números. As diferenças entre as categorias são iguais. Nessa escala há distância entre os dados, no entanto, não se pode concluir nada a respeito da magnitude absoluta das medições, pois a escala intervalar não apresenta zero absoluto. Portanto, temos uma escala intervalar quando a variação entre um e outro ponto é constante e os números não apenas dão ordem às diferentes categorias, mas permitem o estabelecimento de diferenças entre duas categorias. Para ser intervalar, uma variável precisa poder ordenar as categorias, portanto, toda variável intervalar é também uma ordinal e, por consequência, uma nominal. **Exemplos**:

Escala de Temperatura, onde o zero não indica qualquer ausência de temperatura;

Escala de teste de QI, onde o valor 120 está 20 acima do valor 100, mas o zero não representa ausência total de inteligência.

Escala de Razão:

Apresenta as mesmas características da escala intervalar, com a vantagem de apresentar zero absoluto. Assim, permite tirar conclusões sobre a magnitude absoluta das relações e diferenças entre os pontos analisados. Nesse tipo de escala, além da distância entre dois valores, é possível multiplicar ou dividir. Quando apresenta números quebrados é também chamada de variável contínua. Se não, é chamada de discreta. **Exemplos**:

População. Se for zero, a população foi extinta.

Idade. Se for zero, não há idade a ser medida.

Renda (uma pessoa pode ganhar o dobro de outra);

Comprimento, altura, peso, volume,

Tempo decorrido e velocidade.

Quadro 2.2 – Resumo dos tipos de variáveis

	Tipo	Características	Exemplos	Funções Formais
Qualitativa	**Nominal ou de atributos**	Apenas para identificar pessoas, objetos ou categorias.	Cor de cabelo, estado civil, nome, marca de carro.	Equivalência, igualdade. "="
	Ordinal ou de ordem	Respostas podem ser ordenadas em uma dimensão própria.	Escolaridade. Desempenho de governo. Ordem de preferência, de chegada, status social, escala de Likert. (tem moda e mediana, mas não a média)	Além de igualdade, identifica relação de superioridade ou inferioridade. Diferenças de grandezas. ">" ; "<"
Quantitativa Discreta: números inteiros; Contínua: números quebrados	**Intervalar**	Além de poder ser ordenada em uma dimensão específica há intervalos com tamanho conhecido que podem ser comparados.	Escala de Likert, estimativa de distâncias. Escala de QI.	Além das anteriores permite operações aritméticas nas diferenças entre os números que representam os eventos. Determina igualdade de intervalos.
	Proporcional ou de Razão	Além das características da escala anterior, existe ainda um ponto zero absoluto.	Salário, altura, tempo gasto em uma tarefa.	Além das anteriores permite operações aritméticas nos próprios códigos/números. Determina igualdade de relações.

Fonte: Autor, a partir de Günther, 2003.

O que diferencia, na prática, um tipo de variável de outro é o volume de informação ou nível de medida que ela oferece. Quanto maior o número de informações obtidas em uma variável, maiores são as possibilidades analíticas. De um lado, existem variáveis que permitem apenas classificar os casos observados, enquanto, de outro, há aquelas que classificam, ordenam e até indicam a distância absoluta entre os casos quando o zero é não arbitrário, conforme mostra o quadro a seguir.

Quadro 2.3 – Comparação entre níveis informativos das variáveis

Nível de medida	Classificação	Ordem	Distância	Zero não arbitrário
Nominal				
Ordinal				
Intervalar				
De razão				

Além da classificação das variáveis segundo tipo (quantitativa ou qualitativa) ou subtipo (nominal/ordinal, discreta/contínua), também é possível classificá-las em função do plano de análise que se pretende realizar. Nesse caso, temos três tipos de variáveis, independentemente da classificação anterior. Aqui, estamos pensando em termos de relações entre as variáveis. Nesse caso, as principais são três: variáveis dependentes, independentes e de controle ou intervenientes.

Variável Dependente:

É a variável que identifica o fenômeno que se pretende explicar. **Por exemplo**, preferência partidária poderia ser identificada por presença ou ausência (sim/não) para resposta à pergunta se tem preferência por algum partido político. Se a dependência partidária depende de outros fatores, então, é preciso estudá-los para saber o que de fato tem influência sobre a variável dependente.

Variável Independente:

É a variável (ou variáveis) usada como possível explicação para o comportamento da variável dependente. **Por exemplo**, ter ou não preferência por partido político poderia ser explicado por grau de interesse em assuntos relacionados à política ou ao grau de escolaridade. Essas duas variáveis serviriam para explicar, diretamente, o comportamento da dependente que é preferência partidária.

Variável de controle ou interveniente:

É a variável (ou variáveis) que se considera ser capaz de interferir na relação entre as duas anteriores, ou seja, interfere de maneira indireta na variável dependente – através do impacto que tem sobre a independente. Quando identificada, ela pode demonstrar a existência de falsas relações entre a variável dependente e independente. **Por exemplo**, a idade do respondente, como variável interveniente, pode controlar grau de interesse por política, que por sua vez tem impacto direto sobre a preferência partidária. Então, a primeira seria interveniente, a segunda independente e a última a variável dependente, ou seja, aquela que se quer explicar. As duas anteriores seriam as explicativas.

2.2 Validade e Confiabilidade de Medidas

Além da tipologia das variáveis, outra diferença entre elas diz respeito à validade e ao nível de confiança que possuem para representar determinada característica analisada. Entende-se por **validade** a congruência entre definição operacional e significado do conceito; por **confiabilidade**, a consistência e estabilidade da medida. Vale lembrar que uma medida não confiável não pode ser válida. Porém, uma medida confiável pode não ser válida. Para ser válida e confiável ao mesmo tempo, a variável só deve apresentar variações verdadeiras entre os casos observados. As **variações verdadeiras** ocorrem entre as unidades de análise em relação ao conceito que se pretende medir. Porém, também existem as variações não verdadeiras (ROSENBERG, 1971). Elas se dividem em dois tipos:

Tendenciosidade na medida:

É inerente à medida usada para o conceito e não diz respeito à variação natural dos casos. Também é chamada de erro sistemático. A tendenciosidade reduz a confiabilidade da medida e pode acontecer até mesmo em função de as pessoas sentirem-se observadas, como uma forma de reação. Por exemplo, ao serem medidas a partir de pesquisas de opinião pública, pois quase ninguém gosta de emitir opiniões que não são socialmente aceitas. Mas também existem tendenciosidades por condução no questionário, quando os indivíduos tendem a concordar mais do que discordar em função do tipo de pergunta que têm que responder. Dizer sim é mais fácil do que dizer não. Para evitar essa tendenciosidade, deve-se tomar a mesma medida em mais de uma pergunta, alternando as formas de resposta.

Erro randômico:

Não está ligado diretamente à operacionalidade do conceito, mas à operação do aplicador, como o cansaço, por exemplo. A presença, extensão e direção desse erro não são previsíveis e quase nunca podem ser identificados. Como esses erros ocorrem em várias direções, há uma tendência em se anularem. O erro randômico pode afetar a confiabilidade das medidas.

Para evitar variações não verdadeiras entre categorias de variáveis:

1) É preciso que haja correlação entre diversos itens ligados a um mesmo conceito, como, por exemplo, no caso de se medir o nível de adesão à democracia.

2) No caso de pesquisas de opinião pública, fazer pré-testes das medidas com indivíduos que não serão pesquisadas posteriormente. Se a população for pequena, o teste pode ser feito com pessoas próximas ou com características parecidas à da população a ser pesquisada.

3) Analisar o conteúdo das medidas.

4) Usar mais de um observador quando coletar dados a partir de observações como, por exemplo, a análise de programas de televisão.

5) Examinar se há correlação esperada entre as variáveis. A ausência de correlação pode ser causada por erro de medida.

Uma vez apresentadas as principais características das variáveis, é possível antecipar algumas estratégias de análise para as relações entre elas. Nos métodos quantitativos, uma restrição ao uso de determinadas técnicas estatísticas para a identificação das relações é o tipo de variável. Partindo de cada característica indica-se uma técnica analítica específica de estudo. O quadro abaixo, adaptado de Pereira (2004), resume as técnicas indicadas para os tipos de variáveis dependente, independente e interveniente.

Quadro 2.4 – Técnicas para análise em função do tipo de variável

Técnica de análise	Variável dependente	Variável independente	Variável interveniente
Correlação de Pearson (r)	Quantitativa contínua	Quantitativa contínua	Qualitativa nominal ou ordinal quando aplicado em correlações parciais
Regressão Linear	Quantitativa contínua	De preferência quantitativa	De preferência qualitativa
Associação (tab. de contingência: qui-quadrado)	Qualitativa nominal ou ordinal	Qualitativa nominal ou ordinal	Qualitativa nominal ou ordinal

Fonte: A partir de Pereira, 2004.

Agora que as características das variáveis já foram apresentadas e discutidas, podemos passar para o próximo passo, que é a transformação de um tipo de variável

em outro. Isso é necessário quando o pesquisador tem um conjunto de dados contínuos, mas pretende categorizá-los para dar continuidade aos estudos. Existem diferentes formas de transformar uma variável contínua em categórica. Aqui, aprenderemos a fazer pela fórmula de *Sturges*. A partir dela, é possível não apenas fazer a categorização, mas também conseguir categorias com limites entre si que não sejam arbitrários, quer dizer, a partir apenas da vontade do pesquisador. Com a fórmula de *Sturges*, usam-se cálculos matemáticos para estabelecer o número de categorias e os limites entre elas.

2.3 Distribuição de frequências em variável categórica (*Sturges*)

Trata-se de uma das formas possíveis de modelagem de dados já coletados para que sejam analisados. Pode ser aplicada à distribuição de uma única variável ou para duas variáveis correlacionadas. Aqui, demonstraremos a aplicação da fórmula para a distribuição de valores dos casos de uma única variável, como demonstrado por *Sturges* (1926). Antes de estarem organizados numericamente, eles são chamados de matriz de dados brutos, que são de difícil estudo. Para obter informações sobre determinado fenômeno, podemos resumir os dados brutos em uma distribuição de frequências, onde os valores observados não aparecem individualmente, mas agrupados em classes ou categorias.

Sturges propõe a distribuição de casos pelo intervalo ótimo entre as classes, ou seja, ela distribuição dos casos dentro de cada categoria, considerando a mesma amplitude entre as categorias, independente do número de casos dentro de cada categoria. Com isso, se mantém as mesmas médias e dispersão entre os casos dentro das categorias. A fórmula proposta para a distribuição de uma única variável é[2]:

$$C = \frac{R}{K} = \frac{(Xmax - Xmin)}{1 + (3{,}32\ x \log N)}$$

2 No caso da distribuição de duas variáveis, a fórmula de Sturges considera o coeficiente de correção (r) entre elas, ficando com a seguinte forma: $C = \frac{R}{K} = \frac{(Xmax - Xmin)}{1 + (1{,}661 + 1{,}661r)x \log N}$

Onde:

C = amplitude de classe;

R = amplitude total;

K = número de classes;

N = número de casos;

No exemplo a seguir, temos uma coleção de dados brutos da produção legislativa por deputado estadual do Paraná. Cada número no quadro a seguir representa o número de projetos de lei apresentados por parlamentar durante a legislatura. Como se pode perceber, é quase impossível encontrar alguma regularidade ou informação que possa ser interpretada na matriz de dados a seguir. Antes de iniciar qualquer análise, vamos fazer a distribuição das frequências pela fórmula de *Sturges*.

Quadro 2.5 - Exemplo: Número de Projetos de Lei apresentados por deputados na ALEP

55	21	41	12	7	85	21	25	34
22	49	32	23	73	41	21	50	5
65	2	21	14	42	11	20	21	11
5	14	28	18	31	9	12	22	24
45	12	39	12	11	3	20	3	34
9	51	120	26	114	27	49	13	2

Para construir uma distribuição de frequência é preciso seguir alguns passos:

1º - Estabelecer o número de classes. Uma das maneiras mais utilizadas para a determinação do número de classes é através da fórmula de *Sturges*, dada por:

$$K = 1 + (3{,}32 \times \log N)$$

Onde:

K = número de classes.

N = número total de observações.

Exemplo: K: 1 + (3,32 x log54) = 1 + (3,32 x 1,732) = 1 + 5,750 = 6,750. Resposta: o número de classes seria de 7 para esta variável, arredondando para cima.

2º - Determinar a amplitude total dos dados.

$$R = Xmax – Xmin$$

Onde:

R = amplitude total.

Xmax = valor máximo encontrado;

Xmin = valor mínimo encontrado;

Exemplo: R = 120 – 2 = 118. Resposta: a amplitude total é de 118 casos.

3º - Estabelecer o intervalo ou amplitude da classe.

$$C = \frac{R}{K}$$

Onde:

C = intervalo de classe.

R = amplitude total.

K = número de classes.

Exemplo: $H = \frac{118}{6{,}75}$. Resposta: a amplitude de classe é de 17,48.

4º - Estabelecer os limites inferiores e superiores dos intervalos de classes. Sendo que o limite inferior do primeiro intervalo deve ser menor ou igual ao menor valor da série.

Limite inferior = Xi;

Limite superior = Xi + H.

O limite inferior da segunda classe é igual ao limite superior da primeira classe.

Quadro 2.6 – Resultados e transformações das categorias

Classes	Amplitude Real	Amplitude Aproximada	Amplitude Aproximada Corrigida (P/ 16)	Categoria (Prod. Legis.)	Categoria Agregada
Classe 1	De 2 a 19,48	De 2 a 19	De 2 a 18	Muito baixa (20)	Baixa (39) 72,22%
Classe 2	De 19,49 a 36,97	De 20 a 37	De 19 a 35	Baixa (19)	
Classe 3	De 36,98 a 54,46	De 38 a 54	De 36 a 52	Média baixa (9)	Média (13) 24,08%
Classe 4	De 54,47 a 71,95	De 55 a 72	De 53 a 69	Média (2)	
Classe 5	De 71,96 a 89,44	De 73 a 89	De 70 a 86	Média alta (2)	
Classe 6	De 89,45 a 106,93	De 90 a 107	De 87 a 103	Alta (0)	Alta (2) 3,70%
Classe 7	De 106,94 a 124,42	De 108 a 120	De 104 a 120	Muito alta (2)	

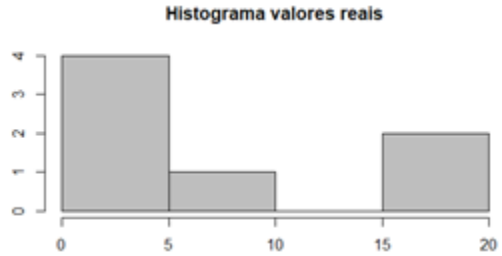


Gráfico 2.1 - Distribuição real

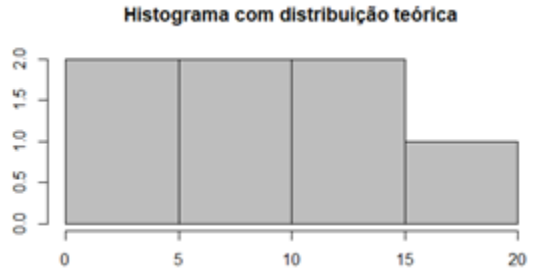


Gráfico 2.2 - Distribuição teórica

Uma forma de visualizar a distribuição acima é através do gráfico histograma. O que interessa aqui não é necessariamente a altura da barra, mas a área delimitada por ela. Aplicando ao caso acima para a distribuição entre as categorias “muito baixa” até “muito alta”, teríamos o que segue nos gráficos abaixo:

Fica evidenciado no histograma 2.1 que não se trata de uma distribuição normal, pois há uma tendência de “acúmulo de casos” nas classes iniciais, reduzindo-se drasticamente a partir da metade da distribuição. Entre outras coisas, o Gráfico 2.1 indica que as chances de encontrar deputados nos primeiros grupos de distribuição são maiores do que nos últimos grupos. Ao contrário do que acontece no histograma do Grafico 2.2, a partir de uma distribuição teórica qualquer. Fica evidente como a curva normal (Curva de Gauss) aproxima-se do formato de um sino, o que é desejável para

análises estatísticas inferenciais - como veremos mais adiante.

Porém, se o objetivo do pesquisador for distribuir as frequências em classes com o mesmo número de casos, para gerar uma igualdade entre as classes ao invés da amplitude de ocorrências, ele pode criar limites das categorias partir dos valores dos quartis. Assim, se forem quatro quartis, a distribuição das frequências será de quatro grupos com aproximadamente 25% dos casos em cada um deles – aqui, a amplitude entre os grupos é que será variável.

Quadro 2.7 - Exemplo: Distribuição frequências em 4 grupos a partir dos quartis

Classe	Amplitude Real	Amplitude Total	Nº De Casos 25% (Aprox.)	Categoria Agregada
Classe 1	De 2 a 12	10	13 (24,1%)	Muito baixa
Classe 2	De 13 a 21	8	14 (25,9%)	Baixa
Classe 3	De 22 a 39	17	13 (24,1%)	Alta
Classe 4	De 41 a 120	69	14 (25,9%)	Muito Alta

São formas distintas de classificação de uma variável contínua que geram consequências também diferentes. No primeiro caso (fórmula de *Sturges*), a amplitude entre as classes é a mesma, o que varia é o número de elementos dentro de cada categoria. No segundo caso (distribuição por quartis), o número de elementos dentro de cada categoria é praticamente o mesmo, ficando em torno de 25% do total, mas o que varia é a amplitude de cada categoria.

2.4 Distribuição em três categorias (ABC) pelo Índice de *Nihans*

Outra forma de agrupar variáveis contínuas em categorias é pelo Índice de *Nihans*, muito utilizado na área de administração para estabelecer diferenças entre fatores em função da importância relativa de cada um deles. O Índice de *Nihans* vem sendo utilizado em substituição à distribuição de Pareto (20:80) por ser mais simples e apresentar praticamente os mesmos resultados (MEIRELES, 2001). A distribuição de Pareto serve para comprovar a hipótese do economista italiano de que 80% dos resul-

tados podem ser explicados por 20% das causas. Para isso ele distribui as causas em um gráfico onde são acumulados os percentuais de explicação em três grupos (A=20%, B=50% e C=80%). No caso do Índice de *Nihans*, faz-se também uma classificação para identificar os pontos de divisão dos distintos subconjuntos das distribuições.

Por princípio, o Índice de *Nihans* não tem número de categorias pré-definido, como no caso da distribuição de Pareto, podendo ter mais que três categorias. Como aqui estamos indicando que *Nihans* pode ser usado para substituir Pareto, apresentaremos como exemplo a distribuição em três categorias, onde A=maior importância (alta); B=importância média (média); C=baixa importância (baixa). Com o índice é possível incluir em cada classe um número harmônico de casos e o resultado é que a classe A reunirá um número maior de casos (independente do número de categorias) que a classe B, que por sua vez terá um número maior de casos que a classe C e assim sucessivamente (caso se queira fazer uma distribuição em mais classes).

Para demonstrar como se calcula o índice de *Nihans*, utilizaremos os dados dos números de eleitos por partido para o Senado Federal em 2014. Ao todo, foram 27 senadores eleitos por 10 partidos diferentes. O índice permite agrupar quais foram os partidos com maior número de eleitos (grupo A), seguido pelo grupo B e depois grupo C. Evidente que a simples organização dos partidos em ordem decrescente de número de eleitos nos mostraria quais elegeram mais e quais elegeram menos. No entanto, se quiséssemos separar a distribuição em três grupos, teríamos que usar um critério *ad-hoc*. A função do índice de *Nihans* é evitar a arbitrariedade na decisão de separar o grupo dos mais importantes, do de importância mediana, daquele que tem menor importância. A fórmula para calcular o índice de *Nihans* é bastante simples:

$$N_p = \frac{\Sigma x^2}{\Sigma x}$$

Onde:

Np = limite da categoria.

x² = o quadrado do valor das ocorrências.

x = o valor das ocorrências.

Assim, o limite entre cada uma das categorias pelo índice de *Nihans* é calculado pela divisão da somatória dos quadrados de todos os valores pela somatória dos valores originais. Definido o primeiro limite, que passa a ser chamado de N_a, excluem-se os valores da Classe A e repete-se a operação para encontrar o limite da classe B (N_b). Se forem apenas três classes, por exclusão, todos os valores que ficarem abaixo do N_b pertencerão à classe C.

1º Passo - ordenar as categorias (no caso do exemplo, partidos) em ordem decrescente:

Quadro 2.8 – Eleitos por partido para Senado Federal em 2014

PARTIDO	Nº ELEITOS 2014
PMDB	5
PDT	4
PSDB	4
DEM	3
PSB	3
PSD	2
PT	2
PTB	2
PR	1
PP	1

2º Passo - calcular o N_a para todos os elementos (partidos) com a elevação ao quadrado do número de eleitos, dividido pela somatória dos eleitos:

$$N_a = \frac{25 + 16 + 16 + 9 + 9 + 4 + 4 + 4 + 1 + 1}{5 + 4 + 4 + 3 + 3 + 2 + 2 + 2 + 1 + 1} = \frac{89}{21} = 4{,}21$$

Assim, o limite da primeira classe é 4,21, o que significa que todos os partidos que tiverem acima desse valor pertencerão à Classe A. No caso do exemplo, pertencem a N_a=PMDB, PDT e PSDB.

3º Passo - refazer o cálculo para N_b, excluindo os valores que pertencem a N_a.

$$N_b = \frac{9+9+4+4+4+1+1}{3+3+2+2+2+1+1} = \frac{30}{14} = 2{,}14$$

O resultado mostra que os partidos que elegeram entre 4,21 senadores e 2,14 senadores devem ser classificados no grupo B (N_b=2,14). No caso, são: N_b=DEM e PSB. Por exclusão, todos os demais partidos pertencem à classe C, conforme o quadro a seguir.

Quadro 2.9 – Resultados finais das categorias por *Nihans*

PARTIDO	Nº ELEITOS 2014	*NIHANS*
PMDB	5	A
PDT	4	A
PSDB	4	A
DEM	3	B
PSB	3	B
PSD	2	C
PT	2	C
PTB	2	C
PR	1	C
PP	1	C
N_a = 4,21 | N_b = 2,14		

Com o índice de *Nihans* podemos distribuir os casos nas três categorias por ordem de importância, mantendo certa homogeneidade entre os grupos. No exemplo acima, a classe A representa 48% dos senadores eleitos e 30% do número de partidos, enquanto a classe B representa 22% dos eleitos e 20% dos partidos, enquanto a classe C totaliza os demais 30% do total de senadores e 50% dos partidos.

Além de permitir a classificação interna, o índice de *Nihans* também possibilita a comparação entre distribuições com números distintos de casos. Por exemplo,

imagine que ao invés que buscarmos apenas categorizar os partidos no Senado por número de leitos, quiséssemos comparar os desempenhos dos partidos nas diferentes eleições legislativas brasileiras. Poderia ser: que partidos encontram-se na classe A de *Nihans* nas eleições para vereador do Brasil, deputado estadual, deputado federal e senador? São os mesmos que têm alto desempenho nas disputas municipais e estaduais? O número de partidos com alto desempenho é maior ou menor nas eleições nacionais? Existem partidos com alto desempenho na disputa municipal (vereador), mas com baixo desempenho nas disputas nacionais? Todas estas perguntas podem ser respondidas comparando as posições dos partidos em distintas disputas pelo índice de *Nihans*, conforme o quadro a seguir:

Quadro 2.10 – Comparações de desempenho por cargo nas eleições de 2014

VEREA12	ELEITOS	*NIHANS*	DEP.EST14	ELEITOS	*NIHANS*	DEP.FED14	ELEITOS	*NIHANS*	SENAD14	ELEITOS	*NIHANS*
PMDB	7.947	A	PMDB	139	A	PT	70	A	PMDB	5	A
PSDB	5.250	A	PT	104	A	PMDB	66	A	PDT	4	A
PT	5.183	A	PSDB	96	A	PSDB	54	A	PSDB	4	A
PP	4.923	A	PSD	73	A	PSD	37	B	DEM	3	B
PSD	4.653	A	PSB	63	B	PP	36	B	PSB	3	B
PDT	3.652	B	PDT	57	B	PSB	34	B	PSD	2	C
PTB	3.570	B	PP	50	B	PR	34	B	PT	2	C
PSB	3.545	B	DEM	45	B	PTB	25	B	PTB	2	C
DEM	3.270	B	PR	45	B	DEM	22	C	PP	1	C
PR	3.175	B	PTB	38	B	PRB	21	C	PR	1	C
PPS	1.855	C	PSC	35	C	PDT	19	C	N_a = 4,21		
PV	1.584	C	PRB	31	C	SD	15	C	N_b = 2,14		
PSC	1.463	C	PROS	30	C	PSC	12	C			
PRB	1.202	C	PV	26	C	PROS	11	C			
PC do B	973	C	PC do B	25	C	PPS	10	C			
PSL	759	C	PPS	22	C	PC do B	10	C			
PMN	603	C	SD	22	C	PV	8	C			
PRP	579	C	PSL	17	C	PSOL	5	C			
PHS	544	C	PTN	17	C	PHS	5	C			
PT do B	534	C	PEN	14	C	PTN	4	C			
PTC	483	C	PT do B	14	C	PRP	3	C			
PSDC	444	C	PRP	12	C	PMN	3	C			
PTN	426	C	PSOL	12	C	PTC	2	C			
PRTB	416	C	PHS	10	C	PSDC	2	C			
PPL	177	C	PTC	10	C	PEN	2	C			
PSOL	49	C	PSDC	9	C	PT do B	1	C			
PCB	5	C	PMN	8	C	PSL	1	C			
PSTU	2	C	PRTB	8	C	PRTB	1	C			
N_a=4.221,9			PPL	3	C	N_a=38,8					
N_b=2.476,2			N_a=64,4			N_b=23,9					
			N_b=35,7								

Ao todo, foram 57.266 vereadores eleitos em 2012 por 28 partidos diferentes. Os cinco partidos da classe A (PMDB, PSDB, PT, PP e PSD) representam, juntos, 49% do total de eleitos. A classe B reúne outros 30% do total de eleitos em outros cinco partidos. Do total de 1.035 deputados estaduais eleitos em 2014 por 30 partidos diferentes, apenas quatro partidos ficaram na classe A (PMDB, PT, PSDB e PSD), com o PP que estava na classe A na eleição para vereadores de 2012 caindo para a classe B. No caso dos deputados estaduais, a Classe A representa 40% do total de eleitos em cindo partidos, a Classe B outros 29% de eleitos em 6 partidos e a Classe C, 31% dos eleitos em todos os demais partidos. Os 29 partidos que elegeram os 513 deputados federais em 2014 distribuíram-se em apenas três deles na Classe A, que representam 37% do total de eleitos, outros cinco partidos ficaram na classe B, com 32% de eleitos e os demais na classe C.

Comparando as eleições para os quatro níveis legislativos no Brasil entre 2012 e 2014, percebe-se que apenas dois partidos conseguem permanecer na Classe A em todas elas (PMDB e PSDB). PT sai da classe A no senado, onde entra o PDT, enquanto PSD mantém-se na classe A apenas nas disputas para vereador e deputado estadual.

Neste capítulo começamos descrevendo os conceitos básicos estatísticos mais usados nos métodos quantitativos aplicados nas ciências sociais. Passamos por uma definição de variável, sua tipologia quanto à origem e volume de informações que elas fornecem sobre o objeto a ser pesquisado e as relações que se dão em uma pesquisa entre pelo menos três tipos de variáveis: dependente (a que se quer explicar), independente (a que explica variações na dependente) e a interveniente (a que interfere de maneira indireta na dependente). Também aprendemos a transformar uma base bruta de dados em distribuições de frequências para obter uma variável categórica. Fizemos isso a partir de três métodos: fórmula de *Sturges*, que mantém fixos os limites entre todas as categorias; a distribuição por quartis, que fixa o número de casos em cada classe; e a fórmula de *Nihans*, que distribui os casos em três grupos hierarquicamente organizados. No próximo capítulo, estudaremos matrizes de dados e como utilizar as variáveis para criar índices e escalas de medidas.

2.5 Exercícios propostos

A partir do texto do Capítulo 2, responda as questões e faça os exercícios a seguir:

a) Qual a diferença entre estatística descritiva e estatística inferencial?

b) Relacione os conceitos “validade”, “confiabilidade”, “tendenciosidade” e “erro randômico”. O que cada um deles representa? Dê exemplos de possíveis ocorrências em bancos de dados.

c) Que tipo de teste é indicado para quando se quer verificar as relações de variações entre duas variáveis categóricas? Por quê?

A partir da sequência de dados a seguir (variável contínua “idade” no banco de dados “B_Exerc_V1” dos respondentes a um questionário), faça:

42	39	20	19	19	20	19	33	24	23	21	29	23	18	37
18	19	19	22	20	19	23	38	18	20	20	49	39	22	32
23	21	27	29	34	45	20	21	24	31	41	31	24	25	27
19	18	19	28	21	22	24	19	20	33	19	26	20	19	20

d) A transformação em variável categórica pela fórmula de *Sturges*.

e) A transformação em variável categórica pela fórmula de *Nihans* e compare com a distribuição categórica anterior.

2.6 Bibliografia citada no capítulo

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

MEIRELES, Manuel. **Ferramentas administrativas para identificar, observar e analisar problemas**. São Paulo: Arte e Ciência, 2001.

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

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

STURGES, Herbert A. **The choice of a class interval. Journal of the American Statistical Associatio**n, v. 21, n.153, p. 65-66, 1926.

Capítulo III

Dados, Indicadores, Índices, Escalas e Formas de Representação

O mundo é caótico, mas, quando as informações são organizadas em bancos de dados e em indicadores, ele começa a ficar mais organizado.

Existem diferentes tratamentos para dados em pesquisas empíricas que utilizam técnicas de análise quantitativas, porém todos eles dependem de uma organização inicial das informações coletadas em campo, chamada de matriz de dados. Na maioria das vezes, uma matriz de dados quantitativos é a forma de organizar as informações disponíveis das unidades pesquisadas em linhas e colunas. A partir daí, já podendo comparar as características de diferentes unidades de análise é possível agregar informações em indicadores mais elaborados.

3.1 Matriz de Dados

A estrutura básica dos dados de pesquisas sociais é composta por:

Unidades – Exemplos: Pessoas, escolas, regiões, cidades, etc.

Variáveis – dimensões a se conhecer das unidades

Valores – categorias que ligam as variações das unidades.

Um conjunto de unidades de análise com respostas para um conjunto de variáveis é o que forma uma matriz de dados.

Quadro 3.1 – Exemplo de matriz de dados

	V1	**V2**	**V3**	**V4**	**Vn...**
O1	R11	R21	R31	R41	R1n...
O2	R21	R22	R32	R42	R2n...
On...	R1n...	R2n...	R3n...	R4n...	Rnn...

Todos os estudos chegam a esse tipo de matriz de dados. A variação entre eles depende do número de variáveis e das unidades de análise. Portanto, as unidades básicas de um banco de dados composto por distintas variáveis são os valores que indicam as categorias dentro de cada característica.

A tabela 3.1 a seguir exemplifica uma base de dados formada a partir de informações dadas por um conjunto de 20 alunos de um curso superior qualquer a partir de oito perguntas. Pode-se perceber que a base de dados é formada por variáveis, casos e valores para cada uma das interseções entre variáveis e casos. Cada linha representa um caso, que aqui é um aluno. As variáveis são as seguintes:

Id: Identificador. Variável contínua e organiza em ordem crescente os respondentes, que aqui tiveram seus nomes substituídos por números para preservar o anonimato.

Idade: Variável quantitativa contínua. É a idade em número de anos completos indicada pelos respondentes.

est_civil: Estado civil do respondente. É uma variável qualitativa nominal. Aqui foram apontadas apenas duas categorias: solteiro ou casado.

not_méd: Nota obtida pelo aluno ao final da disciplina. Variável quantitativa contínua.

num_falt: Número de faltas. Outra variável quantitativa contínua, indica o total de faltas do aluno ao longo do semestre.

grau_interesse: Grau de interesse na disciplina. Variável categórica ordinal,

com duas categorias: interesse principal ou interesse secundário.

temp_ocup: Tempo de ocupação além da faculdade. Qualitativa ordinal, com categorias: não tem ocupação; até 2 horas; de 2 a 4 horas; de 4 a 6 horas; acima de 6 horas.

tip_escola: Tipo de escola em que cursos o segundo grau. Variável categórica ordinal, pois indica desde quem estudou só em escola pública até quem fez todos os estudos secundários em escola privada. Divide-se nas categorias: Tudo em privada; maior parte em privada; maior parte em pública; tudo em pública.

Tabela 3.1 – Banco de dados de Alunos da graduação

id	Idade	est_civil	not_méd	num_falt	grau_interesse	temp_ocup	tip_escola
1	18	Solteiro	70	2	Interes. principal	de 4 a 6 horas	Mais em privada
2	19	Solteiro	65	4	Interes. secundário	+ de 6 horas	Mais em privada
3	19	Casado	80	2	Interes. secundário	+ de 6 horas	Mais em privada
4	22	Solteiro	95	2	Interes. Principal	de 4 a 6 horas	Mais em pública
5	20	Casado	70	2	Interes. secundário	de 4 a 6 horas	Mais em privada
6	19	Solteiro	45	2		de 4 a 6 horas	Mais em pública
7	23	Solteiro	80	2	Interes. Principal	+ de 6 horas	Tudo em privada
8	38	Casado	85	4	Intere secundário	+ de 6 horas	Mais em privada
9	18	Solteiro	95	2	Interes. Principal	Não tem	Tudo em privada
10	20	Solteiro	80	6	Intere secundário	Não tem	Mais em privada
11	20	Solteiro	80	4		+ de 6 horas	Mais em privada
12	49	Casado	75	0	Interes. secundário	+ de 6 horas	Mais em pública
13	39	Casado	70	0	Interes. principal	de 4 a 6 horas	Tudo em pública
14	22	Solteiro	85	6	Interes. Principal	+ de 6 horas	Mais em privada
15	32	Solteiro	85	2	Interes. Principal	+ de 6 horas	Mais em pública
16	19	Solteiro	80	4	Interes. Principal	Não tem	Tudo em privada
17	18	Solteiro	90	0	Interes. secundário	+ de 6 horas	Tudo em pública
18	19	Solteiro	85	6	Interes. secundário	de 2 a 4 horas	Mais em pública
19	28	Solteiro	80	4	Interes. secundário	+ de 6 horas	Mais em pública
20	21	Solteiro	85	0	Interes. Principal	+ de 6 horas	Tudo em pública

Toda matriz de dados é formada por variáveis separadas. No entanto, a partir de uma matriz, o pesquisador pode reunir duas ou mais variáveis para criar uma nova, que seja mais rica na representação dos fenômenos e permita uma redução das dimensões analisadas. Assim, a relação matemática entre a nota obtida pelo aluno e o seu percentual de frequência nas aulas pode ser transformada em um único indicador,

denominado Índice de Rendimento Acadêmico (IRA). Esse índice, portanto, passa a ser uma nova variável que substituirá as duas anteriores. A seguir, discutimos a criação de índices de escalas de valores.

3.2 Índices e escalas

Índices podem ser definidos como medidas mais sofisticadas das variáveis criadas a partir de dados primários. O índice é econômico porque reduz para uma única dimensão um conjunto de variáveis ligadas a determinado conceito. Além disso, alguns conceitos são dificilmente medidos de maneira completa por variáveis isoladas. Em muitos casos, apenas um índice que leva em conta distintos indicadores pode fornecer as informações necessárias. Por isso, trabalhar com índices torna a pesquisa mais rica em informações. No entanto, é preciso ter em mente que dificilmente consegue-se chegar a uma medida inquestionável. Apesar disso, os índices e escalas são usados por várias razões:

- É raro ser possível chegar a um único item que represente uma variável complexa. Como o índice reúne vários itens, esse problema é solucionado.
- Há variáveis dicotômicas que são transformadas em ordinais através de índices.
- A análise pode ser mais eficiente.

Diferenças entre índices e escalas:

Ambos são ordinais e compostos por diversos itens tomados como indicadores de uma mesma variável. O índice é constituído pela soma dos códigos de cada item seguida da atribuição de pontos, sem que haja uma estrutura de intensidade entre os pontos. A escala se dá através da atribuição de pontos ao padrão da resposta, seguida de estruturação e intensidade. Portanto, as escalas são formas de medidas mais sofisticadas que os índices.

Critérios de seleção de itens para formação de índices e escalas:

Existem quatro critérios principais para a seleção de variáveis que irão compor

os índices ou escalas:

- Validade lógica: cada item tem que estar na mesma dimensão do conceito. Uma variável que tem como unidade o indivíduo não pode entrar no mesmo índice de outra variável cuja unidade é o município em que ele vive. Este seria um exemplo de falta de validade lógica.

- Variância do item: é preciso variar para entrar no índice. Item que não varia, não é variável e, portanto, não acrescentará nada ao índice ou escala.

- Existência de relação bivariada entre os diversos itens do índice. É preciso que exista alguma relação (que pode ser positiva ou negativa) entre cada par de variáveis que irá compor o índice ou escala. Assim como se a correlação entre duas variáveis de um mesmo índice for perfeita (1), pode-se dispensar uma das variáveis, pois o conjunto delas não acrescenta nada para o índice que será construído.

- Existência de relação multivariada quando se tratar de índice ou escala formada por mais de duas variáveis. Se houver três itens, eles precisam manter a relação entre si. Se alguma das variáveis não tiver essa relação, deve ser abandonada.

Além disso, é importante para a construção de índices e escalas considerar os seguintes fatores: período que se quer trabalhar; população que se está trabalhando; tipo de ocorrência a medir.

Atribuição de pontos em escalas de respostas:

Antes da atribuição de pontos para as respostas em um questionário, é preciso tomar duas decisões preliminares: 1) Decidir sobre a amplitude da variação dos índices. O melhor é ter maior amplitude possível. Mas, deve haver um número adequado de casos em cada valor do índice. 2) Decidir sobre que valor atribuir aos diversos itens. É possível ter itens com pesos diferentes no índice.

Também é preciso pensar, no momento da definição de pontuações em uma escala, em como tratar não respostas (NR, NS, etc.). Os casos de *missing* dificultam a construção do índice. Porém, existem diversas alternativas para resolver esse problema: se forem poucos os casos, podem ser excluídos, ou pode-se incluí-los analisando o conteúdo da resposta.

Exemplos:

a) Interpretar casos de Não Sei e Não Respondeu, atribuindo valores a eles. Não Sei em resposta à participação em associações em geral pode ser tratado como Não Participação.

b) Dar um valor conservador a uma resposta (Não Sei) se as respostas em outras variáveis forem no sentido de conservador.

c) Atribuir valor intermediário. Variável concorda 1 e não concorda 0 é transformada em concorda 2, indiferente 1 e não concorda 0.

d) Atribuir valores aleatórios, de forma randômica aos casos sem resposta.

Será que um índice é mesmo capaz de medir determinado conceito da forma esperada? Para responder esta pergunta, deve-se fazer: 1) Análise interna do índice, para saber se há correlação entre os itens do próprio índice; 2) Análise externa do índice, para saber se há correlação do índice com outras variáveis da mesma dimensão.

3.2.1 Tipos de Escalas

Escala de Likert (mais comum)

São escalas de cinco pontos que variam de 1. Concorda muito; 2. Concorda; 3. Indeciso/indiferente; 4. Discorda; e 5. Discorda muito. Há várias formas de construir escalas. Elas oferecem mais segurança na ordenação porque leva em conta a intensidade das respostas, as oposições e ainda apresenta um ponto médio ou neutro.

Exemplo: Construção de uma escala de visibilidade dos candidatos na cobertura de quatro jornais nacionais (Folha de S. Paulo, O Estado de S. Paulo, O Globo e Jornal do Brasil) das eleições presidenciais de 2002, a partir de Cervi (2003):

> As variáveis envolvidas nessa escala são: número da página, posição na página e formato de matéria, considerando que essas três variáveis reunidas são capazes de indicar maior ou menor visibilidade dos textos em um jornal, os códigos das diferentes ocorrências de cada uma delas receberam pesos que variam de 1 a 6. Logo, cada uma das variáveis integrantes da escala tem peso de 33,33% na composição final do índice. A variável Número da Página é importante para a visibilidade do texto porque se sabe que as páginas ímpares de um jornal são mais visadas que as páginas pares, e que a primeira página tem a maior visibilidade. As páginas do primeiro caderno têm uma visibili-

dade maior que as dos outros cadernos. Assim, no primeiro caderno, as ocorrências em primeira página receberam peso 6, as ímpares receberam peso 5 e as das páginas pares tiveram peso 4. No segundo caderno, a ocorrência em primeira página teve peso 3, as páginas ímpares receberam peso 2 e as páginas pares, peso 1. Considerando o método de leitura ocidental da esquerda para a direita e de cima para baixo, a posição de um texto na página do jornal indica se ele tem potencial para ser lido por um número maior ou menor de pessoas. Sendo assim, as ocorrências de página inteira receberam peso 6, as que estavam em qualquer quadrante da metade superior da página tiveram peso 3 e as da metade inferior ficaram com peso 1. Isso para manter a proporcionalidade de 1/3 de influência no índice final de cada um dos três fatores de visibilidade. O formato da entrada também é importante para a definição da visibilidade de um texto jornalístico. Na editoria política, as colunas assinadas e reportagens são mais procuradas pelos leitores que os artigos assinados e editoriais, assim como as chamadas de primeira página são mais lidas. As chamadas receberam peso 6; as colunas assinadas, peso 5; as reportagens, peso 4; as charges, fotos e infográficos tiveram peso 3; os artigos assinados receberam peso 2 e os editoriais tiveram peso 1. Somando os três pesos para cada entrada teremos um índice que varia teoricamente de 3 a 18. Porém, o valor 18 é apenas teórico, pois para alcançá-lo seria necessário uma chamada de primeira página, no primeiro caderno e de página inteira, o que é na prática impossível de acontecer. Também não houve nenhum somatório 3, ou seja, um editorial em página par de segundos cadernos e na metade inferior da página. A divisão das categorias ficou entre os valores 3 e 7 para visibilidade baixa; de 8 a 10 para visibilidade média; de 11 a 13 para visibilidade alta e de 14 a 16 para visibilidade muito alta. Cruzando essa nova variável com as aparições e valência é possível identificar possíveis distorções no tratamento dado aos diferentes candidatos pelos jornais, no que diz respeito à maior ou menor visibilidade em relação à cobertura positiva ou negativa dos concorrentes (CERVI, 2003, p.48-49).

Resultado:

Tabela 3.2 - Índice de visibilidade

Categoria	N	% Válido
Muito baixa	30	0,1
Baixa	1520	6,8
Média	8482	38
Alta	11416	51,1
Muito alta	875	3,9
Total	**22323**	**100**

Escala de distância social de Bogartus

Usada para medir o grau de sociabilidade. **Exemplo:** discriminação a estrangeiros: 1. Você aceitaria que albaneses morassem no seu país; 2. Você aceitaria que albaneses morassem na sua cidade; 3. Você aceitaria que albaneses morassem no seu bairro; 4. Você aceitaria que albaneses fossem seus vizinhos; 5. Você aceitaria que um albanês se casasse com sua filha/filho.

Escala de Thurstone

Com grupos a partir de uma lista de itens. Em seguida, são atribuídos pontos para cada item, utilizando-se os itens em que houver concordância de valores dados por juízes.

Escala de Guttman

Baseada nas ideias de que alguns itens têm mais peso que outros. Inicia do mais intenso e segue até o menos intenso.

Sempre haverá casos que contrariam a escala. Por isso, só deve haver escala quando entre 90% e 95% dos casos puderem ser enquadrados nela. Caso contrário deve ser tratado como índice e não como escala.

Tipologias

São medidas multidimensionais que levam em conta vários itens, mas não são escalas. **Exemplo:** Conservadorismo na política externa e interna.

Se combinadas com outras variáveis pode resultar em novas tipologias. **Exemplo:** Sexo e nível de informação podem resultar em homens informados e homens não informados, mulheres informadas e mulheres não informadas.

As vantagens das escalas e índices é que elas reduzem tendenciosidades. Podem fornecer amplitude da variação maior que as variáveis simples e também reduzem o número de variáveis.

3.3 Criação de Escalas por Soma de Valores e teste de Confiabilidade (α de *Cronbach*)

A criação de escalas a partir da soma de valores individuais é uma das ferramentas estatísticas mais usadas para a sumarização de valores individuais nas ciências sociais por apresentar uma série de vantagens. A primeira delas é que uma escala contém valores de diferentes variáveis ou itens que normalmente são somados, tornando a análise mais rica e simples ao mesmo tempo (SPECTOR, 1992). A escala individual mais usada foi proposta por Rensis Likert, em 1932, e apresenta cinco possibilidades de respostas. **Por exemplo**: 1. Discordo fortemente; 2. Discordo; 3. Nem discordo nem concordo; 4. Concordo e 5. Concordo fortemente.

Imagine se em um *survey* são feitas 6 perguntas com respostas em escalas que vão de máxima discordância a máxima concordância a respeito de itens que são relacionados. Nesse caso, o pesquisador pode tentar reunir as informações, somando as respostas, em uma Escala por Soma de Valores. **Por exemplo**, em um *survey* aplicado a alunos da Universidade Federal do Paraná em 2013 foram feitas seis perguntas a respeito do "papel do Estado", como segue:

B.11. O papel do Estado é: Garantir educação universitária universal e gratuita
B.12. O papel do Estado é: Garantir acesso à saúde pública universal e gratuita
B.13. O papel do Estado é: Garantia de renda mínima para cidadãos em situação de pobreza
B.14. O papel do Estado é: Assegurar a oferta de cotas para minorias em universidades públicas
B.15. O papel do Estado é: Promover Reforma agrária com desapropriação de grandes propriedades
B.16. O papel do Estado é: Promoção de fontes alternativas de informação.

As alternativas de respostas cada uma das perguntas são:
1. Concordo fortemente
2. Concordo
3. Discordo
4. Discordo fortemente

Como se pode perceber, optou-se por uma transformação da escala de Likert, com a exclusão do ponto intermediário.

Se somarmos os valores obtidos nas seis variáveis, encontraremos que os valores mais altos indicam maior discordância em relação ao papel do Estado para garantir bens sociais de maneira universal e os valores mais indicam maior concordância com a participação do Estado para garantir bens sociais universais. Para criar a escala por soma basta somar todos os valores indicados por respondente. Nesse caso, o valor mais baixo possível é 6, para os que concordam fortemente com todas as afirmações, e o valor mais alto da escala seria 24, para os que discordam fortemente de todas as afirmações. Portanto, o Índice por soma original variaria em 18 pontos (24 - 6 = 18). Por se tratar de uma escala contínua, esse índice poderia ser usado em testes estatísticos, tais como regressões, para verificar a força da adesão às afirmações de intervenção do Estado na sociedade para distribuição de bens sociais para explicar outros comportamentos ou opiniões. Ou, então, o Índice de Soma pode ter seus valores agregados em cinco categorias, voltando a ser uma escala de Likert, porém, agora com informações agregadas de seis variáveis que antes estavam sendo analisadas individualmente. Por exemplo, ficaria:

Quadro 3.1 – Agregação de valores para categorias do índice

Valor	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Categoria	Muito alta adesão				Alta adesão				Média adesão			Baixa adesão				Muito baixa adesão			

Aplicando ao questionário citado acima, ao todo 445 alunos responderam as seis perguntas e a distribuição da Escala por Soma é a seguinte:

Quadro 3.2 – Distribuição dos casos por valor da soma e por categoria

SOMA	Nº CASOS	% CASOS	CATEGORIA	% CASOS
6	77	17,3	Adesão muito alta	47%
7	33	7,42		
8	55	12,36		
9	44	9,89		
10	48	10,79	Adesão alta	41,54%
11	52	11,69		
12	49	11,01		
13	36	8,09		
14	18	4,04	Adesão média	9,21%
15	12	2,7		
16	11	2,47		
17	5	1,12	Adesão baixa	2,25%
18	2	0,45		
19	2	0,45		
20	1	0,22		
21	0	0	Adesão muito baixa	0,00%
22	0	0		
23	0	0		
24	0	0		

Percebe-se uma concentração de casos na parte inicial do Índice, com quase 90% dos casos entre Adesão muito alta e Adesão alta. Apenas 2,25% dos casos não adere às afirmações feitas no questionário. Nenhum respondente discordou fortemente de todas as afirmações. Como disse anteriormente, a escala contínua pode ser usada em correlações ou regressões com outras variáveis contínuas. **Por exemplo**, pode-se verificar se o Índice de Adesão às Ações Sociais está correlacionado com a Posição Ideológica dos respondentes (que no questionário vai de 1 - extrema esquerda até 7 - extrema direita).

O teste de correlação linear de Pearson para essas duas variáveis mostra um coeficiente r=0,517 e sig.=0,000, indicando correlação significativa entre as duas variáveis. Como o coeficiente é positivo, isso indica que quanto mais à direita o respondente se posiciona no espectro ideológico, menor é a sua adesão à participação do Estado na distribuição de bens sociais. Se quisermos usar o Índice por soma de valores ca-

tegóricos, podemos cruzá-lo em uma tabela de contingência com a variável “Sexo do respondente” para identificar se homens e mulheres distribuem-se de maneira similar ou diferente quanto às categorias de adesão à distribuição de bens públicos, como segue.

Tabela 3.4 – Distribuição das categorias de adesão por sexo do respondente

CATEGORIA	HOMEM	MULHER
Muito alta	85 (52,47%)	124 (43,82%)
Alta	57 (35,19%)	128 (45,23%)
Média	16 (9,88%)	25 (8,83%)
Baixa	4 (2,46%)	6 (2,12%)
Muito baixa	0	0
Total	162 (100%)	283 (64,6%)

Percebe-se que os homens estão mais concentrados na categoria muito alta, com 52,47%, do que as mulheres, que apresentam 43,82% nessa categoria. Essa diferença inverte-se na categoria “alta” e, nas demais, as distribuições ficam muito próximas. Isso indica que homens e mulheres têm uma opinião sobre a presença do Estado para universalização de bens sociais muito parecida.

Antes de construir um Índice por Soma de Valores, é preciso tomar o cuidado de testar a confiabilidade dos valores que compõem o índice. Isso porque nem todas as variáveis podem ser inseridas em um índice, ainda que exista coerência lógica entre elas. As variações das respostas é que indicam se um índice por soma de valores está confiável o suficiente ou não. Para verificar a confiabilidade do índice, recomenda-se usar o coeficiente α de *Cronbach*. O coeficiente α normalmente é positivo e varia de zero a 1,00. Quanto mais próximo de zero, menor a confiabilidade dos itens que compõem o Índice. Portanto, se esperam valores de α mais próximos de um para que os componentes do Índice sejam confiáveis. Esse coeficiente usa a variância total e as variâncias de respostas para cada componente do índice no cálculo. Se o α for negativo, isso significa que uma ou mais variáveis do Índice está com escala invertida, com valores que aumentam em uma direção enquanto as demais diminuem na mesma direção. Nesse caso, é preciso corrigir os códigos da variável que está com escala invertida.

Os princípios teóricos que norteiam um teste de confiabilidade são apresentados por Carmine e Zeller (1979) ao definirem que o α de *Cronbach* é um teste utilizado para validar uma proposta teórica de índice a partir da reunião de diferentes variáveis. Ainda que os componentes pretendidos para o índices tenham validade teórica, sem a validação, não há nenhuma garantia que as variações totais das características apresentem alguma consistência empírica. De acordo com Cronbach (1971), o teste de validação verifica se um instrumento de medida é válido para testar um tipo específico de fenômeno assim como poderia testar outro fenômeno. A validação não identifica as relações do instrumento de medida com outras variáveis, mas sim os instrumentos em relação a si mesmo. No caso, o instrumento é o índice.

Para identificar a força de um instrumento em relação a si mesmo, parte-se do princípio que as variações internas devam ser consistentes, ou seja, estarem indicando a mesma direção (CARMINE; ZELLER, 1979). Por isso, o teste de α de *Cronbach* é calculado a partir das variações individuais e totais de todos os itens de um índice. As validades a serem testadas podem ser de três tipos:

i) *Criterion-related validity*: identifica a importância de um critérios ou conceito para a medição de um comportamento ou resultado que é externo ao próprio instrumento de medida. **Por exemplo**, esse tipo de validade é encontrado na composição de um teste para medir o conhecimento das pessoas sobre a direção de veículos. Se o teste tiver validade, seus critérios deverão representar o mais fielmente possível as diferentes capacidades daqueles que se submeteram a ele. Essa validade depende da correspondência entre os critérios teóricos e a operacionalização dos critérios.

ii) *Content-validity*: é uma validade muito comum em áreas como psicologia e ciências sociais. Está relacionada a quanto uma medida empírica reflete o domínio de um conteúdo específico. **Por exemplo**, quando as operações aritméticas de adição, subtração, multiplicação e divisão identificam diferentes relações aritméticas, a capacidade de diferenciá-las indica uma validade de conteúdo, que é diferente de um teste que mede apenas uma das operações. A validação de conteúdo envolve diferentes passos que são interconectados e é capaz de explicar, por exemplo, quanto que o conceito de alienação explica de comportamentos como isolamento e autodistanciamento social.

iii) *Construct-validity*: está ligada ao tipo de relação existente em ter conceitos e

teorias, não aos objetos empíricos. Ela indica o quanto um conceito teórico/abstrato é entendido como válido para determinado contexto teórico. Esse tipo de validação parte da relação entre os conceitos para chegar à aplicação em uma evidência e a interpretação do processo identifica quão explicada foi determinada medida em um contexto teórico a partir do conceito usado. Para as ciências sociais, esse tipo de validação pode também ser usado para verificar se uma medida empírica pode ou não ser apresentada como válida em determinado contexto teórico. Seu objetivo é explicar a performance de determinada medida empírica em função das expectativas teóricas. Caso sejam identificadas inconsistências entre as expectativas teóricas e as medidas empíricas, significa que as medidas não representam da forma esperada o conceito teórico. Em outras palavras, a media empírica apresenta falhas de construção para um conceito particular (CARMINE; ZELLER, 1979).

A lógica de um teste de validação sustenta-se na medida de confiabilidade daquilo que é representado pelos resultados obtidos. Quanto menor a variação não explicada, mais confiável será a medida e, portanto, maior a validade para representar determinado conceito. A equação abaixo representa essa relação:

$$x = t + e$$

Onde “x” é a medida empírica a ser estudada, que é composta por dois elementos: “t” que é a parte explicada da variação e “e” que é o erro randômico, ou seja, a parte não explicada e que não apresenta nenhuma tendência. Em toda medida a ser validada, os valores esperados são hipotéticos e não podem ser medidos diretamente. Na possibilidade de medir diretamente a relação entre um conceito teórico e uma variável empírica, o teste de validação perde o sentido de existir. Os valores apresentados em um teste como esse são aqueles obtidos em um conjunto de medições realizadas em determinado período de tempo. Como nenhuma medida individual pode ser representativa do valor real, usa-se a estimação de confiabilidade para indicar quanto podem ser confiáveis os resultados obtidos em determinado conjunto de observações.

O erro randômico representa um conjunto de distúrbios ou variações não explicadas no momento em que a medida é tomada. Existem erros positivos e negativos

e, se forem randômicos, terão a mesma magnitude, ou seja, ao final eles se anularão. Se isso não acontecer, os erros não serão randômicos e passarão a ser tendenciosos. Assim, a primeira hipótese de um teste de validação é que a soma dos erros equivale a zero, a média dos erros deve ser zero e a correlação entre os erros de uma variável e os erros de outra também deve ser zero.

Sendo assim, um teste de confiabilidade parte da razão de variâncias. Quanto maior a variância real, menor o erro randômico e maior a confiabilidade da medida. Em outras palavras, uma medida de confiabilidade é a correlação entre medidas paralelas de variâncias (CARMINE; ZELLER, 1979). A tabela a seguir indica os valores padronizados de α de *Cronbach* a partir do número de itens (variáveis) que compõem o indicador e a partir da média da correlação entre os itens. Com essas duas informações é possível identificar qual o α de *Cronbach* esperado em um teste de confiabilidade.

Tabela 3.5 – Valores esperados para correlações entre itens

N. de itens	Correlação média entre itens					
	0	0,2	0,4	0,6	0,8	1
2	0,000	0,333	0,572	0,750	0,889	1,000
4	0,000	0,500	0,727	0,857	0,941	1,000
6	0,000	0,600	0,800	0,900	0,960	1,000
8	0,000	0,666	0,842	0,924	0,970	1,000
10	0,000	0,714	0,870	0,938	0,976	1,000

Fonte: CARMINES, Edward G & ZELLER, Richard A. Reliability and validity assessment. California, USA: Sage University Paper 1979.

A tabela acima mostra como há duas tendências nos testes de validação. A primeira, óbvia, é que quanto maior a média da correlação entre os itens, maior o coeficiente α de *Cronbach*, pois essa correlação média explica o quanto existe de variação real entre os itens de um indicador. A segunda tendência, não tão óbvia, é a de que quanto maior o número de itens ou variáveis, maior tende a ser o coeficiente α de *Cronbach*. Isso porque normalmente as variações de uma nova variável tendem a se relacionar com as variações das anteriores, reduzindo os erros randômicos. Porém, nem sempre é assim.

A fórmula para o cálculo do α de *Cronbach* é a seguinte:

$$\alpha = \frac{k}{k-1}\ x\ \frac{St^2 - \sum Si^2}{St^2}$$

Onde:

K = número de variáveis que compõem o índice;

St^2 = Variância Total (a partir da soma das variâncias dos respondentes);

$\sum Si^2$ = Soma das Variâncias das variáveis ou itens que compõem o índice.

Para calcular St^2 utiliza-se a seguinte fórmula:

$$\mathrm{St}^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

Onde:

x = soma das variâncias das respostas por respondente;

n = número total de respondentes;

Vamos calcular o α de *Cronbach* para o exemplo das respostas às questões que compõem o Índice de Adesão à universalização de bens sociais, descrito acima. Primeiro, temos que encontrar a Variância Total. Os valores necessários para o cálculo são: n = 475 e x = 4675. Então, temos que:

$$\mathrm{St}^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = \frac{50.253 - \frac{21.855.625}{475}}{474} = \frac{4.241{,}15}{474} = 8{,}947$$

A soma das variâncias dos itens é: $\sum Si^2 = 3{,}26$. Agora, substituindo os valores na fórmula de α de *Cronbach*, temos que:

$$\alpha = \frac{k}{k-1}\ x\ \frac{St^2 - \sum Si^2}{St^2} = \frac{6}{6-1}\ x\ \frac{8{,}947 - 3{,}260}{8{,}947} = \frac{6}{5}\ x\ \frac{5{,}687}{8{,}947} = 1{,}2\ x\ 0{,}635 = \mathbf{0{,}762}$$

O resultado é que para o índice por somatória de valores dos 6 itens apresentados aqui, o α = 0,762. Considera-se como valor mínimo de α para ser confiável um coeficiente de 0,70. No nosso caso, ficamos um pouco acima desse limite, portanto, podemos considerar o índice composto pelas variáveis como confiável. A escala mais usada para identificar o nível de confiabilidade a partir do α é a seguinte:

Quadro 3.3 – Limites para confiabilidade do Índice de Cronbach

Coeficiente α	Nível de confiabilidade
Acima de 0,90	Excelente
De 0,80 a 0,89	Bom
De 0,70 a 0,79	Aceitável
De 0,60 a 0,69	Questionável
De 0,50 a 0,59	Pobre
Abaixo de 0,50	Inaceitável

Seguindo o quadro acima, nosso α ficou na faixa de “aceitável”. Para alterar e ver se uma variável está reduzindo o α, podemos fazer testes excluindo uma ou mais variáveis do Índice para refazer o *Cronbach*. Assim, podemos verificar qual a melhor composição para o índice no que diz respeito à confiabilidade dos resultados. O quadro a seguir compara os resultados do teste de *Cronbach* para o exemplo do Índice formado pelas variáveis discutidas acima. Para fazer as comparações, retiramos uma variável de cada vez para gerar um novo α. Os resultados são:

Quadro 3.4 – Comparação entre Cronbachs por composição do índice

Composição do índice	α de Cronbach
B11 + B12 + B13 + B14 + B15 + B16	0,762
B11 + B12 + B13 + B14 + B15	0,756
B11 + B12 + B13 + B14 + B16	0,676
B11 + B12 + B13 + B15 + B16	0,694
B11 + B12 + B14 + B15 + B16	0,702
B11 + B13 + B14 + B15 + B16	0,754
B12 + B13 + B14 + B15 + B16	0,736

Nesse caso, o melhor índice quanto a confiabilidade foi aquele composto pelas seis variáveis, visto que todas as vezes em que excluímos uma delas, o α diminuiu. Nem sempre é assim. Muitas vezes uma variável a menos faz o α subir.

3.4 Principais Indicadores Estatísticos

Além dos índices e escalas, outra forma de fazer análise estatística de dados é a partir dos indicadores. Um indicador é obtido a partir de cálculos matemáticos entre duas ou mais variáveis, podendo ser considerada, ainda, a mudança ao longo do tempo. Os mais comuns são:

Razão: Usada para grupos distintos (Razão Z = X/Y). Aplica-se quando um valor não faz parte de outro, ou seja, uma razão representa o resultado da interação entre outros dois números distintos. É muito usada para posição de grupos e posição de população. Sua maior limitação é que a razão não é dinâmica. Serve para descrever os eventos de uma população estática. **Exemplos:** razão de sexo, razão dependência e razão de densidade populacional.

Proporção: É um tipo específico de Razão onde o denominar inclui o numerador (Proporção Z = X/X+Y). **Exemplo:** proporção de mulheres no mercado de trabalho formal é obtida a partir da divisão do número de mulheres que trabalha pelo total de mulheres (número de mulheres que trabalha mais as que não trabalham).

Percentagem: É um tipo de proporção, onde o resultado do cálculo de proporção é multiplicado por 100.

Taxa: As taxas são os indicadores mais usados nas pesquisas sociais. Ela reflete melhor a dinâmica dos eventos sociais em um intervalo de tempo qualquer. Taxa é o número de eventos ocorridos em um intervalo dividido pelo número de indivíduos dispostos ao risco de enfrentar o evento no mesmo intervalo. Leva em conta

o conceito de exposição ao risco, o que não acontece nos anteriores. **Exemplo:** taxa de fecundidade exclui homens e mulheres fora da idade reprodutiva, ou seja, o número de nascimentos no período é dividido pelo número de mulheres entre 15 e 45 anos no período. A forma correta de evitar o viés em função dos desvios da exposição ao risco, nos casos de fecundidade e mortalidade, é usar o total de indivíduos expostos ao evento no meio do período (por exemplo, no meio do ano). Com isso busca-se cancelar possíveis erros. **Importante**: Muitas taxas na verdade são razões e recebem apenas o nome de taxa.

Probabilidade: São os indicadores mais populares, onde o numerador é o número de eventos ocorridos em um intervalo específico dividido pelo número de indivíduos expostos ao risco de experimentar o evento no início do período. Serve para superar as dificuldades da falta de informações no período. **Exemplo:** carros roubados no ano dividido pelo número total de carros no início do ano. Veremos as probabilidades em maiores detalhes mais adiante.

A seguir, serão apresentados alguns dos indicadores eleitorais, educacionais e socioeconômicos mais comuns em estudos e na literatura das ciências sociais. Os primeiros são indicadores parlamentares e eleitorais, retirados de Santos (2002).

3.5 Principais indicadores Parlamentares e Eleitorais

Indicadores Parlamentares

- Índice de Fracionalização: $1 - \sum pe_2$

Onde:

pe = percentual de cadeiras ocupadas por partido político (RAE,1971).

Mede a dispersão partidária no parlamento e mostra a possibilidade de dois parlamentares tomados ao acaso pertencerem a partidos diferentes.

- Fracionalização máxima: $\frac{N(n-1)}{n(N-1)}$

Onde:

N = número de cadeiras.

n = número de partidos parlamentares.

Serve para medir a relação entre o número máximo de partidos no parlamento, tendo por base o número de partidos parlamentares existentes.

- Fragmentação: $\frac{ÍndiceFracionalização}{ÍndiceFracionalizaçãoMáxima}$

A partir dos dois índices anteriores, a fragmentação no parlamento pode variar entre eles.

- Número efetivo de partidos: $\frac{1}{\sum pe_2}$

Onde,

Pe = proporção de cadeiras ocupadas por partido político (LAAKSO E RAIN, 1979)

Serve para medir o número efetivo de partidos políticos com importância numérica no parlamento.

Indicadores de Renovação das Bancadas

- Renovação Bruta: $\frac{(Desistentes + Derrotados)}{Total} x100$

Mede o número total de representantes novos em uma legislatura, comparando à anterior.

- Renovação Compulsória: $\frac{(Desistentes)}{(Total)} x100$

É o percentual de novos parlamentares que substituíram os representantes da

legislatura anterior que não se candidataram à reeleição.

- Renovação Líquida: $\frac{(Derrotados)}{(\mathrm{Re}eleitos + Derrotados)} x100$

Indica o número de candidatos à reeleição que foram derrotados divido pelo total de candidatos reeleitos e derrotados.

- Taxa de Conservação: $\frac{(reeleitos)}{(derrotados + reeleitos)} x100$

Mostra o percentual de reeleitos em relação àqueles que se recandidataram.

Quando há mudança no tamanho das bancadas ou da representação legislativa, todas as fórmulas anteriores precisam ter seus valores ponderados pela diferença no número de vagas entre as eleições em análise.

- Renovação Vegetativa:

$$\frac{(diferençaNúmeroVagasEntreEleições)}{(Desistent. + Derrot. + Dif.Núm.VagasEleições)} x100$$

É a renovação que acontece quando há um aumento no número de representantes com crescimento de vagas nos legislativos, por efeito da legislação eleitoral.

- Índice de Desproporcionalidade de Loosemore e Hanby:

$$\frac{1}{2}\sum(V - C)$$

Onde:

V = percentual de cadeiras obtidas pelo partido.

C = percentual de votos obtidos pelo partido.

Mais conhecido como índice D, mostra a diferença entre o número de cadeiras e de votos obtidos em determinada eleição.

Indicadores Eleitorais:

- Índice de Avanço Eleitoral:

$$\frac{VotoPartidoT_1 - VotoPartidoT_0 / VotoPartidoT_0}{VotoTotalT_1 - VotoTotalT_0 / VotoTotalT_0}$$

Mede o crescimento real de um partido político no eleitorado entre duas disputas, sem a influência das variações no tamanho do colégio eleitoral (SANTOS, 2002).

- Cociente Legal: $\frac{Eleitorado}{NúmerodeVagas}$

Indica o número de eleitores que cada cadeira no parlamento representa.

- Cociente Operacional ou Eleitoral: $\frac{VotosVálidos}{NúmeroVagas}$

Indica o número de votos necessários para se obter uma cadeira no parlamento. Até 1997, no Brasil, os votos brancos eram somados aos válidos no cálculo do cociente operacional. Isso foi alterado com a Lei 9.504/97.

- Cociente Partidário: $\frac{VotosPartido}{CocienteEleitoral}$

Onde:

VotosPartido = soma entre votos na legenda e votos dados aos candidatos do partido. Indica o número de lugares obtido pelo partido na distribuição dos votos disputados na eleição. Deve-se desconsiderar as frações superiores ou inferiores. Elas serão tratadas na distribuição das sobras.

- Distribuição das Sobras: $MédiaPartido = \frac{VotosPartido}{CocientePartidário + 1}$

As vagas remanescentes são distribuídas pelo critério de maiores médias, obtidas a partir da fórmula anterior de distribuição das sobras. O cálculo deve ser repetido a todos os partidos até que sejam ocupadas todas as vagas remanescentes.

- Índice de Competitividade: $IC = \frac{N}{2W} - 1$

Onde:

N = número real de candidatos.

W = tamanho da bancada (número de vagas).

Normativamente, é preciso que haja um número mínimo de candidato, pelo menos duas vezes superior ao número de vagas, para que haja competitividade eleitoral. Quanto maior for o resultado da equação, maior será a competição entre os candidatos na eleição.

Categorias de competitividade (SANTOS, 2002):	
Eleições Não competitivas:	
- Sub-competitiva (negativa)	IC < 0
- Competitividade zero	IC = 0
- Competitividade quase zero	0 < IC <0,1
- Quase-competitiva	0,1≤ IC < 0,6
Eleições Competitivas:	
- Baixa competitividade	0,6≤ IC ≤ 1
- Alta competitividade	IC > 1

3.6 O Índice de Presença Eleitoral (*Ipe*)

Proposto para análises ecológicas, ou seja, a partir de dados agregados (dos partidos políticos) e não individuais (de integrantes da elite política ou do público), o *Ipe* é um índice aditivo de proporções, variando entre limites teóricos zero a um (CERVI, 2016). Quanto mais próximo de um, maior será a presença eleitoral de um partido em uma circunscrição eleitoral em determinada eleição. Em sua composição, estão as proporções de candidatos e proporções de eleitos em disputas majoritárias e proporcionais (como o sistema eleitoral brasileiro é misto, é necessário distinguir entre os dois tipos de disputa

por conta das diferenças nos números de candidatos e eleitos); as proporções de votos nos candidatos e votos nos partidos (quando em disputas proporcionais); e as proporções de doadores e proporções de recursos doados ao partido. No limite, se um partido apresenta todos os candidatos, tem todos os eleitos, faz todos os votos para candidatos e de legenda, obtêm todas as doações e doadores de campanha, seu *Ipe* nesse município será 1,0 (um), ou seja, valor máximo – o que indicaria completa ausência de competição eleitoral. Ao contrário, se não apresenta candidato, não elege, não tem votos pessoais ou de legenda e não recebe nenhuma doação, sua presença nesse município será zero.

Dada a particularidade do sistema eleitoral misto no Brasil, onde em um mesmo dia são eleitos representantes para cargos Executivos pelo sistema majoritário e para cargos Legislativos pelo sistema proporcional, o índice proposto considera as duas dimensões separadamente. A justificativa para isso é que um partido pode optar estar presente em uma circunscrição eleitoral na disputa majoritária, porém, não na proporcional ou vice-versa. Nas eleições proporcionais, os partidos podem apresentar até 1,5[1] vezes o número de vagas em disputa nos legislativos municipais, estaduais e para Câmara Federal. Partidos também podem fazer coalizões eleitorais e, em função do tamanho das coligações, registrar apenas uma parte do total de candidatos, deixando espaço para outras siglas integrantes da coligação. Já nas eleições majoritárias, para cargos de prefeito, governador, senador ou presidente, cada partido pode apresentar apenas um candidato[2].

Em eleições concomitantes, os eleitores votam no mesmo dia para prefeito e vereador – em disputas municipais – e para deputado estadual, federal, senador, governador e presidente – quando em disputas regionais e nacionais, que ocorrem ao mesmo tempo. Em relação à dimensão “recursos de campanha”, mantém-se o financiamento de disputas proporcionais e majoritárias na mesma dimensão, formando um único componente por distrito eleitoral. Embora parte seja destinada diretamente aos candidatos, em todas as eleições há comitês financeiros partidários para administrar a maior parte das doações recebidas. Tanto comitês financeiros quanto os diretórios partidários rece-

1 Há exceções para distritos eleitorais pequenos, onde é possível apresentar até duas vezes o número de vagas em disputa no distrito.

2 A exceção das eleições parciais para o Senado, quando são preenchidos 2/3 das cadeiras, com duas vagas por distrito eleitoral. Nesse caso o partido pode optar em apresentar um ou dois candidatos por distrito eleitoral.

bem doações e redistribuem para todos os seus candidatos em disputas majoritárias e proporcionais, produzindo ao final uma única prestação final de contas, conjunta, pelo diretório partidário à justiça eleitoral. Assim, a fórmula para representação do Índice de Presença Eleitoral é a seguinte:

$$Ipe = \frac{(Dmaj + Dpro + DR\$)}{3}$$

Onde:

Ipe = Índice de Presença Eleitoral;

D_{maj} = Dimensão disputas majoritárias;

D_{pro} = Dimensão disputas proporcionais;

DR$ = Dimensão Recursos de Campanha.

O *Ipe* é o resultado da média aritmética das proporções das três dimensões: eleição majoritária, eleição proporcional e recursos de campanha. Cada uma das dimensões é composta por distintas proporções. Em todas elas, há normalização em proporções, com variações teóricas indo de zero a um para cada dimensão. Para os cálculos das dimensões, são utilizados diferentes números de indicadores.

O valor da dimensão Disputas Majoritárias (D_{maj}) é obtido da seguinte forma:

$$Dmaj\,\frac{(Cgov + Pgov + Csen + Psen\ \ )}{4}$$

Onde:

DV_{maj} = dimensão disputas majoritárias;

C_{gov} = candidato ao governo (onde 0,5=vitorioso e 0,25=derrotado);

P_{gov} = proporção de votos do candidato ao governo;

C_{sen} = candidato ao senado (onde 0,5=vitorioso e 0,25=derrotado);

P_{sen}= proporção de votos do candidato ao senado;

A dimensão “disputas majoritárias” é o resultado da média aritmética de qua-

tro variáveis no caso de eleições regionais. Se o partido tem candidato ao governo, ele obtém peso de 0,50 se o candidato é vitorioso e 0,25 se derrotado. Se o partido não apresenta candidato ao governo, o indicador é zero. Isso por considerar que a eleição de um candidato majoritário indica maior presença do partido no distrito eleitoral. Presentes apenas na dimensão majoritária, os pesos fixos de 0,50 e 0,25, para vitorioso e derrotado, são os valores que apresentam as menores variabilidades em relação ao conjunto de candidatos de cada eleição[3]. Equivale a dizer que, se o partido se apresenta, já terá um "crédito" de 0,25 e se ganha, sua presença será maior, por isso, o valor de 0,50. O objetivo é dar ao partido que apresenta candidato às disputas majoritárias um peso que o distingue dos que não conseguem apresentá-los.

Esses valores não têm influência direta sobre a dimensão, pois eles serão ponderados pelos outros fatores que são levados em conta na dimensão majoritária. Por exemplo, a viabilidade eleitoral do candidato será medida no fator seguinte, que é a proporção de votos obtidos pelo partido. O segundo fator da dimensão é a proporção de votos que o partido obteve para a disputa ao governo do Estado. As mesmas duas variáveis aplicam-se às disputas para o senado, com peso para candidato eleito ou derrotado e a proporção de votos ao senado. Para as eleições municipais, o cálculo é feito utilizando apenas as duas variáveis da disputa para prefeito: situação do candidato à prefeitura e proporção de votos obtidos pelo partido a prefeito, divididos por dois. Com isso, temos uma proporção da dimensão "disputas majoritárias".

A segunda dimensão é a das "disputas proporcionais", que segue o mesmo princípio anterior, porém, com maior número de variáveis dada a possibilidade de votos nominais e votos em partido. A fórmula para a dimensão "disputas proporcionais" (D_{pro}) é a seguinte:

$$Dpro = \frac{((Pcfed)^2 + Pefed + Pvnfed + Pvpfed + (Pcest)^2 + Pest + Pvnest + Pvlest))}{8}$$

3 Os valores de 0,5 para eleito e 0,25 para candidato, diferenciando da ausência de concorrentes foram obtidos a partir de testes de variância, optando pelos valores que permitam diferenciar partidos com maior presença (candidato eleito) do que não tem nenhuma presença na circunscrição eleitoral, sem que isso gere uma diferença com impacto maior que as demais dimensões.

Onde:

D_{pro}= dimensão de votos proporcionais;

P_{cfed}= proporção de candidatos a deputado federal;

P_{efed}= proporção de eleitos a deputado federal;

P_{vnfed}= proporção de votos nominais a deputado federal;

P_{vlfed}= proporção de votos de legenda a deputado federal;

P_{est}= proporção de candidatos a deputado estadual;

P_{eest}= proporção de eleitos a deputado estadual;

P_{vnest}= proporção de votos nominais a deputado estadual;

P_{vlest}= proporção de votos de legenda a deputado estadual;

Para que a proporção de candidatos não tenha o mesmo peso que a proporção de eleitos na dimensão, dado que ter muitos candidatos não significa necessariamente presença eleitoral, na fórmula, a variável proporção de candidatos sempre é elevada ao quadrado. Como se trata de uma proporção, isso fará com que o valor seja reduzido. Então, as quatro variáveis para deputado federal são: o quadrado da proporção de candidatos a deputado federal; a proporção de eleitos para deputado federal; a proporção de votos nominais a deputado federal; e a proporção de votos de legenda ao partido para deputado federal. As mesmas variáveis valem para a disputa de deputado estadual. Como para estes cargos o distrito eleitoral é o estado e a unidade de análise é o município, mantemos sempre a mesma proporção de candidatos e eleitos por partido em todos os municípios de um mesmo estado. Então, elas se transformam em constantes dentro do estado, porém variam entre diferentes estados. Já as votações nominais e por partido são registradas por município, portanto, apresentam variações em todas as unidades de análise.

A opção de usar as proporções de votos nominais e para partido separado no cálculo da dimensão dá maior peso para os votos partidários. Isso porque, em média, os votos de legenda representam entre 10% e 20% do total de votos para os cargos em disputas proporcionais. Outros 90% ou 80% são votos nominais. Porém, quando usamos as proporções das duas modalidades de maneira independente, estamos equiparando os pesos delas para o cálculo da dimensão. Assim, por exemplo, em um município com

100 mil votos para um cargo em disputa proporcional qualquer, dos quais 80% nominais, isso dará 80 mil votos para candidatos e 20 mil votos de legenda.

Se os candidatos do partido "A" fazem juntos 24 mil votos nominais e o partido consegue seis mil votos de legenda, as proporções serão as mesmas, de 0,30 para nominal e partidário, embora os valores absolutos sejam distintos. Enquanto que, se o partido B obtém 32 mil votos nominais e dois mil votos de legenda, apenas, terá 0,4 de proporção de votos nominais e 0,1 de votos partidários. Ou seja, embora em números absolutos o total de votos do partido "A" tenha sido menor que o do partido B, as somas das proporções no cálculo da dimensão resulta em valores invertidos, com vantagem na presença eleitoral do partido B sobre o partido A. Isso se justifica por considerarmos que os votos de legenda representam melhor a presença eleitoral de um partido em um município do que os votos nominais (tanto que a variação de votos nominais dos partidos nos municípios sempre é maior que a de votos de legenda). Como o objetivo do índice é produzir um indicador de presença partidária, dá-se mais importância relativa aos votos de legenda do que aos nominais[4].

A terceira e última dimensão do índice diz respeito à capacidade que os partidos têm de mobilizar doadores e arrecadar fundos para suas campanhas. Devido a não disponibilidade de dados sobre o financiamento de campanhas por municípios até a eleição de 2000, essa dimensão só se aplica ao índice a partir de 2002. A dimensão "recursos de campanha" (DR$) é calculada a partir da seguinte fórmula:

$$DR\$ = \frac{(Pr\$maj + Pomaj + Pr\$prp + Poprp)}{4}$$

Onde:

DR$= Dimensão Recursos de Campanha;

$P_{r\$maj}$= Proporção de recursos em relação ao total declarado nas eleições majoritárias;

4 No caso das eleições municipais, ao invés de oito variáveis a dimensão é calculada para quatro relativas ao cargo de vereador: quadrado da proporção de candidatos, proporção de eleitos, proporção de votos nominais e proporção de votos partidários divididos por quatro. No caso das eleições para vereador, as proporções de candidatos e eleitos variam entre todos os municípios, dado que nessas disputas o distrito eleitoral é o próprio município.

P_{omaj}= Proporção de operações em relação ao total das eleições majoritárias;

$P_{r\$prp}$= Proporção de recursos em relação ao total das eleições proporcionais;

P_{oprp}= Proporção de operações em relação ao total das eleições proporcionais.

Essa dimensão é formada por dois fatores principais, proporção de recursos e proporção de doadores por eleição majoritária ou proporcional [5]. Como a legislação brasileira, no período analisado, era muito permissiva quanto ao tipo de doador (poderiam doar tanto empresas quando eleitores diretamente a partidos ou candidatos) e volume de doações (praticamente sem limites nominais), a utilização apenas dos valores arrecadados geraria grandes distorções. Isso porque um candidato poderia receber a maior parte das doações realizadas em um município, feitas por uma minoria de doadores. Enquanto, por outro lado, um candidato poderia receber um volume grande de pequenas doações, porém, com valores muito baixos[6].

Como o objetivo do índice é medir a presença eleitoral do partido no município, optou-se por dar o mesmo peso para as duas variáveis. Então, a proporção de recursos arrecadados por um partido para a disputa majoritária é somada à proporção de doadores que contribuíram para as finanças de campanha desse partido na disputa majoritária. O mesmo é feito para as eleições proporcionais e o resultado da soma é dividido por quatro para se ter uma média da proporção de “recursos de campanha”. Em resumo, o que se defende aqui é que a identificação da proporção de doadores em relação ao total para medir a presença eleitoral do partido na dimensão “recursos de campanha” é tão importante quanto a proporção de recursos arrecadados para indicar a presença eleitoral do partido em uma circunscrição eleitoral. Assim, muitos doadores de valores baixos se equiparam a poucos doadores de valores altos. No caso das eleições regionais, os

5 Para análises de desempenho partidário e resultados eleitorais, os recursos de campanha não podem ser considerados dimensão própria, pois são usados como meio para organizar as campanhas em busca do convencimento do eleitor, que é o fim. Aqui nos filiamos à idéia de que o que importa nas eleições são os partidos, seus candidatos e os eleitores. Reconhecemos que dinheiro é uma condição necessária, porém, não suficiente para explicar o resultado das disputas eleitorais e é isso o que justifica uma dimensão própria para os recursos de campanha. A dimensão recursos de campanha permite contrastá-la com as outras duas dimensões, desempenho majoritário e proporcional, de maneira independente para explicar os resultados eleitorais.

6 Em 2015 o congresso fez uma minirreforma nas leis eleitorais e proibiu as doações de empresas a candidatos ou partidos durante os períodos eleitorais. Um provável efeito é que já nas eleições de 2016 exista um maior equilíbrio entre as arrecadações dos candidatos, que só poderão ser feitas por eleitores, pessoas físicas.

"recursos de campanha" para disputas majoritárias são a somatória dos recursos para candidato a governador e candidato a senador[7]. As finanças de disputas proporcionais são a somatória de doações para deputado federal e para deputado estadual. No caso de eleições municipais, os recursos majoritários são os das campanhas para prefeito e os proporcionais, para vereador.

O Ipe apresentado aqui utiliza os dados disponíveis para a representação da presença eleitoral dos partidos nos distritos eleitorais. Na ausência de algumas informações ou em casos onde existam mais dados que ajudem a representar a presença dos partidos, o índice pode ser adaptado, desde que se mantenham os princípios de uso de proporções e da normalização entre os extremos zero e um para cada dimensão e para o indicador final.

3.7 Indicadores Educacionais mais Comuns

- Taxa de analfabetismo:

$$\frac{Número de pessoas analfabetas}{total de pessoas}$$

Considera apenas pessoas com 15 anos ou mais de idade. Considera o espaço geográfico (País, Estado, Região).

- Taxa de escolarização:

$$\frac{Número de pessoas\ matriculadas\ nas\ escolas}{total de pessoas}$$

Considera o grupo de idade e o espaço geográfico.

7 As doações para campanha de presidente não entram porque o distrito eleitoral da campanha presidencial é o País e muitos doadores (empresas) se encontram em vários estados. A opção técnica mais lógica seria manter uma constante para todos os municípios, o que teria efeito nulo.

- Média de anos de estudo:

$$\frac{anos\ de\ escolaridade}{idade}$$

Considera-se a escolaridade na população por grupo de idade.

- Razão Bruta de Escolarização (para ensino fundamental e médio a faixa etária é de 7 a 17 anos no Brasil).

$$\frac{Total\ de\ matriculados\ no\ ensino\ fundamental\ e\ médio}{totaldepessoas\ com\ idade\ correspondente\ ao\ ensino\ fundamental\ e\ médio}$$

Seria 100% se todas as crianças na idade esperada fossem todos os matriculados. Quando houver um número elevado de adultos matriculados nesses graus, a taxa pode ser maior que 100%.

3.8 Indicadores sócio-demográficos

O Brasil experimentou, entre os anos 70 e 80 grandes transformações demográficas. Nesse período, foi constatada uma acelerada queda na mortalidade da população, o que promoveu um crescimento no número de indivíduos na sociedade em um primeiro momento. Logo em seguida, começou a cair a taxa de natalidade (número de nascidos vivos para cada 1.000 habitantes), o que equilibrou o crescimento, criando o que se chama de Transição Demográfica. Essa transição começou das áreas urbanas para as rurais e das regiões sul e sudeste passando para as outras regiões do País. O Brasil urbano colaborou para a transição por questões sociais, de valores e econômicas. Os países mais desenvolvidos passaram por essa transição na primeira metade do século XX. Os países africanos estão entrando na transição agora. Vejamos alguns indicadores sociodemográficos:

- Taxa de fecundidade geral: $\frac{Nj}{Qfj}$

Nj = nascidos vivos em determinado ano

Qfj = população feminina em idade reprodutiva em determinado ano (pode ser 15 a 44 ou 15 a 49anos)

j = ano específico. N e Q são uma população específica de um lugar específico.

Ela é uma taxa pouco precisa do ponto de vista geográfico. A crítica a ela é que não consegue identificar os processos migratórios. Além disso, os imigrantes para determinada região podem ter comportamento reprodutivo diferente do comportamento da população local, o que termina influenciando os números finais. **Exemplo:** para calcular a taxa da população feminina em idade reprodutiva em 1981, deve-se somar a população feminina em idade reprodutiva em dezembro de 1981 e de 1982, dividindo por dois.

- Taxa específica de fecundidade: $\frac{Nj+}{Qfj}$

Onde:

Nj+ = nascidos vivos em determinado ano em um grupo específico. Normalmente as faixas de idade das mães são 15-19, 20-24, 25-29, 30-34.

Qfj = total da população feminina de determinado ano no mesmo grupo específico do numerador. **Exemplo:** total de nascidos em 1990 de mães com idade entre 20 e 24 anos, dividido pelo total da população feminina com idade entre 20 e 24 anos no mesmo ano.

- Taxa de fecundidade total: Soma das taxas específicas x 5 (mais usada)

O resultado é o número de crianças por mulher em determinado período e local.

3.9 Representações Tabulares de Dados

Uma das formas mais comuns de representar um conjunto de informações que relacionam números e palavras é através da construção de tabelas. O objetivo de toda tabela é resumir um conjunto de dados observados. Uma tabela é composta por linhas e colunas. Normalmente as colunas indicam as características analisadas e as linhas descrevem as mesmas. Toda tabela deve ter pelo menos os seguintes elementos:

- Título: deve sumarizar o conjunto de informações contidas na tabela. Localizado no topa, deve responder pelo menos as seguintes perguntas: o quê? Onde? e, se possível, quando?

- Cabeçalho: especifica o conteúdo de cada coluna, localiza-se abaixo do título.

- Coluna inicial ou indicadora: é onde aparecem as indicações de conteúdos para cada linha.

- Corpo: é onde se encontram as células com informações a serem sumarizadas e posteriormente analisadas. É composto por linhas e colunas.

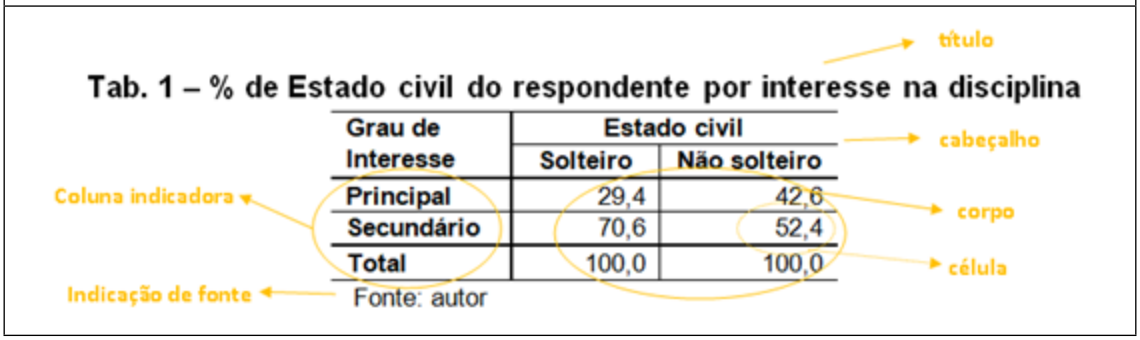
- Células ou casas: são os locais destinados ao posicionamento de informações a serem analisadas. Cada célula deve conter uma única informação.

- Indicação de fonte: na parte inferior da tabela, abaixo do corpo, a última linha deve conter informações sobre a fonte das informações e se é original ou se as informações foram modificadas pelo autor.

As partes indispensáveis em uma tabela são apresentadas no exemplo a seguir (Figura 3.1). Trata-se de uma tabela de cruzamentos de duas variáveis do banco de dados “B_Exerc_V1”. A variável “estado civil” é cruzada com a variável “grau de interesse na disciplina”. Os valores são percentuais de distribuição do estado civil, para indicar a distribuição nas categorias de grau de interesse.

Figura 3.1 – Exemplo de tabela

```
Linha de comando - tabela de contingência
local({
  .Table <- xtabs(~interess+estcivil, data=B_Exerc_V1)
  cat("\nFrequency table:\n")
  print(.Table)
  cat("\nColumn percentages:\n")
  print(colPercents(.Table))
})
```



Cabe lembrar que o *layout* final de uma tabela pode variar em função das escolhas de normatização a partir do que está previsto na normatização geral da ABNT (Associação Brasileira de Normas Técnicas). Na Universidade Federal do Paraná (UFPR), o manual de normas estabelece que as laterais da tabela devem ser "abertas" e que a linha que separa o cabeçalho do corpo deve ser mais acentuada do que as linhas que dividem as células. A finalidade da normatização de tabelas não é simplesmente estética, mas serve para facilitar a compreensão e leitura das informações.

3.10 Representações Gráficas de Dados

Além das escalas, índices e indicadores, outra forma de representar os comportamentos de variáveis é através da construção de gráficos. Além de transmitir informações contidas nos dados de maneira simples, eles permitem compreender os aspectos das relações numéricas entre categorias de uma mesma variável ou entre diferentes variáveis. É importante entender que um gráfico é apenas uma forma de transmitir uma ideia. Logo, é preciso ter a ideia antes do gráfico. Uma teoria fraca, uma coleta de informações falha e uma análise deficiente resultarão em gráficos ingênuos e dispensáveis.

Todo gráfico deve permitir o seguinte:

- Apresentar de maneira simples os dados, comunicando ideias complexas com clareza, precisão e eficiência.

- Focar na substância das informações e não na metodologia ou tecnologia para apresentação.

- Condensar muitos números em pequeno espaço.

- Propiciar a comparação visual.

Por outro lado, o que não deve constar em gráficos:

- Um gráfico não serve para mostrar informações óbvias para não iniciados no assunto.

- Gráfico não deve ser exageradamente dinâmico e decorado.

Existem muitos tipos de gráficos que permitem a comparação entre uma variável qualitativa e uma quantitativa, duas qualitativas ou duas quantitativas. Aqui, serão apresentados apenas os gráficos mais usados nas análises estatísticas e que permitem reunir diferentes tipos de variáveis: histograma, gráfico de dispersão, boxplot e gráfico de setores ou gráfico de pizza.

Todo gráfico permite uma representação de séries de dados a partir de ilustrações, resultando em uma visão geral ao mesmo tempo em que torna possível a caracterização de algum elemento particular pela correspondência entre número de casos e categorias de uma variável ou entre categorias de diferentes variáveis. Apesar da vantagem em facilitar a visualização de dados, deve-se levar em consideração que em todos os gráficos há alguma perda de informação, pois o conjunto de valores originais é pelo menos parcialmente omitido para permitir a concisão e facilidade na interpretação dos resultados plotados na figura.

Histograma

O histograma é um dos gráficos mais úteis e simples para análises estatísticas. Ele representa as distribuições de valores de uma única variável. Sua imagem representa a distribuição absoluta ou relativa de valores da variável analisada. É importante não confundir o histograma com um gráfico de barras. No caso do gráfico de barras, o que se pretende é identificar as diferenças entre as categorias a partir das

diferentes alturas das barras. Já no histograma o que importa é a área ocupada pelas barras e a distribuição dela.

Histograma é uma representação das distribuições de variáveis contínuas ou categóricas na forma de barras. A superfície de cada barra indica a frequência dos valores na distribuição. No eixo horizontal (x), encontram-se os valores originais da variável. No eixo vertical (y), estão os indicadores das frequências de cada intervalo representado pelas barras. As barras com maiores superfícies indicam os valores que mais se repetem, ou seja, as medidas de tendência central. A forma de distribuição, se mais concentrada ou mais dispersa, indica a heterogeneidade dos casos.

As análises que começam pelo histograma conseguem identificar tendenciosidade na distribuição dos casos em relação a uma informação sumária, por exemplo, a média. Ele serve para análise de uma única variável, que pode ser de qualquer tipo – quantitativa ou qualitativa. No entanto, as variáveis quantitativas deverão ser agrupadas para facilitar a leitura do gráfico. A representação da frequência de casos no histograma se dá a partir do tamanho de barras verticais.

Uma distribuição normal segue o formato de um sino, simetricamente distribuído ao redor da média, ou seja, o mesmo número de caso à esquerda e à direita da média. Quando existem mais casos de um dos lados da média, diz-se que a curva não segue uma distribuição normal. Ele permite identificar informação não apresentadas pelas estatísticas. Importante saber que nem toda distribuição é normal e que algumas vezes o que queremos é identificar em que direção segue a tendenciosidade. Essa seria uma distribuição normal com tendência.

Por outro lado, uma distribuição não normal é aquela que apresenta dois ápices em pontos distintos, também chamada de distribuição bimodal, onde não é possível encontrar um único ponto de convergência. Para saber se uma distribuição é ou não normal devem-se usar as estatísticas. Os histogramas são usados para apresentar distribuições normais, mas muitas distribuições não são normais (bimodal, tendência positiva, tendência negativa, uniforme – sem pico – ou leptocúrtica – com um pico muito diferente dos demais pontos).

O exemplo a seguir é de um histograma da variável “média” do banco de dados “B_Exerc_V1”. A imagem mostra uma distribuição normal com tendência. Percebe-se

que o maior número de ocorrências de notas fica no intervalo entre 70 e 80, com mais de 25 casos. Também é possível notar que não houve nenhuma nota abaixo de 40. A distribuição dos casos mostra que uma tendência para a direita, com mais frequência a partir de 70 do que abaixo dele. Trata-se de uma distribuição positiva.

```
Linha de Comando - histograma

with(B_Exerc_V1, Hist(media, scale="frequency", breaks=4,
col="gray",xlab="notas", ylab="freq", main="Histograma"))
```

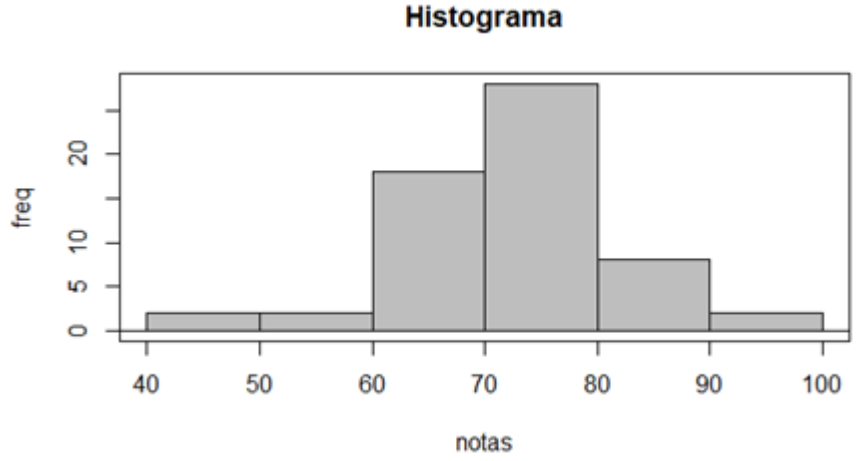


Gráfico 3.1 – distribuição de notas obtidas pelos alunos

Ao histograma também pode ser acrescentada uma curva de distribuição, também chamada de Curva de Gauss, que serve para indicar a forma de distribuição das frequências. As funções dessa curva serão estudadas mais adiante, por isso ela não foi incluída no gráfico acima. Como o histograma só permite o estudo de variáveis isoladas, ele não pode ser usado quando o pesquisador quer comparar duas variáveis. Para isso, existem outros tipos de representações.

Gráfico de Dispersão

O gráfico de dispersão é indicado para quando se quer relacionar duas variáveis quantitativas contínuas ou discretas. Ele é um dos mais importantes na estatística, pois permite visualizar a existência de alguma tendência na relação de variáveis numéricas. A distribuição dos valores é representada pelos pontos de encontro entre a variável

que está no eixo X e a do eixo Y. Se os pontos apresentarem alguma direção, diz-se que há tendência nas relações entre os valores das duas variáveis. Essa tendência pode ser positiva, quando os valores mudam em uma mesma direção, ou negativa, quando as mudanças se dão em direções opostas. Por isso, o gráfico de dispersão é apresentado junto com testes de correlação e regressão (que serão estudados mais adiante). O histograma abaixo representa a relação entre a variável “idade” e a variável “média” dos alunos, do banco de dados B_Exerc_V1.

```
Linha de comando – scatterplot

scatterplot(media~idade, reg.line=lm, smooth=FALSE, spread=FALSE,
boxplots=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9),
xlab=”idade”, ylab=”media”, main=”Gráfico de Dispersão”, data=B_
Exerc_V1)
```

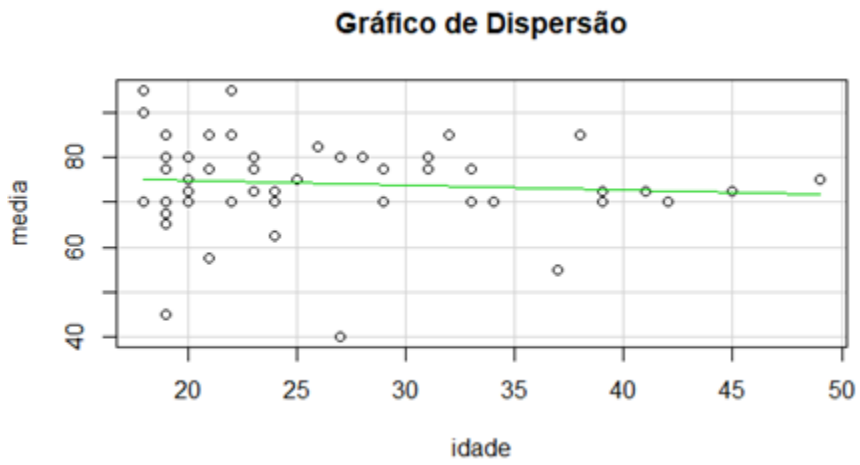


Gráfico 3.2 – Relação entre idade e nota média dos alunos

No eixo X, estão distribuídos os valores da variável “idade” do aluno. Ela varia entre menos de 20 até 50 anos. Percebe-se que há maior concentração de casos entre menos de 20 e até 25 anos. A partir de então, o número de alunos por idade diminui. No eixo Y, estão os valores das “médias” obtidas pelos alunos. Percebe-se uma concentração na faixa acima de 70 e abaixo de 80 pontos de média. A relação entre as duas variáveis é apresentada pelo diagrama formado a partir da distribuição dos dados em um plano cartesiano de coordenadas X e Y, onde cada eixo representa uma variável e os casos são plotados no local onde se dá a interseção dos dois eixos. Para apoiar a

interpretação, é possível plotar uma reta de regressão linear entre os valores (que será discutida em detalhes mais adiante). A direção da reta e o ângulo que ela forma indicam o tipo e intensidade de relação entre as duas variáveis. No exemplo acima, percebe-se que a reta apresenta um pequeno declive, ou seja, com uma relação negativa leve. Quanto maior a idade, menor tende a ser a média do aluno. Porém, como o ângulo da reta é muito pequeno, não é possível afirmar que a relação é estatisticamente significativa. Ou seja, apenas pela distribuição dos pontos e com a direção da reta é possível dizer que não existe relação evidente entre idade e média. Portanto, com o gráfico de dispersão, é possível visualizar possíveis relações entre duas variáveis de maneira muito rápida. No entanto, ele apresenta como limitação o fato de só poder ser usado quando se trata de duas variáveis quantitativas.

Gráfico de setores ou Gráfico de Pizza

O gráfico de setores, mais conhecido como gráfico de pizza, é obtido dividindo um círculo em fatias e estabelecendo os tamanhos das fatias em função da participação proporcional de cada uma das partes em relação ao todo. O problema é que esse tipo de representação gráfica não cumpre alguns dos pressupostos estabelecidos para a boa visualização de dados quando existem muitas categorias a serem representadas. Ou seja, quanto maior o número de fatias, mais difícil a interpretação das proporções relativas. Além disso, por definição, é um gráfico que só pode ser usado para representar variáveis categóricas. Se forem consideradas essas restrições, que são: limitado a variáveis categóricas e limitado a um número reduzido de categorias (sugiro entre duas e, no máximo, quatro categorias), o gráfico de setores pode ser útil para a identificação das diferenças proporcionais de categorias dentro de uma variável.

O exemplo a seguir é de um gráfico de setores para a variável categórica “estado civil” do banco de dados “B_Exerc_V1”. Esta variável está dividida em apenas duas categorias, que são “ser solteiro” ou “não ser solteiro”. Com isso, é possível comparar diretamente as proporções dos dois setores. Mesmo com pouca informação na imagem é possível perceber que o grupo dos solteiros representa algo em torno de 2/3 do total da imagem. Essa primeira aproximação geral é a principal vantagem do gráfico de setores, o que é favorecido quando há poucos setores para serem comparados.

```
Linha de comando – gráfico de pizza

tr <- table (B_Exerc_V1$estcivil)
> nr <- 'Estado Civil'
> ltr <-length(tr)
> lnr <- levels(B_Exerc_V1$estcivil)
> pie (tr)
> pie (tr, main=nr, col = rainbow(ltr))
```

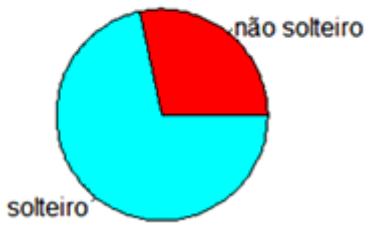


Gráfico 3.3 – Proporção de solteiros e não solteiros

O gráfico de setores, apesar das limitações, pode ser útil para uma representação inicial de diferenças proporcionais de categorias. No entanto, quando se quer representar graficamente a distribuição de uma variável contínua ou comparar as distribuições de quantidades entre distintas categorias, ele não é adequado. A alternativa para esse caso é o gráfico de caixa ou boxplot.

Boxplot

O boxplot é um tipo de gráfico usado para a representação visual das relações existentes entre uma variável quantitativa (contínua ou discreta) e uma qualitativa (nominal ou ordinal). A qualitativa ocupa o lugar do eixo X e serve para diferenciar os grupos que estão agregando os valores da variável quantitativa que se distribuem no eixo Y. Um boxplot é capaz de fornecer informações sobre distribuição dos casos comparativamente entre as categorias, além de dar indicações sobre as medidas de posição.

Entre as informações que a imagem fornece, estão os limites entre os quartis (25% dos casos em cada quartil), valor mínimo, valor máximo e a mediana. Esta última é uma medida de tendência central que será apresentada em detalhes no próximo capítu-

lo. Por enquanto, para visualização do boxplot, basta saber que a mediana é o valor que divide a distribuição em duas partes iguais, ou seja, em metades com o mesmo número de casos cada uma.

O gráfico é formado por figuras que representam as distribuições contínuas nas categorias. Todas as caixas são compostas por uma haste inferior, a caixa propriamente dita – dividida em duas partes não necessariamente simétricas – e uma haste superior. A haste inferior representa o primeiro quartil, indicando o ponto de menor valor para a categoria até a obtenção de 25% dos casos. A caixa indica a distribuição de 50% dos casos, ficando 25% (segundo quartil) do início da caixa até a linha que a divide e outros 25% dessa linha até o limite superior da caixa (terceiro quartil). Por fim, a haste superior indica a distribuição dos casos no quarto quartil (25% superiores). A linha que divide a caixa em duas partes indica a posição da mediana, ou seja, ela mostra o ponto a partir do qual se encontra metade dos casos acima e outra metade abaixo. O boxplot a seguir representa a distribuição das idades dos alunos dividida nos dois grupos de estado civil (solteiros e não solteiros).

```
Linha de comando – boxplot

Boxplot(idade~estcivil, col = 'gray', main = 'Gráfico
de Caixa', data=B_Exerc_V1, id.method="y")
```

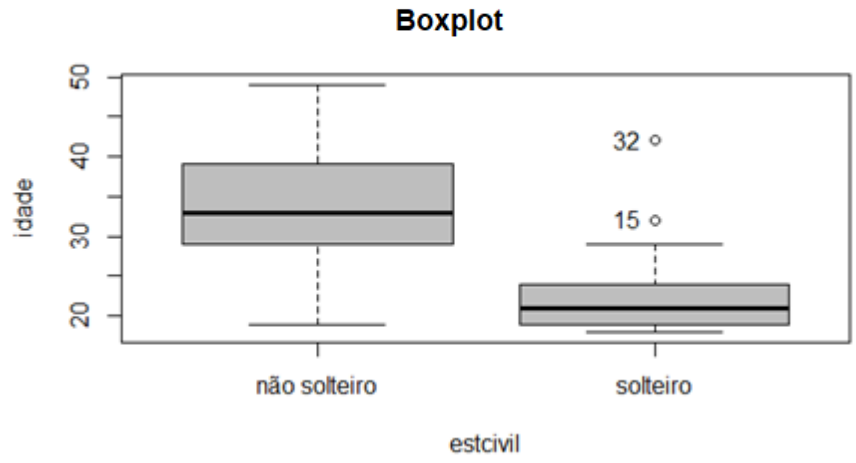


Gráfico 3.4 – Distribuição das idades por estado civil do aluno

A figura fornece várias informações. Cada caixa mostra a distribuição de idades por grupos de estado civil. As distribuições são distintas. Entre os solteiros a variação é menor, ficando em torno de 21 anos e com baixa heterogeneidade. Já para o outro grupo, a distribuição é maior. Os valores medianos giram em torno de 33 anos e os limites de cada quartil estão bem distribuídos. Ou seja, para os solteiros, as idades giram em torno de valores próximos. Já o outro grupo tem uma variação de idade maior. A linha que divide a caixa em duas partes indica a mediana da distribuição, ou seja, a idade que divide o número de casos em duas partes iguais, acima e abaixo dela.

Há uma informação na caixa dos solteiros que não aparece no outro grupo: dois pontos acima da haste superior, com números 15 e 32. Estes são *outliers*, ou seja, valores que estão muito fora da distribuição e que se fossem incluídos na caixa distorceriam artificialmente as dimensões da caixa. Os números representam as posições dos casos no banco de dados. A presença de *outliers* indica que alguns casos ficam fora da norma da distribuição. No exemplo, há dois alunos solteiros com idades muitos diferentes das apresentadas pelos demais pertencentes ao grupo. Já entre os não solteiros não há nenhum caso *outlier*. Esses casos extremos podem ser tanto superiores, quanto inferiores. Como podemos perceber, muitas informações comparativas podem ser extraídas de um boxplot. A seguir, são apresentados os passos para identificar *outliers* de uma distribuição.

Como identificar outliers em boxplots

Nem toda distribuição de casos em um boxplot pode ser considerada livre de valores divergentes, os casos extremos, chamados de *outliers*. Sem a identificação dos limites, possíveis casos extremos podem distorcer muito o formato de um gráfico boxplot. Para calcular os limites acima e abaixo dos quais os casos serão considerados *outliers* deve-se fazer o que segue:

a) Identificar a mediana;

b) Identificar o limite entre primeiro e segundo quartis (Q1) e limite entre terceiro e quarto quartis (Q3).

c) Subtrair o limite superior do limite inferior. O resultado será chamado de L.

d) Para encontrar o limite para os *outliers* superiores: Q3 + (1,5 x L). Para en-

contrar o limite para os casos extremos superiores: Q3 + (3 x L).

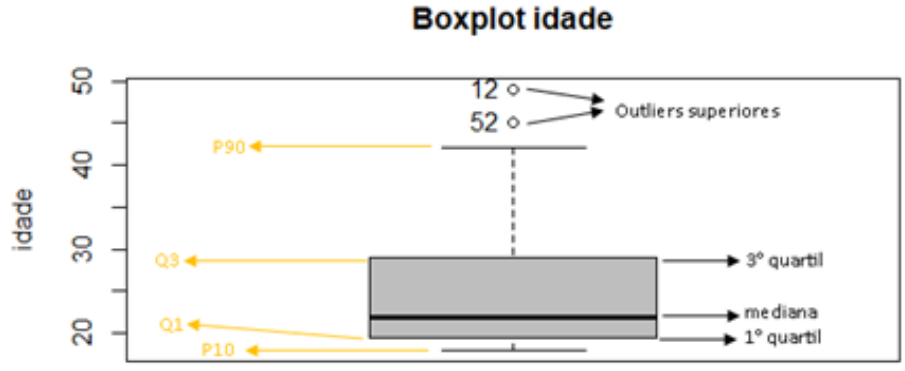
e) Para encontrar o limite para os *outliers* inferiores: Q1 – (1,5 x L). Para encontrar o limite para os casos extremos inferiores: Q1 – (3 x L).

Os casos que ficarem acima ou abaixo dos limites encontrados são considerados *outliers* ou casos extremos, conforme demonstrado no exemplo a seguir.

O gráfico mostra as distribuições das idades dos alunos no banco de dados "B_Exerc_V1". É possível identificar que existem duas idades indicadas como *outliers* superiores, ou seja, estão a +1,5L da mediana e não existem *outliers* inferiores. Do lado esquerdo do gráfico estão indicados os limites de cada parte do gráfico. O primeiro quartil (Q1), início do box, é onde se encontram 25% dos casos. O terceiro quartil (Q3), parte superior do box, é onde se encontram 75% dos casos. A parte inferior do gráfico começa no percentil 10 (10% dos casos) e a superior vai até o percentil 90 (90% dos casos). Identificar esses limites é uma das formas de encontrar os *outliers*. Outra é calculando os valores pela fórmula apresentada acima.

Linha de comando – boxplot *outliers*

```
Boxplot( ~ idade, data=B_Exerc_V1, main = 'Boxplot idade', col =
'gray', id.method="y")
```



Mediana = 22
Q1 = 19
Q3 = 29
P10 = 18
P90 = 42
N = 60

Gráfico 3.5 – Exemplo de outliers em boxplot da variável idade

Vejamos como são calculados limites para o caso acima:

Dos 60 alunos, a mediana ficou em 22 anos. O limite do primeiro quartil foi de 19,75 anos e o limite do terceiro quartil ficou em 29 anos. Então, o L fica sendo 29 - 19 = 10. Assim, os cálculos são:

- Limite superior = 29 + (1,5 x 10) = 29 + 15 = 44.
- Limite inferior = 19 - (1,5 x 10) = 19 – 15 = 4.

Então, qualquer valor acima de 44 anos será considerado pelo menos *outlier* superior e qualquer valor abaixo de 4 será considerado *outlier* inferior. No caso do exemplo do gráfico 3.5, são *outliers* superiores ou casos extremos da idade dois casos que se encontram acima de 44 anos. Por outro lado, como os dados são de estudantes universitários, e o *outlier* inferior começa em oito anos para baixo, isso explica porque eles não existem, pois não há estudantes universitários abaixo dessa idade. O valor mais baixo da distribuição foi de 18 anos, localizado no P10. São considerados casos extremos aqueles que estão acima do dobro do limite de *outliers*, ou seja, para identificá-los basta refazer os cálculos acima, substituindo o 1,5L por 3,0L. No nosso exemplo seria:

- Limite superior = 29 + (3 x 10) = 29 + 30 = 49.
- Limite inferior = 29 - (3 x 10) = 29 – 30 = -1.

Como podemos perceber no gráfico 3.4, não existe nenhum caso abaixo, porém, percebemos que dos dois casos acima, um é *outlier* e outro está muito próximo do limite entre *outlier* e extremo. No próximo capítulo, quando forem apresentadas as principais medidas de tendência central, dispersão e de posição será discutida a fórmula para o cálculo dos limites dos quartis (Q1 para primeiro e segundo quartil e Q3 para terceiro e quarto quartil).

3.11 Exercícios propostos

a) O que compõe uma matriz de dados? Exemplifique.

b) O que diferencia os índices das escalas?

c) Considere um indicador formado por 6 itens e com correlação média entre itens de 0,800, se obtém um coeficiente de alfa de *Cronbach* de 0,750, trata-se de um valor acima ou abaixo do esperado? Por que isso importa? (ver tab. 3.5)

d) Calcule o índice de fracionalização máxima, considerando a existência de 30 vagas no parlamento e nove partidos parlamentares. Inteprete o resultado.

e) Qual a diferença entre taxa específica de fecundidade e taxa de fecundidade total?

f) Descreva os principais elementos de uma representação tabular.

g) Que objetivos devem ser cumpridos na representação gráfica de dados?

h) Calcule os limites para *outliers* e casos extremos de uma distribuição representada por boxplot a partir dos seguintes valores: Mediana = 32; Q1 = 29; Q3 = 39; P10 = 20; P90 = 49.

3.12 Bibliografia citada no capítulo

CARMINE, Edward G.; ZELLER, Richard A. **Reliability and validity assessment**. California, USA: Sage University Paper: 1979.

CERVI, Emerson U. **A cobertura da imprensa e as eleições presidenciais 2002**. Beira Interior: BOCC - Biblioteca Online das Ciências da Comunicação, 2003. Disponível em: http://www.bocc.ubi.pt/pag/cervi-emerson-imprensa-eleicoes-2002.pdf. Acesso em out/2017.

CERVI, Emerson U. **PSDB & PT em eleições nacionais**. Curitiba: CPOP/UFPR, 2016.

CRONBACH, Lee Joseph. Test Validation. In: THORNDIKE, R. L. (ed.), **Educational Measurement**, Washington, DC: American Council on Education, 1971, pp. 443-507.

LAAKSO, M; TAAGEPERA, R. Effective Number of Parties: a measure with application

to west Europe. **Comparative Political Studies**, n.12, p. 3-27, 1979.

RAE, Douglas W. **The political consequences of electoral laws.** New Haven: Yale University Press, 1971.

SANTOS, Wanderlei Guilherme. **Votos e partidos, almanaque de dados eleitorais**. Rio de Janeiro: Editora FGV, 2002.

SPECTOR, Paul E. **Summated Rating Scale Construction**: an introduction. Newbury: Sage University Paper, 1992.

Capítulo IV

Medidas Estatísticas

Para ser estudada, uma variável precisa ser caracterizada por medidas e testes estatísticos identificam as relações entre elas.

4.1 Medidas de tendência central

As medidas de tendência central são números únicos que representam o conjunto de valores obtidos em uma variável contínua. Portanto, é um tipo de sumarização de dados, pois descreve o meio ou o ponto central da distribuição, representando-a. Elas possibilitam a caracterização das frequências a partir de um conjunto de valores que tendem a se concentrar no centro da série. Aqui, busca-se uma informação que possa representar "o que mais ocorre" em uma distribuição. A medida de tendência central fornece uma descrição compacta das ocorrências, focalizando atenção na natureza daquilo que é mais comum ou que mais se repete nos dados medidos. O problema é que isso implica em perder informações, ou seja, há uma redução na complexidade dos dados (BUNCHAFT; KELLNER, 1991). Por isso, sempre é necessário complementar uma medida de tendência central com outro tipo de medida. As medidas de tendência central mais usadas e discutidas aqui são: média (aritmética e ponderada), mediana (indica o meio exato da distribuição) e moda (valor que mais se repete). Cada uma delas representa uma forma distinta para identificar um número

único que seja representativo em uma série de dados. Raramente, há uma coincidência exata entre elas.

Média aritmética

A média é a melhor medida de tendência central quando a distribuição é normal. Também é a que nós estamos mais familiarizados. Trata-se do centro de gravidade da distribuição das massas de frequências. É definida pelo somatório dos resultados dividido pelo número total das ocorrências. Trata-se do valor que representa a média da distribuição. Para a definição dessa medida, o ponto em que os valores encontram-se faz diferença e interfere no resultado final. A média também pode ser definida como o valor que divide as áreas de determinada distribuição em duas partes iguais. Isso é diferente de dizer que a média divide a distribuição em duas partes iguais (de mesmo formato e tamanho). A divisão só acontece em duas partes iguais quando a distribuição é normal, aproximando-se da curva de Gauss, como demonstra o gráfico 4.1 a seguir. Já o gráfico 4.2 mostra que a média não se encontra no centro da distribuição. O segundo gráfico mostra uma distribuição com tendência positiva, ou seja, os valores à direita da média estão distribuídos de forma distinta dos valores à esquerda dela.

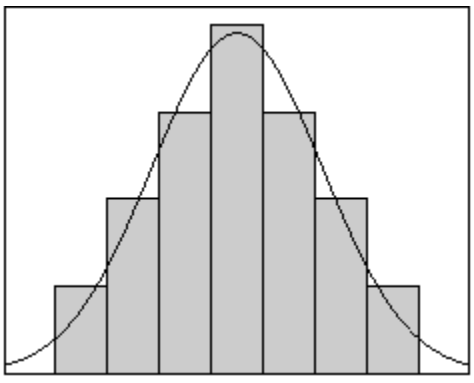


Gráfico 4.1 - Média no centro

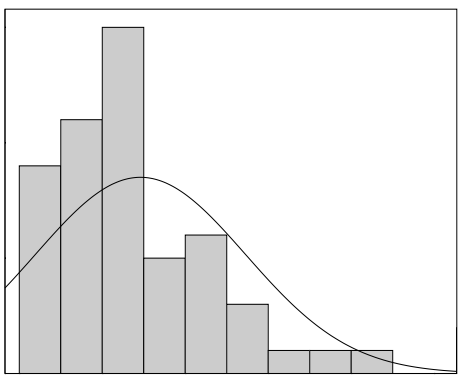


Gráfico 4.2 - Média fora do centro

Para exemplificar, retomamos o quadro apresentado no capítulo anterior sobre o número de projetos de lei por deputado estadual na Assembleia Legislativa do Estado do Paraná (ALEP). São 1.577 projetos de lei apresentados por 54 deputados.

Quadro 4.1 – Número de projetos de lei apresentados por parlamentar

55	21	41	12	7	85	21	25	34
22	49	32	23	73	41	21	50	5
65	2	21	14	42	11	20	21	11
5	14	28	18	31	9	12	22	24
45	12	39	12	11	3	20	3	34
9	51	120	26	114	27	49	13	2

Considerando os dados, para o exemplo, a fórmula da média aritmética é:

$$\mu = \frac{\sum_{i=1}^{n} x}{n} = \frac{1577}{54} = 29{,}20$$

Com isso, podemos dizer que o número médio de projetos de lei apresentados por deputado na ALEP no período analisado foi de 29,20. A média aritmética é empregada principalmente quando:

- Deseja-se obter um indicador médio estável e significativo que inclua no seu cálculo todos os valores. Tem muita utilidade, devido sua significância, como parâmetro em amostras e para estimações estatísticas.

- Na construção de índices de grande importância, tais como desvio padrão, coeficiente de variação e escore reduzido (z).

- Para obtenção de maior precisão na determinação de uma medida, realizam-se várias medições e toma-se como resultado a média aritmética delas.

- Quer-se calcular uma média a partir de valores brutos, sem a necessidade de recorrer a qualquer agrupamento ou ordenação.

Média ponderada

Quando a média a ser extraída leva em consideração as diferenças de tamanho dos subgrupos que compõem o universo pesquisado. Se os grupos têm um número diferente de indivíduos, é preciso fazer uma equiparação, ponderando-os. Para isso, extrai-se a média aritmética de cada grupo, depois se multiplica o valor obtido pelo número de indivíduos do grupo, dividindo o resultado pela somatória dos indivíduos. Assim estaremos ponderando a média. A partir do exemplo anterior, digamos que tivéssemos

os dados brutos do número de projetos de lei apresentados pelos deputados do sexo masculino e feminino, com a seguinte matriz de dados brutos:

Quadro 4.2 – Distribuição dos projetos de lei por sexo do parlamentar

Masculino				Feminino
22	28	31	20	55
65	39	11	12	45
5	12	114	20	49
9	23	85	25	14
21	14	41	50	51
2	18	11	21	120
12	12	9	22	73
41	26	3	3	27
32	7	21	13	49
21	42	21	5	34
2	34	24	11	

Agrupando a produção legislativa dos deputados por sexo teríamos o seguinte:

Tabela 4.1 – Resultados da produção legislativa por sexo do parlamentar

Sexo	Número de deputados	Número de projetos (Frequência Absoluta)	Percentual de projetos (Frequência relativa)
Homens	44	1060	0,67
Mulheres	10	517	0,33

Calculando as médias por sexo do parlamentar teremos que:

Média dos homens = 1060 / 44 = 24,09.

Média das mulheres = 517 / 10 = 51,70.

Como encontramos médias aritméticas bastante distintas entre homens e mulheres, o valor inicial da média de 29,20 pode não representar nada para o conjunto todo já que existe uma diferença significativa entre o número de homens e mulheres no grupo. A alternativa para ponderar seria simplesmente somar as médias dos grupos para obter uma nova média, dessa vez, a partir das médias de cada grupo.

Cálculo a partir da média aritmética dos subgrupos:

$$\mu = \frac{24{,}09 + 51{,}70}{2} = 37{,}89$$

Ou então, pode-se ponderar pelo tamanho de cada subgrupo de casos. A ponderação nos oferece resultados mais precisos. O cálculo é o seguinte:

$$\mu = \frac{(51{,}70\ x\ 10) + \ (24{,}09x44)}{54} = 38{,}77$$

Agora, digamos que tivéssemos que dar pesos diferentes para cada grupo. Então, teríamos que multiplicar o número de casos pelo peso. **Por exemplo**, digamos que uma mulher tenha peso 2 na assembleia legislativa por ter dupla jornada de trabalho, enquanto os homens ficam com peso um. Assim, para encontrar a média ponderada com pesos em um dos subgrupos, a fórmula ficaria:

$$\mu = \frac{(24{,}09x44) + ((51{,}70x(\mathbf{10}\boldsymbol{x}\mathbf{2}))}{44 + (\mathbf{10}\boldsymbol{x}\mathbf{2})} = \mathbf{32,71}$$

Perceba que na fórmula demos peso para as mulheres, multiplicando o número do grupo pelo peso relativo antes de somar com o resultado do grupo dos homens. A média ponderada, nesse caso, fica um pouco diferente da anterior, pois aqui os grupos receberam pesos, enquanto que no exemplo anterior eles foram ponderados apenas pelo tamanho relativo de cada um. Se, por outro lado, ao invés de darmos peso 2, déssemos peso 4,4, o que equipararia matematicamente o grupos das mulheres em relação ao dos homens, o resultado da média ponderara seria de 37,89, muito próxima do resultado da fórmula anterior.

Mediana

É o valor que ocupa o lugar central em uma série de valores ordenados. Trata-se do ponto médio da distribuição, isto é, deixa de cada lado 50% dos casos. Deve-se

preferir a mediana quando a distribuição apresenta valores extremos. Se o número de indivíduos for ímpar, a mediana coincidirá com o valor que ocupa o lugar central. Se o número for par, a mediana será a média dos que ocupam os lugares centrais. Sendo assim, além de constituir um valor representativo da distribuição, a mediana estabelece um limite que separa a metade superior da metade inferior. Quando se quer aplicar um teste para um grupo e pretende-se aprovar exatamente 50% deles, usa-se a mediana como nota mínima a ser atingida pelos indivíduos que serão aprovados. Para calcular a mediana, é preciso organizar os dados em ordem crescente de valores. Apesar de menos conhecida, em muitas pesquisas, a mediana fornece informações mais ricas do que a média sobre a respeito da tendência central.

A comparação entre média e mediana de uma mesma distribuição de frequências permite indicar se a distribuição é simétrica ou assimétrica. No primeiro caso, as ocorrências distribuem-se de maneira igualitária nos dois lados da curva. No segundo, há uma tendência em favor de determinado lado. Se a mediana e a média estiverem muito próximas, então, a distribuição é simétrica, com distribuição equitativa para ambos os lados da curva. Se o valor da média for maior que o valor da mediana, significa que a distribuição é assimétrica para a direita, ou seja, há “casos desgarrados” para cima. Se, ao contrário, o valor da média for menor que o valor da mediana, a assimetria é para a esquerda.

Exemplo: Digamos que em uma classe com 13 alunos, as notas semestrais obtidas em determinada disciplina foram as que constam no quadro abaixo, relacionando o número do aluno com a nota obtida. Se desejássemos obter apenas a média, seria necessário somar as notas de dividir pelo número de alunos da turma:

$$\mu = \frac{962{,}5}{5134} = 74{,}03$$

Mas, para conseguirmos uma mediana é preciso organizar os valores por ordem crescente, que é o que consta na coluna “Nota decrescente”. Por se tratar de um número ímpar, são 13 alunos, uma vez organizados por ordem crescente de valor, a nota que ficar na linha sete dividirá o banco em duas partes iguais, sendo este o valor

mediano (Md=75).

Como neste caso a mediana (75) ficou um pouco acima da média (74,03), significa que existem alguns alunos que estão mais distantes dos demais do grupo na ponta direita da escala, ou seja, entre as notas mais elevadas, isso significa que por outro lado há uma concentração de valores abaixo da Mediana.

Tabela 4.2 – Notas dos alunos da Turma A

Aluno	Nota decrescente	
1	90	
2	85	
3	80	
4	80	
5	77,5	
6	**75**	← *Mediana*
7	**75**	
8	72,5	
9	70	
10	***65***	
11	***65***	← *Moda*
12	***65***	
13	62,5	

Utiliza-se mediana quando:

- Deseja-se encontrar o valor exato que divide a distribuição em duas metades.

- Os resultados extremos são tão díspares que podem afetar sensivelmente o valor da média aritmética.

- Em distribuições de frequências onde ocorram limites superiores ou inferiores abertos, tais como "*menos de...*" ou "*mais de...*".

Moda

Como a própria denominação indica, trata-se do valor que ocorre com maior frequência na variável analisada. A moda é uma medida de tendência central pouco usada, embora possa ser visualizada facilmente em um histograma, pois é o ponto em que se encontra o maior pico entre as colunas do gráfico. Pode não existir uma única

moda em uma distribuição de frequências (o que caracteriza uma série multimodal). Ela pode ser bimodal, quando tem duas modas, ou polimodal, quando possuir três ou mais modas. É preciso ressaltar que embora seja a frequência que mais ocorre, a moda não representa necessariamente a maioria dos resultados. Uma característica importante da moda é a sua aplicabilidade a todos os níveis de medida (nominal, ordinal ou intervalar). No exemplo das notas dos alunos (tabela 4.2), citado acima, a moda ficaria no valor 65 que ocorre três vezes. Perceba que reunindo as três principais medidas de tendência central, teremos uma riqueza maior de informações sobre a distribuição das notas:

Média aritmética: 74,03
Mediana: 75
Moda: 65

A moda é empregada principalmente quando:

- Deseja-se conhecer apenas o valor mais frequente, como, por exemplo, em uma distribuição de ocorrências em uma variável categórica.

- Deseja-se ter uma noção imediata e aproximada de qual será o valor da tendência central.

- Em distribuição de frequências onde ocorram classes com limites indefinidos "*menos de... ou mais de...*". Nesses casos o valor da média aritmética não pode ser determinado com exatidão.

Interpretando os dados obtidos nas três estatísticas do exemplo acima temos que a mediana superior à média indica que há alguns casos "desgarrados" de notas baixas. A mediana em 75 mostra que embora alguns alunos tenham obtido desempenho abaixo da média, metade deles ficou com nota acima de 75. Ao passo que a moda, em 65, indica que esse valor foi o que mais se repetiu em todos os casos da distribuição. Graficamente, teríamos o que segue:

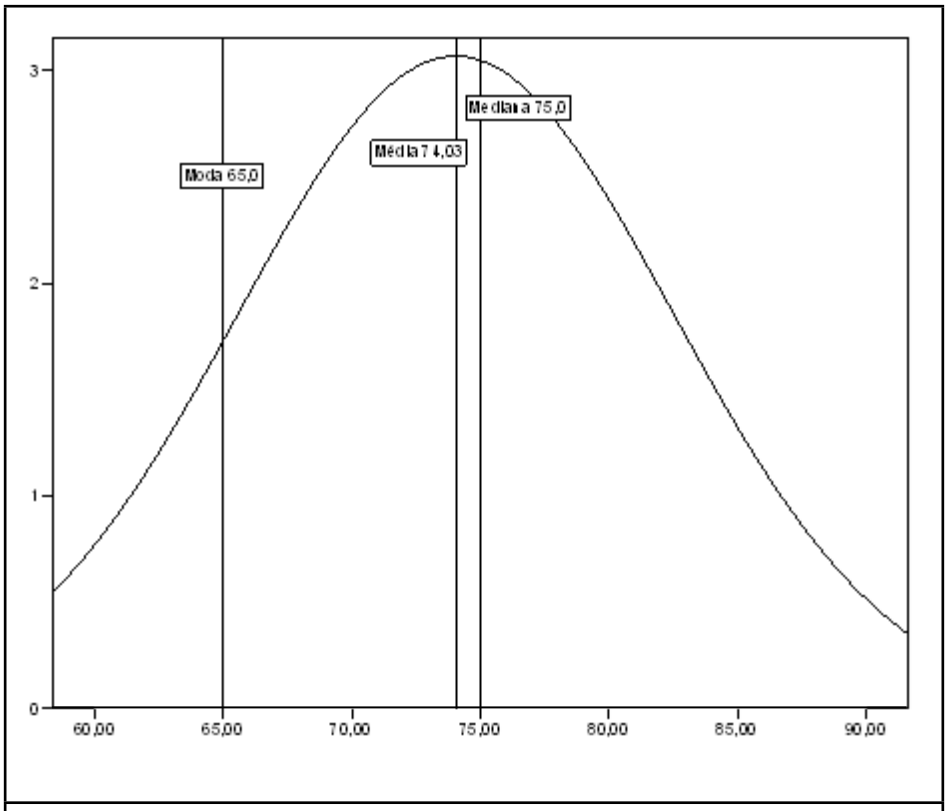


Gráfico 4.3 – Distribuição das medidas de tendência central

Que medida de tendência central usar?

A decisão sobre empregar a média, mediana ou moda envolve uma série de fatores, tal como nível de mensuração, estabilidade da medida e manipulação, entre outras (OVALLE, 1991).

Quanto ao nível de mensuração, a moda é a única que pode ser calculada para qualquer conjunto de dados. Sendo assim, só é possível medir a tendência central de dados nominais com a moda. A mediana exige pelo menos o nível ordinal, pois envolve a ordenação das categorias. A média fica restrita a dados intervalares que se encontrem em escalas, razões ou proporções.

Com respeito à estabilidade da medida, as três tendências centrais diferem quanto ao grau de consistência de amostra para amostra. Se coletarmos um grande número de amostras de uma população e determinarmos a média, mediana e moda para cada uma, observaremos que a média é o valor que difere menos entre as amostras e a moda é a que varia mais. Como em amostras o que interessa é a possibilidade de fazer inferências estatísticas para toda a população, a estabilidade é importante.

Quanto à manipulação subsequente dos dados, como a média aritmética permite operações estatísticas posteriores como a determinação da média combinada do grupo a partir da média de subgrupos, ela é a medida preferencial para obter informações adicionais. Ao contrário, a mediana e a moda são estatísticas terminais.

4.2 Medidas de dispersão, variabilidade ou assimetria

Em toda população grande, as variações são normais e desejadas. O objetivo de uma medida de variabilidade é identificar a amplitude e a diversidade dos valores de uma distribuição (BARBETTA, 2003). Digamos que, ao invés de uma única turma como no exemplo anterior, houvesse duas turmas de um mesmo professor. Uma medida de tendência central, tal como a média, poderia ser usada para resumir o desempenho das notas dos alunos de cada uma delas, para sabermos se os alunos tiveram desempenhos parecidos. Caso o número de alunos por turma fosse o mesmo e as notas as que constam na tabela 4.3 abaixo, teríamos duas médias de notas muito parecidas nas turmas A e B, como segue:

Tabela 4.3 – Notas dos alunos das turmas A e B

Aluno	Turma A	Turma B
1	70	55
2	85	100
3	90	50
4	80	65
5	72,5	58
6	75	55
7	70	62
8	65	90
9	75	95
10	62,5	94
11	77,5	85
12	65	60
13	80	95
Média	**74,42**	**74,15**

Ao olhar os resultados individualmente, percebe-se que o desempenho dos alunos da primeira turma fica mais próximo da tendência central do que os da turma B, que têm uma maior dispersão em direção às extremidades. Essa diferença não pode ser identificada pela medida de tendência central, pois esta aponta apenas os valores que se encontram no centro da distribuição. Como vimos antes, elas são úteis quando é necessário reduzir um grande volume de informações de um conjunto de indivíduos, facilitando a identificação de algumas características fundamentais desse grupo. Porém, a consequência dessa redução é uma limitação de informações sobre a distribuição.

As curvas de Gauss representadas no gráfico a seguir mostram distintas formas de distribuição de notas da turma A e da turma B. Percebe-se, nas imagens, que a distribuição da turma A aproxima-se de uma curva normal mais acentuada, ou seja, há um maior número de valores próximos da tendência central, enquanto que na turma B há uma distribuição maior, afastando-se da média. Visualmente, é possível identificar que as notas da turma B são mais heterogêneas do que as da turma A. Porém, o problema permanece: como representar numericamente as distintas variações de distribuições de valores que não são objeto de atenção das medidas de tendência central e não podem ser aferidas a partir delas?

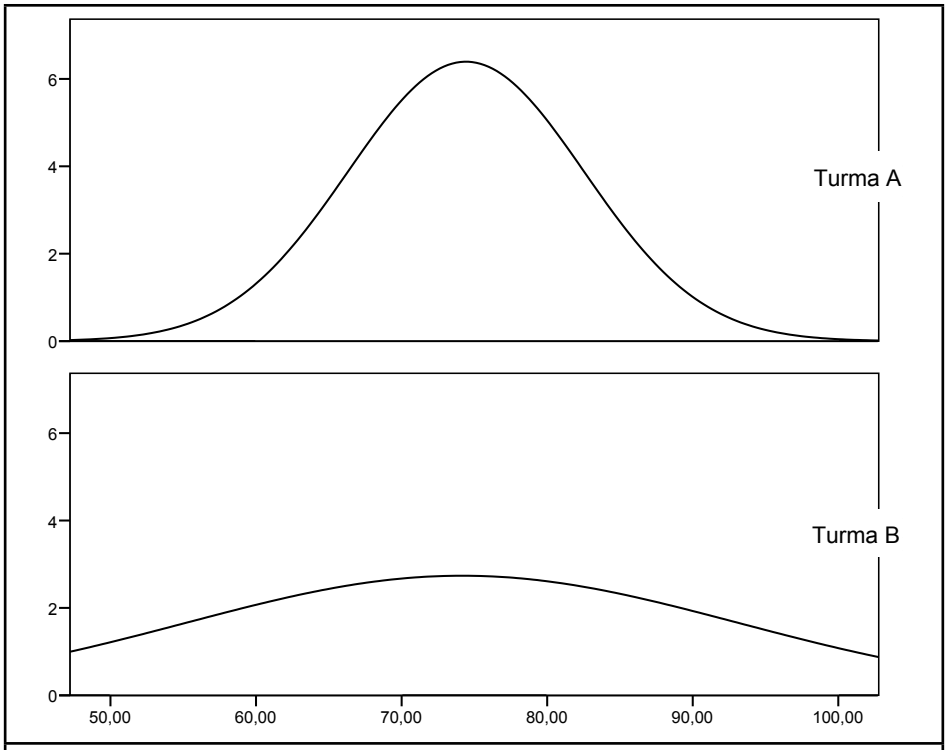


Gráfico 4.4 – Distribuição das notas por turma

Para resolver essa questão existe um grupo de estatísticas, chamadas de medidas de dispersão, variabilidade ou assimetria. Elas acompanham as de tendência central e têm o objetivo de complementar as informações fornecidas por estas últimas. Em outras palavras, enquanto as medidas de tendência central mostram os valores estáveis ou normais, as de dispersão indicam os valores relativos ao afastamento da norma.

Para que haja uma descrição estatística adequada de um fenômeno qualquer é necessário não apenas a apresentação de medidas de tendência central, que mostram valores típicos, mas também medidas de dispersão, que indicam oscilações em torno dos valores centrais, ou seja, os desvios em relação a uma norma. Vejamos como calcular algumas das principais medidas de dispersão utilizadas em testes estatísticos.

Amplitude total

A amplitude total dá uma ideia clara do campo de variação dos valores em qualquer série. Ela também é chamada de limite ou percurso (*range*), pois representa a distância entre dois valores extremos de uma distribuição. O cálculo da amplitude total se dá pela diferença entre o maior e menor valor da distribuição.

$$AT = X_{Máx} - X_{Min}$$

Onde:

AT = amplitude total;

$X_{Máx}$ = valor máximo da distribuição de frequências;

X_{Min} = valor mínimo da distribuição de frequências.

Como medida de variabilidade, a Amplitude total tem o mesmo “status” que a Moda (mo) na tendência central. Seu cálculo é rápido e fácil, porém, por depender de apenas dois valores (máximo e mínimo) é instável. Por isso, é considerada uma medida grosseira de variabilidade, devendo ser usada como índice preliminar. Voltemos ao exemplo anterior das duas turmas de alunos para identificar a AT de cada uma delas:

Tabela 4.4 – Estabelecimento de valores de amplitude para as turmas A e B

TURMA	VALOR MÁXIMO	VALOR MÍNIMO	CÁLCULO	AMPLITUDE TOTAL
A	90	62,5	90 – 62,5	**27,5**
B	100	50	100 - 50	**50**

Também pode ser calculada a Amplitude semi-interquartílica, que é a metade da diferença entre o terceiro e o primeiro quartil. Como o primeiro e o terceiro quartis representam valores entre 25% e 75% respectivamente dos dados da série, a diferença entre eles abrange 50% dos dados centrais da distribuição, como demonstra a tabela 4.5:

Tabela 4.5 – Cálculo da amplitude semi-interquartílica para as duas turmas

TURMA	CÁLCULO	AMPLITUDE SEMI-INTERQUARTÍLICA
A	Terceiro quartil = 80. Primeiro quartil = 67,5 (80 – 67,5) / 2	6,25
B	Terceiro quartil = 94,5. Primeiro quartil = 56,5 (94,5 – 56,5) / 2	19

A amplitude semi-interquartílica é indicada como medida de dispersão quando se usa a mediana como medida de tendência central e para os casos em que há valores extremos muito distantes dos demais, enviesando o máximo ou mínimo da distribuição de frequência.

Desvio padrão

O desvio padrão, representado pela letra “S” ou “SD”, é a medida de variabilidade mais usada como índice de dispersão, sendo também a mais confiável no que diz respeito à generalização da amostra para a população da qual ela foi retirada. Por definição, o desvio padrão é a raiz média quadrática dos desvios calculados em relação à média aritmética da série. Representado pela fórmula quando se pretende fazer a descrição para a população:

$$SD = \sqrt{\frac{[\sum (\mathrm{X} - \mathrm{M})^2]}{N}}$$

Ou a seguinte fórmula quando se pretende fazer inferências estatísticas:

$$SD = \sqrt{\frac{[\sum (\mathrm{X} - \mathrm{M})^2]}{(N-1)}}$$

Vejamos como calcular o desvio padrão no exemplo acima para a população dos alunos das duas turmas.

Tabela 4.6 – Cálculo do desvio padrão para notas dos alunos das turmas A e B

Aluno	Turma A			Turma B		
	x	x	x²	x	x	x²
1	70	-4,42	19,54	55	-19,15	366,72
2	85	10,58	111,94	100	25,85	668,22
3	90	15,58	242,74	50	-24,15	583,22
4	80	5,58	31,14	65	-9,15	83,72
5	72,5	-1,92	3,69	58	-16,15	260,82
6	75	0,58	0,34	55	-19,15	366,72
7	70	-4,42	19,54	62	-12,15	147,62
8	65	-9,42	88,74	90	15,85	251,22
9	75	0,58	0,34	95	20,85	434,72
10	62,5	-11,92	142,09	94	19,85	394,02
11	77,5	3,08	9,49	85	10,85	117,72
12	65	-9,42	88,74	60	-14,15	200,22
13	80	5,58	31,14	95	20,85	434,72
Média	74,42	Soma	789,42	74,15	Soma	4309,69

TURMA	CÁLCULO	DESVIO PADRÃO
A	$\sqrt{789{,}42/13}$	7,79
B	$\sqrt{4309{,}69/13}$	18,2

O desvio padrão apresenta como propriedade inicial o fato de medir o grau de dispersão em torno da média na mesma medida que a média. Se a distribuição que está sendo medida for simétrica (com Curva Normal), o valor de até um desvio padrão para cima e para baixo da média incluirá 68,20% dos casos. Até 1,96 desvio padrão incluirá 95,45% dos casos. Veremos o que é uma distribuição simétrica mais adiante.

Além disso, o desvio padrão apresenta outras propriedades, tais como o fato de a soma ou subtração de uma constante em relação a todos os valores da série tornar o desvio padrão inalterado, ao contrário do que aconteceria com a média aritmética. Já no

caso de multiplicação ou divisão dos valores da série por uma constante, o desvio padrão também será multiplicado ou dividido pela constante, assim como acontece com a média.

Variância

A variância (S^2) é a média aritmética dos quadrados dos desvios de cada valor em referência à média. É o mesmo cálculo para se obter o desvio padrão, sem submeter o resultado à raiz quadrada. Seguindo esses passos para a obtenção do desvio padrão, calculamos o X, x e x^2. Depois, dividimos os resultados da somatória pelo número de casos e teremos a variância, como segue:

$$S^2 = \frac{[\sum (\mathrm{X} - \mathrm{M})^2]}{N}$$

Tabela 4.7 – Cálculo do desvio padrão para notas dos alunos das turmas A e B

TURMA	CÁLCULO	VARIÂNCIA
A	$\frac{789{,}42}{13}$	60,68
B	$\frac{4309{,}69}{13}$	331,51

A variância não é muito utilizada como medida de dispersão pela inconveniência de ser expressa pelo quadrado da unidade da variável em estudo. Como a variância equivale ao quadrado do desvio padrão, utiliza-se mais este último por se encontrar na unidade da variável em análise.

Que medida de dispersão usar?

Antes de decidir que medida de dispersão usar, é preciso levar em conta a estabilidade, forma, tipo de distribuição, medida de tendência central usada e nível de mensuração.

Quando a amostra é selecionada aleatoriamente, a estabilidade da medida segue a seguinte ordem, da mais para a menos confiável: desvio padrão, variância e amplitude. Como geralmente as amostras servem para estabelecer conclusões gerais à população, o

desvio padrão é a medida preferida. Porém, quando o desvio padrão é usado para distribuições assimétricas, ele tende a fornecer uma impressão distorcida da variabilidade da distribuição, pois dá peso maior aos casos extremos. A amplitude total também é afetada pela ocorrência de valores extremos muito afastados da média. A amplitude semi-interquartílica, por dar importância aos valores centrais, é a medida de variação mais indicada para esses casos, assim como a medida de tendência central preferencial para ela é a mediana. Por esse motivo, quando a única medida de tendência central utilizada no estudo é a mediana, a medida de variabilidade indicada é a amplitude semi-interquartílica. Para os casos em que se utiliza a média, todas as demais medidas de variabilidade podem ser adotadas, dando preferência ao desvio padrão por estar na unidade de análise original.

4.3 Coeficiente de variação de Pearson (CV)

Medidas de dispersão como Amplitude, Desvio Padrão e Variância são absolutas, indicadas para variáveis em que todos os casos são submetidos aos mesmos testes ou condições. Porém, quando queremos comparar a variação de dois ou mais grupos que diferem substancialmente quanto à variável investigada ou à média aritmética, é mais adequado usar uma medida de dispersão relativa no lugar de uma dispersão absoluta. Nesses casos, uma alternativa é o Coeficiente de Variação de Pearson (CV) ou simplesmente Coeficiente de Variação. Ele expressa que proporção da média é o desvio-padrão, podendo ser apresentado em porcentagens, quando multiplicado por 100.

Trata-se de uma medida adimensional de dispersão, sendo definida como o cociente entre o desvio padrão (S) e a média aritmética. Portanto, Coeficiente de Variação é indicado como medida de dispersão quando forem comparadas variáveis que não possuem a mesma unidade de medida ou nos casos em que há grandes diferenças de quantidades nos números de observações, pois a variável com mais elementos terá maior probabilidade de ser mais heterogênea. Por esse motivo, o CV tem uso restrito praticamente ao nível de razão, podendo ficar instável quando aplicado a variáveis que atingem o nível intervalar.

A partir do exemplo anterior, digamos que fôssemos comparar medidas de ten-

dência central e de dispersão não mais entre as turmas A e B, porém, entre as turmas A e C. No entanto, a turma C não possui 13 alunos, como as duas anteriores, mas tem 35 alunos. Como o número de integrantes em cada conjunto é diferente, não devemos usar o desvio padrão, amplitude ou variância como medida de dispersão na comparação entre elas, mas sim o Coeficiente de Variação. Para tanto, basta dividir o desvio padrão pela média de cada grupo. Depois podemos multiplicar o resultado por 100, para termos o valor em percentual. Como não temos os valores da Turma C, assumiremos que ela possui um desvio padrão de 13,20 e uma média 71,50.

Tabela 4.8 – Cálculo do CV para distribuições de notas de duas turmas de alunos

TURMA	CÁLCULO	CV
A	$\frac{7,79}{74,42}$ x 100	10,46
C	$\frac{13,20}{71,50}$ x 100	18,46

Agora se pode afirmar que o Coeficiente de Variação da turma C é maior que o da turma A, ou seja, a dispersão relativa da turma C é maior do que o da turma A.Quando o Coeficiente de variação é menor que 20%, diz-se que a distribuição é homogênea, como no caso acima. Quando fica entre 20% e 30%, diz-se que a distribuição é mais ou menos homogênea. Se o CV ficar acima de 30%, diz-se que a distribuição é heterogênea.

4.4 Medidas de posição

Para um conjunto de valores ordenados por grandeza, o que divide o grupo em duas partes iguais é a mediana. A partir daí, pode-se pensar em valores que subdividem o conjunto em outras partes iguais. Quando essa divisão dá-se em quatro partes iguais, elas são chamadas de Quartis, com 25% dos casos em cada grupo. Da mesma forma, valores que dividem o conjunto em dez partes iguais chamam-se Decis (10% agregado no grupo) e valores que dividem em cem partes iguais são chamados de Percentis (com 1% por grupo).

As subdivisões dos valores servem para indicar as posições dentro da distribuição. Uma das medidas de posição mais usadas é o quartil. Para calcular os limites dos quartis existem diferentes fórmulas. A mais simples delas é a que segue abaixo e divide-se em três etapas. Primeiro, organiza-se a série em ordem crescente de valores. Na segunda etapa, calcula-se a posição do limite na série com a fórmula:

$$Q_i = \frac{i}{4} x (N + 1)$$

Onde:

Q_i = limite do quartil desejado;

i = número do quartil desejado [1 (0,25), 2 (0,5) ou 3 (0,75)];

N = número de casos da série.

Depois, encontra-se a média aritmética entre os dois valores mais próximos da posição encontrada acima para identificação do valor do limite do quartil. Calcula-se sempre o limite do quartil 1 e do 3, pois o quartil 2 equivale ao valor da mediana. **Exemplo:** calcular os quartis da seguinte distribuição aleatória de 21 casos:

0,2	0,2	0,6	1,2	1,4	1,5	2,4	3,1	3,4	3,6	3,6	3,9	4,1	4,3	5,4	5,8	7,6	8,2	9,4	9,6	9,9

Como os casos já estão organizados em ordem crescente, basta encontrar o ponto central da distribuição para termos a mediana: Q2 = 3,6. Vamos calcular os outros dois limites de quartis Q1 e Q3. Para calcular o limite do primeiro quartil:

$$Q_1 = \frac{i}{4}\ x\ (N + 1) = \frac{1}{4}\ x\ (21 + 1) = 0{,}25\ x\ 22 = \ \ 5{,}5$$

Com isso concluímos o segundo passo que foi identificar o valor correspondente (5,5). Então, é preciso encontrar a média entre as posições 5 e 6 da série. Este valor será a o limite do primeiro quartil. $\frac{(1{,}4 + 1{,}5)}{2}$ = 145. O limite do primeiro quartil para a série acima é 1,45. Agora, para encontrar o valor do limite do terceiro quartil se replica a fórmula acima, substituindo o valor de “i” por 0,75.

$$Q_1 = \frac{i}{4}x(N+1) = \frac{3}{4}x(21+1) = 0{,}75\ x\ 22 = \ \ 16{,}5$$

O limite do terceiro quartil será a média dos valores dos casos 16 e 17: Q_1= $\frac{(5{,}8+7{,}6)}{2}$ = 6,7. Então, os limites para os quartis da distribuição acima são:

- Limite do primeiro quartil Q_1 = 1,45
- Segundo quartil/mediana Q_2 = 3,60
- Limite do terceiro quartil Q_3 = 6,70

Conhecendo os limites dos quartis, é possível identificar a existência ou não de casos *outliers* e de valores extremos na distribuição. A fórmula para identificação de *outliers* e casos extremos já foi apresentada no capítulo anterior, quando discutimos a apresentação gráfica de dados, especificamente o gráfico boxplot.

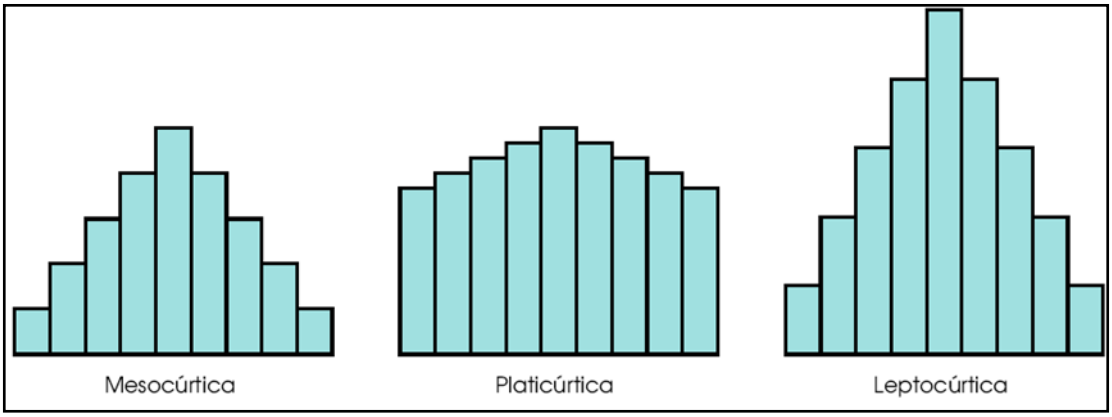
4.5 Medidas de assimetria e curtose

Além das medidas de tendência central e de dispersão, as medidas de assimetria proporcionam uma melhor descrição e compreensão das distribuições de frequências dos fenômenos estudados. Isso porque essas distribuições não diferem apenas quanto ao valor médio e à variabilidade dos casos, mas também quanto à forma. A assimetria de uma distribuição é definida como o grau de afastamento de uma distribuição de frequências da unidade de simetria e pode ser determinada numericamente pelos índices de assimetria (BARBETTA, 2003). Os principais são: Coeficiente de Assimetria de Pearson e Coeficiente Quartílico de Assimetria.

Já a curtose é definida como o grau de achatamento de uma distribuição normal. As distribuições podem ser mesocúrticas (distribuição normal), platicúrticas (quando mais achatadas) ou leptocúrticas (quando menos achatadas). A maneira mais prática de identificar a forma de distribuição dos valores de uma variável qualquer é através de um gráfico de histograma ou através do cálculo do coeficiente momento de curtose

como mostram as figuras a seguir:

Figura 4.1 – Tipos de distribuições por grau de achatamento



Quando a curva apresenta assimetria, ou seja, os dois lados não apresentam as mesmas formas, a média fica sempre do lado mais denso da distribuição. Podemos extrair uma medida de assimetria (positiva ou negativa) a partir da distância entre a média e a moda, que é chamada de Coeficiente de Assimetria de Pearson. Nesse caso, o resultado será na unidade da medida.

Coeficiente de Assimetria de Pearson = média - moda

No exemplo das notas dos alunos da Turma A, teríamos: Assimetria de Pearson = 74,03 – 65 = **9,03**. O valor de 9,03 pontos é a medida de assimetria desta distribuição. Ela torna-se útil quando se comparam medidas de tendência central em diversas distribuições de frequências. Nesse caso, pode surgir um novo problema, que é a comparação entre unidades distintas, por exemplo, entre “nota” e “número de faltas” dos alunos. Como não estão na mesma unidade, torna-se impossível a comparação direta. Para resolver essa questão, deve-se subtrair o resultado da equação acima pelo **desvio padrão** de cada distribuição de valores. Assim, o resultado torna-se adimensional, sendo indicado por número de desvios, como mostra a fórmula a seguir:

$$\text{Assimetria Adimensional} = \frac{\text{Média} - \text{Moda}}{\text{Desvio Padrão}}$$

No nosso exemplo, teríamos para a turma A: 9,03 / 7,79 = **1,15** desvio-padrão de assimetria. Outro coeficiente que indica a forma de distribuição dos valores é o Coeficiente Quartílico de Assimetria, calculado multiplicando-se por três o resultado da subtração da Média pela Mediana e dividido pelo desvio padrão.

$$Coeficiente\ Quartílico\ de\ Assimetria = 3\ x\ (\frac{média - mediana}{Desvio\ padrão})$$

Para a turma A, o cálculo seria: $3\ x\left(\frac{74{,}03 - 75{,}0}{7{,}79}\right) = -0{,}373$

Nos casos de distribuições unimodais (apenas um valor se repete), o mais indicado é o Coeficiente de Assimetria de Pearson, pois a moda fornece um índice de concentração dos dados. Já nos casos de distribuições multimodais, apenas o Coeficiente Quartílico de Assimetria deve ser usado. Uma vez aplicada a fórmula, o resultado indicará a assimetria do conjunto. A distribuição só será Simétrica, quando o resultado for zero (nesse caso, a média será igual à mediana e à moda). Se o resultado for positivo, a distribuição será positivamente assimétrica, pois a média será maior que a mediana. Se o resultado for negativo, a distribuição será negativamente assimétrica e nesse caso a média será menor que a mediana, como no exemplo acima. Para que a distribuição seja considerada Normal, ela precisa ser simétrica e mesocúrtica.

Neste capítulo, conhecemos as principais medidas que sumarizam as características relevantes das distribuições de frequências. As medidas de tendência central, que indicam os valores normais; as medidas de dispersão, que mostram a variabilidade; medidas de posição e medidas de assimetria. Conhecendo esses valores é possível descrever os elementos mais importantes do comportamento de uma variável. Com isso, encerramos a parte do manual que trata da descrição e análise de variáveis e das relações entre elas. No próximo capítulo, vamos introduzir os conceitos básicos da teoria da amostragem, aplicada às pesquisas que produzem resultados inferenciais, ou seja, a partir de amostras.

4.6 Exercícios propostos

A partir da variável idade do banco de dados “B_Exerc_V1”:

Est. Civil	Idade														
Solteiro	18	19	22	19	23	18	20	20	22	32	19	18	19	28	21
	22	24	19	20	19	26	19	20	27	42	20	19	19	20	19
Casado	19	20	38	49	39	33	39	33	37	29	34	45	20	21	31

a) Calcule a média aritmética para todas as idades.

b) Calcule a média ponderada pelas categorias de estado civil. Explique a diferença.

c) Acrescente os pesos 1,5 para solteiros e 3,0 para casados e calcule a média ponderada com pesos. Discuta os resultados em relação ao cálculo anterior.

d) Qual a mediana para todos os valores da distribuição de idades?

e) A distribuição é unimodal ou multimodal? Qual(is) a(s) moda(s) das idades?

f) Qual a amplitude total da distribuição de idades?

g) Calcule o desvio padrão para a série com todas as idades, Calcule o desvio padrão das idades para solteiros e para casados e calcule o Coeficiente de Variação de Pearson. Analise os resultados.

h) Calcule a variância para todas as idades.

i) Identifique os limites dos quartis para todas as idades.

j) Calcule o coeficiente de assimetria de Pearson e o coeficiente quartílico de assimetria. Explique as diferenças entre os resultados.

4.7 Bibliografia citada no capítulo

BARBETTA, Pedro Alberto. **Estatística Aplicada às Ciências Sociais** (5ª edição revisada). Florianópolis: Editora UFSC, 2003.

BUNCHAFT, Guenia; KELLNER, Sheilah Ribno de Oliveira. **Estatística Sem Mistérios** (2ª edição corrigida) vol. I, II, III e IV. Petrópolis - RJ: Editora Vozes, 1999.

TOLEDO, G. L.; OVALLE, I. **Estatística Básica**. São Paulo: Atlas, 1991.

Capítulo V

As Amostras

O ideal, sempre, é ter informações da população.
Se não der, produza uma amostra que seja mais
representativa possível do todo.

5.1 Conceito de amostragem

Em pesquisas empíricas onde não é possível coletar informações de todos os indivíduos estudados, o tamanho e a forma como a amostra é definida importa. Uma amostra muito pequena não gera informações confiáveis sobre a população, produz fofocas. Uma amostra muito grande gera muitos custos e pode inviabilizar a pesquisa, não produzindo resultados. Uma amostra com unidades erradas ou que não tenham o que dizer sobre o objeto da pesquisa gera resultados enganosos. Sempre, o que se busca com uma amostra, é fazer inferências a partir de um subgrupo para o todo. O principal aspecto da estatística inferencial é que a partir da análise de uma amostra é possível generalizar resultados para a população. Ou seja, a pesquisa por amostragem tem por objetivo fazer afirmações válidas para o todo, entrevistando uma parcela pequena da população da qual se deseja retirar as informações. Essa pequena parcela é chamada de amostra e deve ser uma réplica mais fiel possível da população. Em outras palavras, a amostra para pesquisas em ciência política é uma parte da população conhecida que serve para fazer análise inferencial sobre as características de todos.

O teorema do limite central mostra que, conforme cresce o tamanho da amostra, a distribuição da média de uma amostra de qualquer população se aproxima de uma distribuição normal. Isso oferece fundamentos para a compreensão do p-*value*, que é o coeficiente que mede a probabilidade de não errar quando afirmar que uma medida amostral deve ser considerada válida para toda a população. Para conhecermos o comportamento de uma distribuição, basta gerar um histograma e verificar o comportamento das distribuições dos casos pela curva normal (Curva de Gauss). As distribuições normais, ou perfeitas, apresentam algumas características que são replicadas nas amostras e nas populações. Por esse motivo, em distribuições normais as estatísticas amostrais podem ser usadas com algum grau de segurança em afirmações a respeito do comportamento da população (DANTAS, 2004). As principais características de uma distribuição normal são:

- Ao padronizarmos os valores na distribuição, a média é zero;

- A curva é simétrica e apresenta a forma de um sino, distribuindo metade dos casos acima e outra metade abaixo da média. Ou seja, temos a proporção de 0,50 dos casos na cauda da direita e 0,50 na cauda da esquerda;

- A maior parte da distribuição de casos (mais de 95% do total) fica no intervalo de dois desvios padrão para cima e dois desvios padrão para baixo da média.

O teorema do limite central mostra que após determinado tamanho de amostra, a média e a distribuição dos casos tendem à normalidade. Portanto, podemos inferir para a população as informações de uma amostra graças a esse teorema. Nunca é demais lembrar que as predições a partir de amostras são efeitos esperados, ou seja, que existe uma probabilidade de que eles aconteçam. Por exemplo, considere uma pesquisa amostral sobre avaliação de governo a partir da nota dada pelos eleitores a um determinado governante, variando de zero a 10. O resultado da distribuição das notas é uma curva normal, com média sete e desvio padrão de dois pontos. Nesse caso, a média padronizada (z) será zero. Podemos dizer com apenas essas informações que as chances de sortear alguém aleatoriamente que tenha dado nota acima de sete para o governante é de 50%, pois a proporção de 0,50 encontra-se acima da média.

Na prática, esse teorema permite inferir a média da população a partir da média da amostra. A questão é que, enquanto a população é estável, as amostras variam, podendo não constituir uma réplica exata do todo. Sendo assim, a estatística inferencial apresenta resultados incertos e se apoia na Teoria das Probabilidades. Então, por que fazer uma pesquisa por amostra? Quando estudamos todos os integrantes de uma população, estamos realizando um censo. Muitas vezes o censo não pode ser utilizado por impossibilidade econômica de pesquisar toda a população ou quando ela é infinita.

Uma importante característica do Teorema do Limite Central é que a partir dele é possível identificar o p-*value*, que é o limite crítico para fazer afirmações a uma população a partir da amostra com um grau de segurança aceitável do ponto de vista estatístico. Para isso, parte-se dos seguintes princípios:

- A média da distribuição da amostra (ou as médias das amostras) é igual à média dos valores na população;

- O desvio padrão da amostra é a raiz quadrada da variância da amostra.

- Quando o tamanho da amostra é > 30 a forma da distribuição na amostra tende à normal, desde que a distribuição na população também seja normal.

Enquanto a população possui parâmetros populacionais, as amostras têm estatísticas amostrais. A partir de informações da amostra pode-se dizer algo da população por inferência estatística. A estimação pode ser pontual ou na forma de intervalo. Na estimação pontual busca-se uma estatística para representar a melhor conjectura do parâmetro; aqui, a questão é encontrar esse estimador para obter a estimativa. Já na estimação por intervalo, buscam-se valores plausíveis e não mais um ponto específico para representar o parâmetro da população. Ela considera a existência de um erro e o nome que se dá à probabilidade associada ao erro é “intervalo de confiança”. Os intervalos de confiança mais usados são de 95% e 99%.

Antes de definir tipo e tamanho da amostra, é preciso encontrar os limites da população que será pesquisada. A população é o grupo total de pessoas, animais, instituições ou coisas sobre as quais se buscam informações. A unidade de amostragem

não é necessariamente um indivíduo, pode ser uma ou mais unidades elementares em conjunto, constituindo a base do processo seletivo (ALMEIDA, 2002). As principais terminologias usadas aqui são:

Unidade de amostragem – pode conter uma ou mais unidades elementares, constituindo a base do processo seletivo.

Unidade de listagem – diz respeito ao sistema de referência ou cadastro utilizado.

Unidade de observação – é aquela que irá fornecer as informações buscadas pela pesquisa.

Variável – característica das unidades que queremos obter informações.

Parâmetro – número que descreve a característica da população. Um valor que existe, mas é desconhecido na maioria das vezes.

Estatística – número que descreve uma amostra. A estatística é obtida quando se tem uma amostra e pode variar de amostra para amostra. Ela é usada para estimar um parâmetro desconhecido.

Viés – desvio consistente, repetido e na mesma direção da estatística amostral em relação ao parâmetro da população.

Precisão ou eficiência – quando os valores de amostras repetidas da mesma população ficam muito próximos do parâmetro da população.

Existem dois grandes tipos de amostras: probabilísticas, onde todas as unidades de uma população têm a mesma chance de ser incluídas na amostra, ou seja, onde se conhece a probabilidade de cada elemento da população em participar da amostra, pois ela é escolhida de forma aleatória e em sorteio. Essa seleção independe do pesquisador e eliminam-se possíveis tendenciosidades. Não probabilísticas, quando todas as unidades da população não têm as mesmas condições de fazer parte da amostra, pois a probabilidade de escolha de determinados elementos da amostra é previamente conhecida. Nesse caso, a seleção não utiliza o sorteio aleatório em todas as fases de construção da amostra, o que leva a uma subjetividade e influência do pesquisador sobre quem fará parte da amostra. Há um intenso debate sobre as possibilidades de uso da teoria probabilística para análise de resultados em amostras não probabilísticas, no entanto, essa apostila não se ocupará de apresentar os argumentos de cada um dos lados. Vamos assumir a viabilidade funcional de se usar amostras não probabilísticas em pesquisas que buscam descrever fenômenos sociais.

5.2 Tipos de amostras

Uma amostra é aleatória quando cada elemento da população tem a mesma probabilidade de ser selecionado. Ela pode ser com reposição, quando o elemento amostrado é reposto à população e pode ser selecionado novamente; ou sem reposição, quando não há retorno à população. Nas amostras com reposição, as chances de cada elemento ser escolhido não estão relacionadas com as dos demais. Já nas amostras sem reposição, há uma mudança na probabilidade de seleção dos elementos seguintes a uma escolha, que passam a ter mais chances de serem escolhido, impossibilitando a aleatoriedade. Se a população for muito grande, os dois processos passam a apresentar, na prática, as mesmas probabilidades de seleção. Assim, quando o processo é com reposição ou a população tende ao infinito, fica mantida a independência dos elementos da amostra. Sempre parte-se da suposição de que haverá independência entre os elementos selecionados para compor a amostra.

Amostra probabilística simples

É o tipo mais básico de amostra, sendo definido por sorteio, o que permite que cada unidade da população tenha uma chance conhecida e igual de ser sorteada. Em geral, ela é representativa da população que se quer pesquisar e expressa características socioeconômicas, demográficas e de opinião da população. Porém, também pode acontecer de uma amostra probabilística não ser representativa da população. Por exemplo, uma população formada por 50% de mulheres e 50% de homens, o sorteio pode gerar 70% de mulheres e 30% de homens na amostra. Nesse caso, a amostra é aleatória, mas não representativa.

As Amostras probabilísticas simples podem ser com ou sem reposição. A primeira acontece quando se podem eleger todos os elementos da população nos sorteios seguintes. A sem reposição é quando isso não acontece. As amostras com reposição têm características estatísticas superiores à amostra sem reposição, porém, na prática, uma unidade pode ser selecionada mais de uma vez. Em caso de populações grandes, elas devem ser tratadas como amostras com reposição, nas quais a chance de sortear duas vezes a mesma unidade é muito pequena. A amostra probabilística simples serve

de base para se estimar as margens de erro e os intervalos de confiança na maioria das pesquisas de opinião sobre temas políticos, por exemplo.

Amostra sistemática

É uma variante da amostra probabilística, mas obriga que o sorteio das unidades a serem selecionadas considere todas as unidades, do início ao fim de uma lista. A amostra sistemática, além de aleatória, tem a vantagem de sempre ser representativa quando o intervalo de seleção não está correlacionado com alguma característica da listagem de unidades da população. Na prática, ela resulta da escolha de um número inicial qualquer, em uma lista com todos os integrantes da população enquanto os demais números são selecionados a partir de um intervalo pré-definido até preencher o total da amostra desejada. **Por exemplo**: quero retirar uma amostra de 100 parlamentares em uma lista com 500 deputados de uma casa legislativa. Considerando que a posição de cada indivíduo na lista é aleatória, faço o sorteio de um número entre um e cinco (500/100=5). Se o sorteio for do número 2, a partir dele eu retiro um número a cada intervalo de cinco na lista (2, 7, 12, 17, etc..) até obter a amostra de 100 indivíduos. Um detalhe interessante é que uma vez estabelecido o ponto de partida e o intervalo, a amostra sistemática não necessita da listagem completa de todos os indivíduos para ser feita, pois, dependendo do caso, podemos entrevistar uma pessoa a cada seis residências de uma rua ou gabinetes de representação. Essa amostra só pode ser considerada probabilística quando o primeiro caso é escolhido aleatoriamente e os indivíduos não estão dispostos por alguma ordem pré-determinada, tal como ordem de desempenho escolar entre alunos de uma escola. Nesse caso, indica-se que a lista seja por ordem alfabética.

Amostra estratificada

Também probabilística, procura assegurar que a amostra seja realmente representativa da população. É realizada quando se divide a população em subconjuntos ou estratos e dentro dos estratos se sorteia uma amostra aleatória, simples ou sistemática. O importante é que os estratos sejam mutuamente excludentes segundo algumas características relevantes, para que uma amostra aleatória simples seja retirada de cada grupo. Com isso, melhora-se a representatividade da amostra, obtendo-se estimativas

mais confiáveis do que as conseguidas em amostras probabilísticas simples. Na maioria das vezes, a seleção de uma amostra estratificada depende do conhecimento prévio sobre a população pesquisada. As amostras aleatórias estratificadas podem ser proporcionais ao tamanho de seus estratos correspondentes. Nesse tipo de amostra, os casos são retirados de cada grupo na mesma proporção em que ocorrem na população. A proporção da população pode ser obtida através de outras fontes, tais como censo do IBGE. Porém, quando há estratos com poucos casos, recomenda-se o uso da amostra estratificada desproporcional, onde os casos são retirados em proporção diferente do que ocorre na população. Com isso, assegura-se a inclusão de indivíduos de estratos reduzidos. Entre as amostras probabilísticas, a estratificada é a que oferece as estimativas mais seguras sobre a população.

Amostra por conglomerados

Permite fazer várias entrevistas numa mesma unidade geográfica. Em geral, é parte de um desenho de amostra em múltiplos estágios. As características dos indivíduos a serem entrevistados não precisam ser conhecidas nem antes nem durante o processo de seleção da amostra. As únicas informações necessárias referem-se às unidades geográficas e são mais fáceis de obter. O princípio da amostra por conglomerados é oposto ao da estratificada. Enquanto, nesta, a população se divide em grupos segundo características apropriadas e homogêneas, nos conglomerados, o que se busca é a heterogeneidade, pois quanto mais variabilidade, maior precisão. O processo começa com a seleção de uma amostra por meio de vários estágios, partindo de unidades maiores para menores, até chegar ao indivíduo que se deseja estudar. **Por exemplo**: se quisermos conhecer determinada característica de pacientes dos setores de emergência de hospitais, podemos selecionar aleatoriamente alguns hospitais, dividindo em dois grupos: públicos e privados. Depois, selecionamos algumas especialidades dentro desses hospitais e, em seguida, selecionamos aleatoriamente os pacientes que serão entrevistados. Esse sistema será mais representativo quando os grupos forem mais heterogêneos. Se houver uma característica que só aparece em determinado hospital ou especialidade e este não for selecionado, terminará não fazendo parte das informações disponíveis na amostra, ainda que faça parte da população.

Amostra por cotas

Todas as modalidades de amostra anteriores pertencem ao grupo das probabilísticas. A amostra por cotas é um exemplo de não probabilística. Na verdade, ela é o seu principal representante, sendo muito utilizada no Brasil, principalmente nas pesquisas de opinião pública ou intenção de voto, feitas por institutos comerciais. Na amostra por cotas, divide-se a população em subgrupos e calcula-se o tamanho proporcional de cada subgrupo. Em seguida, calcula-se o número total de entrevistas a serem feitas e divide-se de acordo com as proporções encontradas para cada subgrupo da população. Na seleção do entrevistado, o entrevistador deve escolher aquele que preenche as características da cota predeterminada e que terá que cumprir quando estiver coletando os dados. A principal vantagem desse tipo de amostra é o baixo custo aliado à rapidez. Porém, a desvantagem é que o entrevistador seleciona o entrevistado, o que pode resultar em viés, apesar de serem cumpridas corretamente as cotas da amostra, pois pode existir uma característica da população que não faça parte da cota e que esteja correlacionada com a informação que se deseja obter. Ou então, quando os entrevistadores sistematicamente entrevistam mais pessoas com essas características. Por outro lado, o método não probabilístico é amplamente usado por acrescentar segurança, diferenciando-se da amostra estratificada pela ausência de escolha aleatória. Se já sabemos que em uma população existe 50% de homens e 50% de mulheres, a pesquisa por cotas entrevistará o mesmo número de homens e de mulheres. Portanto, a amostragem por cotas depende da seleção de características da população consideradas relevantes para o estudo. Estas características devem estar associadas ao que se pretende investigar.

5.3 Tamanho das amostras

O tamanho da amostra é determinado antes de se sair a campo para a coleta de dados. Ela independe do tamanho da população, quando se trata de grandes populações. Se for utilizado o método probabilístico de amostra, realiza-se um pré-teste que permite estimar algumas informações da população e a partir de procedimentos estatísticos determina-se o tamanho ideal da amostra. No caso de amostras não probabilísti-

cas, é impossível calcular o tamanho desejável da amostra estimada de forma subjetiva. Além disso, a aleatoriedade gera a possibilidade de estimarmos erros de amostragem, o que não acontece em amostras não aleatórias. No entanto, é possível aproximar-se do tamanho ideal da amostra a partir do grau de heterogeneidade da população (desvio padrão) e do tipo de análise que se quer fazer. Para os casos de pequenas populações, o indicado é que existam pelo menos dez casos em cada casela das tabelas que serão construídas posteriormente, para a análise dos dados; o problema é que essa informação nunca pode ser obtida com antecedência.

Margem de erro

Mostra o quão perto a estatística da amostra deve cair ou deve estar em relação ao parâmetro da população. Se em uma amostra uma estatística qualquer for de 25%, e a margem de erro de quatro pontos amostrais, na população, o parâmetro deverá estar entre 21% e 29%.

Intervalo de confiança

Mostra qual percentual de todas as amostras possíveis satisfaz a margem de erro. Um intervalo de confiança de 95% indica ser esperado uma em cada 20 amostras com a mesma metodologia irá apresentar um resultado fora da margem de erro. Portanto, intervalo de confiança é uma medida usada para identificar os limites com máximo de precisão possível e a menor chance de erro ao fazer qualquer afirmação para os resultados a partir da amostra. Toda estatística é uma estimativa a partir de uma amostra. Se ela for pontual, ou seja, indicar um único valor, a chance de errar é grande. A lógica de um intervalo de confiança é a de que em determinado intervalo de valores sempre há uma chance maior de acerto do que em um único ponto. Em outras palavras, um intervalo estimado tem mais chance de acerto do que um ponto estimado. Pretende-se que, dentro do intervalo estimado a partir da amostra, encontre-se o parâmetro populacional. Assim, o intervalo de confiança de 95% é aquele em que há 95% de chance de que o parâmetro da população esteja contido na amplitude indicada a partir da amostra. A amplitude do Intervalo de Confiança depende:

- do tamanho da amostra, pois quanto maior o número de casos, menor a am-

plitude para tornar os resultados confiáveis, e;

- da variância na população e na amostra, pois quanto maior a variação dos casos, maior deve ver o intervalo de confiança para conseguir capturar o parâmetro populacional.

A fórmula para calcular a amplitude de um intervalo de confiança depende, portanto, da média e dos valores que indiquem a heterogeneidade da distribuição. Em termos matemáticos, um intervalo de confiança fica entre os seguintes limites:

Limite Superior: μ + t x (SE)

Limite Inferior: μ - t x (SE)

Onde:

μ = média da distribuição amostral;

t = é o valor encontrado na tabela padronizada de t e dos graus de liberdade da amostra (n – 1) e do nível de significância pretendido. No exemplo acima seria 95%.

SE = é o erro padrão, calculado a partir de: $SE = \frac{SD}{\sqrt{(n)}}$

Onde:

SD = desvio padrão da distribuição;

n = é o tamanho da amostra.

Uma rápida análise das fórmulas é suficiente para mostrar como tanto a homogeneidade quanto o tamanho da amostra interferem diretamente sobre a amplitude do intervalo de confiança. O tamanho da amostra (N) entra no denominador do cálculo do Erro Padrão (SE). Assim, quanto maior o N, menor tenderá a ser o SE. Esse Erro Padrão será multiplicado por um valor pré-estabelecido (t) para definir a amplitude do intervalo. Assim, quando maior a amostra e menor o desvio padrão (heterogeneidade), menor o Erro Padrão. E quanto menor o SE, menor a amplitude do intervalo de confiança em torno da média, dentro do qual se espera que esteja o parâmetro para a população.

Por exemplo: calcular os limites do intervalo de confiança para duas amostras hipotéticas, ambas com μ=25,00, Desvio Padrão de 8,00 e Intervalo de Confiança de 95%. Porém, a amostra a) tem N=50 e a amostra b) tem N=500. Vejamos os cálculos:

$$a) \qquad SEa = \frac{8{,}00}{\sqrt{(50)}} = \frac{8{,}00}{7{,}07} = 1{,}13$$

Limite Superior A= μ + tx(SE) = 25,00 + t x (SE) = 25+(1,96 x 1,13) = 25+2,21= 27,21
Limite Inferior A = μ - tx(SE) = 25,00-t x (SE) = 25-(1,96 x 1,13) = 25-2,21 = 22,79

$$b) \qquad SEb = \frac{8{,}00}{\sqrt{(500)}} = \frac{8{,}00}{22{,}36} = 0{,}35$$

Limite Superior B= μ + tx(SE) = 25,00 + t x (SE) = 25+(1,96 x 0,35) = 25+0,68= 25,68
Limite Inferior B = μ - t x (SE) = 25,00-t x (SE) = 25-(1,96 x 0,35) = 25-0,35 = 24,65

O resultado é que a amplitude do intervalo de confiança da amostra A está variando 4,58, de 22,79 a 22,21, enquanto o intervalo de confiança da amostra B varia apenas 1,03, de 24,65 a 25,68. A diferença entre elas é explicada pela mudança no tamanho da amostra, pois todos os demais valores do exemplo foram mantidos. A mesma lógica do intervalo de confiança para a média amostral é usada em coeficientes de regressão, para estabelecer o intervalo para estimação do coeficiente Beta.

Erro amostral e intervalo de confiança são instrumentos que permitem fazer uma estimativa dos valores populacionais, pois o parâmetro na população sempre será desconhecido. Além disso, a existência de intervalos de confiança é uma prova de que o método pode falhar e que a probabilidade de ocorrência dessa falha é estimada, mesmo que as pesquisas realizadas sejam totalmente corretas. Sempre existirá erro algum amostral, por maior que sejam as amostras. No entanto, outro tipo de erro, o não amostral, pode levar a resultados mais distantes do parâmetro da população, sem que haja condições de serem detectados ou controlados. Ou seja, o erro amostral é apenas um dos erros possíveis de uma pesquisa.

Variância ou heterogeneidade

Quando se trabalha com amostras aleatórias simples, o tamanho da amostra depende da variância das médias, também chamada de erro amostral. Quanto maior for

a variância na população – que é desconhecida – maior terá que ser a amostra. Para obter a variância da amostra utilizam-se as fórmulas a seguir:

a) $\sigma_x = \frac{\sigma}{\sqrt{n}}$ b) $sx = \frac{s}{\sqrt{n}}$

Na fórmula (a), obtêm-se o erro padrão a partir do desvio-padrão da população (que raramente é conhecido); enquanto na fórmula (b) a variância é obtida a partir do desvio-padrão estimado. Conhecendo o erro-padrão, o tamanho da amostra é fornecido pela relação entre o nível de confiança com que o pesquisador quer trabalhar, a precisão da sua estimativa e o desvio-padrão da população, que pode ser real ou estimado, como nas fórmulas a seguir:

a) $n = \left(\frac{z.\sigma}{d}\right)^2$ b) $n = \left(\frac{z.s}{d}\right)^2$

Onde:

n = tamanho da amostra;

z = desvio encontrado na tabela da curva normal que corresponde ao nível de confiança estabelecido;

d = margem de erro aceita pelo pesquisador.

Na fórmula (a), o tamanho da amostra é definido a partir do erro-amostral da população, quando conhecido. Na amostra (b), é pelo erro-padrão estimado.

Exemplo: Definir tamanho da amostra para uma pesquisa a ser realizada com nível de confiança de 95% (valor correspondente na tabela = 1,96), margem de erro aceitável de 4% e variância de 0,5.

$$n = (\frac{(1{,}96\ x\ 0{,}5)}{0{,}04})^2 = 600{,}25\ entrevistas$$

Se não houver uma estimativa para a variância, o tamanho da amostra cresce exponencialmente, como se vê a seguir:

$$n = (\frac{1{,}96}{0.04})^2 = 2401\ entrevistas$$

Quando não se conhece a variância da população, costuma-se considerar a pior situação, que é a de que a característica a ser analisada apresenta uma variação muito grande, a maior delas em uma população é de 50%, ou seja, metade da população possui a característica, enquanto a outra metade não a possui. Sendo assim, a variância para a fórmula passa a ser de 0,5. Com isso, o erro real aferido na amostra tenderá a ser menor que a margem de erro inicial, que no exemplo foi de 4%.

Existem outras variantes da fórmula para definição do tamanho da amostra, que dependem também do nível de confiança, da variância e da precisão que se pretende ter, tanto em cálculos para amostras em grandes populações quando para em pequenas populações.

c) $n = \frac{z^2.(p.q)}{e^2}$ d) $n = \left(\frac{t.s}{e}\right)^2$

A fórmula (c) é usada para calcular o tamanho da amostra para pequenas populações, enquanto a fórmula (d) serve para grandes populações.

Onde:

z ou t = desvio encontrado na tabela da curva normal que corresponde ao nível de confiança estabelecido, dependendo do tamanho da população (estatística t serve para pequenas amostras);

p = proporção de determinada característica presente na população;

q = 1 – p

e = margem de erro aceita pelo pesquisador.

Também é possível obter-se a margem de erro real de uma pesquisa a partir da variância apresentada pela amostra e pelo tamanho da amostra, a partir da fórmula a seguir:

$$e = z.\sqrt{\frac{p.q}{n}}$$

Onde:

e = margem de erro obtida;

z = desvio correspondente ao nível de confiança;

p = proporção de determinada característica presente na população;

q = 1 – p;

n = tamanho da amostra;

Existe ainda uma fórmula para fazer a correção do tamanho da amostra quando se está trabalhando com populações pequenas. Trata-se do Fator de Correção da População (FCP) e deve ser multiplicado pelo número de indivíduos encontrado na amostra para grandes populações, como segue:

$$FCP = n.\sqrt{\frac{N-n}{N-1}}$$

Onde:

n = tamanho da amostra;

N = tamanho da população;

Por exemplo, uma amostra de 600 indivíduos deve ser corrigida para aplicação em uma população de 1000 pessoas.

$$FCP = n.\sqrt{\frac{1000-600}{1000-1}} = 600x0{,}632 = 379{,}6$$

Resposta = aproximadamente 380 entrevistas

A tabela 5.1 a seguir indica os tamanhos de amostras em função do erro máximo aceitável pelo pesquisador, da heterogeneidade esperada da população, via desvio padrão, e do intervalo de confiança, que pode ser de 95% ou 99% para cada uma das faixas de desvio padrão. Para cada uma das combinações de erro, heteroge-

neidade e intervalo de confiança há um valor que representa o tamanho da amostra. A faixa de erros mais comum, entre 3 e 4 pontos percentuais, e os intervalos de confiança de 95% e 99% para um desvio padrão de 0,50 estão em negrito. São os tamanhos de amostras mais utilizados em pesquisas de opinião. A tabela permite identificar as diferenças nos tamanhos das amostras em função da variação de um dos três fatores. Por exemplo, ao passar de 95% para 99% de intervalo de confiança para um erro de 3 pontos percentuais e um desvio padrão de 0,50, o tamanho da amostra cresce de 1.100 para 1.800 indivíduos. Já ao passar de 3 pontos percentuais para 3,5 de erro, mantendo 95% de intervalo de confiança e desvio padrão de 0,50, o tamanho da amostra cai de 1.100 para 820 indivíduos. Por fim, se mantemos o erro de 3 pontos percentuais para intervalo de confiança de 95%, mas mudamos de desvio padrão de 0,50 para o outro extremo da tabela, de 0,05 ou 0,95, o tamanho da amostra cai de 1.100 para 210 indivíduos.

Tabela 5.1 – Tamanho da Amostra Com Base na Homogeneidade da População e Margem de Erro

	Dp. 05 ou 95		Dp. 10 ou 90		Dp 20 ou 80		Dp 30 ou 70		Dp 40 ou 60		Dp. 50 ou 50	
Erro aceitável	Intervalo de Confiança											
	95%	99%	95%	99%	95%	99%	95%	99%	95%	99%	95%	99%
1	1.900	3.000	3.600	5.700	6.400	10.100	8.400	13.270	9.600	15.160	10.000	15.800
1,5	960	1.520	1.850	2.900	3.300	5.200	3.800	6.000	4.300	6.800	4.500	7.100
2	480	760	900	1.400	1.600	2.500	2.100	3.300	2.400	3.800	2.500	4.000
2,5	340	540	650	1.030	1.160	1.830	1.350	2.130	1.540	2.430	1.600	2.530
3	210	330	400	630	710	1.120	930	1.470	1.060	1.680	**1.100**	**1.800**
3,5	170	270	330	520	590	930	690	1.090	790	1.250	**820**	**1.300**
4	120	190	225	350	400	630	520	820	600	950	**600**	**990**
4,5	105	165	200	315	360	570	420	660	480	760	500	790
5	75	120	150	235	260	410	340	540	370	580	400	640
5,5	70	110	130	205	235	370	275	435	315	500	330	520
6	----	---	100	160	180	280	230	360	270	430	280	440
6,5	---	---	95	150	170	260	200	260	230	365	240	380
7	---	---	75	120	130	200	170	270	190	300	200	330
7,5	---	---	65	100	115	180	150	235	170	270	180	285
8	---	---	---	---	100	160	130	200	150	240	160	250
8,5	---	---	---	---	95	150	120	190	135	210	140	220
9	---	---	---	---	80	125	105	165	115	180	120	200
9,5	---	---	---	---	75	115	95	150	105	165	110	175
10	---	---	---	---	---	---	85	135	95	150	100	160

5.4 Exercícios propostos

a) Escreva sobre as diferenças entre as amostras por conglomerado e por cotas. Dê exemplos.

b) Calcule o tamanho da amostra para uma população infinita, com intervalo de confiança de 95% (1,96), margem de erro aceitável de 3% e desvio padrão de 0,5. Refaça o cálculo, mudando o erro aceitável para 5%. Interprete os resultados.

c) Calcule o erro real de uma amostra a partir de p=0,30, q=0,70, intervalo de confiança de 95% (1,96) e 1.100 questionários aplicados.

d) Considere uma população de 800 indivíduos, para a qual foi calculado o tamanho de uma amostra para populações infinitas, e faça a correção da população aplicando a fórmula FCP para uma amostra de 500 questionários.

5.5 Bibliografia citada no capítulo

ALMEIDA, Alberto C. **Como São Feitas as Pesquisas Eleitorais e de Opinião.** Rio de Janeiro: Editora FGV, 2002.

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

Capítulo VI

Surveys para Coleta de Dados Individuais

Um questionário serve para que as pessoas
digam o que esperam que os outros
descubram sobre o que elas pensam.
Não o que elas pensam.

Quando a pesquisa busca informações no nível microssociológico, ou seja, individual, existem pelo menos três grandes maneiras para se compreender o comportamento humano empiricamente:

a) Observar o comportamento que ocorre naturalmente na realidade. Seu ponto forte é o realismo da situação estudada. Seu ponto fraco é a necessidade de acompanhamento da realidade estudada, o que demanda tempo e recursos que muitas vezes não estão disponíveis para o pesquisador.

b) Criar situações artificiais e observar o comportamento diante de tarefas definidas para essas situações; através do experimento. Possibilita a randomização e total controle de características das pessoas estudadas quanto a inferências causais. Por outro lado, ao isolar o objeto de análise da influência de fatores não controlados, os resultados, ainda que bastante precisos para laboratório, podem se distanciar da realidade.

c) Perguntar às pessoas sobre o que fazem ou pensam, através de um questionário. Aqui, a vantagem é que se assegura melhor representatividade e permite generalizações para uma população mais ampla, pois os indivíduos continuam inseridos na

realidade enquanto são pesquisados. Um *survey* pode ser um censo (grande *survey*), uma pesquisa de opinião pública, de mercado, acadêmica, etc.

6.1 Questionário

Lessler e Kalsbeek (1992) definem *survey* como um estudo científico de unidades de uma população existente, formada por pessoas, instituições ou qualquer objeto físico. A partir de um *survey,* é possível conhecer e observar a população em suas condições naturais (essa é uma diferença em relação a experimentos) para fazer afirmações quantitativas a respeito das características agregadas da população. O importante nessa definição de Lessler e Kasbeek (1992) é notar que estamos tratando sempre de informações agregadas em pesquisas por *survey*s, não permitindo fazer afirmações sobre o nível individual dos que compõem a população pesquisada. Ou seja, partimos da coleta individual de informações para agrega-las e, a partir daí, inferirmos conclusões sobre o todo. Um *survey* pode ser usado em estudos censitários, quando inclui todos os membros da população ou em pesquisas amostrais, quando se utiliza uma amostra (sobre amostras, ver Capítulo V) que seja representativa da população estudada.

No caso de *survey*s, a formulação do questionário tem grande efeito sobre as conclusões da pesquisa. Perguntas distintas sobre o mesmo assunto podem levar a resultados completamente diferentes. Como o questionário tem por objetivo medir as ocorrências de variáveis ou conceitos, para cada variável existe uma pergunta correspondente que a mensura. **Exemplo:** variável “intenção de voto”. Pergunta: se a eleição fosse hoje e os candidatos estes, em quem o(a) senhor(a) votaria?

Existem dois tipos de perguntas em questionários: abertas e fechadas. Perguntas abertas são aquelas em que o entrevistado pode dar sua própria resposta. Nas perguntas fechadas, as respostas possíveis são pré-definidas e o entrevistado tem que escolher entre elas. A primeira gera uma resposta espontânea enquanto a segunda promove uma resposta estimulada. Em geral, consideram-se as seguintes regras para a formulação de um questionário:

- Sempre perguntar o que se deseja saber.
- Elaborar perguntas claras.
- Usar redação balanceada, apresentando os dois lados da questão.
- Evitar jargão ou lugar-comum.
- Levar em consideração o nível de instrução formal dos entrevistados.
- Elaborar perguntas de modo a evitar as respostas socialmente aceitáveis.
- Testar as opções de respostas das perguntas fechadas.
- Avaliar que perguntas inserir no início e no fim do questionário.

A atenção aos princípios técnicos de um questionário é importante para que os resultados de um *survey* sejam confiáveis para quando se quer produzir afirmações quantitativas sobre uma população estudada qualquer. Uma vez respeitados esses princípios, o *survey* serve para descrever e analisar relações entre subgrupos da população ou para interpretar a natureza de fenômenos sociais e de processos econômicos. O desrespeito a esses princípios leva a erros nas pesquisas.

Existem dois grandes tipos de erros em *survey*. O primeiro tipo já foi discutido no Capítulo V, dedicado ao cálculo de amostragem, e chama-se “erro amostral”. Ele acontece quando há distorções na amostra pesquisada em relação às características da população. Não se trata de um erro em si, mas de um erro anterior à aplicação do questionário, que se dá no momento de definição do desenho da amostra. Aqui, a atenção ficará no segundo tipo de erro, o chamado erro não amostral.

6.2 Erros não amostrais

Boa parte do debate sobre problemas em pesquisas sobrevaloriza o erro amostral e negligencia outras fontes de erro, tal como os erros não amostrais. Esse tipo de erro é causado por equívocos durante a execução dos procedimentos de *survey*. São gerados por definições mal utilizadas nas perguntas, imperfeições na tabulação dos dados, falhas na obtenção das respostas, entre outras. Como defendido por Lessler e Kalsbeek (1992), apenas um desenho de pesquisa perfeito e perfeitamente implementado poderá estar livre de erros não amostrais.

Como um trabalho de campo de má qualidade pode comprometer toda a pes-

quisa, é importante ter cuidado especial nessa fase. As equipes de campo responsáveis pela coleta de dados são compostas por dois tipos de profissionais: os entrevistadores e os supervisores, que podem trabalhar juntos ou separados. Para essa fase da pesquisa existem pelo menos quatro fontes possíveis de erros não amostrais:

i) Quando a informação está incorreta;

ii) Quando a definição de parâmetros usados na pesquisa estiver errada ou for insuficiente;

iii) Quando as unidades de pesquisa selecionadas na amostra não possuírem representação na população, ou;

iv) Quando o erro identificado for maior que o estimado (LESSLER; KALSBEEK, 1992).

Em outras palavras, variáveis definidas corretamente podem ter suas informações colhidas de maneira incorreta. Informações podem ser perdidas entre os integrantes da amostra. A tabulação dos dados pode apresentar erros e a população a ser estudada pode não ser a melhor. Existem vários fatores que podem causar uma diferença entre os resultados de um *survey* e os valores na população. O treinamento é fundamental para esse tipo de trabalho. Lembre-se que entrevistador sem treinamento influencia as respostas.

Uma boa análise dos resultados de uma pesquisa de opinião começa com algumas delimitações. Em primeiro lugar, é preciso delimitar o escopo da análise, buscando apenas o que se deseja saber. Também é preciso considerar que a interpretação dos resultados de uma só pesquisa apresenta limitações no fornecimento de dados contextuais. O ideal é ter condições para fazer uma análise temporal, utilizando mais de um *survey* aplicado em momentos distintos do tempo. Outra alternativa para quando se trabalha com uma única pesquisa é a produção de índices (ver Capítulo III) que reúnem variáveis. Índices são importantes ferramentas para a análise de dados quantitativos. Do ponto de vista geral, o *survey* tem por objetivo descrever, explicar ou explorar uma questão. A unidade de análise do *survey* normalmente é o indivíduo, mas pode ser famílias, cidades, etc. É possível fazer *survey* com dados de pessoas e de outras unidades de análise. Nesse caso, é preciso tomar cuidado com a falácia ecológica.

6.3 Principais tipos de *surveys*

***Survey* de corte transversal**

É o mais comum. São coletadas informações de uma amostra selecionada com intenção de explicar a relação de variáveis em uma população em dado momento.

***Survey* de corte longitudinal**

Permite a análise temporal, pois os dados são coletados em diferentes pontos do tempo. Ele pode ser:

- Estudos de tendências – a mesma população é amostrada e estudada em diferentes momentos. Em geral envolve longos períodos de tempo.

- Estudos de coortes – foca uma população específica para cada análise feita. A amostra não precisa ser a mesma. Trata-se de um recorte específico da população. **Exemplo:** acompanhar uma turma de formandos por ano ou acompanhar os nascidos logo após uma guerra.

- Painel – são dados recolhidos em momentos diferentes de uma mesma amostra. Permite fazer afirmações a respeito de mudanças na população com maior segurança.

O corte transversal pode se aproximar do longitudinal quando se coletam informações sobre o passado dos respondentes. Também se pode usar a idade como indicador de mudanças no tempo. Todo estudo contextual enriquece o *survey*. Ele é feito anotando dados de algumas variáveis de contexto junto com a entrevista.

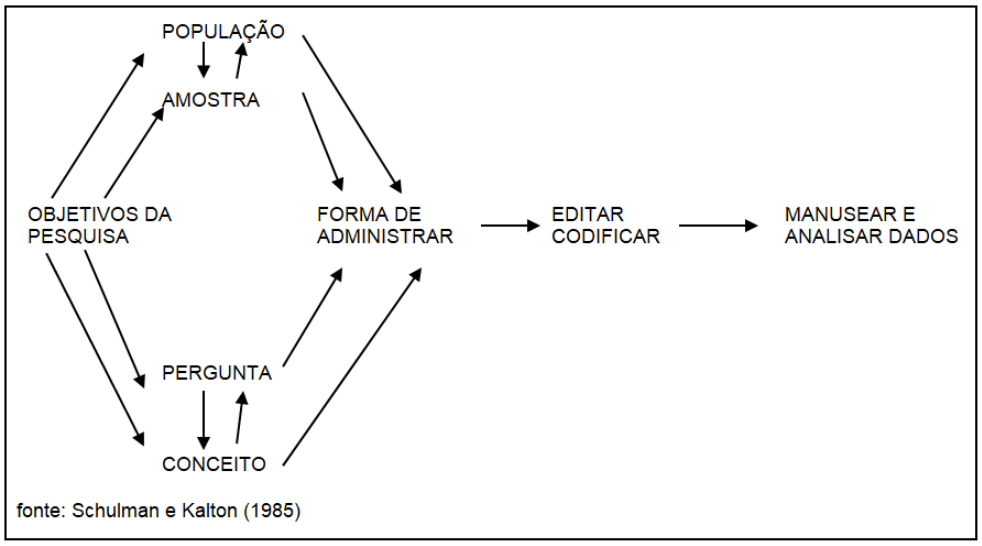
Em um *survey*, o questionário é o centro da pesquisa. Perguntas precisam ter como características básicas a objetividade, a brevidade e a clareza. As questões devem focar diretamente o assunto que se quer pesquisar, serem curtas, claras e simples. O questionário é definido como “um conjunto de perguntas sobre determinado tópico que não testa a habilidade do respondente, mas mede a opinião, interesses, aspectos da personalidade e informação biográfica” (YARENKO; HARRISON; LYNN, 1986). A forma como um questionário é preparado tem grande efeito sobre os resultados da pesquisa. Perguntas distintas sobre o mesmo assunto podem levar a resultados completamente diferentes.

As pesquisas realizadas no período de uma campanha eleitoral, por exemplo,

têm de ser tratadas como outra fonte de evidências ou um conjunto de indicadores e não como determinantes exclusivas de decisões. Ou seja, ainda que as pesquisas de opinião exerçam um papel fundamental dentro de uma campanha política, pois auxiliam na tomada de decisões, não devem ser responsáveis pela última palavra da equipe de campanha. Apesar do consumo de pesquisas ter aumentado nas últimas décadas, a análise dos dados ainda é limitada, proporcionando vantagem competitiva para quem sabe interpretar e analisar as informações extraídas dos questionários.

Antes de elaborar um questionário é preciso saber qual o objetivo da pesquisa em termos de conceitos a serem pesquisados e qual é a população-alvo? Em uma pesquisa, os objetivos levam necessariamente à relação conceito/pergunta e a relação população/amostra, como indica a figura a seguir:

Figura 6.1 – Esquema de organização da análise de Schulman e Kalton



Conceito/Pergunta

O objetivo de cada estudo é o que determina os conceitos a serem investigados. Podemos diferenciar, conceitualmente, uma avaliação de algo existente, de um levantamento de necessidades de algo existente. Assim como distinguir existência de um objeto externo ao indivíduo, de um estado de espírito interno. A partir daí, o objetivo de um estudo é fundamental para determinar as perguntas concretas a serem apresentadas. Dependendo dos conceitos pesquisados, o conteúdo das perguntas varia. No qua-

dro 6.1, abaixo, são apresentados alguns exemplos de conceitos que se transformam em variáveis e a forma como elas se transformam em perguntas:

Quadro 6.1 – Conceitos transformados em perguntas de questionário

Conceito	Pergunta
Intenção de voto	Se a eleição para prefeito fosse hoje e os candidatos os seguintes, em quem o senhor votaria?
Religião	Qual a sua religião?
Escolaridade	Até que série o senhor estudou?
Renda	Qual a renda mensal da sua família?

População/Amostra

Dependendo do objetivo da pesquisa, do tamanho e acessibilidade à população alvo, serão definidos diferentes tipos de amostras. Os critérios relacionados à população-alvo para definir a amostra normalmente são nível educacional, idade, gênero, etc.

Um questionário tem por objetivo medir as ocorrências de determinadas variáveis e, por consequência, permitir a análise de conceitos relacionados a elas. Para cada variável, existe uma pergunta correspondente que a mensura (ver quadro 6.1). É possível identificar essa interdependência entre a elaboração de um questionário e a estratégia de aplicação com o grau de complexidade dos conceitos que determina o número de itens e a forma de apresentação deles. Existe uma relação recíproca entre as características da população-alvo e a complexidade dos conceitos a serem investigados, pois eles definem a maneira de transformação dos conceitos em perguntas. O tamanho da amostra influencia o tamanho do questionário, por exemplo. A definição final do questionário dependerá da população-alvo, do tamanho da amostra e dos conceitos a serem explorados, assim como dos recursos disponíveis para sua aplicação.

Em uma pesquisa por questionário, a disposição do respondente em revelar algo sobre si mesmo varia conforme a situação. Por isso, o respondente precisa ser convencido de “gastar” seu tempo e fazer um esforço mental para responder as perguntas. O pesquisador deve fazer algumas perguntas antes de pensar no questionário: até que ponto se aceita ser indagado por um estranho sobre assuntos pessoais? Quais são os assuntos privados e públicos?

Bingham e Moore (1934) definem entrevista como uma conversa com um objetivo e, assim como qualquer outra interação social, é composta por um cumprimento, na relação em si e na despedida. Assim, a estrutura do questionário deve possuir essas três fases. Considerando os problemas inerentes da relação entre entrevistador e entrevistado, há três ações operacionais que o pesquisador deve adotar para obter os melhores resultados de seu questionário:

1) No cumprimento, reconhece-se o outro e estabelece-se o nível de confiança necessário oferecendo um sinal de apreciação antecipadamente; identificando-se com uma instituição conhecida e legitimada; aproveitando outros relacionamentos de troca.

2) A segunda parte da entrevista é a interação entre perguntas e respostas. Aqui, é importante reduzir o custo das respostas fazendo com que a entrevista pareça breve; diminuindo os esforços mentais exigidos, eliminando a possibilidade de embaraços, qualquer implicação de subordinação ou custos financeiros imediatos.

3) Na despedida, reforça-se qualquer sinalização de benefícios futuros através da demonstração de consideração, oferecendo apreciação verbal; apoiando seus valores; oferecendo recompensas concretas ou tornando o questionário interessante (GÜNTER, 2003).

Para estabelecer confiança, o entrevistador precisa capturar o interesse do respondente pelo tema e sua importância para quem está respondendo. Na prática, trata-se da apresentação do instrumento e da pessoa que irá aplicá-lo. O entrevistador, devidamente treinado, deve explicar e tirar dúvidas sobre quem é o responsável pela pesquisa, quais os objetivos e o que o respondente deve fazer. Na interação entre as perguntas e as respostas, a principal meta é reduzir os custos para quem está respondendo. Como o respondente pode desistir a qualquer momento, sempre é preciso convencê-lo a manter seu interesse fazendo com que a tarefa pareça breve, ao reduzir o esforço físico e mental, eliminando a possibilidade de embaraços e qualquer ideia de subordinação e custos financeiros. Por fim, na despedida, é preciso demonstração de cortesia, que consiste em um agradecimento pela colaboração do entrevistado. Porém, não se deve exagerar, fazendo promessas irreais, tais como “sua participação é importante e isso irá melhorar sua vida”, etc. Comunicar aos entrevistados os resultados é uma maneira interessante de recompensá-los.

6.4 A ESTRUTURA DO QUESTIONÁRIO

Pensar antecipadamente a estrutura do questionário contribui para a redução do esforço físico e mental do respondente e assegura que todos os temas de interesse do pesquisador sejam tratados numa ordem sugerida pela conversa com objetivo de manter o interesse do respondente. Para isso, o primeiro passo é focalizar o objetivo da pesquisa nas perguntas. Em seguida, saber claramente o que está sendo incluído em cada item no questionário. Depois saber o que as respostas implicam para o andamento da pesquisa.

Uma forma inicial de estruturar o questionário é direcionar as perguntas do geral para o específico e do menos delicado e pessoal para o mais delicado e pessoal. Sendo assim, o primeiro conjunto de perguntas precisa ser mais geral e menos sensível. As primeiras perguntas servem mais para estabelecer a confiança do que para obter informação do entrevistado. Por isso não é recomendado começar a entrevista com perguntas do sobre o nome, sexo, idade ou renda familiar. Se alguém concorda em responder um questionário é porque acha o tema interessante. As primeiras perguntas devem tratar dessa temática. Somente no último bloco de perguntas é que se deve tratar das características socioeconômicas dos entrevistados.

Quando forem tratados aspectos que, ao serem reunidos em conjunto, constituem-se em escala, as perguntas precisam ser misturadas para evitar que dois ou mais sejam apresentados um após o outro. Se o questionário tiver uma ordem lógica, progredir do geral para o específico, fazendo perguntas pessoais só depois de se estabelecer um bom nível de confiança, o pesquisador terá melhores condições para obter respostas mais autênticas.

As Perguntas

A parte central do questionário são as perguntas através das quais se pretende obter determinada informação. Fowler (1998) define uma boa pergunta como aquela que gera respostas fidedignas e válidas. Uma medição válida é aquela que mede exatamente o que se quer medir. Já as respostas fidedignas ou confiáveis são as obtidas em pesquisas diferentes, desde que feitas com a mesma metodologia. Para tanto, as

perguntas devem apresentar cinco características básicas:

- Precisam ser compreendida de maneira consistente;
- Precisam ser comunicada de maneira consistente;
- As expectativas quanto à resposta precisam ser claras para o entrevistado;
- Se não estiver verificando conhecimento, o entrevistado precisa ter toda informação necessária para a resposta;
- O entrevistado precisa estar disposto a responder.

A garantia da presença de todos os atributos em uma pergunta só se consegue se ela for específica, breve, clara, escrita em vocabulário apropriado e correto. O princípio básico é realizar pelo menos um pré-teste ou estudo piloto com as perguntas para um público com as mesmas características do público-alvo, evitando suposições *a priori*. Para escrever boas perguntas em um *survey,* Sudman e Bradburn (1982) dão três sugestões:

1ª - controle o impulso de escrever perguntas específicas sem ter refletido sobre as respostas às perguntas.

2ª - anotar as questões da pesquisa e tenha-as sempre por perto no momento da produção do questionário.

3ª - cada vez que fizer uma pergunta, indague porque você quer saber isso.

Com relação à linguagem usada na formulação da pergunta é preciso atenção especial com as ambiguidades. Ela deve ser bem compreendida pelo público-alvo da pesquisa. Evitar elementos linguísticos como abreviações, gírias ou termos regionais, assim como termos específicos ou sofisticados que estejam além da compreensão do respondente. O viés é outro problema gerado pelo mau uso da linguagem na formulação das perguntas, pois a escolha de palavras pode direcionar as respostas. Os questionários e suas perguntas são feitos para realizar medições não enviesadas, ou seja, não devem favorecer sistematicamente determinados resultados. Não basta mensurar as variáveis, é preciso mensurá-las corretamente. Convêm fazer um pré-teste para verificar se as perguntas estão sendo compreendidas pelo público-alvo, sempre que houver um novo questionário para um novo público.

Em resumo, as regras gerais para preparação do questionário são perguntar

o que se deseja saber; elaborar perguntas claras; usar uma redação balanceada que apresente os dois lados da questão; evitar jargões ou lugares-comuns; levar em consideração o nível de instrução formal dos entrevistados; elaborar perguntas de modo a evitar as respostas socialmente aceitáveis; testar as opções de resposta das perguntas fechadas; avaliar que perguntas inserir no início e que perguntas inserir no fim do questionário.

Perguntas abertas X Perguntas fechadas

Existem dois tipos de perguntas: abertas e fechadas. Perguntas abertas são aquelas em que o entrevistado pode dar sua própria resposta, havendo liberdade para qualquer resposta possível. Nas perguntas fechadas, as respostas possíveis são pré-definidas e apresentadas ao entrevistado, que tem que escolher entre elas. O primeiro tipo gera uma resposta espontânea, enquanto o segundo promove uma resposta estimulada. É indicado, em uma pesquisa inicial e exploratória, onde não se conhece a abrangência e variabilidade das possíveis respostas, que as perguntas sejam abertas. Porém, quando já se conhecem os tópicos geralmente mencionados pelo entrevistado sobre determinado tema, quando existem muitos respondentes ou pouco tempo, deve-se optar pelas perguntas fechadas para não correr o risco de inviabilizar o trabalho. No entanto, é possível usar as duas modalidades de perguntas para obter a mesma informação.

É importante notar que existem duas formas de apresentar perguntas abertas aos entrevistados. Em uma delas, há respostas escritas no questionário que devem ser assinaladas pelo entrevistador e que não são apresentadas ao entrevistado, embora já estejam codificadas. Na outra, existe um campo aberto para que o entrevistador ou o próprio entrevistado anote exatamente a resposta que ele pretende fornecer à pergunta. Nesse caso, após o trabalho de campo, uma equipe de codificadores deverá fazer o trabalho de organização das respostas escritas nos questionários.

Conteúdo das perguntas

Uma pergunta nunca pode gerar ameaça ao entrevistado. Se o tema for sensível para o respondente ou se envolver comportamentos considerados socialmente

inaceitáveis é necessário verificar a melhor maneira de obter a informação sem provocar constrangimento, o que na maioria das vezes significa substituir uma pergunta direta por várias indiretas. Pior do que perder um respondente que se irrita com uma pergunta constrangedora é, a partir de então, receber respostas não-autênticas. Pode-se evitar o constrangimento em função da falta de conhecimento, deixando claro que as perguntas não são testes, sendo natural que as pessoas não tenham respostas para todos os itens. Perguntas que testam conhecimento são invariavelmente constrangedoras, porém podem ser importantes em uma pesquisa social por servirem como filtro das questões sobre atitudes. Para reduzir o nível de ameaça, pode-se iniciar uma pergunta com "você sabe por acaso" ou "a propósito". Também é necessário evitar as adivinhações dos entrevistados que não querem admitir falta de conhecimento. Faz-se isso com mais de uma pergunta sobre o mesmo tema, o que reduz a possibilidade de acertar a resposta correta por acaso, principalmente nas perguntas cuja resposta é sim ou não.

Vale lembrar que é importante perguntar apenas o que será utilizado na pesquisa. Na última seção, deve-se buscar informações factuais, antecipando-as com a justificativa de que essas questões são necessárias para caracterizar melhor o grupo de pessoas que participam da pesquisa. Só então são feitas perguntas sobre gênero, religião, idade, escolaridade, moradia, renda, etc. No final, deve-se encerrar recordando que todas as declarações serão tratadas de maneira confidenciais e a apresentação dos resultados será de tal forma a não permitir a identificação individual dos participantes.

Tipos de perguntas fechadas

Perguntas dicotômicas: são direcionadas a respostas que se limitam a duas opções. Normalmente afirmativa ou negativa. **Exemplo:**

A1. O senhor concorda ou discorda com a seguinte afirmação? "De maneira geral, no último ano a economia brasileira tem melhorado".
1. Concorda
2. Discorda
8. NS
9. NR

Perguntas de Múltiplas Escolhas: são aquelas que permitem ao respondente escolher uma ou mais alternativas. Aqui, as respostas não seguem uma ordem de preferência. **Exemplo:**

A1. Quando o assunto é política, com quem o senhor costuma conversar?
A11. Parentes mais próximos (pai, mãe, tios, filhos, avós).
A12. Parentes mais distantes (primos, tios de segundo grau).
A13. Amigos ou pessoas que mantenha contato permanente.
A14. Colegas de trabalho, escola ou do bairro.
A15. Chefe imediato no trabalho, professor ou líder religioso.
A16. Pessoas conhecidas em redes sociais
A17. Outra
A18. NS
A19. NR

Perguntas em ordem de preferência: além de permitir mais de uma resposta, aqui o respondente deve indicar a ordem de preferência das alternativas. Essa preferência pode ser por critério de avaliação ou por ordem de lembrança, para se medir a saliência na memória. **Exemplo:**

A1. Responda qual a ordem de maior importância que o senhor dá para as opiniões políticas das seguintes pessoas. O 1 significa opinião mais importante e o 6 é opinião menos importante. (USE 8 PARA NÃO SABE E 9 PARA NÃO RESPONDEU)

ORIGEM DA OPINIÃO	ORDEM
A11. Familiares	
A12. Amigos	
A13. Colegas	
A14. Chefe no trabalho, professor ou líder religioso	
A15. Pessoas conhecidas em redes sociais	
A16. Apresentadores de TV, rádio.	

Perguntas com Escala de graduação qualitativa (Likert): São perguntas cujas respostas permitem a organização de escalas de intensidade. **Exemplo:**

A1. Qual a sua posição a respeito da seguinte afirmação: “É sempre preferível a igualdade social, ainda que para tanto seja necessário restringir a liberdade individual”.
1. Concordo totalmente.
2. Concordo.
3. Nem concordo, nem discordo.
4. Discordo.
5. Discordo totalmente.
8. NS
9. NR

Perguntas com escala de valores de Osgood: Permite classificar as opiniões sobre diferentes dimensões do objeto pesquisado. **Exemplo:**

A1. Em uma escala de 1 a 7, onde você posicionaria o seu curso de graduação no que diz respeito a:

		1	2	3	4	5	6	7	
A11. Currículo:	Atrasado								Atualizado
A12. Corpo docente:	Despreparado								Preparado
A13. Infraestrutura:	Precária								Adequada
A14. Alunos:	Desinteressados								Interessados

Perguntas encadeadas: São aquelas feitas em sequência, onde a segunda pergunta é condicionada à primeira, pois depende das informações já fornecidas ou complementa a resposta anterior. **Exemplo:**

A1. Você está matriculado em algum curso universitário no momento?
1. Sim
2. Não
8. NS
9. NR

A2. Se sim, em qual das seguintes áreas?
1. Humanas
2. Exatas
3. Saúde
4. Sociais aplicadas
5. Agrárias
6. Outra
8. NS
9. NR

Além da preocupação natural com a qualidade das perguntas, um bom questionário também depende da qualidade das respostas obtidas, ou seja, com a forma possível de responder e com as alternativas apresentadas ao entrevistado. Isso é o que será discutido a seguir.

6.5 Características técnicas das respostas

Existem basicamente quatro níveis de medição ou escalas de respostas a questionários de *survey*s. A escala nominal é onde os números são usados apenas para identificar objetos, pessoas ou categorias. Escala ordinal ocorre quando além da identificação, os números ordenam numa dimensão específica as diferentes respostas. Na Escala intervalar, as características, além de ordenadas conforme uma dimensão específica, permitem o estabelecimento de intervalos ou distâncias entre as alternativas, pois têm tamanho conhecido e podem ser comparados. Na escala de Razão, reúnem-se as possibilidades de uma mediação nominal à escala intervalar, gerando uma escala de razão, ou seja, trata-se de duas respostas, uma de presença ou ausência, e outra de volume. Normalmente, ela está relacionada a perguntas abertas.

A distinção entre os quatro níveis de escala é importante por ter relação direta com a complexidade da análise posterior. As informações de qualquer uma das escalas podem ser apresentadas em estatísticas descritivas, tabelas ou gráficos. Porém, para usar estatísticas inferenciais, que permitem verificar se determinadas relações são sistemáticas ou não, é necessário que os dados estejam em escalas nominais e ordinais, pois assim podem ser analisados com testes não paramétricos.

Quadro 6.2 – Caracterização por tipo de variável

Tipo	Características	Exemplos	Características Formais
Nominal	Apenas para identificar pessoas, objetos ou categorias.	Cor de cabelo, estado civil, nome, marca de carro.	Equivalência, igualdade. "="
Ordinal	Respostas podem ser ordenadas em uma dimensão própria.	Ordem de preferência, de chegada, status social, escala de Likert (aqui os valores modais e medianos podem ser calculados, mas não a média).	Além de igualdade, identifica relação de superioridade ou inferioridade. ">" "<"
Intervalar	Além de poder ser ordenada em uma dimensão específica há intervalos com tamanho conhecido que podem ser comparados.	Escala de Likert, estimativa de distâncias.	Além das anteriores permite operações aritméticas nas diferenças entre os números que representam os eventos.
De Razão	Além das características da escala anterior, existe ainda um ponto zero absoluto.	Salário, tamanho, tempo gasto em uma tarefa.	Além das anteriores permite operações aritméticas nos próprios códigos/números.

Fonte: Günther, 2003

As escalas intervalares e de razão permitem, além das estatísticas não paramétricas, os testes paramétricos. Todas as perguntas que oferecerem mais de duas opções de resposta podem ser convertidas em uma série binária de alternativas através do processo de codificação *dummy*. **Exemplo**: estado civil, religião, etc. Com isso, é possível uma série de operações estatísticas reservadas a escalas intervalares e de razão.

Respostas "não sabe (NS)" e "não respondeu (NR)"

As chamadas não respostas são um dos grandes problemas das perguntas em *survey*s e requerem um tratamento específico para o não conhecimento ou falta de vontade de responder à pergunta. A maioria dos questionários traz as opções "não sabe" (NS) e "não respondeu" (NR) separadas, mas, dependendo do que se busca na pergunta, elas podem vir juntas, pois não há diferença entre não saber e não querer responder. Porém, em pesquisas de intenção de voto, por exemplo, elas não podem estar reunidas, pois não saber em quem vai votar é qualitativamente distinto de ter um candidato, mas não querer responder qual o nome dele. Nesse caso, é preciso separar os dois tipos de não respostas.

Outra questão é decidir se a alternativa "Não Sabe" será ou não apresentada ao entrevistado. Existe muita controvérsia sobre se o entrevistador deve apresentar ao entrevistado a possibilidade de não resposta. Quando isso é feito, pode-se gerar o que se chama de "falso negativo". Ou seja, apesar de ter uma opinião, o entrevistado, por cansaço ou pressa, pode optar pela resposta mais rápida e cômoda: "não sei". Por outro lado, quando a alternativa não é apresentada, o efeito pode ser contrário e chama-se "falso negativo". Nesse caso, apesar de não ter opinião formada sobre o tema, e por desconhecer a possibilidade de resposta "não sei", ele pode optar, aleatoriamente, por uma das alternativas apresentadas para não passar pelo constrangimento de parecer que não tem opinião sobre o tema. Como se vê, há pontos positivos e negativos em cada uma das possibilidades. O pesquisador deve decidir em função dos objetivos da pesquisa, da pergunta específica e das possibilidades de resposta.

6.6 Organização do Questionário na Prática

Como as perguntas que integram um questionário são agrupadas em conjuntos temáticos, chamados módulos, cada módulo deve ter um código alfanumérico próprio, com início e fim. Ou seja, a numeração das questões recomeça a cada novo módulo. Assim, se for necessário incluir ou retirar uma pergunta depois que o questionário estiver pronto, ficará mais fácil readequar a numeração. **Exemplo**: para um questionário que tenha dois blocos: Opiniões sobre Democracia e Carreira Política, as questões poderiam ser identificadas da seguinte forma:

Bloco Opiniões sobre Democracia.
D01. (primeira pergunta deste bloco)
D02. (segundo pergunta deste bloco), etc.

Bloco Carreira Política
C01. (primeira pergunta deste bloco)
C02. (segunda pergunta deste bloco), etc.

A numeração das respostas deve ser feita diretamente com os códigos que serão usados para os testes estatísticos depois da coleta de informações. **Exemplo**:

A01. Qual a opinião do senhor sobre o desempenho, de maneira geral, do governador do Estado?
1. Péssimo
2. Ruim
3. Regular
4. Bom
5. Ótimo
8. NS
9. NR

Quando a pergunta permite mais de uma resposta, cada uma destas últimas irá se transformar em uma nova variável, *dummy*, que no banco será codificado como presença ou ausência 1 ou 0. **Exemplo**:

D01. Vou ler os nomes de algumas drogas e gostaria que me dissesse qual você já viu:

D010. Maconha	D011. Cocaína	D012. Heroína
D013. Haxixe	D014. Cola de Sapateiro	D015. nenhuma
D016. Outra	D017. NS	D018. NR

Aqui, o entrevistador irá marcar os códigos das drogas indicadas pelo entrevistado e, no momento da transposição de informações para o banco de dados, cada alternativa se transformará em uma variável que, se no questionário estiver marcada a resposta, receberá código 1; senão, receberá código 0. Essa questão pode ser construída na forma de tabela no questionário:

D01.Vou ler os nomes de algumas drogas e gostaria que me dissesse qual você já viu:

	SIM	NÃO
D010. Maconha		
D011. Cocaína		
D012. Heroína		
D013. Haxixe		
D014. Cola de Sapateiro		
D015. Nenhuma		

D016. OUTRA. Qual? _____

D017. NS D018. NR

Além dos códigos de perguntas e respostas, o questionário também deve "dialogar" com o entrevistador. Sempre que houver uma indicação ou recomendação ao entrevistador que não deve ser falada ao entrevistado, esta deve estar em CAIXA ALTA e entre parênteses. **Exemplo**:

Q01. Qual a sua opinião sobre os alimentos transgênicos?
1. A Favor
2. Contra
3. Não tem opinião formada (NÃO LER)
8. NS
9. NR

Ou então:

Q02. O senhor acredita na existência de vida inteligente em outros planetas?
1. Sim
2. Não (PULE PARA A Q09)
8. NS
9. NR

No *layout* do questionário, deve constar um espaço ao lado de cada pergunta ou categoria de resposta que se transformará em nova variável para ser preenchido pelo coordenador de campo. O coordenador, após receber os questionários do entrevistador, faz a revisão das questões, verifica se não está faltando nenhuma anotação e depois

preenche o campo reservado a ele com a resposta para cada pergunta. Esse campo será usado pelo digitador no momento da transferência das informações do questionário para a base de dados. Isso facilita o trabalho de digitação e evita o erro involuntário. **Exemplo**:

Q01. Sexo do entrevistado [_____] 1. Masculino 2. Feminino

O anexo 6.1, ao final deste capítulo, apresenta um exemplo de *layout* completo de questionário para *survey*. Nele, consta ao final de cada questão, uma barra, chamada de barra de *probe*, onde o entrevistador anota se o entrevistado demonstrou ter algum problema com os significados dos enunciados, das opiniões ou com as escalas. Essa barra serve para orientar os responsáveis pela pesquisa sobre possíveis necessidades de mudanças para a versão final do questionário. São muito úteis nos pré-testes.

6.7 O PRÉ-TESTE

Depois de definir a primeira versão do questionário é preciso aplicá-lo a um conjunto de indivíduos com proximidade ao público-alvo da pesquisa para testar as questões, os enunciados e as possibilidades de respostas. Esse momento é chamado de pré-teste. No anexo 6.2, ao final deste capítulo, é apresentado um modelo de relatório de pré-teste com vários tipos de avaliações (avaliação da entrevista, das questões, das escalas e do comportamento). Ao final, há um exemplo de formatação do "caderno do entrevistado". Este caderno é utilizado quando não se pretende ler todas as alternativas ao entrevistado. Ao invés disso, ele recebe um caderno com os números das perguntas e as alternativas. Cabe ao entrevistador apenas indicar em qual página o entrevistado deve olhar as alternativas disponíveis para cada pergunta. O uso do "caderno do entrevistado" é recomendado para evitar o viés de respostas cômodas. Por exemplo, o respondente sempre optar pelas primeiras alternativas lidas pelo entrevistador.

6.8 Exercícios propostos

a) Qual a diferença entre *survey* transversal e longitudinal? Dentre os longitudinais, o que caracteriza o estudo de tendências e a pesquisa por painel? Exemplifique.

b) Quais as três principais ações operacionais no momento de uma entrevista para garantia de melhores resultados?

c) Quais as cinco características de uma pergunta para garantir a validade da medição?

d) Escreva sobre as vantagens e desvantagens das perguntas abertas e fechadas em um questionário.

e) Descreva as principais características de cada um dos quatro tipos de variáveis obtidas por diferentes escalas de respostas em um questionário.

f) O que é “falso positivo” e “falso negativo” em um questionário? Quais os possíveis tratamentos a eles?

g) Para que serve e como deve ser feito o pré-teste de um questionário?

6.9 Bibliografia citada no capítulo

BINGHAM, W. V. D.; MOORE, B. V. **How to interview.** New York: Harper Collins, 1934.

FOWLER, F. J. Design and evaluation of *survey* questions. In: L. Bickman & D. J. Rog (Eds.), **Handbook of applied social research methods**. Thousand Oaks, CA: Sage, 1998, pp. 343-374.

GÜNTHER, H. **Como Elaborar um Questionário** (Série: Planejamento de Pesquisa nas Ciências Sociais, Nº 01). Brasília - DF: UnB, Laboratório de Psicologia Ambiental, 2003.

LESSLER, J. T.; KALSBEEK, W. D. **Nonsampling Errors in Surveys.** New York: John Wiley & Sons Inc, 1992.

SCHUMAN, H., & KALTON, G. Survey methods. In: G. Lindzey & E. Aronson (eds.), **Handbook of social psychology** (3rd ed.), v.1, pp. 635-697. New York:

Random House, 1985.

SUDMAN, S., & BRADBURN, N. M. **Asking questions**. San Francisco - CA: Jossey-Bass, 1982.

YAREMKO, R. K., Harari, H.; HARRISON, R. C.; LYNN, E. **Handbook of research and quantitative methods in psychology.** Hillsdale - NJ: Lawrence Erlbaum, 1986.

ANEXOS DO CAPÍTULO VI

ANEXO 6.1 – EXEMPLO DE *LAYOUT* DE QUESTIONÁRIO COM BARRA DE *PROBE*

Hora exata de início: _____ horas _____ minutos

Estou participando de uma pesquisa acadêmica e gostaria de lhe fazer algumas perguntas. É muito importante para nós sabermos o que você pensa sobre a qualidade de vida no/a seu/sua (bairro, vila).

Q1. São várias as atitudes que as pessoas podem tomar para resolver problemas no/a seu/sua (bairro, vila) ou vizinhança. No último ano você:.

Q1.a. Foi membro de associação comunitária ou de (bairro, vila) |___| Q1.a

1. Sim
2. Não
8. NS
9. NR

1. R. Enun.	3. Sig. Enun.

1. Por favor me dê alguns exemplos de associações no seu (bairro, vila)

_____

_____

Q1.b. No último ano, você procurou algum vereador para resolver algum assunto relacionado ao/a seu/sua (bairro, vila) ou vizinhança? |___| Q1.b

1. Sim
2. Não (VÁ PARA A Q1d.)
8. NS
9. NR

Q1.c. Você fez isso sozinho, através de um grupo ou associação, ou de ambas as formas? |___| Q1.c

1. Sozinho
2. Através de grupo ou associação
3. Ambas as formas
8. NS
9. NR

1. R. Enun.	2. R. Op.	3. Sig. Enun.	4. Sig. Op.

1. As pessoas no seu/sua (bairro, vila) geralmente procuram os vereadores? Se sim, como normalmente o fazem?

_____

2. Porque você acha que as pessoas no seu (bairro,vila) (não) procuram os vereadores?

_____

Q1.d. No último ano, você procurou a prefeitura para resolver algum assunto relacionado ao/a seu/sua (bairro, vila) ou vizinhança? |___| Q1.d

1. Sim
2. Não (VÁ PARA Q2)
8. NS
9. NR

Q1.e. Você fez isso sozinho(a), através de um grupo ou associação, ou de ambas as formas? |___| Q1.e

1. Sozinho
2. Através de grupo ou associação
3. Ambas as formas
8. NS
9. NR

1. Foi difícil ou fácil responder estas duas últimas perguntas?

_____

2.O que você entende por “procurar a prefeitura” ?

_____

3.As pessoas no seu/sua (bairro, vila) geralmente procuram os vereadores? Se sim, como normalmente o fazem?

_____

4. Porque você acha que as pessoas no seu (bairro,vila) (não) procuram os vereadores?

_____

Q2. Excluindo as reuniões do Orçamento Participativo, você participou de alguma reunião com representantes ou técnicos da prefeitura no último ano? |___| Q2

1. Sim
2. Não
8. NS
9. NR

1. R. Enun.	3. Sig. Enun.

Q3. No último ano, você se reuniu com vizinhos para discutir problemas do/da (bairro/vila)? |___| Q3

1. Sim
2. Não
8. NS
9. NR

1. Quem você considera “seus vizinhos” ?

_____

ATENÇÃO ENTREVISTADOR:
PERGUNTAR A PRÓXIMA QUESTÃO APENAS PARA OS QUE MORAM EM PRÉDIO OU CONDOMÍNIO, PARA OS DEMAIS VÁ PARA Q5.

Q4. (CE. p. 1) Com que frequência você vai às reuniões do seu prédio ou condomínio? Você vai frequentemente, algumas vezes, raramente ou nunca? |___| Q4

1. Frequentemente
2. Algumas vezes
3. Raramente
4. Nunca
8. NS
9. NR

1. R. Enun.	2. R. Op.	3. Sig. Enun.	4. Sig. Op.	5. Escala

1. Em que tipos de reuniões você pensou?

_____

2. Há outros tipos de reuniões no seu (prédio, condomínio)? Quais? Você comparece a elas? Se não, por quê?

_____

Q7. (CE, p.2) Agora eu vou ler uma lista de problemas que existem em alguns (bairros, vilas). Para cada um deles, diga se esse é um problema grave, é um problema ou se não é um problema no seu (a) (bairro, vila).

	Problema grave	É um problema	Não é problema	NS	NR	
						|___| Q7.a
Q7.a. Meninos de rua	1	2	3	8	9	|___| Q7.b
Q7.b. Gangues	1	2	3	8	9	|___| Q7.c
Q7.c. Tráfico de drogas	1	2	3	8	9	|___| Q7.d
Q7.d. Muita discussão ou briga entre vizinhos	1	2	3	8	9	

1. R. Enun.	2. R. Op.	3. Sig. Enun.	4. Sig. Op.	5. Escala

1. Em quem você pensa quando falamos em meninos de rua?

_____

2. Há muitos no seu bairro?

_____

3. Em que você pensa quando falamos em gangues? Por favor, dê exemplos.

_____

4. Há muitas no seu bairro?

_____

5. O que você considera como sendo discussões ou brigas entre vizinhos? Por favor, dê exemplos na sua vizinhança.

_____

Q12b. (CE, p.3) Com que frequência você usa a internet para enviar e-mail ou mensagens ou para participar de grupos de conversa? Você o faz (LER OPÇÕES) |___| Q12b

1. Diariamente ou quase todos os dias,
2. Algumas vezes por semana,
3. Algumas vezes por mês, ou
4. Raramente.
7. NSA
8. NS
9. NR

1. R. Enun.	2. R. Op.	3. Sig. Enun.	4. Sig. Op.	5. Escala

Q14. (CE, p. 5) Há muitas maneiras das pessoas sentirem que fazem parte de um grupo. Eu vou ler uma lista de grupos e gostaria que você me dissesse qual deles lhe dá o sentimento mais forte de fazer parte do grupo. E em segundo lugar? E o que menos lhe dá esse sentimento? (LER OPÇÕES)

Q14.a. O que lhe dá o sentimento mais forte de grupo:_____ |___| Q14.a

Q14.b. Em segundo lugar: _____ |___| Q14.b

Q14.c. E o que lhe dá menos o sentimento de grupo:_____ |___| Q14.c

1. Seu/sua (bairro, vila)
2. Seus amigos
3. Seus vizinhos
4. Sua igreja
5. Seus colegas de trabalho
6. Sua família

7. As associações ou outros grupos dos quais você participa ou é membro
8. NS
9. NR

1. R. Enun.	2. R. Op.	3. Sig. Enun.	4. Sig. Op.	5. Escala

1. Para você o que é " ter um sentimento forte de grupo?"

_____

2. Por que (CITAR RESPOSTA DO ENTREVISTADO PARA Q14a) lhe dá o sentimento mais forte de fazer parte de um grupo?

_____

3. Por que (CITAR RESPOSTA DO ENTREVISTADO PARA Q14c) lhe dá o sentimento menos forte de fazer parte de um grupo?

_____

Q15.a. Agora eu vou ler uma lista de afirmativas sobre vizinhos e vizinhança e gostaria de saber se você concorda ou discorda de cada uma delas. |___| Q15.a
Você tem pouco a ver com as pessoas que moram neste/nesta (bairro, vila). Você concorda ou discorda?
(SE CONCORDA, PERGUNTE): Concorda totalmente ou concorda em parte?
1. Concorda totalmente
2. Concorda em parte
(SE DISCORDA, PERGUNTE): Discorda em parte ou discorda totalmente?
3- Discorda em parte
4- Discorda totalmente
8- NS
9- NR

1. R. Enun.	2. R. Op.	3. Sig. Enun.	4. Sig. Op.	5. Escala

1. O que você considera como sendo o seu/sua (bairro, vila)?

_____

2. O que você entende por "ter pouco a ver" com as pessoas do seu/sua (bairro, vila)?

_____

Q15.b. Nós podemos confiar nas pessoas que moram neste/nesta (bairro, vila). Você concorda ou discorda? |___| Q15.b
(SE CONCORDA, PERGUNTAR): Concorda totalmente ou concorda em parte?
1. Concorda totalmente
2. Concorda em parte
(SE DISCORDA, PERGUNTAR): Discorda em parte ou discorda totalmente?
3. Discorda em parte
4. Discorda totalmente
8. NS
9. NR

1. R. Enun.	2. R. Op.	3. Sig. Enun.	4. Sig. Op.	5. Escala

1. O que você entende por confiar nas pessoas do seu/sua (bairro,vila) ?

_____

Hora exata do final do questionário: _____horas _____ Minutos

Anexo 6.2 – Exemplo de Formulário para Pré-teste de Questionário

AVALIAÇÃO DA ENTREVISTA

Questionário: !____!____!

Entrevistador: _____

1. AVALIAÇÃO DAS QUESTÕES

Numa escala de zero a 10, sendo zero Nenhuma dificuldade e 10 muita dificuldade, qual o grau de dificuldade quanto a:

Tipos de dificuldades:

1. Enunciado (compreensão, longo, ambíguo, etc.)
2. Terminologia (termos difíceis, ambíguos, etc.)
3. Compreensão do objetivo da questão
4. Problemas relativos à memória
5. Formato da questão (lista, ordenação, escalas, etc.)
6. Opções de respostas
7. Constrangimento
8. Desconhecimento do assunto
9. Entendimento dos temas
10. Não é sensível a questão de gênero
11. Outras (especificar)

1.A. Avaliação por bateria: (Classificar o grau de dificuldade de 0 a 10) e indicar as questões onde aparecem problemas e os tipos de dificuldades:

BATERIA	GRAU DE DIFICULDADE	QUESTÕES COM PROBLEMAS	TIPOS DE DIFICULDADE
B1. INTRODUÇÃO			
B2. QUALIDADE DE VIDA E CAPITAL SOCIAL			
B3. POLÍCIA E CRIMINALIDADE			

2. AVALIAÇÃO DAS ESCALAS

Numa escala de 0 a 10, sendo 0 Nenhuma dificuldade e 10 muita dificuldade, em sua opinião, qual o grau de dificuldade em relação às escalas?

Tipos de dificuldades:

1. Distinção dos pontos
2. Nº de pontos
3. Outras (especificar)

ESCALAS/QUESTÕES	GRAU DE DIFICULDADE	TIPOS DE DIFICULDADE
1. Frequentemente		
2. Algumas vezes		
3. Raramente		
4. Nunca		
1. Problema grave		
2. É um problema		
3. Não é problema		
1. Diariamente ou quase todos os dias		
2. Algumas vezes por semana		
4. Algumas vezes por mês		
5. Raramente		
6. Nunca		
1. Frequentemente		
2. Às vezes		
3. Raramente		
4. Nunca		
1. Concorda totalmente		
2. Concorda e parte		
3. Discorda e parte		
4. Discorda totalmente		
1. Muito boa		
2. Boa		
3. Ruim		
4. Péssima		
1. Sempre		
2. De vez em quando		
3. Raramente		
4. Nunca		
1. Diariamente ou quase todos os dias		
2. De quatro a seis vezes por semana		
3. De uma a três vezes por semana		
4. Uma vez por semana		
5. Raramente		
6. Nunca		
1. Diariamente, mais de 5 horas		
2. Diariamente, de 3 a 5 horas		
3. Diariamente, até 2 horas		
4. Algumas vezes por semana		
5. Muito raramente		
6. Nunca		
1. Muito seguro		
2. Seguro		
3. Inseguro		
4. Muito inseguro		
Diagramática de 1 a 10;		
“completamente insatisfeito”/		
”completamente satisfeito”		
Diagramática de 1 a 10;		
“nenhum crime”/”muitos crimes”		

1. Aumentou		
2. Diminuiu		
3. Continua o mesmo		
1. Muito democrática		
2. Razoavelmente democrática		
3. Pouco democrática		
4. Nada democrática		

3. AVALIAÇÃO DA ENTREVISTA E COMPORTAMENTO DO ENTREVISTADO

3. A. Interesse pelo assunto por bateria: Classificar o grau de interesse do entrevistado em uma escala de 0 a 10, sendo 0 Nenhum interesse e 10 muito interessado:

BATERIA	**GRAU DE INTERESSE**
B1. INTRODUÇÃO	
B2. QUALIDADE DE VIDA/CAPITAL SOCIAL	
B3. POLÍCIA E CRIME	
B4. PARTCIPAÇÃO E ASSOCIATIVISMO	
B5. VALORES	
B6. RELIGIÃO	
B7. RAÇA E COR	
B8. TRABALHO	
B9. ESTRATIFICAÇÃO	

3.B. Receptividade (gentil, hostil, etc.?): Classificar o grau de receptividade do entrevistado em uma escala de 0 a 10, sendo 0 pouco receptivo e 10 muito receptivo:

_____

3.C. Comportamento do entrevistado quanto ao tempo de duração da entrevista (impaciente? Em quais momentos da entrevista?)

_____

_____

3.D. Outros comentários sobre as atitudes dos entrevistados em relação à entrevista:

_____

_____

4. OUTROS COMENTÁRIOS

4.A. Comentários sobre a situação da entrevista (outras pessoas presentes? local da entrevista, etc.)

_____

_____

4.B. Comentários sobre a situação socioeconômica do entrevistado, sua família e vizinhança:

_____

_____

5. CADERNO DO ENTREVISTADO

Página 01	**Página 02**	**Página 03**
Nos comitês eleitorais Conversando com os familiares Na igreja Através do rádio e televisão De outra forma	Muita atenção Pouca atenção Nenhuma atenção	Discorda muito Discorda Não discorda nem concorda Concorda Concorda muito.
Página 04	**Página 05**	**Página 06**
Sim, venderia o voto com certeza. Sim, poderia vender o voto em Não venderia o voto nunca.	Sentiria muito medo Sentiria pouco medo Talvez sentisse medo Não sentiria medo.	Concorda muito Concorda pouco Discorda pouco Discorda muito.

Capítulo VII

Testes de Hipóteses de Diferenças de Médias

Um teste estatístico é uma das formas de identificar quão iguais ou diferentes são as variações de características entre indivíduos.

7.1 Conceito de probabilidade

A teoria da probabilidade é fundamental para a estatística inferencial. Uma probabilidade é um número entre zero (0) e um (1) usado para exprimir o grau de certeza sobre a ocorrência de um evento associado a um experimento aleatório. Temos probabilidade zero quando a acerto é impossível e 1 quando o acerto for seguro. Cabe ressaltar que de acordo com a própria definição, probabilidade não é certeza, portanto, o que caracteriza a teoria da probabilidade é justamente a incerteza da ocorrência de eventos estudados. Apesar dessa incerteza, a regularidade no longo prazo, com a repetição do experimento, possibilita a previsão da probabilidade de que eventos ocorram de maneira uniforme em uma experiência aleatória. Assim, aleatoriedade é fundamental para o conceito de probabilidade (BARBETA, 2003). Para ser considerada uma experiência aleatória, o evento precisa poder ser repetido indefinidamente; sermos capazes de descrever todos os resultados do experimento, embora não seja possível predizer qual ocorrerá e, ainda, obedecer a uma regularidade estatística.

Podem ser considerados experimentos aleatórios o lançamento de um dado e

a observação do número obtido na face de cima; a verificação do número de pedidos de informações numa repartição pública em um dia de trabalho ou a verificação do tempo de vida de uma lâmpada. Os modelos probabilísticos vinculam-se a certas hipóteses sobre a questão que se quer estudar e constituem-se de duas partes:

1 – dos possíveis resultados;

2 – de uma lei que mostra a probabilidade da ocorrência de cada resultado.

Por exemplo, ao lançar uma moeda e observar a face voltada para cima, os possíveis resultados são cara ou coroa. Admitindo que a moeda não seja viciada e se o lançamento for imparcial, pode-se dizer que a probabilidade de ocorrer cara é a mesma de ocorrer coroa. A fórmula da probabilidade, também chamada de fórmula de Laplace é:

$$p = \frac{n}{N}$$

Onde:

n = número de casos favoráveis;

N = total de casos possíveis.

No caso do lançamento de uma moeda, a probabilidade de dar cara é de p: 1/2 = 0,50. Se um dado for lançado, a probabilidade de acertar uma face é de p: 1/6 = 0,167. Isso ocorre sempre que for cumprido o princípio da **equiprobabilidade**, ou seja, desde que todos os casos sejam igualmente prováveis. Avançando, podemos considerar p como sendo a probabilidade de êxito e q a probabilidade de não êxito, que é a diferença entre a probabilidade de êxito e um (1). Portanto, **p + q = 1**.

As estatísticas descritivas são pontuais, ou seja, para cada dado é obtido um valor que representa seu parâmetro. Porém, na estimação, trata-se da inferência de um parâmetro através de um valor estatístico representativo (estimador). Portanto, o estimador é uma estatística que serve para estimar parâmetros. Sempre é mais seguro que a estimativa não seja pontual, mas por intervalo, o que garantirá uma probabilidade maior de acerto. Assim, dizemos que as chances de ocorrer determinado evento em um experimento está dentro de um dado intervalo de valores obtido a partir de testes estatísticos (IGNÁCIO, 2003). Para ser possível a estimação de resultados prováveis é preciso que os experimentos não sejam tendenciosos, ou seja, que sigam uma distribuição normal. Em

função de sua importância para a teoria da probabilidade, discutiremos um pouco mais o conceito de “distribuição normal” no próximo tópico.

7.2 Distribuição normal

Entende-se por Distribuição Normal aquela em que as distribuições dos valores nas variáveis são aleatórias porque apresentam um componente que varia ao acaso. **Por exemplo**, altura de alunos matriculados em determinado curso superior de uma Universidade. Ou a idade das pessoas que se encontram em determinado ônibus urbano em dado momento do dia. Quando se utiliza grandes amostras com variáveis aleatórias, podemos construir gráficos que apresentam uma característica típica, permitindo observar uma variação contínua e gradativa dos valores na população entre os dois extremos – o máximo e mínimo. Se a maioria dos valores encontra-se em torno da média, no centro, e a distribuição para os dois lados é equivalente (TOLEDO, 1991), ela é chamada de “curva de Gauss” ou Curva Normal, como demonstrado no gráfico 7.1 a seguir.

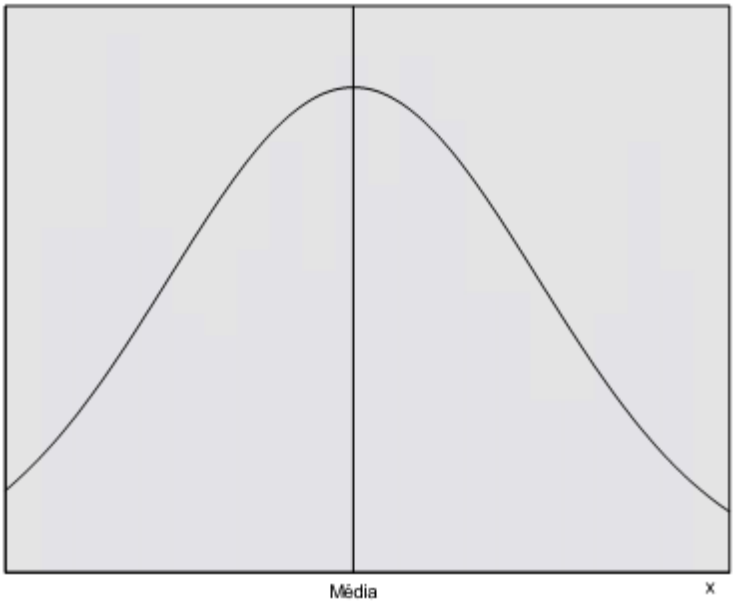


Gráfico 7.1 – Representação da Curva Normal

Quando as medidas dão origem a gráficos como anterior (Gráfico 7.1), diz-se que possuem distribuições que se aproximam de uma “normal”, com área total da curva com valor um (1). Ou seja, todos os valores reais apresentam alguma probabilidade de ocorrência. As principais características das variáveis representadas por essa imagem são as seguintes:

a) A variável pode assumir qualquer valor real;

b) O gráfico da distribuição normal – curva de Gauss – apresenta uma forma de sino simétrico em torno da média (μ);

c) Como a área total da curva de Gauss equivale a um (1), a probabilidade de ocorrer qualquer valor real é menor que um (1).

d) Por se tratar de uma curva simétrica em torno da média, valores acima da média apresentam a mesma probabilidade de ocorrência do que valores menores que ela. Em outras palavras, a probabilidade de ocorrer um valor acima da média é de 0,5 e abaixo da média é de 0,5.

A principal propriedade associada à distribuição normal é a sua capacidade para predizer as probabilidades de encontrar um valor entre dois números. Para aplicação dessa propriedade, considera-se que qualquer variável pode ter seu valor transformado a partir de um processo de padronização. A transformação de escala de uma variável qualquer para uma escala padronizada tem como vantagem a maior comparabilidade entre casos e entre variáveis.

A escala de valores padronizados é representada pela letra Z e chamada de Z-*scores*. Ou seja, todo valor pode ser transformado em seu padronizado Z-*score* aumentando a capacidade de interpretação dos comportamentos da distribuição. Para isso, utiliza-se a tabela de valor Z para distribuição normal. No exemplo representado no Gráfico 7.2 a seguir, queremos encontrar a probabilidade de ocorrência de um valor de 150 para uma distribuição que apresenta uma média de 90.

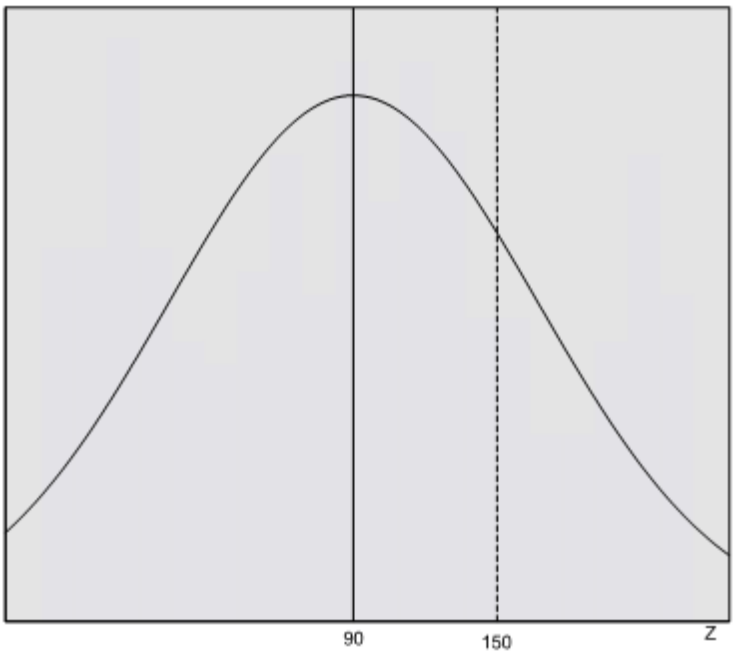


Gráfico 7.2 – Aplicação do Cálculo Distribuição Normal

Imagine que a curva indique os valores dos batimentos cardíacos por minuto encontrados de pacientes sorteados aleatoriamente. A questão formulada seria: qual a probabilidade de encontrar um valor entre a média e 150 para a distribuição normal com média de 90? Para calcular a probabilidade associada à distribuição normal, é preciso encontrar o valor padronizado (z) a partir da seguinte fórmula:

$$Z = \frac{x - \mu}{\sigma}$$

Onde:

Z = valor da distância em relação à média;

X = valor pesquisado;

μ = média;

σ = desvio padrão.

Para calcular o valor do exemplo acima, é preciso dispor do desvio padrão da distribuição, que será considerado como 36 aqui. Substituindo na fórmula, temos:

$$Z = \frac{150 - 90}{36} = 1{,}66$$

O valor Z-*score* baseado na curva normal indica quanto um determinado valor afasta-se da média em unidades de desvio-padrão. Assim, se o valor for igual à média, o Z será zero, pois não há distância entre eles. Conhecendo a distância do valor em relação à média, é possível encontrar qual a probabilidade de termos valores nesse intervalo para a população. O primeiro passo é consultar a tabela de distribuição normal (ver anexo 7.1, ao final deste capítulo).

Como o valor encontrado no nosso exemplo para Z é de 1,66, basta encontrarmos a probabilidade relativa na tabela padronizada de valores de Z (ver anexo 7.1). Lê-se a tabela da seguinte forma: na coluna da direita encontra-se a linha referente a 1,6; e busca-se o valor onde se encontra o dígito 0,06 nas colunas. Nesse caso, temos que a probabilidade de Z assumir valor entre 90 e 150 é de 0,4515, ou seja, de 45,15%. No exemplo dos batimentos cardíacos de um conjunto de indivíduos com distribuição normal em torno da média de 90 batimentos por minuto e desvio padrão de 36, pode-se dizer que há uma probabilidade de 45,15% de encontrar alguém dentro do grupo com batimentos entre 90 e 150 por minuto. Ou, em outras palavras, 45,15% dos casos encontram-se no intervalo entre 90 e 150 batimentos por minuto. Considerando tratar-se de uma distribuição normal, existem 50% dos casos acima da média. Portanto, o valor de 150 está muito próximo do limite extremo de batimentos cardíacos encontrados na distribuição.

Em um valor padronizado Z para uma distribuição normal, a média sempre será zero. Os valores positivos são aqueles acima da média e os negativos, ficam abaixo da média. No caso, o valor 1,66 significa que essa medição fica acima da média. A teoria da probabilidade permite, a partir da previsibilidade de ocorrências de valores em relação à média, estabelecer inferências de uma amostra a toda população. Para isso, são utilizadas as chamadas hipóteses estatísticas. Inferências estatísticas são importantes, pois, na maioria das vezes, o pesquisador precisa tomar uma decisão a respeito do compor-

tamento de determinada população obtendo informações apenas de uma de uma amostra, ou seja, examinando uma parte representativa do todo. Também é possível usar testes inferenciais para populações com o objetivo de medir a força das similaridades ou diferenças de variações. Para entender como isso é possível, precisamos estudar as hipóteses estatísticas nos testes inferenciais.

7.3 Testes de Hipóteses Estatísticas (NHST)

O teste de hipótese estatística é representado pela sigla para o nome em inglês NHST (*Null Hipothesis Significance Test*). Na estatística inferencial, é comum a aplicação de provas de significação ou significância estatística, também chamada de prova de contraste (TRIOLA, 1999). Elas têm a função de determinar se existem diferenças entre grupos em relação a uma variável ou dependência entre variáveis de um mesmo grupo. As provas de significância partem de uma hipótese estatística, que é submetida à comprovação através de testes. As hipóteses estatísticas são:

Hipótese Nula (H_0)

Indica que a diferença entre os valores é estatisticamente nula. Nesse caso, as diferenças devem-se às oscilações do acaso. Ela sempre é representada por uma relação matemática de igualdade:

Média 1 = Média 2; Média da Amostra = Média da População; Proporção 1 = Proporção 2.

O que equivale dizer que:

Média 1 - Média 2 = 0; Média da Amostra - Média da População = 0.

Hipótese Alternativa (H_1)

Indica que as diferenças observadas não podem ser explicadas por oscilações do acaso, ou seja, as diferenças são estatisticamente significativas. Portanto, H_1 sempre será a negação de H_0.

Média 1 ≠ Média 2; Média da Amostra ≠ Média da População.

Citando Fisher, Fonseca, Martins e Toledo (1995) lembram que todo experimento estatístico existe para dar oportunidade de rejeitar a hipótese nula. Ao afastar a possibilidade de aceitação da hipótese nula estamos validando a hipótese alternativa. Na prova de significância, a hipótese nula é submetida ao teste, pois se considera que todas as diferenças devem-se ao acaso até que se prove o contrário. Rejeitar a hipótese nula após o teste de significância representa aceitar automaticamente a hipótese alternativa. Para rejeitar a hipótese nula, é preciso que o resultado do teste fique na região crítica da zona de distribuição dos resultados, ou seja, fora do intervalo esperado.

Importante: Hipóteses estatísticas não devem ser confundidas com hipóteses científicas. Estas são sugestões de solução para o problema de pesquisa e estão baseadas em fundamentação teórica, podendo ser comprovadas por experiências. Por outro lado, a hipótese estatística é uma proposição sobre o comportamento de alguns dados e constitui um passo na análise dos resultados de pesquisas. Portanto, a aceitação de uma hipótese nula não significa a confirmação de uma hipótese científica automaticamente, pois disso dependerá da colocação do problema de pesquisa e de como as hipóteses foram formuladas. As conclusões a partir dos resultados de uma prova de significância devem seguir o raciocínio lógico.

7.4 O ERRO NOS TESTES DE HIPÓTESES

A prova de significância serve para que se tome uma decisão a respeito da hipótese nula. Essa decisão está sempre associada a um risco de erro. Existem dois erros possíveis de se cometer em uma prova de significância.

Erro de tipo 1 α (Alfa)

Quando se rejeita a hipótese nula, mesmo ela sendo verdadeira. A probabilidade de cometer um erro tipo 1 é representada por Alfa, denominado de nível de significância. Esse nível de significância é fixado *a priori* e convencionalmente costuma-se adotar os valores de 0,050 (95%) e 0,010 (99%). Esse tipo de erro ocorre principalmente quando os testes são feitos a partir de pequenas amostras.

Erro de tipo 2 β (Beta)

Quando se aceita a hipótese nula sendo ela na realidade falsa. A probabilidade de cometer o erro de tipo 2 é chamada de Beta e este risco sempre é desconhecido, assim como os parâmetros da população. Do ponto de vista estatístico, nunca se tem certeza de estar acertando ou errando ao rejeitar a hipótese nula. O que se conhece é a probabilidade de cometer um erro de tipo 1 ao rejeitar a hipótese nula sendo ela verdadeira. A esta probabilidade dá-se o nome de nível de significância do teste (FROTA, 1999). Se a probabilidade de erro for muito pequena, a decisão de rejeitar a hipótese nula está bem fundamentada. Para saber se a possibilidade de erro é ou não pequena, identifica-se o limite crítico do nível de significância, ou seja, o limite até o qual se aceita que a probabilidade de erro é pequena o suficiente para rejeitar a hipótese nula. Podemos resumir os resultados dos testes de hipóteses no quadro 7.1 abaixo.

Quadro 7.1 – Possíveis resultados do teste de independência de médias

	Assume H_0	**Rejeita H_0**
H_0 verdadeiro	Decisão correta	Erro de tipo I (alarme falso)
H_0 falso	Erro de tipo II (não reconhecido)	Decisão correta

Os maiores problemas dos testes de hipóteses são causados por:

a) Viés por tamanho da amostra. Considerando que o teste usa coeficientes de diferença como o (t) e o cálculo de t é t = $\frac{\beta}{SE}$ e erro padrão (SE) é obtido a partir de SE = $\sqrt{\frac{SS.residual}{(N-2)}}$. Se o N for muito pequeno, o valor de SE será grande, o que reduzirá o coeficiente t ao final. Com isso, teremos uma redução das diferenças não por conta da independência entre as variáveis, mas como resultado de um N baixo. Os melhores resultados são obtidos em testes aplicados a grandes amostras.

b) Outro problema é a arbitrariedade no estabelecimento do limite crítico. Usado internacionalmente como limite de 0,050 para definir se deve ou não rejeitar uma hipótese nula, esse limite é arbitrário. O problema cresce quando os coeficientes obtidos ficam muito próximos do limite crítico.

7.5 Nível de Significância no Teste de Hipóteses

O nível de significância é o risco de erro que se está disposto a assumir em caso de se rejeitar a hipótese nula. Ele se dá pelo estabelecimento do valor da probabilidade tolerável de incorrer no erro de rejeitar a hipótese nula sendo ela verdadeira.

Os valores mais comuns adotados para níveis de significância são 0,05 e 0,01, ou seja, aceita 5% ou 1% de erros possíveis quando rejeitar a hipótese nula. O valor a ser escolhido dependerá dos objetivos de cada pesquisa. Como demonstra o gráfico 7.3 a seguir, indicando a área de rejeição de valores ao nível de significância de 0,05.

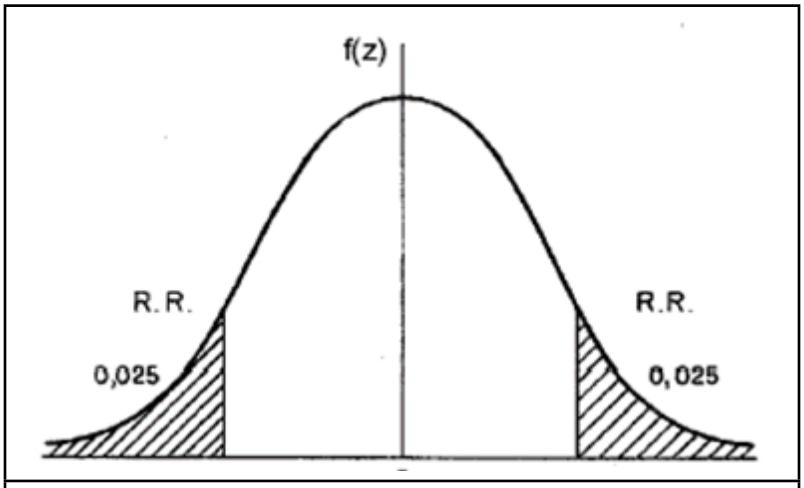


Gráfico 7.3 – Exemplo de área de rejeição da hipótese nula

Se o pesquisador opta por rejeitar a hipótese nula, estando os valores fora da área crítica ele terá chances maiores que as aceitáveis em cometer um erro do tipo 1, portanto, nesse caso não se deve rejeitar a hipótese nula, aceitando que as diferenças são consequência de oscilações do acaso. A seguir, serão apresentadas algumas provas estatísticas para verificar o risco de erro ao se rejeitar ou não a hipótese nula.

7.6 Prova t de *Student*

Os testes t existem para comparar duas médias quaisquer, permitindo concluir se elas são originárias da mesma população ou se as diferenças são estatisticamente

significativas. É um teste de diferenças de médias que considera o p-*value* para estabelecer o limite crítico de aceitação ou rejeição da hipótese nula. Foi desenvolvido por William Gossett, em 1908, para monitorar a qualidade da cerveja produzida pela cervejaria *Guiness*, em Dublin, na Irlanda. No entanto, o teste foi publicado em periódico científico sob o pseudônimo de "t de *Student*" porque os donos da cervejaria convenceram Gosset de que aquilo era um segredo industrial que não poderia ser divulgado. Com o teste t, o controle de qualidade da *Guiness* era mais preciso e com menores custos do que o das cervejarias concorrentes (BUCHAFT; KELLNER, 1999).

A prova t de *Student* é usada para comparar médias entre uma variável contínua e outra categórica dicotômica, sendo que a hipótese nula sempre será que "não existem diferenças significativas das médias da variável contínua para os dois grupos da variável categórica analisada". Assim, as duas amostras podem ser consideradas como procedentes da mesma população e as diferenças observadas entre eles são ocasionais. Em comparação com a distribuição normal, seu emprego é específico para o que se denomina de amostras pequenas, que variam entre 30 e 120 casos. A partir de n = 120, considera-se que há uma superposição da distribuição t e a normal.

Na prova, t é o valor da tabela para t de *Student* em "n" graus de liberdade e nível de significância Alfa. Ela é muito empregada em estatísticas inferenciais para o estabelecimento de intervalos de confiança e em testes de hipóteses, sendo útil para verificação de diferenças entre médias de dois grupos. A partir desses testes, é possível identificar a existência ou não de independência de médias, o tamanho do efeito de uma amostra sobre a outra e, ainda, o intervalo de confiança das médias de amostras para rejeição da hipótese nula de independência de médias. Existem quatro tipos de teste t:

- Teste z e t para uma amostra – quando se compara a média de uma amostra com a média da população, desde que o desvio padrão da população seja conhecido;

- Teste t para amostras dependentes ou pareadas – quando se avalia a diferença entre duas amostras relacionadas ou com casos pareados;

- Teste t para amostras independentes – para quando a avaliação é feita a partir de duas amostras independentes. O quadro 7.2 a seguir sumariza os pressupostos de cada um dos testes:

Quadro 7.2 – Tipos de testes t de *Student*

Teste	Freq. Observada	Freq. Esperada	Erro Padrão (SE)
Z ou T para uma amostra	Média da amostra	Média da população	SE da média
T para amostras independentes	Diferença entre médias de duas amostras	Diferença entre médias de duas populações	SE da diferença entre as médias
T para amostras dependentes/pareadas	Média das diferenças dos valores da amostra	Média das diferenças dos valores da população	SE da diferença de médias

A partir daqui, são discutidos separadamente cada um dos testes apresentados no quadro 7.2 acima.

Teste Z

O teste Z é um teste realizado para verificar se a distribuição de casos em uma única variável fica próxima da distribuição esperada ou se ela apresenta concentração em alguns valores. A fórmula é:

$$z = \frac{\mu}{SE}$$

Onde:

μ = média das observações;

SE = Erro Padrão.

O Erro Padrão (SE) é obtido a partir de:

$$SE = \frac{SD}{\sqrt{N}}$$

Onde:

SD = Desvio Padrão;

N = tamanho da amostra.

Percebe-se que o Erro Padrão é função do tamanho da amostra. Quando maior a amostra, menor tende a ser o erro padrão. O teste Z é pouco utilizado na área de ciências sociais, pois ele depende que Desvio Padrão (SD) da população seja conhecido, o que nem sempre é possível. Para saber se o valor de Z é estatisticamente significativo, calcula-se o p-*value* a fim de identificar o limite crítico. Esse limite depende do tipo de teste, se direcional ou não direcional e dos graus de liberdade da amostra. Os graus de liberdade são obtidos com N - 1.

Vejamos como exemplo a distribuição de notas de uma turma da disciplina de métodos quantitativos. São 23 alunos (N=23) e as notas variam de 5 a 50 pontos. As notas são as seguintes:

30	45	05	45	35	05	30	10	45	35	50	30	50	20	05	40	45	35	20	35	40	50	45

O cálculo é:

$$Z = \frac{\mu}{SE} = \frac{32{,}60}{3{,}12} = 10{,}43$$

O valor de Z é 10,43. Agora, para saber se podemos ou não rejeitar a hipótese nula é preciso encontrar o limite crítico na tabela padronizada (ver anexo 7.1). Para isso, precisamos usar os graus de liberdade (gl) que é N-1 = 23-1 = DF=22. Precisamos saber se a distribuição é unicaudal ou bicaudal. Como a média é 32,6 e temos valores acima e abaixo dela, não restam dúvidas que se trata de distribuição bicaudal. Por fim, precisamos determinar o Intervalo de Confiança. Usaremos o padrão que é de 95%. No *RCommander* o teste Z equivale ao teste t para uma amostra. Os resultados são os que seguem:

```
Linha de Comando – teste t para uma amostra (z)

Rcmdr>  with(Pasta1, (t.test(notas, alternative='two.sided', mu=0.0,
Rcmdr+ conf.level=.95)))
```

```
Resultados – teste t para uma amostra (z)

t = 10.433, df = 22, p-value = 5.555e-10
Intervalo de confiança de 95%: 26.12648 39.09091
Média das notas: 32.6087
```

Como se percebe nos resultados acima, a estatística t =10,433 é a mesma que a obtida pela fórmula anterior. Além disso, ele fornece o número de graus de liberdade (df=22) e o p-*value*, que no caso está em notação científica para indicar dez casas após a vírgula com zero. O intervalo de confiança de 95% é outro indicador da significância estatística do teste. Se o intervalo entre os valores máximo e mínimo não passa por zero, podemos, então, rejeitar a hipótese nula.

A verificação da significância do resultado se dá pelos graus de liberdade e intervalo de confiança que identificam o valor do limite crítico na tabela dos valores padronizados para o teste t. Essa tabela encontra-se no anexo 7.2 ao final do capítulo. Nela, a coluna de verificação do limite crítico para distribuição bicaudal e 95% de intervalo de confiança é sexta da esquerda para a direita. Como temos 22 graus de liberdade, olhamos para o valor da casa da linha 22 para a sexta coluna. O limite crítico é de 2,09. Como o coeficiente Z encontrado fica muito acima do limite crítico, podemos rejeitar a hipótese nula e aceitar que as diferenças de notas entre os alunos são estatisticamente significativas, não sendo resultado de variações ao acaso. A tabela de valores padronizados para o teste t (Anexo 7.2) vale para todos os demais tipos de teste t que serão apresentados a seguir.

Uma aplicação comum do teste Z-score na ciência política é a utilização do valor padronizado para identificar que categorias de casos apresentam-se mais distantes da média da distribuição. Aqui, o objetivo é verificar se existem categorias concentrando mais casos que o esperado ou não. Para calcular o Z-score de cada categoria, usa-se a fórmula:

$$Z - score = \frac{x - \mu}{DP}$$

Onde:

x = valor de cada categoria;

μ = média da distribuição;

DP = desvio padrão da distribuição.

Vamos aplicar o cálculo dos valores padronizados para o seguinte caso: dis-

tribuições das médias de IDH (Índice de Desenvolvimento Humano) em 2012 por continente. Nosso objetivo é, a partir do valor padronizado Z-*score*, identificar que continentes apresentam valores acima e abaixo da média. A média dos IDHs dos países agregados por continente em 2012 ficou em 0,687, com desvio padrão de 0,127. Com essas informações e com os valores por continente, é possível calcular o Z-*score* de cada um deles, da seguinte forma:

$$Z - score\ África = \frac{x - \mu}{DP} = \frac{0{,}486 - 0{,}687}{0{,}127} = -1{,}58$$

$$Z - score\ América = \frac{x - \mu}{DP} = \frac{0{,}733 - 0{,}687}{0{,}127} = +0{,}36$$

$$Z - score\ Ásia = \frac{x - \mu}{DP} = \frac{0{,}686 - 0{,}687}{0{,}127} = -0{,}01$$

$$Z - score\ Europa = \frac{x - \mu}{DP} = \frac{0{,}835 - 0{,}687}{0{,}127} = +1{,}17$$

$$Z - score\ Oceania = \frac{x - \mu}{DP} = \frac{0{,}696 - 0{,}687}{0{,}127} = +0{,}07$$

Os resultados estão sumarizados na tabela a seguir:

Tabela 7.1 – Valores originais e Z-score para média de IDH/2012 por continente

Continente	IDH_12	Z-score
África	0,486	-1,58
América	0,733	0,36
Ásia	0,686	-0,01
Europa	0,835	1,17
Oceania	0,696	0,07

A tabela 7.1 mostra que a África apresenta IDH bem abaixo da média, enquanto a Europa e a América ficam acima (valores positivos). Já a Ásia e a Oceania apresentam valores de z-*score* muito próximos da média, ficando em torno de zero, tanto negativo quanto positivo.

Identificação de outliers com Z-score

A partir do Z-score também é possível identificar casos desviantes, muito fora da norma, chamados de *outliers*. Na maioria dos casos, quando os *outliers* são desconsiderados, eles exercem forte influência sobre os testes, podendo alterar significativamente os resultados. Para identificar o valor limite para *outliers* nas distribuições de Z-score basta conhecer o número de casos. O quadro a seguir sumariza as distribuições:

Quadro 7.3 – Limites para estabelecimento de *outliers* para estatística Z

Limites para considerar *outliers*	Número de casos
-2,5 a +2,5	Baixo (menos que 50 casos)
-3,0 a +3,0	Grande (entre 50 e 1.000 casos)
-3,3 a +3,3	Muito Grande (acima de 1.000 casos)

Como podemos perceber, para o quadro do exemplo acima não existem *outliers*, pois se trata de número pequeno de casos (apenas cinco categorias) e os valores de Z-*score* ficaram todos entre os limites de *outliers* (-2,5 e +2,5). Como o Z-*score* é aplicado a casos onde se conhece o desvio padrão da população, não faz sentido verificar o intervalo de confiança para rejeição da hipótese nula. Aqui, estamos falando para toda a população. Os testes a seguir são indicados para quando não se conhece o desvio padrão da população e, portanto, devem ser acrescidos das verificações dos intervalos de confiança.

Teste t para amostras independentes

O teste de médias para amostras independentes serve para comparar os valores de amostras extraídas de universos distintos. **Por exemplo**, de valores entre homens X mulheres; indivíduos do grupo de controle X do experimento; pacientes X indivíduos saudáveis, etc. Inclui os mesmos resultados que os testes para amostras pareadas que serão apresentados a seguir:

- t-*value*;
- p-*value*;
- d de Cohen (para medir o tamanho do efeito);
- intervalo de confiança.

Aqui, a frequência esperada (F_e) é diferente de zero, dado que são amostras distintas. A fórmula para o cálculo é:

$$t = \frac{(\mu_1 - \mu_2)}{SE_m}$$

Onde:

μ_1 = média da amostra 1;

μ_2 = média da amostra 2;

SE_m = diferença do erro padrão médio. $\frac{(SE_1 + SE_2)}{2}$

O “d de Cohen” é um coeficiente que usa a média para identificar o efeito prático da significância estatística da diferença entre duas médias. Ele equivale a um teste de correlação de Pearson, que usa as variâncias, porém, o coeficiente “d de Cohen” é calculado a partir das diferenças das médias. Para o tamanho do efeito, a fórmula do “d” é a que segue:

$$d = \frac{(\mu_1 - \mu_2)}{SD_m}$$

Onde:

SD_m = $\frac{((N_1 \, x \, SD_1) + (N_2 \, x \, SD_2))}{N_1 + N_2 - 2}$

Convenciona-se usa os seguintes limites para a interpretação dos resultados do teste de Cohen:

Quadro 7.4 – Tamanho do efeito pelo coeficiente d de Cohen

Tamanho do efeito	d
Muito pequeno	< 0,20
Pequeno	0,20 a 0,39
Médio	0,40 a 0,70
Grande	> 0,70

O intervalo de confiança também é calculado por:

Limite Superior: dif.μ + t.(SE)

Limite Inferior: dif.μ - t.(SE)

Nos casos em que duas amostras foram observadas com dados independentes e os indivíduos escolhidos ao acaso entre os dois grupos, supõe-se que a associação das variações deles seja nula e que, portanto, haja independência estatística. Para poder aplicar essa prova, é preciso que exista:

1 – homogeneidade de variâncias;

2 – os dois grupos devem proceder de uma população que segue a lei da distribuição normal;

3 – a variável dependente deve ser medida em escala, intervalo ou razão.

Exemplo: comparar as médias dos alunos do banco de dados "B_Exerc_V1" com o estado civil deles, se solteiro ou não solteiro. Para isso, usaremos o pacote *RCommander* para obter as estatísticas. A hipótese é que os alunos solteiros, por terem menos atividades paralelas à faculdade, apresentem médias estatisticamente distintas das obtidas pelos não solteiros. O teste t para amostras independentes é indicado para verificar a independência das variações, pois temos a variável dependente contínua (média) e a explicativa é binária (solteiro ou não solteiro). Se o resultado do teste ficar acima do limite crítico do nível de significância, não se pode rejeitar a hipótese nula e, portanto, devemos assumir que não há diferenças estatisticamente significativas na avaliação dos docentes por homens e mulheres. Se o nível de significância ficar abaixo do limite crítico, significa que há diferenças estatisticamente significativas. Os resultados são os que seguem:

```
Linha de Comando – teste t para amostras independentes

Rcmdr>  t.test(media~estcivil, alternative='two.sided', conf.
level=.95,
Rcmdr+    var.equal=FALSE, data=B_Exerc_V1)
```

```
Resultados – teste t para amostras independentes

Variáveis: media por est_civil
t = -0.98505, df = 39.733, p-value = 0.3306
Intervalo de confiança para 95%: -7.359038  2.536877
Estimativas da amostra:
Média do grupo não solteiro       Média do grupo solteiro
                  72.64706                     75.05814

SE média do grupo não solteiro    SE média do grupo solteiro
                    2.169                       0.691
```

O pacote estatístico *RCommander* retorna os resultados acima a partir da linha de comando apresentada na primeira parte do quadro. Dos resultados, o que mais importa é o valor de t=-0,985 e um p-*value*=0,330. Como estamos usando o limite crítico de 0,050 e o p-*value* ficou acima dele, se rejeitássemos a hipótese nula estaríamos correndo um risco de errar acima do tolerado. Portanto, não podemos rejeitar a hipótese nula e temos que assumir que as diferenças de médias entre os dois grupos são frutos do acaso, ou seja, do ponto de vista estatístico não há diferenças de médias entre alunos solteiros e não solteiros.

Outra forma de verificar a impossibilidade de rejeição da hipótese de independência das variações é pelo intervalo de confiança. Os limites do intervalo para 95% são -7,359 e +2,536, ou seja, o intervalo passa por zero, pois um valor é negativo e outro é positivo. Como o intervalo passa por zero, é possível que a diferença entre as médias seja zero, portanto, não se pode rejeitar a hipótese nula. Rejeita-se quando os dois limites são negativos ou são positivos, portanto, sem que o intervalo passe por zero. Assim, o resultado é que as médias dos solteiros (72,64) e dos não solteiros (75,05) não são estatisticamente distintas ou independentes.

Podemos aplicar os resultados das estatísticas sumarizadas na fórmula do teste t para amostras independentes e obteremos resultados um pouco distintos. Isso porque o teste realizado pelo *RCommander* considerou que as variações das duas variáveis não são as mesmas. A fórmula abaixo desconsidera a independência de variações. Leva-se em conta apenas as médias e os erros padrões (SE) de cada grupo:

$$t = \frac{(\mu_1 - \mu_2)}{SE_m} = \frac{-2{,}41}{1{,}72} = -1{,}40$$

Ao desconsiderar que as variâncias são distintas, o resultado fica menos preciso, ou seja, mais próximo de ser estatisticamente significativo (t = -1,40). No teste realizado pelo *RCommander,* como pode-se perceber na linha de comando ("var. equal=FALSE"), não se assumem mesmas variâncias. Já no cálculo com a fórmula acima, essa informação é desconsiderada, tornando-o menos robusto. Com o valor de t, recorremos à tabela de valores padronizados para identificar se ele está acima ou

abaixo do valor para o n=60, que é o número de casos no banco de dados. O anexo 7.2 indica que para um n=60, distribuição bicaudal e 95% de intervalo de confiança, o valor padronizado é de 2,00. Como a estatística obtida é de -1,40, ficando abaixo, portanto, não podemos rejeitar a hipótese nula.

Perceba que o sinal negativo deve-se apenas ao fato de a média dos não solteiros ter sido incluída como segundo elemento na subtração. Se fosse invertido, o resultado seria o mesmo, porém, com sinal positivo. O importante aqui é perceber a diferença nos resultados quando não se consideram variâncias iguais (no teste do *RCommander*) e quando se consideram mesmas variâncias (na aplicação da fórmula acima). No segundo caso o resultado quase ficou acima do limite crítico, invertendo a interpretação dos coeficientes obtidos no primeiro teste.

Como os resultados do teste não indicaram diferenças estatisticamente significativas, não faz sentido testar o tamanho do efeito com o coeficiente “d de Cohen”. Se existissem diferenças estatisticamente significativas, o teste “d de Cohen” seria o próximo passo da análise, pois ele permitiria indicar a intensidade do efeito da mudança de categoria na variável categórica, no caso estado civil, para a mudança nas médias das notas, a variável contínua.

O teste a seguir é aplicado para quando as amostras não são independentes. Ou seja, as informações são extraídas dos mesmos indivíduos.

Teste t para amostras relacionadas (Comparação de dados pareados)

Nesse caso, o objetivo é comparar dados de um mesmo grupo em dois momentos distintos, antes e depois de determinada ocorrência de interesse. Um exemplo seria comparar as pontuações de um grupo de indivíduos em um pré-teste e depois de uma experiência educativa com o desempenho deles em um pós-teste. Outro exemplo seria comparar os pesos de um grupo de indivíduos que se submetem a determinada dieta antes e depois da experiência. Neste teste, o objetivo sempre é saber se haverá diferença significativa entre os resultados nos dois momentos distintos.

As principais características dos testes t entre dados independentes e relacionados são que:

1– com os dados independentes, comparam-se grupos;

2– com os dados pareados, comparam-se variáveis.

É o teste para verificar se determinado fenômeno interveniente tem efeitos sobre as médias obtidas dos mesmos indivíduos. Para isso, obtêm-se as médias antes e depois do fenômeno a fim de verificar se as diferenças nos valores são estatisticamente significativas ou não. O teste permite que o montante de variação dos mesmos indivíduos seja testado. A análise inclui:

- t-*value*;

- p-*value*;

- d de Cohen (para medir o tamanho do efeito);

- intervalo de confiança.

O cálculo do valor de t é feito a partir da fórmula:

$$t = \sum \frac{(\mu_1 - \mu_2)}{SE} \; como\ a\ hipótese\ nula\ é\ a\ de\ que\ não\ há\ diferenças,$$

$$então, Fe = 0. Assim, a\ fórmula\ fica: t = \frac{dif\mu}{SE_m}$$

Onde:

difμ = diferença das médias da distribuição;

SE_m = Erro Padrão Médio.

Para calcular o intervalo de confiança, é preciso identificar o limite superior e o inferior com a seguinte fórmula:

Limite Superior: dif.μ + t.(SE_m)

Limite Inferior: dif.μ - t.(SE_m)

Se o intervalo não passar por zero, significa que a hipótese nula pode ser rejeitada e os valores das diferenças considerados significativos.

Vamos aplicar o teste com amostras pareadas para o seguinte caso: tempo na tribuna da Assembleia Legislativa do Paraná (ALEP) para discursos realizados por deputados estaduais do antes e depois da implantação do sistema de transmissão ao vivo das sessões pela televisão. A hipótese é que os parlamentares fiquem mais tempo na tribuna após o início da transmissão das sessões ao vivo pela emissora de televisão própria da ALEP. São usados os totais de tempo (em minutos) ocupados em discursos

realizados pelos deputados no ano anterior à implantação da tevê (2007) legislativa e no primeiro ano após o início de seu funcionamento (2008). Os dados são:

Tabela 7.2 – Diferença de tempo de discursos entre 2007 e 2008

		Média	N	Desvio padrão	Erro médio padrão
Variáveis	Tempo em discursos (2007)	228,13	43	269,58	41,11
	Tempo em discursos (2008)	275	43	288,3	43,96

Aplicando a fórmula para encontrar o coeficiente t:

$$t = \frac{dif\mu}{SE_m} = \frac{-46{,}86}{20{,}23} = -2{,}316$$

Já sabemos que t=-2,316. Resta saber se esse valor é estatisticamente significativo para as diferença ou não. Podemos recorrer aos valores padronizados da tabela 7.2. Neles, perceberemos que para intervalo de confiança de 95% e n=40 (o mais próximo de 43), o limite crítico é 2,02. Como a estatística fica acima do limite crítico, já podemos rejeitar a hipótese nula e dizer que as diferenças de médias não são aleatórias. Porém, também podemos aplicar as fórmulas dos limites superior e inferior. Se o intervalo passar por zero, não podemos rejeitar a hipótese nula. Vejamos:

Limite Superior: dif.μ + t.(SE_m) = -46,86 + (-2,316*20,23) = - 93,712

Limite Inferior: dif.μ - t.(SE_m) = -46,86 - (-2,316*20,23) = - 0,007

Embora o valor do limite superior fique muito próximo de zero, ainda assim ele é negativo, portanto, podemos aceitar que a estatística t é significativa e rejeitar a hipótese nula porque o intervalo dos limites não passa por zero. A afirmação é que há diferenças estatisticamente significativas entre as médias de tempo de discurso dos parlamentares antes e depois do início das transmissões das sessões plenárias.

Como aqui o teste t para amostras pareadas apresentou coeficiente estatisticamente significativo, permitindo a rejeição da hipótese nula, podemos afirmar que o risco de erro é baixo ao considerar que as médias são distintas. Assim, o próximo passo é medir o tamanho do efeito da variável categórica "transmissões das sessões plenárias" sobre as médias da variável contínua "duração dos discursos dos parlamentares". A fórmula é a que segue:

$$d = \frac{\mu_m}{SD_m} = \frac{46{,}87}{285{,}58} = 0{,}164$$

Considerando os valores apresentados anteriormente no quadro 7.4, podemos interpretar o resultado como muito fraco o tamanho do efeito das transmissões ao vivo das sessões da assembleia legislativa para a diferença na duração dos discursos dos parlamentares, pois fica abaixo de 0,20.

Todos os testes t de diferenças de médias partem do pressuposto de homogeneidade da variância. No caso de uma amostra, espera-se a homogeneidade de distribuição dos casos nessa amostra. Quando estão sendo comparadas duas amostras (pareadas ou independentes), espera-se uma homogeneidade das variâncias conjuntas. A inserção do Desvio Padrão Médio (SDm) no cálculo só faz sentido se a hipótese de homogeneidade de variâncias não for violada. Caso o pressuposto da homogeneidade de variâncias seja quebrado, é possível incorrer em erro de tipo 1 quando se rejeita a hipótese nula.

Para testar a violação do pressuposto da homogeneidade de variância, usa-se o teste de "*Levene*". Caso o resultado seja estatisticamente significativo, com p-*value* abaixo de 0,050 para 95% de intervalo de confiança, interpretam-se os resultados do teste t, pois há homogeneidade de variância. Caso o pressuposto tenha sido quebrado, é preciso ajustar os graus de liberdade (df) e o p-*value* através do procedimento de *Welch*. Se ainda assim, não for alcançada a homogeneidade de variância, deve-se optar por um teste não paramétrico em substituição ao teste t.

7.7 Análise de Variância (ANOVA)

A análise de variância é uma generalização da prova t de *Student*, sendo utilizada para comparar as médias de mais de dois grupos simultaneamente, ou seja, mede a independência de variações entre uma variável contínua e uma categórica com mais de duas categorias. Sua forma abreviada é chamada de ANOVA (*Analysis of Variance*). A hipótese nula é a de que não existem diferenças significativas entre as médias observadas para cada categoria. O principal objetivo da ANOVA é comparar as diferenças entre os grupos com as diferenças entre os indivíduos dentro dos grupos (inter e intra-variabilidade).

A Análise de Variância (ANOVA) é indicada para quando a variável independente é categórica e a dependente é contínua, sendo uma das formas mais comuns para análise de dados randomizados em experimentos controlados. Sua utilidade cresce quando são gerados mais que dois grupos com médias diferentes para serem testadas. Se houver apenas dois grupos de médias, o indicado é usar um dos testes t. O teste ANOVA mais comum é entre grupos, mas também existe a ANOVA para medidas repetidas. Qualquer uma delas é um teste NHST de diferença de médias. Aqui, a estatística medida é a F de Fisher (F-*ratio*).

$$F = \frac{variância\ entre\ grupos}{variância\ dentro\ dos\ grupos}$$

Assim como no caso do teste t, também existe uma família de distribuições F. Essas distribuições dependem do número de casos inseridos na análise e do número de grupos da distribuição. Porém, uma diferença em relação ao teste t é que no caso da estatística F a curva não apresenta a forma de Curva Normal, pois não existem resultados negativos de F, ou seja, as distribuições sempre são unicaudais. A fórmula é:

$$F = \frac{variância\ entre\ grupos}{variância\ dentro\ dos\ grupos} = \frac{MS_{entre}}{MS_{dentro}} = \frac{MS_a}{MS_{s/a}}$$

Onde:

$$MS_a = \frac{SS_a}{DF_a}$$

$$MS_{s/a} = \frac{SS_{s/a}}{DF_{s/a}}$$

Onde:

$$SS_a = Nx\sum(Yi - Yt)^2$$

Yi : médias dos grupos.
Yt : média geral.
DF_a : a - 1
$DF_{s/a}$: ax(n-1)
DF_{total} : N - 1

Para identificar o tamanho do efeito calcula-se o $r^2 = n^2$ (chamado de ETA^2) com:

$$n^2 = \frac{SS_a}{SS_{total}}$$

Todos os cálculos da ANOVA partem dos seguintes pressupostos:

- a variável dependente é contínua (intervalar ou de razão);

- a variável dependente apresenta distribuição normal;

- existe homogeneidade da variância;

Para verificar que o pressuposto da homogeneidade não foi quebrado, analisa-se o resultado teste de *Levene*. Se for estatisticamente significativo, significa que a homogeneidade da variância foi violada. Caso não tenha havido violação de nenhum pressuposto e a estatística F mostre-se significativa, isso indica que pelos menos a média de um dos grupos é estatisticamente diferente da dos demais. Não é possível saber quantos são e qual a magnitude das diferenças de médias entre cada um dos grupos. Para saber isso, é necessário usar um dos testes *post hoc* para fazer comparações entre pares com termo de erro restrito (*pairwise*).

Para que se possa aplicar a ANOVA devem ser cumpridos os supostos paramé-

tricos. Nesses casos, sabe-se que a melhor forma de comparar um conjunto de média é através da análise de suas variâncias. Com a ANOVA, a variabilidade total fica dividida em duas componentes, a que se deve às diferenças entre grupos e a que se deve a diferenças individuais dentro de cada grupo. Para exemplificar a realização de um teste ANOVA com um fator apenas com o *RCommander,* aplicaremos ao banco de dados B_Exerc_V1 como fator "tempo_ocupação" e a variável a ser testada é o número de faltas do aluno.

A pergunta aqui é se existem diferenças estatisticamente significativas entre as médias das faltas dos alunos em relação ao tempo que ele dedica a outras atividades que não a faculdade diariamente. A hipótese de pesquisa a ser testada é que quanto maior a ocupação do aluno, mais cresce a média de faltas dele na faculdade. Se as faltas apresentarem diferenças estatisticamente significativas, o coeficiente F ficará acima do limite crítico para os graus de liberdade e intervalo de confiança de 95% e o p-*value* estará abaixo de 0,050. Os resultados do teste são os que seguem:

```
Linha de comando – Teste ANOVA

Rcmdr>  ANOVAModel.9 <- aov(faltas ~ tempocup, data=B_Exerc_V1)

Rcmdr>  with(B_Exerc_V1, numSummary(faltas, groups=tempocup,
statistics=c("mean", "sd")))
```

```
Resultados – Teste ANOVA

Rcmdr>  summary(ANOVAModel.9)
            Df Sum Sq Mean Sq F value Pr(>F)
tempocup     3   37.2  12.417   1.696  0.178
Residuals   56  410.0   7.321

                    mean       sd data:n
acima 6 horas  4.333333 2.746765     42
de 2 a 4 horas 6.666667 3.055050      3
de 4 a 6 horas 3.000000 2.489980     11
não tem outra  5.000000 2.581989      4
```

Os resultados acima mostram que a estatística F é de 1,696, bem abaixo do limite crítico da significância estatística, e que o p-*value* = 0,178 está acima de 0,050, portanto, não é possível afirmar que as diferenças de faltas dos alunos dos quatro grupos de tempo de ocupação são estatisticamente distintas entre si. Se olharmos a tabela de médias na última parte do resultado, perceberemos que quem não tem ocupação apresenta média

de 5 faltas, quem trabalha de 2 a 4 horas, com média de 6,6 faltas e depois, conforme cresce o tempo de ocupação, cai a média de faltas. O desvio padrão da média de cada categoria gira em torno de 2,581 e 3,055, o que indica uma similaridade em variações em torno da média em todas as categorias de tempo de ocupação. Assim, a hipótese inicial de que a ocupação do aluno explicaria as faltas não se aplicada ao grupo de estudantes que integra a pesquisa analisada aqui. A ANOVA mais comum é a análise de variância unidirecional, ou seja, com uma única variável independente e uma variável dependente, como a apresentada acima. Mas existem outras possibilidades, como a ANOVA bidirecional, análise fatorial da variância e a análise multivariada da variância. A seguir, serão apresentados os testes de comparações múltiplas e depois a análise fatorial.

Comparações múltiplas com ANOVA

O teste de ANOVA permite concluir se a hipótese nula é aceitável ou não. No caso de rejeitá-la, sabe-se que existe pelo menos uma diferença significativa entre algum dos conjuntos de médias. Porém, ela não permite saber que conjunto ou quais conjuntos de médias encontra-se a diferença estatisticamente significativa. Para conhecer esse detalhe é preciso recorrer a um dos vários testes de comparações múltiplas, também conhecidas como testes *post hoc*. Os mais comuns são o teste de *Scheffé, Tukey, Duncan* e *SNK*.

Os testes *post hoc* mostram quais pares apresentam comparações de médias com diferenças estatisticamente significativas. Os testes *post hoc* evitam aumentar a probabilidade de cometer o erro de tipo 1 ao rejeitar a hipótese nula. O mais comum é o teste Tukey, que permite múltiplas comparações. Existem outros testes *post hoc* que consideram os graus de liberdade e outros procedimentos. Os mais liberais não fazem ajustes do modelo. Os mais conservadores fazem ajustes e tornam mais difícil a rejeição da hipótese nula. Um exemplo de teste *post hoc* conservador é o Bonferroni. É bom lembrar que só faz sentido um teste *post hoc* depois que a estatística F mostra-se significativa, caso contrário, todas as comparações entre pares ficarão abaixo do limite crítico (p-*value*).

Como exemplo de teste ANOVA com *post hoc* Tukey no *RCommander*[1], usa-

1 Para realizar os testes com fatores, no caso da ANOVA e do teste ***post hoc*** Tukey no *RCommander* é preciso transformar a variável numérica categórica em fator no próprio banco de dados. O adendo 1 explica como fazer isso no pacote *RCommander.*

remos o banco de dados B_Exerc_V1 para testar as seguintes variáveis: como fator, a variável “turma_num”. São três turmas, logo, três categorias, o que impede a utilização do teste t. Como variável contínua, as notas dos alunos de cada turma. Assim, a pergunta a ser respondida é: existem diferenças estatisticamente significativas entre os desempenhos dos alunos nas turmas e entre as turmas? Se houver diferença no teste ANOVA, o *post hoc* Tukey mostrará se ela é entre as três turmas, se em duas ou para apenas uma delas.

Na caixa de resultados abaixo as linhas de comando são apenas ilustrativas do comando efetuado no paco estatístico *RCommander*. O que importa são os resultados, tanto para o teste ANOVA, quanto para o Tukey. No Caso da ANOVA, os resultados mostram os graus de liberdade (gl = n-1), a soma quadrada e a média quadrada dos valores das turmas. Também é indicado o valor da estatística F=3,802 e o p-*value* do teste de diferenças de médias (Pr(>F)=0,0282*. Conforme os códigos de significância indicados abaixo, a presença de (*) mostra significância estatística de 0,010, ou seja, ao nível de 99%. Portanto, é possível afirmar que existe diferenças estatisticamente significativas entre as médias dos alunos de pelo menos uma das turmas em relação às demais. A questão é que só com a estatística F não é possível dizer qual ou quais as turmas apresentam médias estatisticamente distintas das demais. Para isso, existe o teste Tukey na sequência. Antes, porém, os resultados da ANOVA mostram as médias por turma, o desvio padrão e o número de alunos. A média da turma 3joad de 79 é a que mais se distingue das outras duas médias. 3joan, 71,44, e 5joan, 72,61. Ao que tudo indica, o teste Tukey deve mostrar que a turma 3joad é a que apresenta média significativamente distinta das outras duas e que não há diferenças estatísticas entre 3joan e 5joan.

Na janela de resultados do teste Tukey, devemos olhar primeiro para a indicação de significância estatística na última coluna da direita dos resultados (Pr(>(>|t|). A indicação é que para o par 3joan – 3joad a significância estatística é de 0,01 (*), ou seja, aceitável a 99% de intervalo de confiança. Para o par 5joan – 3joad a significância estatística é de 0,05 (.), aceitável a 95% de intervalo de confiança. Já o par 5joan – 3joan não apresenta diferenças estatisticamente significativas, podendo-se dizer que as médias dos alunos entre essas duas turmas não são distintas. Portanto, é mesmo o 3joad que apresenta as médias mais distintas em relação às outras duas turmas, ainda que o 5joan se aproxime

dela. Os resultados também mostram as estatísticas do teste t para cada par e a estimativa do valor, que é a subtração da média do primeiro fator pelo segundo fator na linha.

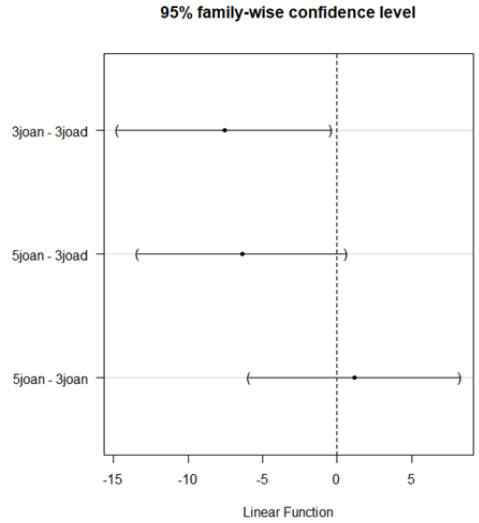
```
Linha de Comando – Teste ANOVA

Rcmdr> ANOVAModel.7 <- aov(media ~ turma_num, data=B_Exerc_V1)
Rcmdr> with(B_Exerc_V1, numSummary(media, groups=turma_num,
statistics=c("mean",
Rcmdr+   "sd")))
```

```
Resultados – Teste ANOVA

Rcmdr> summary(ANOVAModel.7)
            Df Sum Sq Mean Sq F value Pr(>F)
turma_num    2    655   327.7   3.801 0.0282 *
Residuals   57   4915    86.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


          mean        sd data:n
3joad 79.00000 11.309056     20
3joan 71.44737 10.518363     19
5joan 72.61905  4.967154     21
```

```
Linha de Comando – Teste post hoc Tukey

Rcmdr> local({
Rcmdr+   .Pairs <- glht(ANOVAModel.7, linfct = mcp(turma_num =
"Tukey"))
Rcmdr+   print(summary(.Pairs)) # pairwise tests
Rcmdr+   print(confint(.Pairs)) # confidence intervals
Rcmdr+   print(cld(.Pairs)) # compact letter display
Rcmdr+   old.oma <- par(oma=c(0,5,0,0))
Rcmdr+   plot(confint(.Pairs))
Rcmdr+   par(old.oma)
Rcmdr+ })
```

```
Resultados – Teste post hoc Tukey

Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = media ~ turma_num, data = B_Exerc_V1)

Linear Hypotheses:
                   Estimate Std. Error t value Pr(>|t|)
3joan - 3joad == 0   -7.553      2.975  -2.539   0.0364 *
5joan - 3joad == 0   -6.381      2.901  -2.199   0.0798 .
5joan - 3joan == 0    1.172      2.940   0.399   0.9163
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

```
Resultados - Gráfico – Intervalos das diferenças de médias de Tukey

Simultaneous Confidence Intervals
Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = media ~ turma_num, data = B_Exerc_V1)

Quantile = 2.406
95% family-wise confidence level

Linear Hypotheses:
                    Estimate lwr      upr
3joan - 3joad == 0  -7.5526 -14.7102  -0.3951
5joan - 3joad == 0  -6.3810 -13.3615   0.5996
5joan - 3joan == 0   1.1717  -5.9024   8.2457

3joad 3joan 5joan
  "b"   "a"  "ab"
```



Entre os resultados do teste de Tukey fornecidos no *RCommander* está o gráfico linear de intervalos de confiança simultâneos. Ele é calculado a partir da estimativa de um intervalo de confiança em função da estimativa de diferença de médias em cada par. Quanto mais próximo de zero estiver o intervalo, menos significância estatística haverá entre as diferenças do par. Se o intervalo passar por zero, não se pode rejeitar a hipótese nula de independência de variações. No caso do exemplo acima, o gráfico reforça as informações anteriores, com maior significância na diferença entre 3joan-3joad, que não passa por zero. Depois, o par 5joan – 3joad fica muito próximo a zero, porém, ainda com significância estatística. Já o par 5joan – 3joan tem um intervalo que passa por zero e estimador de 1,717, ficando muito próximo de zero. Assim, o teste de

Tukey mostra que não há diferenças de médias entre as duas turmas e que a turma 3joad é a que apresenta as maiores médias, cujas diferenças são estatisticamente significativas em relação às outras duas turmas.

O teste ANOVA para um fator que realizamos até aqui se aplica a análises descritivas das relações entre médias identificadas em diferentes grupos. A ANOVA também pode ser usada para análises fatoriais exploratórias, quando o objetivo é conhecer como se dá a relação entre os as unidades de análise que pertencem a grupos distintos. É o que trataremos a seguir.

Análise Fatorial com ANOVA

Até aqui, aplicamos o teste ANOVA para verificar a relação entre as categorias de duas variáveis apenas, uma categórica, o fator - com mais de duas categorias -, e uma contínua. Imagine que seja necessário acrescentar uma segunda variável explicativa que não tenha nenhuma relação com a anterior. Nesse caso, teríamos que analisar três tipos de efeitos: o efeito da variável independente 1 (VI1); o efeito da variável independente 2 (VI2); e o efeito conjunto das duas variáveis. (VI1 x VI2). O objetivo aqui é medir os efeitos individuais e a interação entre os efeitos das variáveis independentes como nova explicação para as variações na variável dependente. Nesse caso, precisamos usar a técnica de Análise Fatorial com ANOVA. Na Fatorial com ANOVA são analisados os efeitos de três fatores:

$Fator_a$ = Efeito da VI_1.

$Fator_b$ = Efeito da VI_2.

$Fator_{axb}$ =Efeito conjunto da VI_1 e da VI_2.

Os testes estatísticos de fatores com ANOVA produzem três resultados distintos.

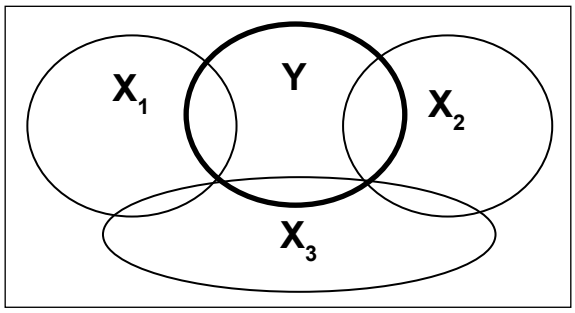
- Principais efeitos: são os efeitos médios de uma variável independente sobre o nível de variação de outra independente.

- Efeito de Interação: é o efeito de uma variável independente que depende de outra independente. Trata-se do efeito simples de uma independente sobre a variação da outra. Essa interação indica o quanto a variação de um preditor depende de outro preditor no teste.

- Efeito Simples: é o efeito de uma variável independente para um nível particular de outra variável independente.

No teste, o pressuposto é que os "principais efeitos" são independentes do "efeito de interação" entre as variáveis. Em outras palavras, só faz sentido testar os fatores com ANOVA se as variáveis independentes não tiverem nenhuma relação entre si. A análise Fatorial com ANOVA nada mais é do que um caso especial de regressão múltipla, onde os preditores são perfeitamente independentes. Esses efeitos independentes podem ser representados pela figura 7.1, abaixo.

Figura 7.1 – Representação as interações em Fatorial em ANOVA



Onde Y é a variável dependente que sofre efeitos de dois fatores, X_1 e X_2, que são variáveis independentes sem nenhuma interação entre si. Além dessas duas, há também um terceiro fator, o X_3, que é o resultado da interação entre X_1 e X_2. Os cálculos fatoriais da ANOVA são baseados em cálculos do GLM, sigla em inglês para Modelo Linear Geral. Sua fórmula é:

$$Y = B_0 + B_1 x\, X_1 + B_2\, x\, X_2 + B_3\, x\, X_3 + e$$

Onde:

Y = variável dependente

$X_1 = Fator_a$ (variável independente 1)

$X_2 = Fator_b$ (variável independente 2)

$X_3 = Fator_{axb}$ (variável independente 1 X variável independente 2)

As fórmulas para cálculos dos efeitos dos fatores (F-*ratios*) são as seguintes:

$$F_a = \frac{MS_a}{MS_{s/ab}} \qquad F_b = \frac{MS_b}{MS_{s/ab}} \qquad F_{axb} = \frac{MS_{axb}}{MS_{s/ab}}$$

Onde:

MS = médias quadradas.

Para calcular as médias quadradas de cada fator, usam-se as seguintes fórmulas:

$$MS_a = \frac{SS_a}{DF_a} \qquad MS_b = \frac{SS_b}{DF_b} \qquad MS_{axb} = \frac{SS_{axb}}{DF_{axb}} \qquad MS_{s/ab} = \frac{SS_{s/ab}}{DF_{s/ab}}$$

E as fórmulas para cálculo dos graus de liberdade são:

$$DF_a = a - 1 \qquad DF_b = b - 1 \qquad DF_{axb} = (a-1)x(b-1)$$

$$DF_{s/ab} = ab\,(n-1) \qquad DF_{total} = abn - 1 = N - 1$$

Os pressupostos para a aplicação da análise Fatorial com ANOVA são três:

- a variável dependente deve ser contínua;

- a variável dependente deve ter distribuição normal;

- deve existir homogeneidade de variâncias.

Uma vez identificados os efeitos individuais dos Fatores, procede-se um cálculo complementar para identificar a magnitude do efeito (*effect size*) das variáveis explicativas. São dois efeitos possíveis: efeito completo e o efeito parcial. As fórmulas são as que seguem:

Efeito completo: $n^2 = \frac{SS_{efeito}}{SS_{total}}$

Efeito parcial: $n^2 = \frac{SS_{efeito}}{(SS_{efeito} + SS_{s/ab})}$

O coeficiente da magnitude do efeito indica quanto de fato cada variável contribui para a variação testada. No caso, podemos ter efeitos significativos para uma variável e não para outra. Por fim, é bom não esquecer que a análise Fatorial com ANOVA só é indicada quando se tem certeza de inexistência de associação das variações das variáveis explicativas. Se as variáveis estiverem associadas, quebra-se o pressuposto inicial que sustenta a variância individual e agregada como indicador de efeito.

Para exemplificar com o *RCommander*, usaremos o banco de dados "B_Exerc_V1". Os fatores do teste serão o tempo de ocupação além da faculdade (possui 4 categorias) e a turma do aluno (possui três categorias). A variável contínua é a nota média do aluno. Assim, a ANOVA Fatorial tem o objetivo de descrever os efeitos individuais de ocupação além da faculdade e turma a que pertence o aluno sobre as notas médias. Além dos efeitos individuais, o teste também indica o efeito conjunto das duas variáveis.

O *RCommander* retorna uma série de resultados para a análise ANOVA Fatorial. Na caixa com os resultados a seguir, estão indicados os coeficientes individuais e para a interação entre tempo de ocupação e turma. A estatística F e o p-*value* (Pr(>F)) devem ser lidos como no teste anterior. O teste mostra que a variável que teve efeito mais forte sobre a média dos alunos é "turma" (p-*value*=0,031) com (*) que representa nível de significância estatística a 99% de intervalo de confiança. A interação entre tempo de ocupação e turma (p-*value* = 0,064) tem significância estatística para 95% de intervalo de confiança. Já o tempo de ocupação apenas não apresenta significância estatística para as variações das médias dos alunos.

O segundo conjunto de resultados do teste mostra as tabelas de cruzamentos das médias e dos desvios para cada par de variáveis transformada em fator. Na tabela das médias é possível perceber que a média do 3joad é maior que a das outras turmas para

três categorias de tempo de ocupação, exceto de 4 a 6 horas de ocupação, em que fica abaixo das outras turmas. A tabela de desvios também mostra variações significativas, em especial nas médias daqueles que têm entre 4 e 6 horas de ocupação além da faculdade.

```
Linha de comando – ANOVA fatorial

Rcmdr>  ANOVAModel.2 <- lm(media ~ tempocu_num*turma_num, data=B_
Exerc_V1, contrasts=list(tempocu_num =”contr.Sum”, turma_num =”contr.
Sum”))
```

```
Resultados – ANOVA Fatorial

Rcmdr>  ANOVA(ANOVAModel.2)
ANOVA Table (Type II tests)

Response: media
                      Sum Sq Df F value  Pr(>F)
tempocupação           221.6  3  0.9366 0.42999
turma                  587.3  2  3.7227 0.03111 *
tempocupação:turma     749.3  4  2.3749 0.06451 .
Residuals             3943.9 50
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Linha de comando – ANOVA Fatorial – média e desvio padrão

Rcmdr>  with(B_Exerc_V1, (tapply(media, list(tempocu_num, turma_num),
mean, na.rm=TRUE))) # means

Rcmdr>  with(B_Exerc_V1, (tapply(media, list(tempocu_num, turma_num),
sd, na.rm=TRUE))) # std. deviations
```

```
Resultados – ANOVA Fatorial – media e desvio padrão

                  3joad    3joan    5joan
Sem ocupação     85.00000       NA 67.500
De 2 a 4 horas   85.00000 60.00000     NA
De 4 a 6 horas   70.00000 77.50000 74.375
Acima de 6 horas 80.90909 72.16667 72.500

                   3joad     3joan    5joan
Sem ocupação      8.660254        NA       NA
De 2 a 4 horas          NA 28.284271       NA
De 4 a 6 horas   17.677670 10.606602 3.750000
Acima de 6 horas  6.640099  7.249795 5.244044
```

```
Linha de comando – ANOVA Fatorial – tabela de cruzamentos

Rcmdr>  xtabs(~ tempocu_num + turma_num, data=B_Exerc_V1) # counts
```

```
Resultados – ANOVA Fatorial – tabela de cruzamentos

                          turma_num
tempocu_num               3joad 3joan  5joan
          Sem ocupação      3     0      1
          De 2 a 4 horas    1     2      0
          De 4 a 6 horas    5     2      4
          Acima de 6 horas 11    15     16
```

Por fim, o último conjunto de resultados é o cruzamento para número de casos por turma e tempo de ocupação. Nessa tabela é possível perceber uma concentração de casos em ocupação acima de 6 horas por dia. Além disso, percebe-se a ausência de casos em algumas casas, como nenhum aluno sem ocupação no 3joan e nenhum de 2 a 4 horas na turma 5joan. Essas informações são importantes, pois sabemos que a ausência de casos em testes de independência de médias pode interferir nas estatísticas inferenciais. Apesar disso, no nosso exemplo, fica claro que o fator que mais explica a variação das médias dos alunos é a turma e não o tempo de ocupação diária além da faculdade.

Nos testes t e nos testes ANOVA, estamos considerando sempre que a variável a ser testada é contínua, enquanto os fatores explicativos são categóricos. Esses testes não servem para quando temos que relacionar as diferenças de variações entre categorias de variáveis categóricas. Para testar duas ou mais variáveis categóricas entre si, são necessários outros conjuntos de testes. A seguir, discutiremos um teste de independência de médias para quando se têm duas variáveis categóricas nominais e não mais uma contínua e outra categórica, como nos testes realizados até aqui.

7.8 Prova de Qui-quadrado *(x^2)*

O teste de qui-quadrado (x^2) foi proposto pelo estatístico Karl Pearson em 1900 e por isso é conhecido por x^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 parte da mesma 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 uma 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).

Assim como a distribuição t, o x^2 depende apenas dos graus de liberdade como parâmetro e os limites críticos variam em função dos graus de liberdade e do intervalo de confiança (ver anexo 7.3, ao final deste capítulo). A sua forma de distribuição aproxima-se da normal à medida que aumentam os graus de liberdade, equiparando-se a esta quando houver 30 ou mais graus de liberdade.

O x^2 é um teste indicado para duas variáveis categóricas, ou seja, quando não há variável para parametrizar as variações das distribuições. Nos testes anteriores, o pressuposto era que pelo menos uma das variáveis, a contínua, seria normalmente distribuída. Agora, quando todas as variáveis são nominais, não é possível parametrizar. Nesses casos, o teste indicado é o x^2. **Por exemplo**, para quando estamos tratando de determinado diagnóstico (positivo ou negativo), o resultado de uma eleição (eleito, derrotado), ou então o desempenho dos candidatos em uma disputa (voto em A, voto em B, voto em C), ou o veredicto de um julgamento (culpado, inocente).

Para testar a hipótese nula nesses casos é preciso verificar a "bondade" (*goodness*) do ajustamento do modelo para atestar a independência das variações das categorias das variáveis. O x^2 para bondade do modelo indica quão boa é a distribuição das proporções modeladas em uma distribuição teórica. Aqui, a hipótese nula é que as proporções de distribuições são equivalentes entre as categorias das variáveis. A hipótese alternativa é que as distribuições das proporções são distintas. Para a realização de um teste x^2 é necessário respeitar os seguintes pressupostos:

- é preciso ter um número mínimo de cinco casos em cada casa de tabelas quádruplas (2x2) ou em mais de 80% das casas em tabelas maiores (se esse pressuposto for quebrado, recomenda-se o teste de Fisher);

- não pode existir nenhuma casa com ausência total de casos (zero na casa);

- deve haver independência entre as observações das variáveis testadas;

- o x^2 não pode ser usado para testar dados correlacionados, como em pesquisas de painel, pois isso quebraria o pressuposto anterior.

O x^2 tem muitas aplicações, a seguir serão discutidas as principais delas.

Comparar uma distribuição observada a uma teórica

Quando se lança um dado 90 vezes, teoricamente, deveria haver 15 vezes cada uma das seis faces como resultado se as distribuições fossem aleatórias e o número de resultados de cada face casual. Trata-se de comparar se as faces observadas no experimento diferem significativamente das esperadas. Se o nível de significância (Alfa) do teste fica acima de 0,050, então, nada se opõe a aceitar a hipótese nula de que não existe diferença estatisticamente significativa e as observadas são frutos do acaso.

Vamos a um exemplo prático de cálculo do coeficiente. Digamos que se queira verificar como se deu a distribuição das mulheres 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 teste de x^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 7.3 – 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
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 x^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 x^2.

Na tabela do anexo 7.3 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 coluna (0,050). 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 7.3 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.

Comparar distribuições observadas com dados independentes

Essa prova é usada para comparar se diferentes médias observadas em dois grupos independentes são estatisticamente significativas. A hipótese nula pode ser formulada como não existem diferenças significativas entre os dois grupos ou que as diferenças observadas são frutos do acaso ou, ainda, as duas amostras procedem da mesma população. A prova mais utilizada para esse tipo de comparação é a U de Mann-Whitney, que prova se dois grupos independentes procedem da mesma população.

Esse teste é usado para comparar se diferentes distribuições observadas em dois grupos independentes são estatisticamente significativas. 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 procedam 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 x^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 x^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:

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

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 a Fe para cada casa, depois calculamos o x^2 para cada uma das categorias e somamos os valores. O x^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 7.4 – 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	0,02	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	57	56,93	0	495
PP	417	410,63	0,099	47	53,37	0,76	464
PSB	389	389,39	0	51	50,61	0,003	440
PDT	284	272,58	0,479	24	35,42	3,685	308
PTB	260	260,19	0	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, $x^2 = 1,083 + 8,331 = 9,414$. 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, nem comparar, coeficiente x^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 x^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 x^2. Nesse caso, estamos usando o x^2 para confirmar a hipótese de independência, ou seja, para saber se duas variáveis categóricas estão ou não relacionadas. A hipótese nula afirma que as duas variáveis são independentes, quer dizer, não apresentam nenhuma relação entre si. Podemos aplicar a prova de independência no seguinte **exemplo**: usando o banco de dados "B_Exerc_V1" e rodando o teste no *RCommander*, produzir uma tabela de cruzamento entre "interesse" e "turma". O objetivo é identificar se as distribuições dos casos variam significativamente de interesse pela disciplina entre as turmas analisadas. Se o coeficiente ficar acima do limite crítico, não podemos rejeitar a hipótese nula e, portanto, devemos considerar que não há diferenças de interesse das disciplinas entre as turmas. Os resultados são os que seguem:

```
Linha de comando – tabela de contingência com chi-quadrado de Pearson

Rcmdr>  local({
Rcmdr+    .Table <- xtabs(~interess+turma, data=B_Exerc_V1)
Rcmdr+    cat("\nFrequency table:\n")
Rcmdr+    print(.Table)
Rcmdr+    .Test <- chisq.test(.Table, correct=FALSE)
Rcmdr+    print(.Test)
Rcmdr+  })
```

```
Resultados – tabela de contingência com chi-quadrado de Pearson

Frequency table:
             turma
interess     3joad 3joan 5joan
  principal     10     6     9
  secundário    10    12    13

Pearson's Chi-squared test

data:  .Table
X-squared = 1.0909, df = 2, p-value = 0.5796
```

Os resultados mostram que a tabela de contingência entre interesse e turma apresenta um coeficiente x^2 = 1,090 e um p-*value* = 0,579, o que significa que não há diferença estatisticamente significativa entre o grau de interesse pela disciplina e a turma do aluno.

Comparar duas distribuições observadas com dados relacionados

Trata-se o teste x^2 para comparar distribuições de dados relacionados ou pareados. Para tanto, são feitas duas medições sobre os mesmos indivíduos que foram previamente igualados. **Por exemplo**, a um grupo de 190 estudantes foi aplicado um exame tradicional e outro tipo teste. No tradicional, foram aprovados 120 alunos e, no outro teste, 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. Esta aplicação não será discutida aqui, apenas citada, por ser pouco usada na ciência política.

Com a apresentação dos testes de independência de média, encerramos o primeiro volume da introdução aos métodos quantitativos aplicados à ciência política. Até aqui, foram apresentadas as formas de representação de informações e os testes de independência de médias. No próximo volume, serão discutidos testes mais avançados para análises descritivas, exploratórias e inferenciais de dados quantitativos.

7.9 Exercícios propostos

a) O que são modelos probabilísticos?

b) Quais são as características de uma Curva Normal ou Curva de Gauss?

c) Defina hipótese nula e hipótese alternativa em um NHST.

d) Diferencie erro de tipo 1 (alfa) do de tipo 2 (beta) em um NHST.

Usando o banco de dados “B_Exerc_V1” e o *RCommander*, faça os seguintes testes:

e) Teste t para uma variável. Variável: “faltas”. Interprete o resultado.

f) Teste t para amostras independentes: fator: “interesse”; variável: “média”. Interprete o resultado.

g) Teste ANOVA com comparações múltiplas (Tukey): fator: “tempo_ocupa”; variável: média. Interprete os resultados.

h) Análise Fatorial ANOVA. Fatores: “tipo_escola” e “tempo_ocupação”; variável: faltas. Interprete os resultados.

i) Teste x^2 em tabela de contingência: variáveis: “est_civil” e “turma”. Interprete o resultado.

7.10 Bibliografia citada no capítulo

BARBETTA, Pedro Alberto. **Estatística Aplicada às Ciências Sociais** (5ª edição revisada). Florianópolis: Editora UFSC, 2003.

BUNCHAFT, Guenia; KELLNER, Sheilah Ribno de Oliveira. **Estatística Sem Mistérios** (2ª edição corrigida) vol. I, II, III e IV. Petrópolis - RJ : Editora Vozes, 1999.

FONSECA, Jairo Simon; MARTINS, Gilberto de Andrade; TOLEDO, Geraldo Luciano. **Estatística Aplicada.** São Paulo: Editora Atlas, 1995.

FROTA, Álvaro. **A Estatística e o Conhecimento Humano**. Apostila, 1999.

IGNÁCIO, Sérgio Aparecido. Estatística Geral e Aplicada. Apostila, 2003.

TOLEDO, G. L.; OVALLE, I. **Estatística Básica.** São Paulo: Editora Atlas, 1991.

TRIOLA, Mario F. **Introdução à Estatística** (7ª edição). Rio de Janeiro: Editora LTC, 1999.

Anexos do Capítulo VII

Anexo 7.1 – Tabela de Valores Padronizados Z

Z	0	1	2	3	4	5	6	7	8	9
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	**0.4515**	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

Anexo 7.2 – Tabela de valores padronizados para teste t

Unicaudal	75,0%	80,0%	85,0%	90,0%	95,0%	97,5%	99,0%	99,5%	99,8%	99,9%	100,0%
Bicaudal	50,0%	60,0%	70,0%	80,0%	90,0%	95,0%	98,0%	99,0%	99,5%	99,8%	99,9%
1	1	1,38	1,96	3,08	6,31	12,71	31,82	63,66	127,3	318,3	636,6
2	0,82	1,06	1,39	1,89	2,92	4,3	6,97	9,93	14,09	22,33	31,6
3	0,77	0,98	1,25	1,64	2,35	3,18	4,54	5,84	7,45	10,21	12,92
4	0,74	0,94	1,19	1,53	2,13	2,78	3,75	4,6	5,6	7,17	8,61
5	0,73	0,92	1,16	1,48	2,02	2,57	3,37	4,03	4,77	5,89	6,87
6	0,72	0,91	1,13	1,44	1,94	2,45	3,14	3,71	4,32	5,21	5,96
7	0,71	0,9	1,12	1,42	1,9	2,37	3	3,5	4,03	4,79	5,41
8	0,71	0,89	1,11	1,4	1,86	2,31	2,9	3,36	3,83	4,5	5,04
9	0,7	0,88	1,1	1,38	1,83	2,26	2,82	3,25	3,69	4,3	4,78
10	0,7	0,88	1,09	1,37	1,81	2,23	2,76	3,17	3,58	4,14	4,59
11	0,7	0,88	1,09	1,36	1,8	2,2	2,72	3,11	3,5	4,03	4,44
12	0,7	0,87	1,08	1,36	1,78	2,18	2,68	3,06	3,43	3,93	4,32
13	0,69	0,87	1,08	1,35	1,77	2,16	2,65	3,01	3,37	3,85	4,22
14	0,69	0,87	1,08	1,35	1,76	2,15	2,62	2,98	3,33	3,79	4,14
15	0,69	0,87	1,07	1,34	1,75	2,13	2,6	2,95	3,29	3,73	4,07
16	0,69	0,87	1,07	1,34	1,75	2,12	2,58	2,92	3,25	3,69	4,02
17	0,69	0,86	1,07	1,33	1,74	2,11	2,57	2,9	3,22	3,65	3,97
18	0,69	0,86	1,07	1,33	1,73	2,1	2,55	2,88	3,2	3,61	3,92
19	0,69	0,86	1,07	1,33	1,73	2,09	2,54	2,86	3,17	3,58	3,88
20	0,69	0,86	1,06	1,33	1,73	2,09	2,53	2,85	3,15	3,55	3,85
21	0,69	0,86	1,06	1,32	1,72	2,08	2,52	2,83	3,14	3,53	3,82
22	0,69	0,86	1,06	1,32	1,72	2,07	2,51	2,82	3,12	3,51	3,79
23	0,69	0,86	1,06	1,32	1,71	2,07	2,5	2,81	3,1	3,49	3,77
24	0,69	0,86	1,06	1,32	1,71	2,06	2,49	2,8	3,09	3,47	3,75
25	0,68	0,86	1,06	1,32	1,71	2,06	2,49	2,79	3,08	3,45	3,73
26	0,68	0,86	1,06	1,32	1,71	2,06	2,48	2,78	3,07	3,44	3,71
27	0,68	0,86	1,06	1,31	1,7	2,05	2,47	2,77	3,06	3,42	3,69
28	0,68	0,86	1,06	1,31	1,7	2,05	2,47	2,76	3,05	3,41	3,67
29	0,68	0,85	1,06	1,31	1,7	2,05	2,46	2,76	3,04	3,4	3,66
30	0,68	0,85	1,06	1,31	1,7	2,04	2,46	2,75	3,03	3,39	3,65
40	0,68	0,85	1,05	1,3	1,68	2,02	2,42	2,7	2,97	3,31	3,55
50	0,68	0,85	1,05	1,3	1,68	2,01	2,4	2,68	2,94	3,26	3,5
60	0,68	0,85	1,05	1,3	1,67	2	2,39	2,66	2,92	3,23	3,46
80	0,68	0,85	1,04	1,29	1,66	1,99	2,37	2,64	2,89	3,2	3,42
100	0,68	0,85	1,04	1,29	1,66	1,98	2,36	2,63	2,87	3,17	3,39
120	0,68	0,85	1,04	1,29	1,66	1,98	2,36	2,62	2,86	3,16	3,37

Anexo 7.3 – Tabela de limites críticos para qui-quadrado

gl	0,005	0,01	0,025	0,05	0,1	0,25	0,5	0,75	0,9	0,95	0,99
1	0	0	0,001	0,004	0,016	0,102	0,455	1,323	2,706	3,841	7,879
2	0,01	0,02	0,051	0,103	0,211	0,575	1,386	2,773	4,605	5,991	10,597
3	0,072	0,115	0,216	0,352	0,584	1,213	2,366	4,108	6,251	7,815	12,838
4	0,207	0,297	0,484	0,711	1,064	1,923	3,357	5,385	7,779	9,488	14,86
5	0,412	0,554	0,831	1,145	1,61	2,675	4,351	6,626	9,236	11,07	16,75
6	0,676	0,872	1,237	1,635	2,204	3,455	5,348	7,841	10,645	12,592	18,548
7	0,989	1,239	1,69	2,167	2,833	4,255	6,346	9,037	12,017	14,067	20,278
8	1,344	1,646	2,18	2,733	3,49	5,071	7,344	10,219	13,362	15,507	21,955
9	1,735	2,088	2,7	3,325	4,168	5,899	8,343	11,389	14,684	16,919	23,589
10	2,156	2,558	3,247	3,94	4,865	6,737	9,342	12,549	15,987	18,307	25,188
11	2,603	3,053	3,816	4,575	5,578	7,584	10,341	13,701	17,275	19,675	26,757
12	3,074	3,571	4,404	5,226	6,304	8,438	11,34	14,845	18,549	21,026	28,3
13	3,565	4,107	5,009	5,892	7,042	9,299	12,34	15,984	19,812	22,362	29,819
14	4,075	4,66	5,629	6,571	7,79	10,17	13,339	17,117	21,064	23,685	31,319
15	4,601	5,229	6,262	7,261	8,547	11,04	14,339	18,245	22,307	24,996	32,801
16	5,142	5,812	6,908	7,962	9,312	11,91	15,338	19,369	23,542	26,296	34,267
17	5,697	6,408	7,564	8,672	10,09	12,79	16,338	20,489	24,769	27,587	35,718
18	6,265	7,015	8,231	9,39	10,87	13,68	17,338	21,605	25,989	28,869	37,156
19	6,844	7,633	8,907	10,12	11,65	14,56	18,338	22,718	27,204	30,144	38,582
20	7,434	8,26	9,591	10,85	12,44	15,45	19,337	23,828	28,412	31,41	39,997
21	8,034	8,897	10,28	11,59	13,24	16,34	20,337	24,935	29,615	32,671	41,401
22	8,643	9,542	10,98	12,34	14,04	17,24	21,337	26,039	30,813	33,924	42,796
23	9,26	10,2	11,69	13,09	14,85	18,14	22,337	27,141	32,007	35,172	44,181
24	9,886	10,86	12,4	13,85	15,66	19,04	23,337	28,241	33,196	36,415	45,559
25	10,52	11,52	13,12	14,61	16,47	19,94	24,337	29,339	34,382	37,652	46,928
26	11,16	12,2	13,84	15,38	17,29	20,84	25,336	30,435	35,563	38,885	48,29
27	11,81	12,88	14,57	16,15	18,11	21,75	26,336	31,528	36,741	40,113	49,645
28	12,46	13,57	15,31	16,93	18,94	22,66	27,336	32,62	37,916	41,337	50,993
29	13,12	14,26	16,05	17,71	19,77	23,57	28,336	33,711	39,087	42,557	52,336
30	13,79	14,95	16,79	18,49	20,6	24,48	29,336	34,8	40,256	43,773	53,672
40	20,71	22,16	24,43	26,51	29,05	33,66	39,335	45,616	51,805	55,758	66,766
50	27,99	29,71	32,36	34,76	37,69	42,94	49,335	56,334	63,167	67,505	79,49
60	35,53	37,49	40,48	43,19	46,46	52,29	59,335	66,981	74,397	79,082	91,952
70	43,28	45,44	48,76	51,74	55,33	61,7	69,334	77,577	85,527	90,531	104,22
80	51,17	53,54	57,15	60,39	64,28	71,15	79,334	88,13	96,578	101,88	116,32
90	59,2	61,75	65,65	69,13	73,29	80,63	89,334	98,65	107,57	113,15	128,3
100	67,33	70,07	74,22	77,93	82,36	90,13	99,334	109,14	118,5	124,34	140,17

Adendo I

Gabarito de respostas aos exercícios propostos

Respostas às Questões da Introdução

a) As quatro etapas são ideias, evidências, análises, representações. As ideias servem como premissas para a pesquisa empírica e normalmente são resultados de transmissões intergeracionais de pesquisadores. Evidências empíricas são dados coletados da realidade com o uso de técnicas de pesquisa que selecionam os elementos que devem ser levados em conta. As análises são feitas com o uso de ferramentas técnicas disponíveis para interpretar os dados coletados da realidade. As representações são construções interpretativas dos pesquisadores a partir das ideias iniciais e após a coleta e análise de dados empíricos.

b) A estratégia de pesquisa qualitativa tem o objetivo de examinar detalhes. Por exemplo, uma pesquisa que busca identificar as opiniões parlamentares sobre um tema público qualquer, usando entrevistas em profundidade para não apenas descrever os posicionamentos individuais, mas também relacioná-los com posições sobre outros assuntos, históricos, visão de mundo e contexto geral. A estratégia comparativa tem como objetivo principal comparar características. Por exemplo, comparar as opiniões sobre determinado tema público de parlamentares de diferentes países de um continente. A estratégia quantitativa apresenta como objetivo a identificação de padrões gerais. Por exemplo, a partir de um questionário identificar, os posicionamentos de parlamentares estaduais de todo o país a respeito de um tema público com a finalidade de descrever os

padrões gerais de posições a respeito desse tema em relação a outras características dos parlamentares.

c) De específico no experimento de campo, há alguma alteração intencional no objeto de análise para medir padrões de comportamento. A pesquisa descritiva usa muitas variáveis para descrever como os fenômenos acontecem sem intervenção humana. Já a pesquisa exploratória é aplicada a objetos de estudo pouco descritos pela literatura, sendo uma primeira abordagem sobre um tema novo. Por exemplo, uma pesquisa exploratória sobre as novas formas de interação no debate sobre temas públicos em redes digitais no início da disponibilidade dessa tecnologia. Já uma pesquisa descritiva seria a ampla descrição das variações de características já identificas sobre as interações em redes digitais. Por fim, o experimento de campo seria testar as reações de comentadores de redes digitais após receberem determinados conteúdos de comentários controlados e destinados a gerar estímulos como parte do experimento.

Respostas às Questões do Capítulo 1

a) Taxonomia é um conjunto de definições de conceitos para identificar algo. Em função das definições, os conceitos podem ser mais ou menos ricos. Teoria é quando há relação entre diferentes proposições conceituais apresentadas. As teorias podem ser verdadeiras ou faltas em relação a determinada realidade em função do tipo de relação que apresenta entre os diferentes conceitos. Por exemplo, é considerada taxonomia o conjunto de definições sobre o que é comportamento democrático. Já uma teoria da democracia relaciona um conjunto de conceitos sobre comportamento e instituições definidas como democráticas para diferenciar a democracia de outros tipos de relações políticas que não se encaixam na teoria apresentada.

b) Problema de pesquisa é a interrogação que o pesquisador faz a respeito de algum aspecto da realidade que merece ser investigado. Por isso, todo problema científico sur-

ge de uma pergunta. Hipóteses são afirmações provisórias na tentativa de dar resposta ao problema científico. Não passam de conjecturas iniciais na tentativa de direcionar os passos seguintes da pesquisa. Por isso, elas sempre são afirmações.

c) Falácias científicas são afirmações conceituais que não podem ser confirmadas cientificamente, por estarem sustentadas em procedimentos inadequados do ponto de vista científico para a explicação da realidade. O primeiro tipo é a falácia ecológica, cometida quando os dados dizem respeito a unidades agregadas, regiões, macro-informações, mas o pesquisador tira conclusões sobre indivíduos a partir delas. O segundo é a falácia individualista que, ao contrário da primeira, parte de dados individuais, micro-informações, para fazer afirmações sobre regiões e unidades agregadas.

d) Os cinco tipos são: Reversíveis e irreversíveis, quando a relação é nas duas direções ou em uma única direção, por exemplo: interesse em política e atenção destinada às campanhas eleitorais ou quando a posição social explica opinião sobre tema público; Determinista ou probabilística, quando a presença de uma variável determina presença ou ausência de outra característica ou quando a presença de uma variável estabelece alguma chance da presença de outra característica no indivíduo, por exemplo: quando o bom desempenho em testes matemáticos está presente apenas em provas de alunos que receberam conteúdos de matemática na escola ou quando determinada metodologia de ensino de matemática aumenta as chances de bom desempenho dos alunos nesse tipo de prova; Sequencial ou coexistente, quando a ocorrência de uma característica em um momento acarreta a aparição de outra no momento seguinte ou quando as duas acontecem ao mesmo tempo, por exemplo: acesso a informações sobre determinado tema público no presente molda opinião sobre esse tema no futuro ou posição sobre um tema público e a tomada de decisão em um referendo sobre o mesmo tema; Suficiente ou contingente, quando a presença de uma característica basta para explicar outra ou quando é preciso contingência de uma terceira variável para explicar o resultado, por exemplo: ter idade mínima já é suficiente para ser eleitor ou ter idade mínima para votar, mas precisa de estímulo e vontade para exercer o voto; Necessárias ou substituíveis, quando a presença de uma característica é necessidade exclusiva para determinada

ocorrência ou quando uma característica é apenas uma das condições para determinada ocorrência, por exemplo: a morte de um presidente é condição necessária para sua substituição ou uma crise de legitimidade pode ter entre suas consequências a substituição do presidente.

e) A diferença é que relações simétricas não envolvem causalidade, podendo ser apenas indicadores alternativos de um conceito, ou seja, as variações estão apenas associadas. Já nas relações assimétricas, existe um efeito de causalidade, com uma variável determinando o comportamento de outra variável. Enquanto que, nas relações espúrias, há uma suposta associação entre ocorrências de duas variáveis. Essas ocorrências podem até ser concomitantes, mas não existe nenhuma explicação lógica que permita uma associação ou relação determinística. Considerar associação ou determinação em relações espúrias é um erro de avaliação do pesquisador.

Respostas às Questões do Capítulo 2

a) Estatística descritiva preocupa-se com a coleta e descrição dos dados observados, sem pretensão de conclusões mais genéricas para além da população descrita. A diferença para a estatística inferencial é que esta produz informações a partir de um grupo reduzido de observações para se aproximar aos valores estimados da população. Faz isso a partir de uma amostra e utilizando técnicas probabilísticas, que não são necessárias na estatística descritiva.

b) Validade diz respeito ao nível de confiança que uma variável apresenta ao representar determinada característica analisada. Uma variável é válida quando congruente com o que pretende representar. Confiabilidade diz respeito à certeza de que o processo de coleta dos dados foi adequado e que as variações indicadas nas variáveis são representativas do que acontece na realidade. Uma medida pode ser válida, mas, não confiável. Porém, uma medida confiável pode não ser válida. Para apresentar as duas

características, a variável precisa apresentar variações reais dos casos observados e estar relacionado a um conceito. Tendenciosidade são variações não reais que tendem sempre na mesma direção, enviesando os resultados das análises. Medições tendenciosas acabam com a confiabilidade da variável. O erro randômico, por sua vez, não está ligado a problemas de coleta das informações e não pode ser antecipado. O erro randômico acontece por diferentes motivos, porém, como ele não indica uma tendenciosidade, normalmente, seus efeitos são minimizados, pois erros em uma direção são anulados pelos cometidos na direção oposta.

c) Indica-se um teste não paramétrico, pois não há uma das medidas para ser parametrizada por média e desvio-padrão. O teste mais comum para quando as duas variáveis são categóricas nominais ou ordinais é o qui-quadrado de Pearson (x^2) aplicado a tabelas de contingência. Existe também o phi e outros testes.

d) Organize os dados em ordem crescente de valores para fazer a categorização da variável pela fórmula de *Sturges*:

$$C = \frac{R}{K} = \frac{(Xmax - Xmin)}{1 + (3{,}32\ x \log N)} = \frac{31}{6{,}90} = 4{,}49$$

O número de classes (k=6,90) arredondado para 7 e o intervalo entre classes é de 4,49. A categorização fica:

Categoria	Intervalo		Final	N	Percentual
1	18	22,49	18 a 22	31	51,66
2	22,49	26,98	23 a 27	12	20
3	26,98	31,47	28 a 31	5	8,34
4	31,47	35,96	32 a 36	4	6,67
5	35,96	40,45	37 a 40	4	6,67
6	40,45	44,94	41 a 45	3	5
7	44,94	49,43	46 a 49	1	1,66

e) Organize os dados em ordem decrescente de valores para fazer a categorização da variável pela fórmula de *Nihans*:

Estabelecem-se os limites das categorias A, B e C individualmente e depois são agregados os casos incluídos entre os limites.

$$N_a = \frac{\Sigma x^2}{\Sigma x} = \frac{41.734}{1.514} = 27{,}56$$

$$N_b = \frac{\Sigma x^2}{\Sigma x} = \frac{17.804}{850} = 20{,}94$$

$$N_c = \frac{\Sigma x^2}{\Sigma x} = \frac{5.267}{281} = 18{,}74$$

Estabelecidos os limites de cada uma das três categorias, o número de casos passa a ser:

Categoria	Intervalo	N	Percentual
A	49 a 27	19	31,66
B	26 a 20	26	43,34
C	19 a 18	15	25

Em comparação com a distribuição de *Sturges*, como *Nihans* produz três categorias, há uma concentração maior de casos nelas em relação às sete categorias originais de *Sturges*. Em *Nihans,* as distribuições foram mais homogêneas, girando entre 43,34% e 25%, enquanto que, em *Sturges,* a primeira categoria concentrou 51,66% dos casos, gerando uma variável categórica mais heterogênea no que diz respeito ao número de casos agregados em cada uma das categorias.

Respostas às Questões do Capítulo 3

a) Uma matriz de dados é composta por variáveis, unidades e valores. As unidades

são aquilo que se estuda, que podem ser pessoas, partidos políticos, municípios, países e outros. Variáveis são as características das unidades que se estuda, podendo ser dados socioeconômicos ou opiniões políticas se a unidade for pessoas. Valores são as categorias das variáveis que se aplicam às unidades analisadas. Por exemplo, se a unidade é a pessoa e a variável é opinião sobre aborto, os valores são as possíveis respostas: favorável ou contrário.

b) Índices são constituídos pela agregação, normalmente por somatória de valores, de variáveis individuais. O índice é mais econômico por reunir informações de diferentes variáveis, sem diferenciar a participação ou intensidade de cada componente individual. Já as escalas se dão por atribuição de pontos a respostas, com diferenciação e intensidade de importância. Por este motivo, escalas são medidas mais sofisticadas do que os índices.

c) Para uma correlação média de 0,800 e seis itens no indicador o valor do coeficiente esperado segundo a tabela 3.5 é 0,970. Como o valor obtido foi 0,750, então o valor abaixo indica que a confiabilidade ficou abaixo do esperado pela tabela padronizada do *Cronbach*. Isso é importante porque mostra que a confiabilidade está abaixo do esperado e isso implica na tomada de alguma decisão pelo pesquisador, que pode ser reduzir ou aumentar o número de itens no indicador.

d) O índice de fracionalização máxima depende apenas do número total de cadeiras no parlamento e do número total de partidos. Um parlamento com 30 cadeiras e nove partidos apresenta fracionalização máxima de 0,919. Como quanto mais próximo de um, maior a fracionalização, o indicador de 0,919 indica alta fracionalização do parlamento.

$$\text{IFM: } \frac{N(n-1)}{n(N-1)} = \frac{30\ x\ (9-1)}{9\ x\ (30-1)} = \frac{240}{261} = 0{,}919$$

e) A taxa específica de fecundidade é calculada por faixas de idade das mães, dividindo os números de nascidos vivos por ano para cada grupo específico de idade das

mães. A taxa total de fecundidade pode ser obtida a partir do total de crianças nascidas em um ano pelo total de mulheres em idade fértil, independente do grupo de idade.

f) Uma representação tabular é composta por linhas e colunas, onde as colunas indicam as características analisadas e as linhas descrevem os valores das mesmas. Toda tabela deve conter pelo menos um título que sumarize o conjunto de informações da tabela, localizado no topo dela. Cabeçalho, que especifica o conteúdo de cada coluna, localizado abaixo do título. Coluna indicadora, onde aparecem as indicações dos conteúdos de cada linha. O corpo é onde se encontram células com informações a serem sumarizadas em linhas e colunas. Células, que são os locais destinados às informações a serem analisadas. Por fim, indicação de fonte, na parte inferior da tabela, abaixo do corpo, com informações sobre a fonte original dos dados.

g) A representação gráfica de dados deve cumprir o objetivo principal de simplificar e sumarizar as informações principais a serem apresentadas. Para isso, é preciso ter em mente que os gráficos devem comunicar ideias complexas com clareza, precisão e eficiência, focando na substância das informações e não no método ou tecnologia para apresentação. O bom gráfico é aquele que condensa muitas informações em pouco espaço visual e não muitos elementos visuais para poucas informações.

h) Para calcular os limites para *outliers*:

$Q_1 = 29$

$Q_3 = 39$

$L = Q_3 – Q_1 = 39 – 29 = 10$

Outlier superior: $Q_3 +(1,5xL) = 39+15 = 54$

Outlier inferior: $Q_1-(1,5xL) = 29 – 15 = 34$

Caso extremo superior: $Q_3+(3xL) = 39+30 = 69$

Caso extremo inferior: $Q_1-(3xL) = 29 -30 = -1$

Respostas às Questões do Capítulo 4

a) Cálculo da média aritmética para todos os casos:

$$\mu = \frac{\sum}{N} = \frac{1.140}{45} = 25{,}33$$

A média aritmética das idades para os 45 casos é de 25,33 anos.

b) Cálculo da média ponderada por estado civil:

$$\mu_{solteiros} = \frac{\sum}{N} = \frac{653}{30} = 21{,}76$$

$$\mu_{casados} = \frac{\sum}{N} = \frac{487}{15} = 32{,}46$$

$$\mu_{ponderada} = \frac{(\mu_{solteiros}\ x\ N_{solteiros}) + (\mu_{casados}\ x\ N_{casados})}{N_{solteiros} + N_{casados}} = \frac{(21{,}76x31) + (32{,}46x15)}{30 + 15}$$

$$= \frac{674{,}56 + 519{,}36}{30 + 15} = \frac{1.193{,}92}{45} = 26{,}53$$

A média ponderada das médias de idade por estado civil é de 26,53 anos. A diferença para as médias aritméticas é que a ponderação considera o maior número de solteiros (30) do que de casados (15). Com isso, a participação relativa da média mais baixa dos solteiros exerce o efeito de "puxar para baixo" a média ponderada do grupo.

c) Para calcular a média com os pesos, partiremos dos valores do exercício acima de 674,56 para solteirões e 519,36 para casados. O cálculo é o que segue:

$$\mu_{com\ pesos} == \frac{(674{,}56x1{,}5) + (519{,}36x3)}{(30x1{,}5) + (15x3)} = \frac{1.011{,}84 + 1.558{,}08}{45 + 45} = \frac{2.569{,}92}{90} = 28{,}55$$

O resultado da média ponderada com pesos é de 28,55 anos. Ao final, o valor não alterou muito porque o dobro de peso foi dado ao grupo com a metade do tamanho.

d) Após organizar todos os valores das idades em ordem crescente, considerando que há 45 casos, o valor que divide a distribuição ao meio é o que se encontra na posição 23. No caso, a mediana é 21 anos, pois ela está na posição 23 da distribuição.

e) A distribuição é unimodal com grande concentração de casos no início. A idade com maior número de repetições é 19 anos, com onze casos.

Respostas às Questões do Capítulo 5

a) A diferença é que na amostra por conglomerados não se conhece e se desconsidera o conjunto de informações sobre as características dos indivíduos pesquisados, enquanto que na amostra por cotas é o conjunto de características que define a participação proporcional dos grupos. Na amostra por conglomerados a distribuição é por área geográfica e o que se espera é a maior heterogeneidade possível. Na amostra por cotas se estabelece previamente quais são as características relevantes dos indivíduos amostrados e se preenchem as cotas proporcionalmente à participação dessas características na população. Por exemplo, em uma amostra por conglomerados de uma pesquisa escolar são distribuídas as áreas geográficas e sorteadas dentro de cada área as escolas existentes a serem pesquisadas. Em uma amostra por cotas, estabelecem-se as características relevantes para a pesquisa, por exemplo, nível escolar (fundamental ou médio) e a partir da participação proporcional dessa característica, calcula-se quantos questionários serão aplicados a alunos de cada um dos tipos de escolas. Ainda se podem fazer cotas por idade, série ou sexo dos alunos, aumentando o número de cotas na amostra.

b) Cálculo do tamanho da amostra para população infinita, com intervalo de confiança de 95%, margem de erro de 3% e desvio padrão de 0,50.

$$n = \left(\frac{z.s}{d}\right)^2 = \left(\frac{1{,}96x0{,}5}{0{,}03}\right)^2 = 1.066{,}89 \cong 1.067\ questionários$$

Com os valores fornecidos, será necessário aplicar 1.067 questionários para a compo-

sição da amostra.

c) O cálculo do erro real é o seguinte:

$$e = z.\sqrt{\frac{p.q}{n}} = 1{,}96\ x\ \sqrt{\frac{0{,}3x0{,}7}{1.100}} = 1{,}96\ x\ 0{,}013 = 0{,}027$$

O erro real da amostra é de 2,7%.

d) Cálculo de ajuste do tamanho da amostra para população de 800 indivíduos:

$$FCP = n.\sqrt{\frac{N-n}{N-1}} = 500\ x\ \sqrt{\frac{800-500}{800-1}} = 500\ x\ 0{,}612 = 306{,}37\ \cong 307\ questionários$$

Com o fator de correção da população, o tamanho da amostra cai de 500 para 307 indivíduos.

Respostas às Questões do Capítulo 6

a) O de corte transversal é o mais comum, com informações coletadas em um único momento do tempo a partir de uma amostra específica. Já a pesquisa longitudinal coleta informações em diferentes momentos do tempo, permitindo a análise temporal. As pesquisas longitudinais que repetem as perguntas em diferentes momentos do tempo, mas não para as mesmas pessoas. Ou seja, para amostras diferentes, são os estudos de tendência. Quando as perguntas são repetidas para as mesmas pessoas em diferentes momentos do tempo, trata-se de estudo tipo painel.

b) As três etapas são: cumprimento, com o estabelecimento da confiança necessária para o início das perguntas; interação entre perguntas e respostas, com redução de custos, sem parecer que é teste; e ao final, na despedida, um reforço de benefícios futuros

com demonstração de consideração e apreciação verbal.

c) Para garantir a validade da medição, uma pergunta precisa ser comunicada pelo entrevistador de maneira consistente e ser compreendida pelos entrevistados de maneira consistente. Além disso, as expectativas quanto às respostas precisam ser claras para o entrevistado, ele deve ter toda informação necessária para a resposta e precisa estar disposto a responder.

d) Nas perguntas abertas, a principal vantagem é capturar um alto nível de detalhamento nas respostas dos entrevistados. A desvantagem é que cresce a dificuldade de interpretação de toda a variedade de respostas. Já nas perguntas fechadas, a vantagem é a facilidade na análise e interpretação devido a padronização de respostas e a principal desvantagem é não permitir captar todas as possíveis variações de posições dos respondentes.

e) As principais características são: variável nominal apresenta apenas igualdade ou equivalência entre as categorias; variável ordinal, além de igualdade, permite a identificação de uma relação de superioridade ou inferioridade entre as categorias; variável intervalar, apresenta igualdade, relação de superioridade e ainda permite operações aritméticas entre os valores; variável de razão, além de todas as características anteriores permite operações aritméticas com os próprios números e apresenta o zero absoluto.

f) O falso positivo acontece quando o entrevistado não tem uma resposta, mas decide escolher uma das opções apresentadas por considerar que não pode deixar sem responder. O falso negativo ocorre quando o entrevista tem uma resposta, porém, decide não apresentá-la por cansaço ou falta de confiança no entrevistador. Os possíveis tratamentos dependem das decisões de pesquisa. Informar as opções de respostas “Não Sei” e “Não Quero Responder” é uma estratégia para evitar o falso positivo, pois as alternativas são apresentadas ao entrevistado. No entanto, isso pode gerar o falso negativo em questionários muito longos ou com perguntas mal formuladas.

g) O pré-teste é a aplicação da primeira versão do questionário a um conjunto de respondentes com as mesmas características da população que será amostrada para a pesquisa. O objetivo é testar a compreensão dos enunciados das perguntas, abrangência das respostas, duração do questionário e possíveis respostas não consideradas antes.

Respostas às Questões do Capítulo 7

a) Modelos probabilísticos são usados para explicar a possibilidade de ocorrência de determinado resultado a partir de experimentos aleatórios. Tais modelos estão vinculados a algumas hipóteses a respeito dos resultados possíveis e de uma lei geral que mostra a probabilidade de ocorrência de cada resultado.

b) A curva de distribuição normal é aquela em que as distribuições dos valores de uma variável qualquer são aleatórias, pois variam ao acaso, com uma média e um desvio padrão definidos. Além disso, a maior parte dos valores encontra-se em torno da média, no centro, com distribuição dos casos equivalente para os dois lados da média. A distribuição apresenta a forma de uma curva em sino. Como a área total dentro da curva equivale à probabilidade um, a probabilidade de ocorrência de qualquer valor dentro da área da curva é menor que um.

c) A hipótese nula (H0) indica que as diferenças de médias são estatisticamente nulas, ou seja, são resultados de variações do acaso, meras oscilações, sendo representadas pelo sinal de igualdade. A hipótese alternativa (H1) só ocorre quando as diferenças de médias observadas não podem ser explicadas por oscilações do acaso, sendo consistente com diferenças estatisticamente significativas. A hipótese alternativa só ocorre depois de rejeitada a hipótese nula. Por isso o objetivo de todo teste de diferença de médias (NHST) é rejeitar a hipótese nula.

d) O erro de tipo 1 (alfa) ocorre quando se rejeita hipótese nula no teste de diferenças de

médias, sendo ela verdadeira. O limite crítico para evitar o erro de tipo 1 é chamado de nível de significância, ou p-*value*, ou alfa. O erro de tipo 2 (beta) ocorre quando se aceita a hipótese nula sendo ela falta na realidade. Não existe como identificar o erro de tipo 2. Por isso, o que se faz nos testes de diferença de médias é considerar o limite crítico para o erro de tipo 1 como uma garantia de que não se incorrerá no único tipo de erro que é conhecido.

e) O teste t aplicado à variável "faltas" dos alunos mostrou um coeficiente t=11,957 e um p-*value* bem abaixo do limite crítico de 0,050. Com isso, é possível rejeitar a hipótese nula de variações aleatórias e afirmar que a chance de errar é pequena se considerarmos que as variações entre as faltas dos alunos são estatisticamente significativas, ou seja, não seguem uma norma. Além do coeficiente t, o teste também mostra que o intervalo de confiança não passa por zero, com os valores superior e inferior do intervalo sendo positivos.

```
Rcmdr>  with(B_Exerc_V1, (t.test(faltas, alternative='two.sided',
mu=0.0, conf.level=.95)))
```

```
One Sample t-test
data:  faltas
t = 11.957, df = 59, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.538754 4.961246
sample estimates:
mean of x
     4.25
```

f) O teste t para amostras independentes visa verificar se as médias dos alunos que dizem ter mais interesse pela disciplina são estatisticamente distintas das médias daqueles que dizem ter menos interesse pela disciplina. Os resultados abaixo mostram que não é possível rejeitar a hipótese nula. O coeficiente t=1,036 é muito baixo e o nível de significância p-*value*=0,304 fica bem acima do limite crítico de 0,050. Isso significa que se rejeitar a hipótese nula corre-se um risco acima do aceitável de cometer erro do tipo 1 (alfa). Outro indicador é que o intervalo de confiança passa por zero, pois o limite inferior é -2,277 e o superior é +7,162. Ou seja, aquela diferença de médias de 75,80 para

o grupo do interesse principal e 73,35 de média para alunos com interesse secundário não é forte o bastante para apresentar significância estatística.

```
Rcmdr>  t.test(media~interess, alternative='two.sided', conf.
level=.95, var.equal=FALSE, data=B_Exerc_V1)
```

```
Welch Two Sample t-test

data:  media by interess
t = 1.0365, df = 56.857, p-value = 0.3044
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.277101  7.162815
sample estimates:
 mean in group principal mean in group secundário
                75.80000                 73.35714
```

g) Os resultados dos testes de ANOVA com comparações múltiplas de Tukey estão divididos em duas partes. Na primeira, a seguir, constam as estatísticas para a análise de variância entre as categorias de tempo de ocupação e as médias dos alunos. A estatística t=1,024 e p-*value*=0,389 mostram que não há diferença estatisticamente significativa entre o tempo de ocupação dos alunos e as médias deles na faculdade. Isso dispensaria a realização do teste de múltiplas comparações de Tukey, pois já sabemos que não haverá nenhum par de comparações com significância estatística. No entanto, o segundo conjunto de resultados está apresentado a seguir para demonstrar o fato.

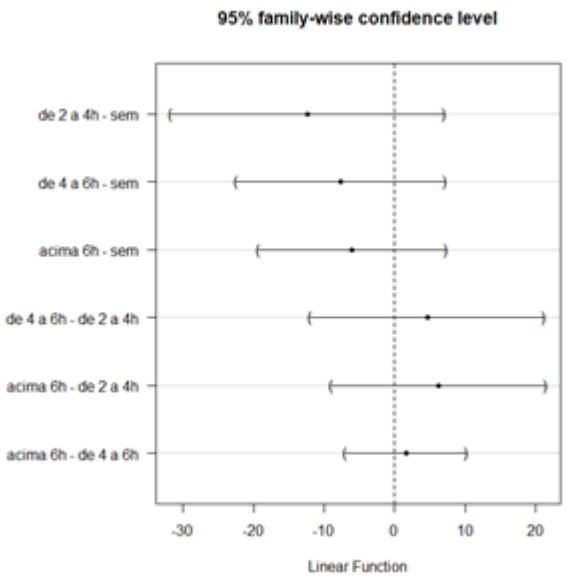
```
Rcmdr>  B_Exerc_V1 <- within(B_Exerc_V1, {tempocu_num <-
factor(tempocu_num, labels=c('sem','de 2 a 4h','de 4 a 6h','acima
6h'))})
Rcmdr>  ANOVAModel.3 <- aov(media ~ tempocu_num, data=B_Exerc_V1)
```

```
Rcmdr>  summary(ANOVAModel.3)
             Df Sum Sq Mean Sq F value Pr(>F)
tempocu_num   3    290   96.59   1.024  0.389
Residuals    56   5281   94.30
```

Nos resultados simultâneos a seguir, as diferenças estatísticas, p-*value*, ficam todas acima do limite crítico de 0,050. Todos os intervalos passam por zero. E ao final, o gráfico de intervalo de confiança mostra as distâncias entre os extremos de cada par de comparações, além do fato de todos passarem pelo zero. Assim, não é possível dizer que nenhuma das quatro categorias de ocupação apresenta impacto estatisticamente significativo sobre as médias dos alunos representados no banco de dados "B_Exerc_V1".)

```
Rcmdr>  local({.Pairs <- glht(ANOVAModel.3, linfct = mcp(tempocu_num
= "Tukey"))
Rcmdr+    print(summary(.Pairs)) # pairwise tests
Rcmdr+    print(confint(.Pairs)) # confidence intervals
Rcmdr+    print(cld(.Pairs)) # compact letter display
Rcmdr+    old.oma <- par(oma=c(0,5,0,0))
Rcmdr+    plot(confint(.Pairs)) par(old.oma)})
```

```
Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = media ~ tempocu_num, data = B_Exerc_V1)

Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
de 2 a 4h - sem == 0        -12.292      7.417  -1.657    0.337
de 4 a 6h - sem == 0         -7.670      5.670  -1.353    0.514
acima 6h - sem == 0          -6.042      5.081  -1.189    0.618
de 4 a 6h - de 2 a 4h == 0    4.621      6.325   0.731    0.876
acima 6h - de 2 a 4h == 0     6.250      5.803   1.077    0.689
acima 6h - de 4 a 6h == 0     1.629      3.289   0.495    0.957
(Adjusted p values reported -- single-step method)

       Simultaneous Confidence Intervals
Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = media ~ tempocu_num, data = B_Exerc_V1)

Quantile = 2.6124
95% family-wise confidence level

Linear Hypotheses:
                           Estimate lwr      upr
de 2 a 4h - sem == 0       -12.2917 -31.6666   7.0833
de 4 a 6h - sem == 0        -7.6705 -22.4820   7.1411
acima 6h - sem == 0         -6.0417 -19.3158   7.2325
de 4 a 6h - de 2 a 4h == 0   4.6212 -11.9018  21.1442
acima 6h - de 2 a 4h == 0    6.2500  -8.9101  21.4101
acima 6h - de 4 a 6h == 0    1.6288  -6.9633  10.2209

      sem de 2 a 4h de 4 a 6h  acima 6h
      "a"       "a"       "a"       "a"
```



h) O teste fatorial ANOVA para tipo de escola e tempo de ocupação para explicarem simultaneamente o número de faltas dos alunos mostra que nem as variáveis individuais, nem os fatores com efeitos simultâneos apresentam significância estatística para explicar as variações das faltas dos alunos. Todos os p-*value*s (Pr(>F)) ficam acima do limite crítico de 0,050, o que não permite rejeitar a hipótese nula. Além disso, curioso como o efeito de interação dos fatores (escola:tempo_ocu) apresenta uma estatística F (1,671) um pouco abaixo do efeito do fator "tempo_ocu" isolado (F=1,882).

```
Rcmdr>  ANOVAModel.4 <- lm(faltas ~ escola*tempocu_num, data=B_Exerc_V1,
Rcmdr+    contrasts=list(escola ="contr.Sum", tempocu_num ="contr.Sum"))
```

```
Rcmdr>  ANOVA(ANOVAModel.4)
ANOVA Table (Type II tests)

Response: faltas
                   Sum Sq Df F value Pr(>F)
escola              32.15  3  1.5985 0.2020
tempocu_num         37.86  3  1.8824 0.1451
escola:tempocu_num  56.03  5  1.6715 0.1596
Residuals          321.82 48

Rcmdr>  with(B_Exerc_V1, (tapply(faltas, list(escola, tempocu_num),
mean,
Rcmdr+    na.rm=TRUE))) # means
                    sem de 2 a 4h de 4 a 6h acima 6h
maior parte privada   7        NA         3 5.166667
maior parte pública  NA         5         2 5.363636
tudo privada          3        NA         4 4.666667
tudo pública         NA        10         3 2.937500

Rcmdr>  with(B_Exerc_V1, (tapply(faltas, list(escola, tempocu_num),
sd,
Rcmdr+    na.rm=TRUE))) # std. deviations
                         sem de 2 a 4h de 4 a 6h acima 6h
maior parte privada 1.414214        NA  2.000000 1.800673
maior parte pública       NA  1.414214  0.000000 2.500909
tudo privada        1.414214        NA  2.828427 2.309401
tudo pública              NA        NA  4.358899 3.151058

Rcmdr>  xtabs(~ escola + tempocu_num, data=B_Exerc_V1) # counts
                   tempocu_num
escola              sem de 2 a 4h de 4 a 6h acima 6h
  maior parte privada   2         0         4       12
  maior parte pública   0         2         2       11
  tudo privada          2         0         2        3
  tudo pública          0         1         3       16
```

i) O teste x^2 para a tabela de contingência entre “est_civil” e “turma” mostrou um coeficiente x^2 = 6,111 e p-*value* = 0,036, abaixo do limite crítico de 0,050. Isso significa que é possível rejeitar a hipótese nula e dizer que as variações de categorias de estado civil não são aleatórias entre as três turmas de alunos. A tabela de contingência do teste mostra que as turmas 3joad e 5joan apresenta um número maior de solteiros, enquanto a 3joan tem o mesmo número de solteiros e não solteiros. A partir dessas informações, é possível afirmar que as principais diferenças estatísticas estão na 3joan em relação às demais.

```
Rcmdr>  local({
Rcmdr+    .Table <- xtabs(~estcivil+turma, data=B_Exerc_V1)
Rcmdr+    cat("\nFrequency table:\n")
Rcmdr+    print(.Table)
Rcmdr+    .Test <- chisq.test(.Table, correct=FALSE)
Rcmdr+    print(.Test)})
```

```
Frequency table:
              turma
estcivil       3joad 3joan 5joan
  não solteiro     5     9     3
  solteiro        15     9    19

      Pearson's Chi-squared test

data:  .Table
X-squared = 6.6111, df = 2, p-value = 0.03668
```

Adendo II

Passo a passo para uso do *RCommander*

Este passo a passo tem o objetivo de apresentar o pacote estatístico *RCommander* (*Rcmdr*), do pacote R, em especial para o uso no material apresentado neste livro. O *Rcmdr* é uma interface que permite realizar os principais testes estatísticos básicos no R e apresenta como principal vantagem a simplicidade para solicitação dos testes. Ainda que apresente uma interface específica, é importante considerar que o *Rcmdr* segue a lógica de pacotes estatísticos que está na origem do R. A diferença é que normalmente os pacotes estatísticos do R são monotemáticos. O *Rcmdr* é uma interface que permite a realização de diferentes testes básicos, além de permitir a produção de tabelas e gráficos. Não é objetivo deste adendo fazer uma apresentação exaustiva do pacote *Rcmdr*. Os interessados devem buscar informações em outras fontes para aprofundar os conhecimentos sobre o pacote estatístico[1].

A apresentação a seguir está dividida em duas partes. A primeira, mais curta, mostra como instalar o pacote *Rcmdr* no ambiente R, carregar o pacote e abrir um banco de dados para ser analisado no *Rcmdr*. Em seguida, são apresentados como carregar o banco de dados para uso no *Rcmdr* e produção das tabelas, gráficos e testes apresentados no livro.

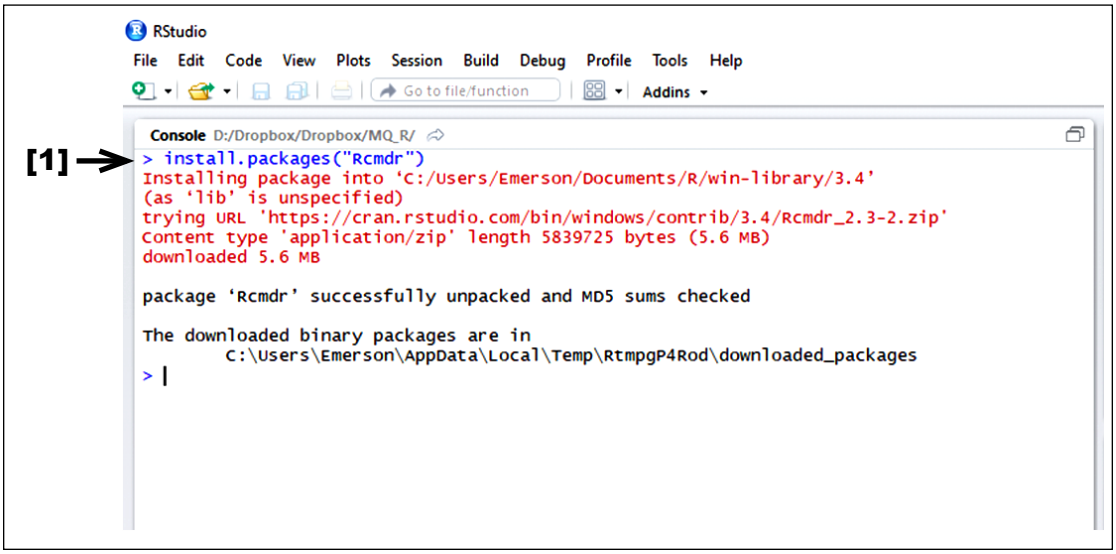
1 São sugeridas duas referências básicas sobre a interface Rcmdr:
FOX, J. **The RCommander**: **a basic statistics graphical user interface to R**, journal of statistical software, n. 14., v. 9, p. 1-42, 2005.
FREIJO, Jesús Bouso. **El paquete estadístico R**. Cuadernos metodológicos 48. Madrid: Centro de Investigaciones Sociologicas (CIS), 2013.

A) Instalação e carregamento de banco de dados no *Rcmdr*

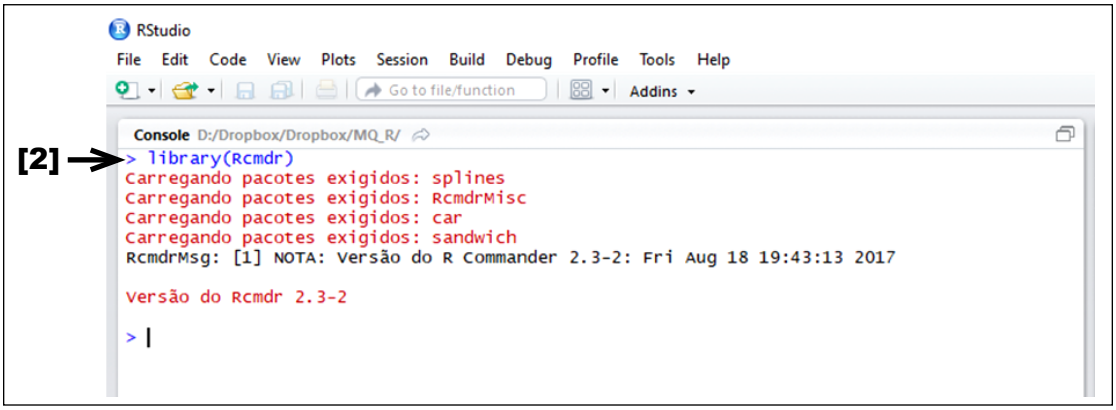
O *Rcmdr* é um pacote estatístico que roda em ambiente R ou *RStudio*. Aqui, será utilizado o *RStudio*. Em primeiro lugar, é preciso instalar o *RStudio* no computador. Isso é feito a partir do link:

https://www.*RStudio*.com/products/*RStudio*/download/

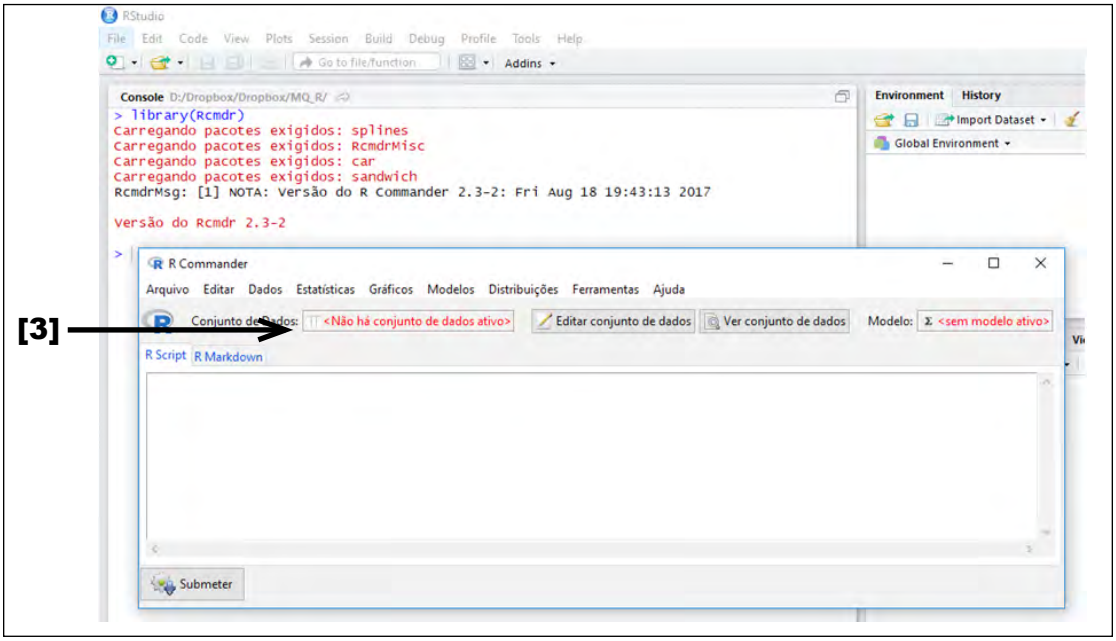
Uma vez instalado o *RStudio* e iniciado o programa, é preciso instalar o pacote *Rcmdr* no ambiente *RStudio*. Como indica a imagem abaixo, a instalação se faz no console do *RStudio* com o comando "install.package(*Rcmdr*)" **[1]**.



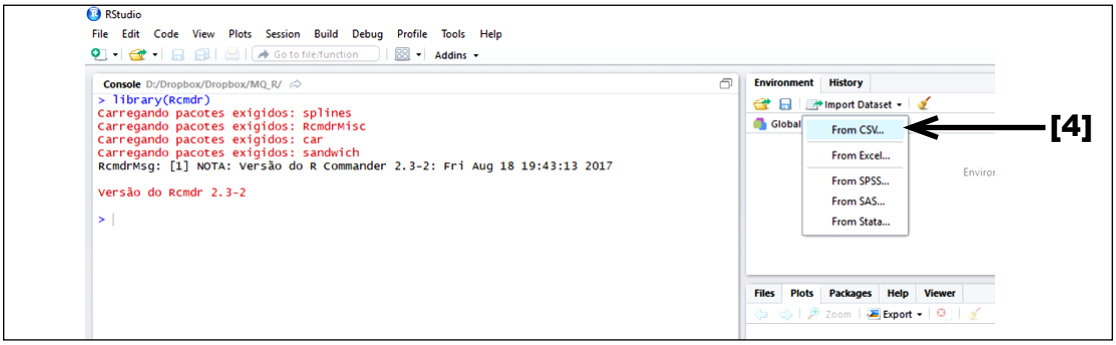
Depois de concluído o processo de instalação, o programa informa a conclusão do *download* e indica o local em que o pacote está instalado no computador. Uma vez concluída a instalação, é preciso carregar o pacote no ambiente de trabalho. Para isso, digita-se o comando "library(*Rcmdr*)" **[2]**. O *RStudio* informará todos os pacotes que foram carregados e, ao final, indicará a versão do *Rcmdr* que foi carregada no ambiente de trabalho, conforme indica a imagem a seguir.



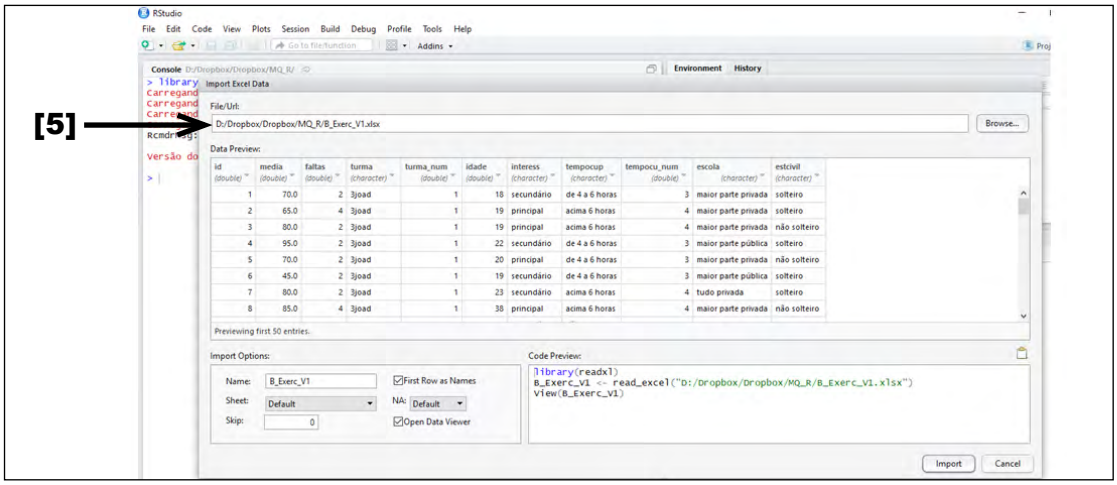
Assim que o carregamento for concluído, o pacote estatístico abrirá uma nova janela, a do *Rcmdr*. A janela do pacote é independente do *RStudio* e nela serão realizadas todas as solicitações de testes. Assim que é aberta, a janela do *Rcmdr* indica que não há nenhum banco de dados carregado nela: "não há conjunto de dados ativos" **[3]**. Para rodar os testes, é preciso carregar o banco de dados na área de trabalho do *RStudio* e depois ser inserido no *Rcmdr*.



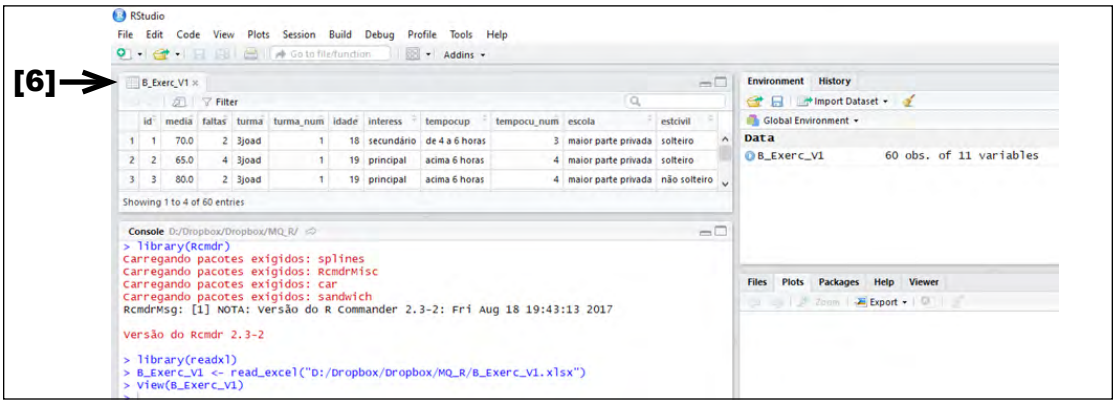
Para carregar um banco de dados no ambiente *RStudio,* a partir do "import dataset" se escolhe o formato do banco de dados que será importado. No nosso caso, o formado é excel (From .CSV) **[4]**. O banco é "B_Exerc_V1.xls".



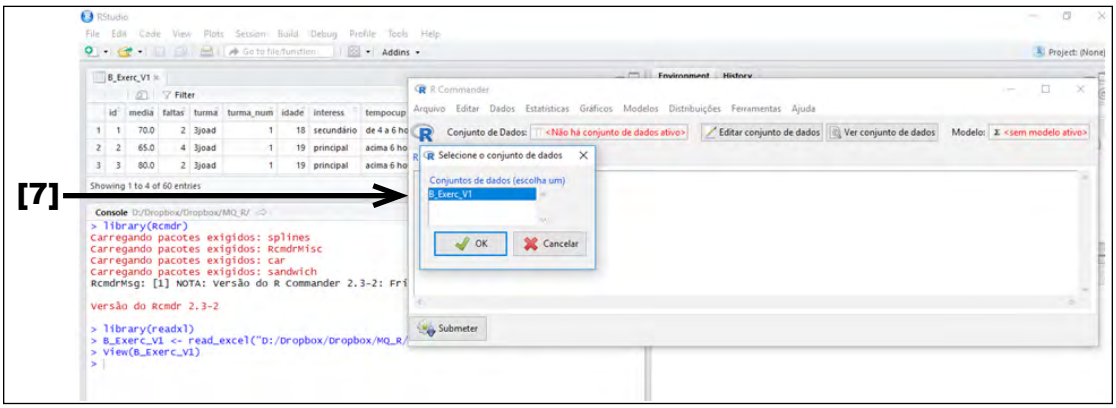
Após selecionar o formato do banco de dados a ser importado, é aberta uma janela de importação. É possível escolher o local/pasta **[5]** em que se encontra o arquivo com os dados. Uma vez selecionado o arquivo, ele será aberto na própria janela, com a apresentação das variáveis e com a linha de código em *preview*.



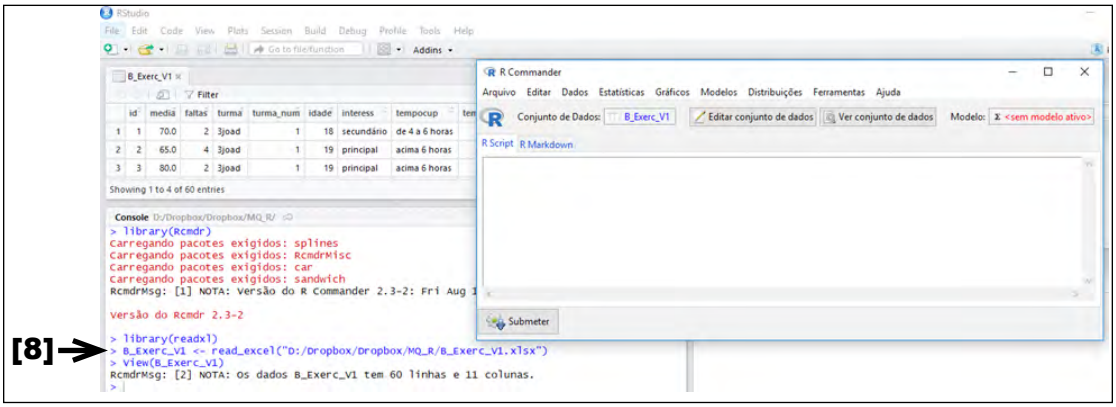
Assim que for escolhida a opção “import”, o banco de dados é importado para o ambiente de trabalho e a linha de códigos é impressa no console **[6]**. Tudo o que for feito no *Rcmdr* também será registrado no console do *RStudio*.



Agora que o banco de dados está carregado no ambiente de trabalho, devemos voltar para a janela do *Rcmdr* e clicar na caixa “não há conjunto de dados ativos”. Ele abrirá uma caixa de diálogo com todos os bancos de dados que estiverem carregados no ambiente de trabalho. No caso, temos apenas o “B_Exerc_V1” **[7]**. É preciso selecionar o nome do arquivo e clicar em ok.



Uma vez selecionado o banco de dados, o nome dele passa a aparecer na janela de dados ativos e tudo o que for processado no *Rcmdr* a partir de então será com os dados ativos. O *RStudio* registra no console que o banco de dados foi carregado corretamente **[8]**.

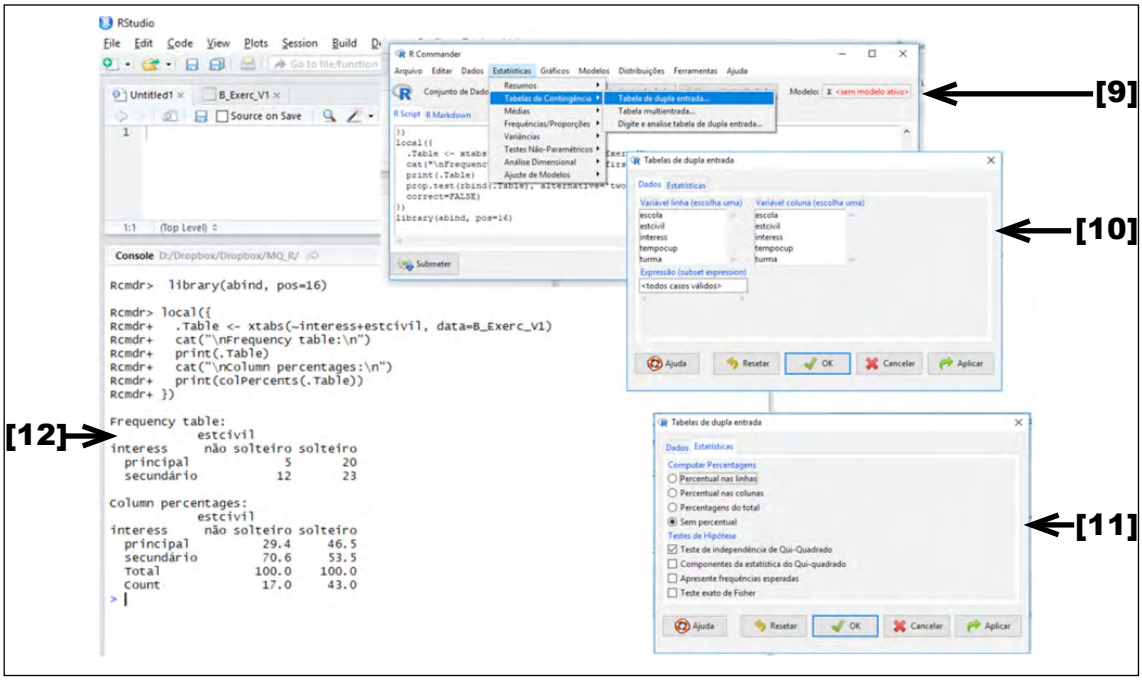


A partir daqui, os comandos para todas as operações serão realizados nos atalhos da linha superior da tela do *Rcmdr*. O console do *RStudio* servirá para o registro de todos os resultados dos comandos solicitados, seja para produção de tabelas, geração de gráficos ou realização de testes estatísticos.

B) Produção de Tabelas, Gráficos e Testes no *Rcmdr*

Os atalhos da linha superior da tela do *Rcmdr* (arquivo/ editar/ dados/ estatística/ gráficos/ modelos/ distribuições/ ferramentas/ ajuda) apresentam todas as opções que o pacote oferece ao usuário. Para usar além das necessidades deste livro, sugiro uma pesquisa aprofundada no manual do próprio pacote, disponível em: http://www.*RCommander*.com/.

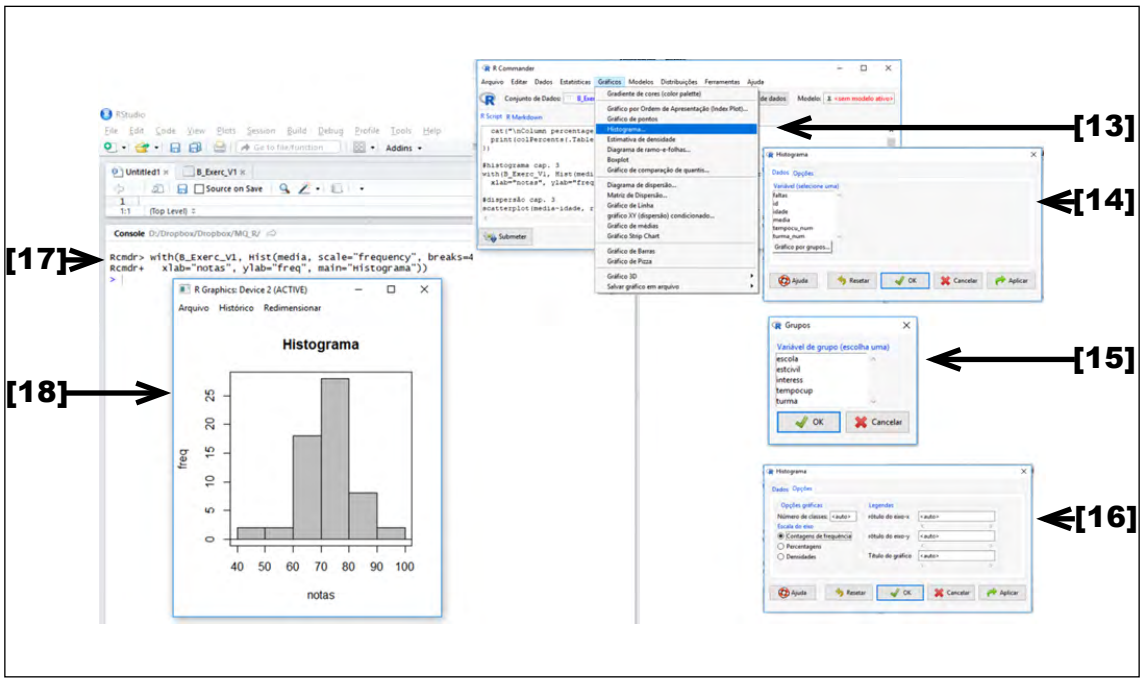
Seguindo a ordem de apresentação dos testes no livro, a primeira solicitação foi a de gerar uma tabela de cruzamentos entre as variáveis "est_civil" e "interesse". O caminho para abrir a caixa de solicitação da tabela de cruzamentos é: Tabela de Contingência / Tabela de dupla entrada **[9]**. Selecionada a opção tabela de dupla entrada, o programa abre uma caixa de diálogos em que apresenta duas listas de variáveis a serem selecionadas **[10]**. Uma para as linhas e outra para as colunas. Dentro dessa caixa é possível clicar em "estatísticas" **[11]**. Nela, é possível selecionar a apresentação de percentual na linha, na coluna, total ou sem percentual (que é o padrão). Também existe a possibilidade de selecionar três testes de hipóteses não paramétricos: teste de independência de chi-quadrado, teste de componentes da estatística de qui-quadrado, teste de Fisher. E é possível solicitar a apresentação de uma tabela de frequências esperadas.



Uma vez selecionadas as estatísticas desejadas, a seleção “ok” gera os resultados no console do *RStudio*. No caso, o exemplo apresenta a tabela dos valores reais e a tabela de percentuais na coluna **[12]**.

O livro apresenta alguns gráficos. O primeiro deles é o histograma da distribuição dos valores da variável “notas” dos alunos. No *Rcmdr* é: Gráficos / Histograma **[13]**.

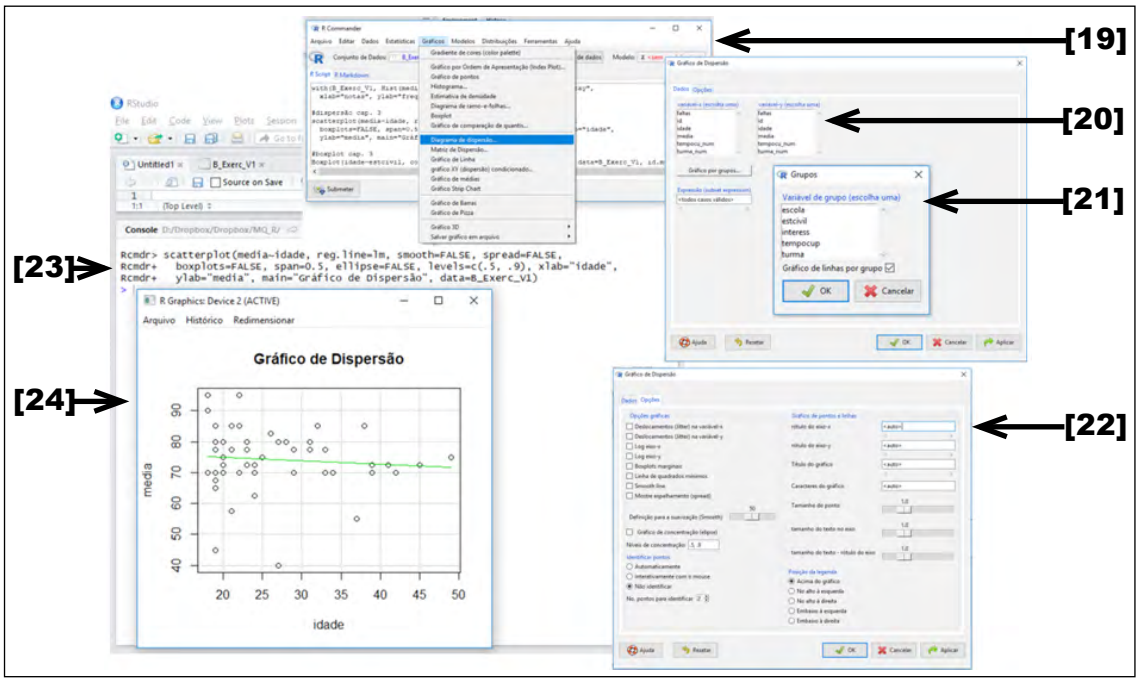
A partir daí, abre uma caixa de diálogos na aba “Dados” com as variáveis do banco de dados que podem gerar histogramas **[14]**. Nessa mesma aba, há uma opção de “gráfico por grupos”. Ao clicar nessa opção, abre uma caixa com as variáveis categóricas do banco **[15]**. Ao selecionar uma delas, será gerado um histograma para cada categoria da variável escolhida. Em uma segunda aba da caixa “Opções”, é possível escolher algumas características do histograma **[16]**. Pode-se selecionar o número de classes do gráfico, se a escala do eixo X será por frequência, percentagem ou densidade, além de digitar as legendas dos eixos X e Y e o título do gráfico.



Uma vez feitas todas as seleções e pressionado “ok”, os resultados produzidos são a linha de comando no console do *RStudio* e uma nova janela com o histograma **[17]**. A partir desta janela, é possível redimensionar o gráfico e, clicando com o botão direito sobre a imagem, copiar a imagem em formato “metalife” ou “bitmap” ou salvá-la em arquivo **[18]**.

O segundo gráfico apresentado no livro foi o gráfico de dispersão entre as variáveis contínuas “idade” dos alunos e “média” das notas deles. Para abrir a caixa de diálogos é preciso ir em: Gráficos / Diagrama de dispersão **[19]**.

A aba “Dados” da caixa de diálogo apresenta duas colunas com as variáveis contínuas do banco de dados. É preciso selecionar a variável que ocupará o eixo X e a variável do eixo Y **[20]**. Nessa aba, é possível selecionar o botão “gráfico por grupos” e indicar uma variável categórica que definirá os grupos de casos na imagem **[21]**. Na aba “Opções” é possível selecionar uma série de características do gráfico. Desde as opções gráficas para cada eixo, os rótulos de eixos e o título do gráfico, até a posição da legenda **[22]**.

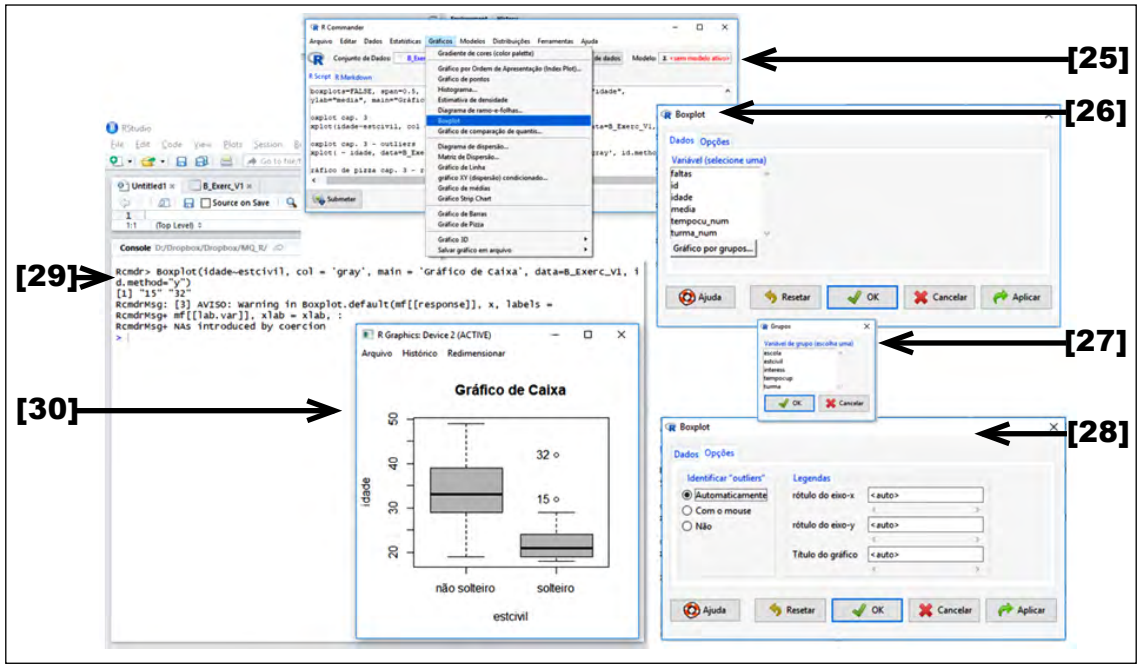


Feitas as opções e escolhido “ok”, a linha de comando é impressa no console do *RStudio* e abre-se uma nova janela com a imagem do gráfico de dispersão selecionado **[23]**. No caso da imagem, constam os *labels* dos eixos e o título do gráfico, além da linha de regressão (reg.line=lm) no comando do gráfico. Estes elementos adicionais devem ser digitados diretamente na linha de comando, caso não exista a possibilidade de optar por eles nas caixas de diálogo do *Rcmdr* **[24]**.

O gráfico boxplot é utilizado para identificar a distribuição de uma variável contínua a partir de uma categórica. Neste livro, foi utilizada a variável “idade” do aluno

como contínua e “est_civil” como categorias para visualizar se as distribuições das idades de alunos solteiros e não solteiros apresentavam diferenças. O caminho para pedir a produção do gráfico é: Gráficos / Boxplot **[25]**.

A partir daí, abre-se a caixa de diálogo. Na aba “Dados” com uma lista de variáveis contínuas para que seja selecionada aquela que será usada no boxplot **[26]**. Nessa caixa, existe a opção “gráfico por grupos”, que apresenta as variáveis categóricas do banco de dados **[27]**, aquelas que podem ser usadas para separar as distribuições da variável contínua. No exemplo do livro é usada a variável “est_civil”. Na aba “Opções” é possível escolher como identificar *outliers*, se de maneira automática ou manualmente (com mouse), inserir as legendas dos eixos e o título do gráfico **[28]**. Feitas as seleções, a opção “ok” determina a geração da imagem.



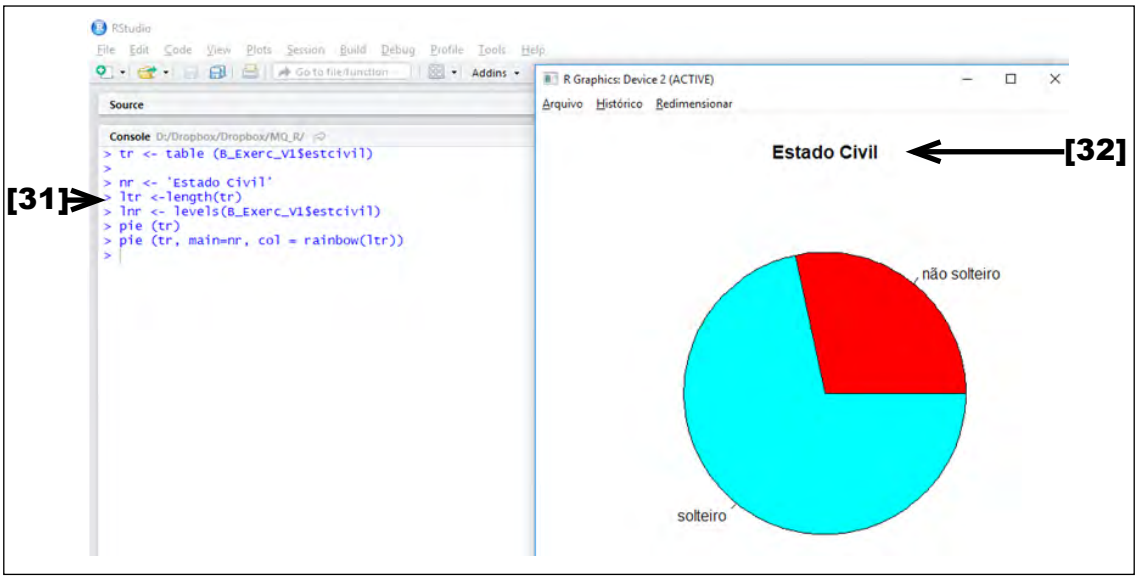
Assim como nos gráficos anteriores, é registrada a linha de comando **[29]** no console do *RStudio* e a imagem é aberta em uma janela própria **[30]**. No exemplo acima, é possível perceber que foi inserido na linha de comando a solicitação de cor cinza para as caixas (col = ‘gray’), o que não é possível fazer nas caixas do *Rcmdr*.

O último gráfico apresentado no livro é o gráfico de setores, também conhecido como gráfico de pizza. O caminho para fazer a solicitação é o mesmo: Gráfico / gráfico de pizza.

No entanto, para o livro, o gráfico de setores foi gerado digitando as linhas de comando no próprio *RStudio* para fins de comparação com o resultado final. As linhas de comando no console são:

```
> tr <- table (B_Exerc_V1$estcivil)
> nr <- 'Estado Civil'
> ltr <-length(tr)
> lnr <- levels(B_Exerc_V1$estcivil)
> pie (tr)
> pie (tr, main=nr, col = rainbow(ltr))
```
[31]

A primeira linha carrega o banco de dados e seleciona a variável "estcivil". As demais estabelecem as categorias a serem representadas em cada setor da imagem. A última indica que cada setor deve ser de uma cor distinta. O resultado está na imagem a seguir:

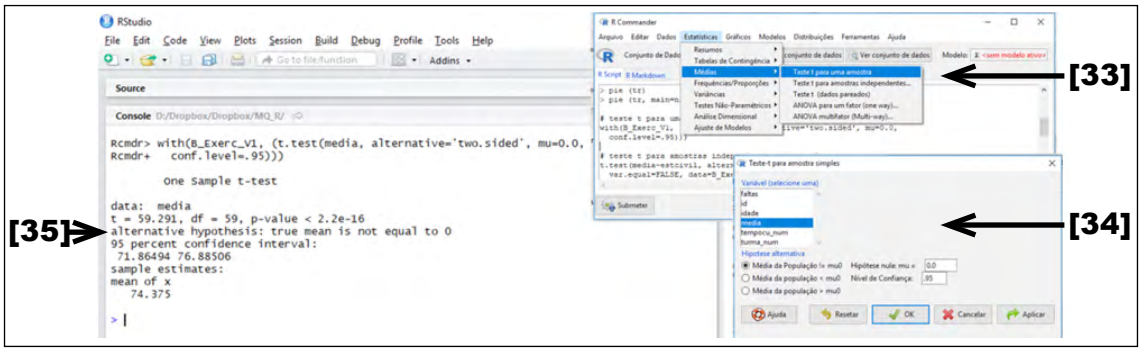


Assim como nos gráficos anteriores, no exemplo acima, a imagem é apresentada em uma janela própria **[32]**, com as características solicitadas na linha de comando digitada no console do *RStudio*.

Além dos gráficos, também usamos o *Rcmdr* para gerar os testes de hipóteses apresentados no livro. O primeiro deles foi o teste t para uma amostra. No capítulo 7 o banco de dados usado para exemplificar o teste não é o "B_Exerc_V1", que estamos usando aqui. Portanto, os resultados não serão iguais. Para mostrar como gerar o teste t para uma amostra, usamos aqui a variável "média" dos alunos no caminho: Estatísticas

/ Médias / Teste t para uma amostra **[33]**.

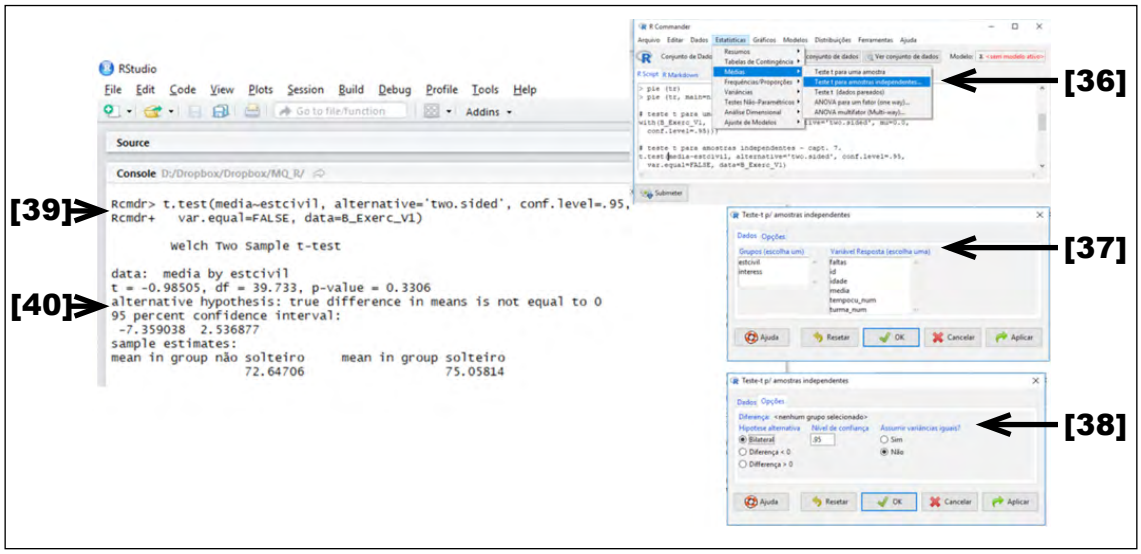
Ao clicar nela, abrirá uma única caixa de diálogo onde há uma lista de variáveis contínuas que podem ser testadas. Escolhe-se uma. Também é possível selecionar se a hipótese alternativa é que a média da população é diferente da média da amostra, é maior ou menor. Outra opção é o valor do intervalo de confiança. Por padrão, ele vem com 95% marcado **[34]**.



O resultado e a linha de comando são apresentados no console do *RStudio*. As informações da saída de resultados são o coeficiente "t", os graus de liberdade do teste "df", o p-*value*. Além dos valores de limites superior e inferior para o intervalo de confiança de 95% (71,86 a 76,88) e, por fim, a média da variável (74,375) **[35]**.

Outro é o teste t para amostras independentes, quando uma variável binária é usada como fator para testar as diferenças de médias de uma variável contínua. O caminho no *Rcmdr* é: Estatísticas / Médias / Teste t para amostras independentes **[36]**.

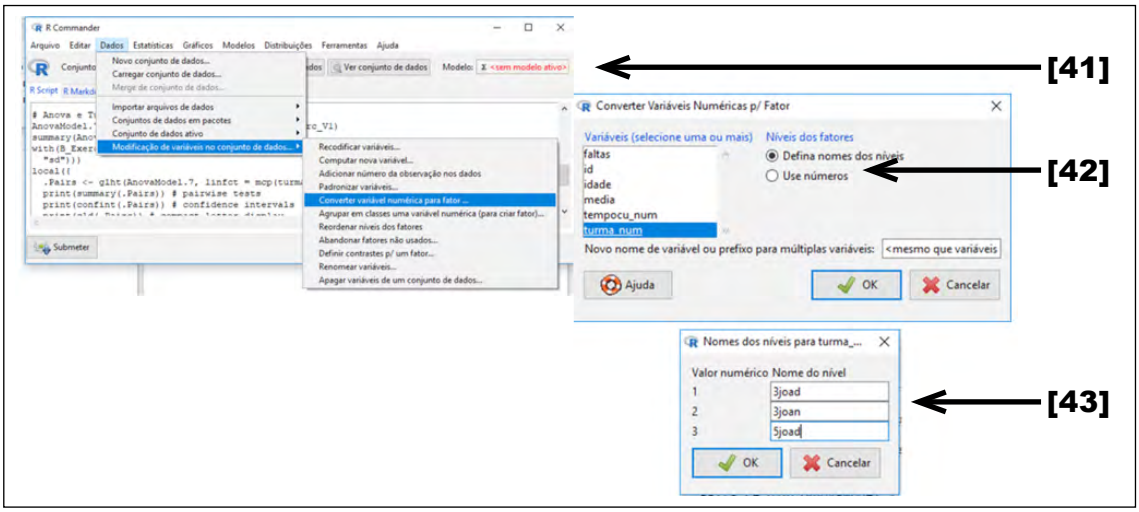
A caixa de diálogo em "Dados" abre duas listas. Uma de variáveis binárias para uma ser escolhida como "grupo" para o teste e outra com variáveis contínuas, para a seleção daquela cujas médias serão testadas **[37]**. No exemplo, a binária é a variável "estcivil" (solteiro ou não solteiro) e a contínua é "média". Com o teste, deseja-se saber se as diferenças de médias de idade entre solteiros e não solteiros apresentam significância estatística. A aba opções apresenta três conjuntos de alternativas: para definir se a hipótese alternativa é bilateral ou não, para definir o intervalo de confiança (o padrão é 95%) e para definir se o teste assumirá variâncias iguais ou não **[38]**. O padrão é não assumir mesmas variâncias.



A saída de resultados do teste t para diferenças de médias é muito parecida com a do teste anterior. Depois da linha de comando **[39]**, ele mostra o coeficiente "t", os graus de liberdade "df", o p-*value* e os valores do intervalo de confiança de 95%. Além disso, também são apresentadas as médias das notas para alunos não solteiros (72,64) e para solteiros (75,05) **[40]**.

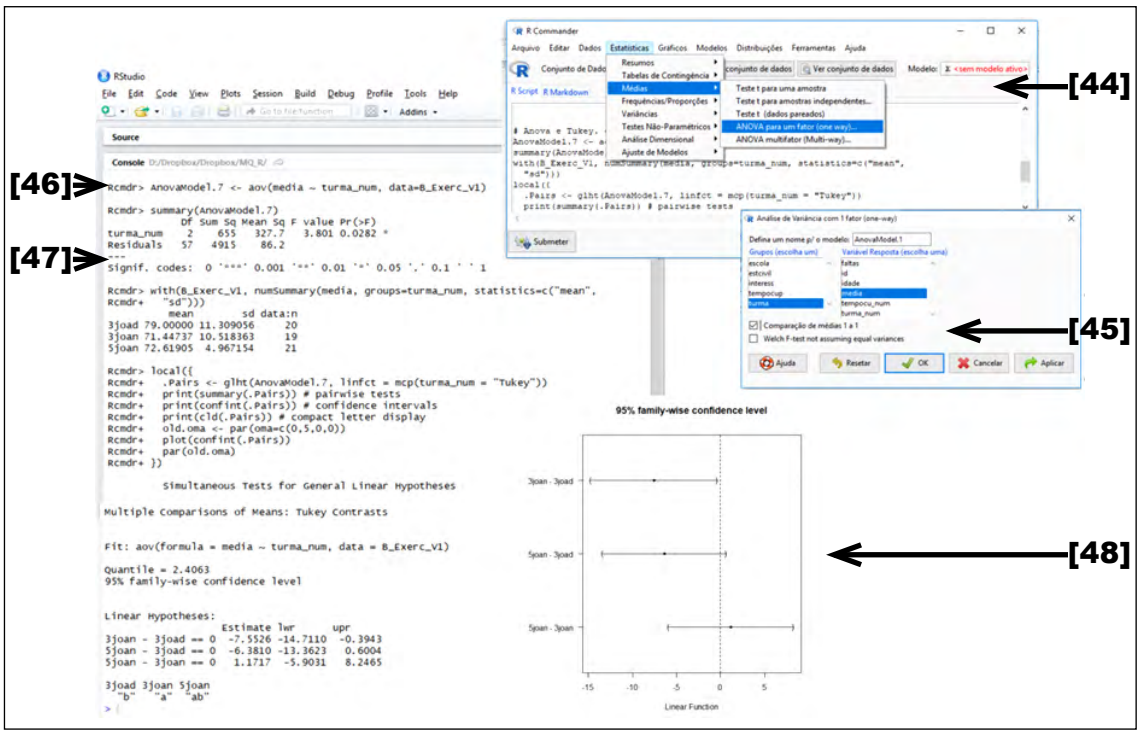
Outro teste de hipóteses, agora para fator com mais de duas categorias, é a análise de variância (ANOVA) e comparações múltiplas de ANOVA para identificar que categorias apresentam diferenças estatisticamente significativas. No exemplo aqui, usaremos a variável "turma_num", identificador da turma, com três categorias, para testar a variável "média" de notas dos alunos. O objetivo é identificar se existem diferenças estatisticamente significativas de médias entre as turmas e, caso existam, entre quais turmas essas diferenças são significativas. No *Rcmdr,* sempre que uma variável é usada como fator em um teste de hipóteses, ela precisa ser transformada em fator antes do início do teste. Essa transformação se faz seguindo o caminho: Dados/ converter variável numérica para fator **[41]**.

Será aberta uma caixa de diálogo com as variáveis do banco de dados que podem ser transformadas em fator **[42]**. No exemplo, é a variável "turma_num". Uma vez selecionada a variável e escolhido "defina nomes dos níveis", ao clicar em "ok" ela abre uma nova caixa para o preenchimento dos nomes das categorias. No caso, 1=3joad, 2=3joan e 3=5joad **[43]**. Após clicar "ok", a variável já estará transformada em fator e com as categorias com os respectivos nomes.



Transformada a variável em fator já é possível solicitar o teste de análise de variância ANOVA com comparações múltiplas no *Rcmdr*. Isso se faz a partir do seguinte caminho: Estatísticas / Médias / ANOA para um fator (*one way*) **[44]**.

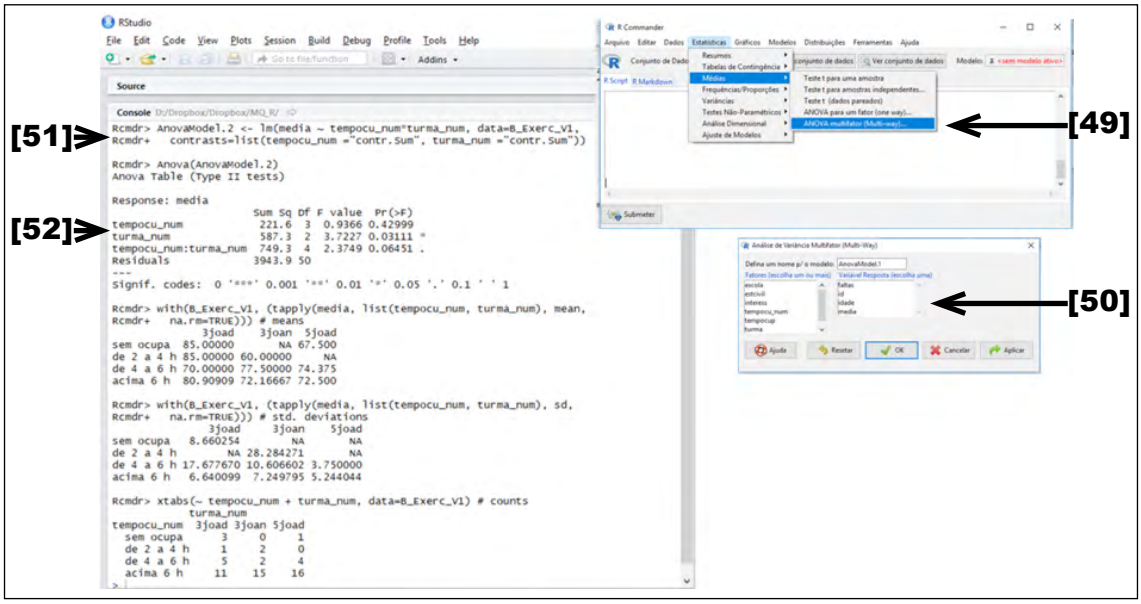
Será aberta uma caixa de diálogo na qual consta uma lista de variáveis categóricas, inclusive a(s) transformada(s) em fator para o teste, e, ao lado, outra lista com variáveis contínuas, a serem testadas. Abaixo das duas listas aparece a opção "Comparação de médias 1 a 1" **[45]**. Este é o teste de comparações múltiplas de Tukey. É preciso selecionar essa opção e clicar em "ok".



Os resultados são apresentados, junto com as linhas de comando, no console do *RStudio* **[46]**. Os primeiros resultados são as sumarizações de valores, graus de liberdade, suma dos quadrados e média quadrada. O mais importante é a estatística F e o p-*value* (Pr(>F). Abaixo, há indicação do nível de significância do teste. Como ele se mostrou significante a 0,01 (0,0282*), a conclusão é que as diferenças são estatisticamente significativas pelo menos entre uma turma com as demais **[47]**. Na sequência, aparecem as médias de notas, o desvio padrão e o número de alunos em cada turma. O teste de comparações simultâneas vem a seguir, com comparações entre pares. As estimativas para cada par indicam se as diferenças são estatisticamente significativas para aqueles dois casos. A maneira mais direta de identificar isso é olhando para os limites inferior e superior em cada linha (lwr / upr). Se o intervalo passar por zero é porque as diferenças de médias para aquele par não são estatisticamente significativas. No exemplo, na comparação 3joan-3joad, o intervalo não passa por zero; 3joan-3joad, passa um pouco por zero; e 5-joan-3joan, passa por zero, o que permite concluir que apenas a diferença de médias entre as turmas 3joan e 3joad é estatisticamente significativa. Além dos limites numéricos, o resultado também apresenta um gráfico com os intervalos de confiança dos pares em janela separada **[48]**. Esse gráfico mostra se os intervalos dos pares passam ou não por zero, permitindo a decisão de rejeição da hipótese nula apenas para a relação entre duas das três categorias e não entre todas elas.

Outro teste de análise de variância é a Análise Fatorial com ANOVA, que permite a utilização de mais de uma variável como fator e ainda verifica os efeitos individuais e integrados dos fatores. Aqui, no exemplo, usaremos as variáveis “turma”, com 3 categorias” e “tempocu_num” com 4 categorias para identificar os efeitos individuais e em conjunto das duas para as variações das “médias” dos alunos. Antes de selecionar as variáveis categóricas, é preciso transformá-las em fatores no *Rcmdr*. O procedimento é o mesmo que o descrito anteriormente para o teste ANOVA com comparações múltiplas. Com a diferença de que aqui transformaremos em fatores duas variáveis categóricas e não apenas uma. Faz-se isso selecionando cada variável e dando os nomes separadamente. Uma vez criado os fatores que serão usados no teste, o caminho para a Análise Fatorial com ANOVA no *Rcmdr* é: Estatísticas / Médias / ANOVA multifactor (multi-*way*) **[49]**.

Será aberta a caixa de diálogo em que se pode definir o nome para o modelo e, em seguida, escolher as variáveis que entrarão no modelo. A primeira lista é das variáveis que podem ser selecionadas como fatores, as categóricas. Aqui, devem-se selecionar duas ou mais variáveis. A segunda lista é das variáveis contínuas, para testar as médias. Aqui, seleciona-se apenas uma variável para ser a teste **[50]**.

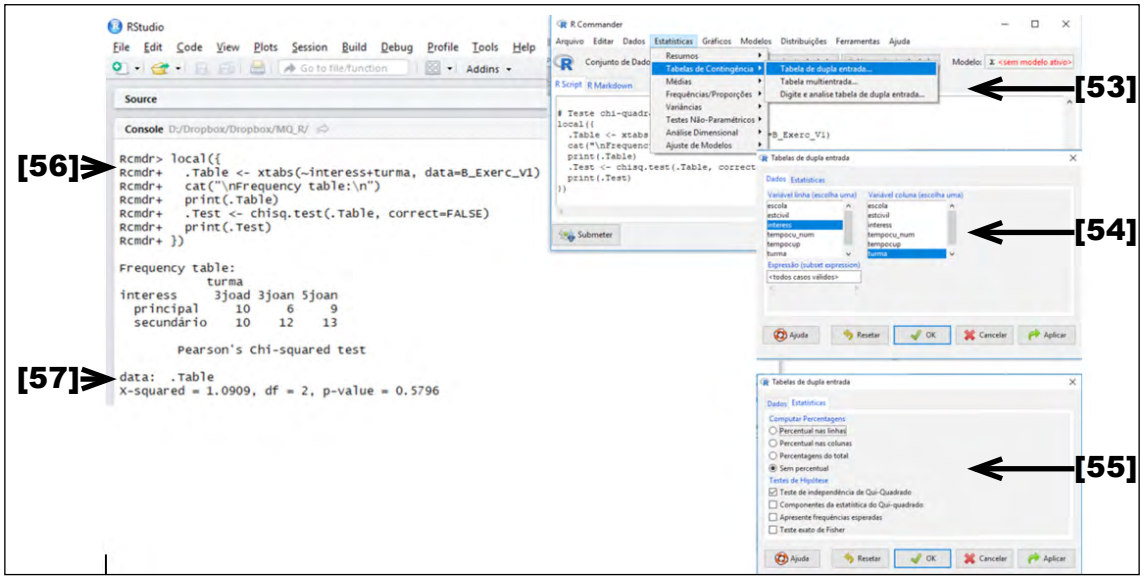


A linha de comando e os resultados da saída do teste são impressos no console do *RStudio* **[51]**. O primeiro conjunto de dados mostra o impacto individual e conjunto de cada variável teste sobre as variações das médias dos alunos. O tempo de ocupação (tempocu_num) não apresenta efeito significativo sobre as variações das notas (F=0,936 e p-*value*=0,429). Já a variável turma (turma_num) sim, apresenta efeito estatisticamente significativo para intervalo de confiança de 99%, com F=3,722 e p-*value*=0,0311 nas médias dos alunos. Já o efeito conjunto da interação entre tempo de ocupação e turma fica entre os dois anteriores, com F=2,374 e p-*value* de 0,064 **[52]**. O teste indica um nível de significância para 95%, embora o valor pontual do coeficiente tenha passado um pouco desse limite. Ou seja, como tempo de ocupação não tem influência sobre a média, quando ele interage com a turma, reduz o efeito desta variável sobre as variações de médias. Além dos coeficientes do teste, a saída também fornece as médias de notas para os cruzamentos das categorias e o desvio padrão para cada casa da tabela.

A última tabela indica o número de casos em cada casa, permitindo identifica a existência de casas com poucos casos, o que reduz a confiabilidade do teste.

O último teste apresentado no capítulo VII deste livro foi o teste de hipóteses não paramétrico x^2 de Pearson, onde não há presença de nenhuma variável contínua, portanto, não sendo possível parametrizar as diferenças por uma média e desvio padrão. Para o exemplo aqui, usaremos um cruzamento entre a variável interesse do aluno pela disciplina (interess) e a variável turma que o aluno pertence (turma). O objetivo é verificar se há diferenças estatísticas para o nível de interesse dos alunos entre as turmas. O procedimento para o teste x^2 de Pearson é o mesmo que para solicitar uma tabela de cruzamentos: Estatísticas / Tabelas de contingência / Tabela de dupla entrada **[53]**.

Na caixa de diálogo, na aba “Dados”, são apresentadas duas listas onde se deve escolher em cada uma delas a variável que entrará nas linhas e nas colunas da tabela de contingência **[54]**. Na aba “Estatísticas”, a opção teste de independência de qui-quadrado já está marcada por padrão **[55]**. Manter e pressionar “ok”.



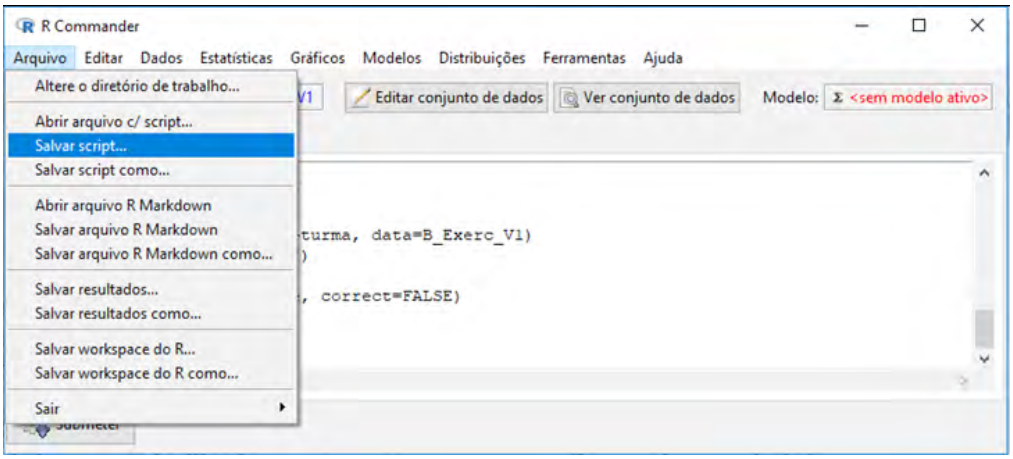
A linha de comandos e os resultados são apresentados na saída no console do *RStudio* **[56]**. Além da tabela com as frequências dos cruzamentos apenas, já que não solicitamos nenhum percentual, há uma última linha com informações do teste de hipóteses x^2 de Pearson. É apresentado o coeficiente x^2 = 1,090, o número de graus de liberdade da tabela (df=2) e o p-*value*=0,579 **[57]**. Como o p-*value* fica muito acima do

limite crítico padrão de 95% de intervalo de confiança, não podemos rejeitar a hipótese nula. Ou seja, não há diferenças estatisticamente significativas entre as turmas no que diz respeito ao nível de interesse pela disciplina.

Por fim, o *Rcmdr* permite que sejam arquivadas, na forma de script, todas as linhas de códigos registradas na janela do pacote estatístico. Para tanto, basta seguir o atalho: Arquivo / Salvar script ou Salvar script como **[58]**.

Na primeira vez, é preciso “salvar script como” para nominar um novo script. A partir de então basta salvar script.

[58]➢



Ainda é possível salvar as linhas de código como R Markdown ou salvar resultados. Neste último caso o que se salva são as saídas dos testes realizados na sessão de *Rcmdr*. Na próxima sessão, basta selecionar “arquivo / abrir arquivo c/ script” e selecionar o script salvo na sessão anterior para dar continuidade ao trabalho e manter toda a memória dos testes realizados em um único arquivo de script do *Rcmdr*.

Copyright @ 2017 do autor

Agência Brasileira do ISBN
ISBN 978-85-915195-4-5

www.cpop.ufpr.br www.cienciapolitica.ufpr.br

ISBN 978-85-915195-4-5