Metodologia para a Investigação Social

Procurámos nos vários capítulos deste livro, em primeiro [illegible] os caminhos que uma investigação desta índole deverá [illegible] componentes e propósitos foram consecutivamente detalhados em consideração a complementaridade e a lógica subjacente [illegible] partes que compõe o *puzzle* intelectual que constitui uma [illegible]

Em segundo lugar, procurámos exemplificar como cada uma destas partes deverá ser compreendida e executada face ao conjunto ao qual deveremos ter como referência num projecto de investigação. Nesta perspectiva o leitor é convidado desde logo a pensar sobre as componentes basilares da ideia que pretende executar, sendo que a exequibilidade e validade das suas opções necessitam ser equacionadas e maturadas em moldes similares àqueles que são aqui propostos.

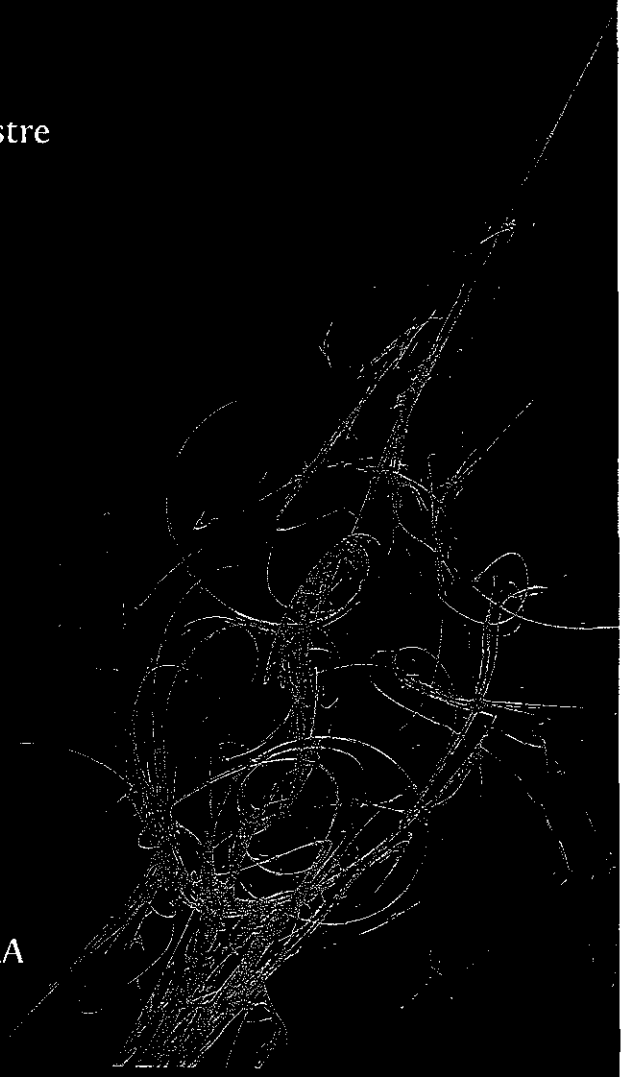
ESCOLAR EDITORA
http://www.escolareditora.com

ISBN 978-972-592-923-0
9 789725 923290

Metodologia para a Investigação Social

Metodologia para a Investigação Social

Coordenação
Hugo Consciência Silvestre
Joaquim Filipe Araújo



ESCOLAR EDITORA

CAPÍTULO IX

O TRATAMENTO E ANÁLISE DE DADOS

Miguel Ângelo Vilela Rodrigues

Objectivos do capítulo: este capítulo tem como principal objectivo auxiliar os investigadores na determinação do tipo de análise e na espécie de mecanismos de tratamento e análise de dados que melhor se adequam à investigação que se deseja levar a cabo. Por outras palavras, até ao momento, o leitor foi convidado a reflectir sobre uma pergunta de investigação. Tudo começou com a definição de um problema que deverá ser suficientemente forte e relevante de maneira a servir de suporte e justificação à investigação decorrente. Do problema decorre uma pergunta de investigação que visa possibilitar a enumeração de hipóteses, com o propósito de serem testadas de maneira a resolver o problema enunciado. Ora, facilmente compreendemos que o tratamento e a análise de dados se revelam de extrema importância para qualquer tipo de investigação. Numa perspectiva dedutiva, permite-nos aferir a veracidade das hipóteses definidas, enquanto, se adoptarmos uma estratégia indutiva, é a ferramenta utilizada na tentativa de generalização das explicações encontradas.

Palavras-chave: recolha de dados, tratamento de dados, tipos de estudos, estatística descritiva, inferência estatística.

1. Introdução: Estratégias Qualitativas e Quantitativas no tratamento e análise de dados

São várias as ferramentas que estão ao dispor do investigador para que este proceda à sua investigação. No entanto, é necessário que se tenha o devido cuidado para seleccionar o método mais indicado no tratamento e análise de dados, face

à estratégia desenhada, bem como ao paradigma assumido. Uma abordagem mais quantitativa usará técnicas relacionadas com o tratamento de um grande número de variáveis e de observações. Terá a necessidade de fazer uma análise focalizada na procura de padrões de relacionamento em variáveis, ou relações de causalidade entre uma variável dependente e (diversas) variáveis independentes. Pelo contrário, uma abordagem mais qualitativa procurará usar técnicas que lhe permitam ter uma percepção mais completa de uma realidade mais restrita. Ou seja, não usando um universo tão vasto como o usado pela abordagens quantitativas, este paradigma de investigação pretende absorver, ao máximo, os valores, crenças e processos do facto social em análise, de maneira a dotar o investigador da visão do mundo através da perspectiva dos actores que visa estudar.

Assim, independentemente da estratégia de investigação adoptada (qualitativa vs. quantitativa) a questão que se põe a este nível gira em torno de saber qual o melhor tratamento que se pode dar às hipóteses definidas, que tipo de informação recolher, que técnicas usar, e sobretudo, como testar estas hipóteses.

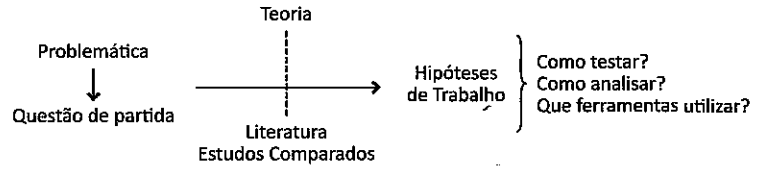


Figura 9.1. O Tratamento e análise de dados

Aquilo que cada investigador procura como ferramenta de tratamento e análise de dados depende muito da natureza e objectivos da investigação que se esteja a realizar. Desta forma, uma análise quantitativa procurará operacionalizar conceitos, estabelecer relações de causalidade, generalizar as conclusões do seu estudo à população e permitir que o estudo realizado seja passível de ser reproduzido. Por sua vez, a abordagem qualitativa centrará a sua atenção na análise exaustiva do fenómeno social e na acumulação de informação que permitirá a generalização empírica das conclusões obtidas.

A operacionalização de conceitos é um aspecto fundamental numa estratégia de investigação. Esta abordagem centra-se numa perspectiva de construção e teste de hipóteses com o objectivo de responder à pergunta de investigação formulada. Os

conceitos desempenham aqui um papel nuclear: na pergunta temos genericamente uma relação entre, pelo menos, dois conceitos; e na(s) hipótese(s) temos a utilização dos mesmos, numa assumpção passível de ser testada. Assim, a operacionalização de conceitos revela-se uma tarefa da maior importância, uma vez que, caso não seja bem feita, poderá pôr em causa todo o trabalho do investigador. Neste ponto, trata-se de reunir um conjunto de indicadores que sejam capazes de preencher, na sua plenitude e extensão, o conceito que queremos utilizar. O problema a este nível está na decisão do investigador em escolher a forma mais adequada de como vai medir o conceito que pretende utilizar.

Exemplo 1

Um conceito bastante utilizado nas Ciências Económicas é o de desenvolvimento económico. Por vezes, existe a tendência para o confundir com crescimento económico, pese embora o primeiro seja muito mais abrangente.

Assim, para operacionalizar o conceito de desenvolvimento económico, poderemos usar como indicadores:

- Número de médicos *per capita*;
- PIB *per capita*;
- Taxa de Alfabetismo;
- Número de diplomados *per capita*;
- Esperança média de vida;
- Etc....

Obviamente, o conceito poderá ser operacionalizado com outros indicadores, tudo dependerá da revisão de literatura feita, bem como dos próprios objectivos do investigador.

O passo seguinte é o de relacionar os conceitos enumerados nas hipóteses, ou seja, evidenciar que existe uma determinada conexão entre ambas. Em cada caso, o investigador deve ter o cuidado de tentar estabelecer relações de associação, dissociação, dependência ou causalidade entre os conceitos. A sua investigação vale pela capacidade explicativa demonstrada. Assim, o investigador deverá assegurar-se que existe uma relação de causalidade/associação existe entre as variáveis seleccionadas e o facto que pretende estudar. O maior mal para o investigador, e consequentemente para o seu estudo, vem do facto das suas variáveis demonstrarem parco poder expli-

cativo. Por isso, o investigador deve assegurar-se que, na sua investigação, não está a ignorar variáveis explicativas[1]. Outro problema frequente é a situação contrária. Isto é, a ânsia de evitar deixar de fora variáveis explicativas, criar situações onde estas se sobreponham, caracterizando a mesma realidade[2]. Portanto, o investigador deve usar, da melhor forma, a revisão de literatura para identificar as variáveis que deve incluir na sua análise.

Exemplo 2

Imagine que um professor deseja testar o uso de novas práticas pedagógicas com os seus alunos. Para tal, selecciona voluntários para integrarem um grupo de intervenção reservando os restantes alunos para as aulas mais convencionais. Obtendo resultados académicos mais favoráveis no grupo onde aplicou as práticas mais inovadoras, não constituiu razão suficiente para determinar que estas tenham sido a verdadeira e maior causa do sucesso escolar. Podem ser avançadas um conjunto de explicações alternativas: pode-se dar o caso de se terem oferecido como voluntários os melhores alunos; pode o grupo, pelo simples facto de saber que estaria a ser monitorizado, ter alterado o seu comportamento para fazer mais e melhor, etc.

Portanto, estabelecer relações de causalidade é um sério desafio para investigadores em ciências sociais.

Após enunciar o problema, definirmos a pergunta de investigação, bem como as hipótese decorrentes, recolhermos os dados necessários e procedermos ao respectivo tratamento, resta ao investigador promover a generalização dos resultados. Esta fase é um dos maiores desafios postos ao investigador. Numa abordagem quantitativa, o investigador procura que os dados conseguidos e as informações produzidas ultrapassem os limites, no espaço e no tempo, da população em que se realizou a investigação. Esta generalização só será plenamente conseguida se o estudo for passível de reaplicação por outros investigadores, em contextos semelhantes, eliminando assim a suspeita de subjectividade na investigação.

Em alternativa, uma perspectiva qualitativa procura a generalização de outra forma. Procura, por seu lado, assumir uma posição de observador participante ou não, procurando absorver a totalidade da realidade em causa, compreender o processo social e definir novas abordagens teóricas.

Usando este tipo de paradigma, a preocupação primária é a de entender os factos sociais como eles são percebidos pelas pessoas que os vivem. Para tal, o investigador

desenvolve formas de adquirir as próprias construções da realidade dos actores investigados, bem como dos seus mecanismos de percepção da realidade. Desta forma, o investigador procura compreender e absorver a plenitude dos conceitos bem como a linguagem usada pelos actores intervenientes. A recolha de dados tem um substrato mais conceptual quando comparada com a abordagem qualitativa. Há uma profunda preocupação em acumular dados com o máximo de pormenor e exactidão possível.

Exemplo 3

Imagine um investigador europeu que procura compreender os comportamentos e padrões sociais de um país islâmico. Tal entendimento, numa perspectiva qualitativa, não poderá resultar do simples tratamento por inquérito. Segundo esta abordagem, o investigador, à distância, não consegue compreender os factos sociais relevantes na cultura islâmica. Não absorve a sua riqueza. Para tal, necessita de ter contacto directo com os actores sociais, participar no seu dia-a-dia. Só assim, conseguirá compreender os processos e padrões comportamentais, isto porque fará uma análise por dentro dos princípios e valores vigentes e aceites.

Portanto, como podemos verificar, a estratégia qualitativa tem um entendimento diferente das relações de causalidade que necessariamente têm de ser estabelecidas entre conceitos. Nesta abordagem, não aparecem pré-definidas à partida. Surgem normalmente após a acumulação e tratamento de dados. Neste ponto de vista, o método quantitativo é mais flexível do que o quantitativo, permitindo a natural construção de relações durante o período de recolha e tratamento de dados. Como sugerido por Blaikie (2003) e Bryman (2008), a teoria resulta, numa posição qualitativa, nas relações conceptuais estabelecidas graças à acumulação e tratamento exaustivo de dados.

2. Os diferentes tipos de estudos

Todos os estudos têm uma referência e uma relação natural com o momento histórico em que são realizados. Podem ser feitos no momento actual, comparando momentos diferentes, analisando series temporais ou tentando extrapolar tendências para o futuro. O modelo clássico pressupõe um estudo comparativo em dois momentos diferentes, de maneira a comparar e aferir as diferenças provocadas por um processo de mudança (Bryman, 2008). Assim, o processo de análise implica a existência de

um momento inicial, onde se identificam dois grupos: grupo de intervenção e outro grupo de controlo. A análise de dados, que ocorre após a acção do investigador, destina-se a comparar as modificações ocorridas no grupo de intervenção em relação ao sucedido no grupo de controlo, que se isola da acção do investigador.

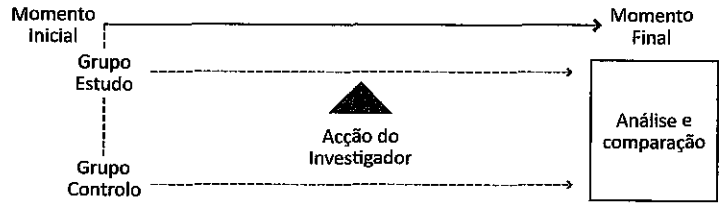


Figura 9.2. O Tratamento e análise de dados

Actualmente, podem ser destacados e identificados outros tipos de estudo, em função do horizonte temporal, estamo-nos a referir a estudos do tipo: *estudo transversal*, *análise temporal*, etc. — *vide* capítulo VI.

Os estudos tipo *cross-section* têm por objectivo captar o momento actual. Isto é, adequam-se a uma estratégia onde o investigador pretende retratar, analisar e dar a conhecer uma realidade social. Há uma preocupação clara em descrever o problema em causa, e estabelecer relações com fenómenos sociais, demográficos, económicos ou políticos, entre outros. O investigador não procura fazer qualquer tipo de avaliação ou de juízo de valor acerca de um processo de mudança, procura sim apresentar uma imagem fiel e apropriada da realidade que está a investigar.

Num estudo longitudinal há uma intenção do investigador em acumular observações de vários anos para estudar e analisar a evolução dos factos em causa, ou seja, ao contrário das investigações do tipo *cross-sectional*, existe uma preocupação pelos factores de mudança ou pelos condicionalismos que conduziram a uma determinada realidade. Este tipo de estudos pode ser ainda dividido em função da sua natureza: avaliação e prospecção da mudança. No primeiro caso, há bastantes semelhanças com o método clássico discutido há pouco. O estudo tem em consideração dois momentos distintos de maneira a permitir que se faça uma avaliação das mudanças ocorridas. É portanto uma estratégia adequada quando pretendemos verificar o impacto de um facto social. O segundo tem uma natureza mais direccionada para a percepção de acontecimentos passados ou mesmo, usando esta fase retrospectiva,

poder servir de base de trabalho para prever acontecimentos futuros. Ou seja, a postura do investigador poderá ser a de querer analisar as mudanças ocorridas num dado momento da história, ou, em alternativa, usar dados relativos a series temporais para poder estabelecer um prognóstico relativamente a uma realidade social futura.

Pelo acumular de uma enorme quantidade de dados, e pelo facto de ter uma maior exactidão na observação, os estudos longitudinais têm maior capacidade do que os *cross-sectional*, no entanto, também podemos entender este tipo de estudos como sendo um mero somatório de vários estudos *cross-section*.

Exemplo 4

Cross-section

Exemplo 1: Uma investigação que vise retratar e analisar a forma como as Autarquias Locais colaboram e/ou se associam para prestarem serviços públicos às suas populações.

Exemplo 2: Uma investigação que vise analisar o actual estado das finanças públicas dos diferentes estados membros da União Europeia.

Longitudinal

Impacto/Exemplo 1: Uma investigação que avalie o impacto que uma crise económico-financeira poderá ter na decisão e realização de projectos de investimentos públicos.

Mudança/Retrospectiva: Uma investigação que pretenda apurar os condicionalismos que levaram à eclosão da revolução francesa.

Mudança/Prospectiva: Uma investigação que, usando os registos médicos das populações, tenta prever o comportamento de um vírus.

3. Técnicas de Recolha de Dados

Como já ficou evidente ao longo destes capítulos, existem duas estratégias alternativas, no que diz respeito à investigação científica. Assim, fruto da sua natureza, é normal que o método qualitativo use técnicas de recolha de dados diferentes das usadas pelo método quantitativo. O primeiro usa técnicas que lhe permitam integrar e compreender os factos sociais que pretende estudar, o segundo está mais preocupado em medir e determinar matematicamente a realidade envolvente.

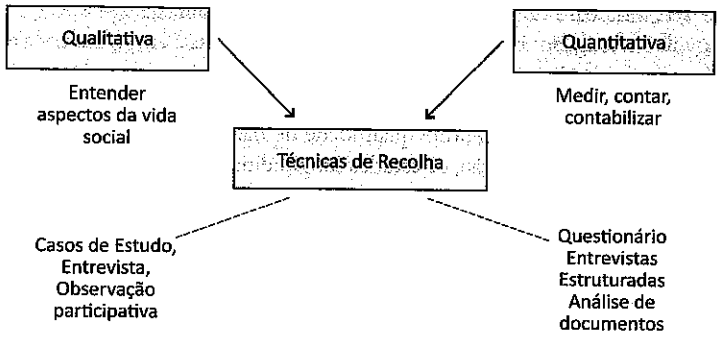


Figura 9.3. O Tratamento e análise de dados

A abordagem quantitativa procura manter-se, a todo tempo, afastada da realidade que pretende analisar, de maneira a garantir objectividade e evitar a própria interpretação do investigador. Como tal, esta estratégia usa inquéritos/questionários como principal meio de recolher informação. A construção destes deve ser feita de maneira a permitir que qualquer pessoa possa responder sem necessidade da intervenção do investigador. Por vezes existe a necessidade de usar conceitos mais ambíguos e menos determinados. Nestes casos deverá o investigador garantir que a pessoa que está a responder seja objectivamente informada da situação que o investigador pretende investigar.

Exemplo 5

Por exemplo, um investigador que use, num inquérito, o conceito de «activos específicos», terá de prestar informações suplementares acerca da sua definição. Assim poderá mencionar que: «O conceito refere-se ao nível de especialização dos recursos humanos e dos equipamentos utilizados na prestação de serviços. Assim, deverão ser consideradas de elevada especificidade a produção de bens e serviços que:

1. Impliquem um elevado custo na sua deslocalização;
2. Usem recursos humanos altamente especializados;
3. Usem ferramentas especificamente desenhadas para o efeito».

No entanto, apesar de todo o esforço do investigador em ser claro e preciso na construção do inquérito, o questionário poderá sempre ser alvo da subjectividade

interpretativa por parte da pessoa que responde. Como tal, a abordagem quantitativa também recorre frequentemente a entrevistas estruturadas, em tudo semelhantes às entrevistas puras usadas pelas estratégias qualitativas, procurando recolher informação com base na vivência do entrevistado. Todavia, não é deixada total liberdade de resposta, antes estão condicionadas às variantes que o investigador tem necessidade de analisar.

Como já referimos, a abordagem qualitativa, para além do método de entrevista, também usa outras ferramentas. A observação participante é, provavelmente, a mais característica. Pressupõe que o investigador «mergulhe» no contexto socioeconómico da realidade que pretende analisar, o que obriga o investigador a ser parte da realidade estudada, aumentando assim a dose de subjectividade inerente à investigação. A técnica da entrevista aberta, permite uma profunda recolha de dados ao nível do significado e da interpretação que o entrevistado tem da realidade social em análise.

Independentemente da estratégia de investigação adoptada, a codificação dos dados recolhidos reveste-se da maior importância. Convém não esquecer que os dados recolhidos serão tratados estatisticamente. Ora, uma errada codificação poderá conduzir à inviabilização da investigação, pelo facto das variáveis não estarem correctamente construídas.

4. Ferramentas de análise de dados

Muitas vezes, o investigador encontra-se numa situação em que, após a recolha de dados, tem algumas dúvidas dentro das ferramentas e técnicas que tem ao seu dispor, questionando-se sobre quais serão as mais adequadas para cumprir o seu objectivo de estudo. Portanto, o objectivo deste capítulo focaliza-se na apresentação das principais alternativas de análise de dados que possibilitam ao investigador testar as suas hipóteses e, como objectivo final, responder à sua pergunta de investigação.

As técnicas de análise de dados variam em função da estratégia de investigação bem como da natureza das variáveis. Desta forma, vamos apresentar as principais medidas de estatística descritiva, de relações de associação, de causalidade, estudos de proporções e comparação de populações. Ao longo das próximas páginas iremos apresentar e exemplificar cada uma destas técnicas.

4.1. Considerações Gerais

Antes de começarmos a apresentação propriamente dita há algumas considerações, bem como alguns conceitos, que devem ser introduzidos.

As variáveis podem ser divididas em variáveis quantitativas e qualitativas. No caso das qualitativas, estas podem ser qualificadas em ordinais, nominais e dicotómicas:

- Ordinais — Aqui as variáveis são compostas por categorias passíveis de serem ordenadas. Ou seja, é possível estabelecer uma relação de preferência em que um é melhor/maior do que outro;
- Nominais — Neste caso, as classes da variável, ao contrário da anterior, não conseguem ser ordenadas;
- Dicotómicas — Aqui estamos na presença de uma variável composta unicamente por duas categorias mutuamente exclusivas.

As variáveis quantitativas serão classificadas em intervalares e de razão:

- Intervalares/Razão: São variáveis que têm observações equidistantes ao longo de toda a categoria que pretendem medir.

Exemplo 6

Tipo de variáveis:

- Ordinais: Questionarmos um indivíduo acerca da frequência com que vai a um ginásio, as respostas poderão variar entre uma única vez, de duas a três ou mais de três vezes. Como podemos verificar, há uma ordenação possível. Ir mais de três vezes é mais do que ir uma única vez.
- Nominais: Questionarmos um indivíduo acerca dos motivos que o levam a frequentar um ginásio. As respostas poderão variar entre preocupações físicas, estéticas, para aliviar os níveis de stress, ou para acompanhar um amigo. Como se pode verificar, não é possível estabelecer uma relação em que se diga que uma razão é superior ou melhor do que outra.
- Dicotómicas: Questionarmos um indivíduo sobre se está ou não inscrito num ginásio.
- Intervalares/Razão: Verificarmos, por exemplo o rendimento de uma população ou a temperatura dos concelhos em Portugal. Em ambas as situações temos unidades de medida equidistantes (cêntimos e graus centigrados) e podemos comparar e estabelecer relações de relação exacta.

Também é necessário saber qual o tipo de testes que pode ser aplicado face às variáveis que construímos, sobretudo face à sua distribuição. Assim, necessitamos

verificar se a distribuição das variáveis se aproxima da distribuição Normal, condição necessária para aplicarmos testes paramétricos (Maroco, 2007; Pestana & Gageiro, 2005; Bryman, 2008). O teste de Kolmogorov-Smirnov (K-S) permite apurar se uma determinada amostra provém de uma população com distribuição normal. Como tal, teremos de nos socorrer do método dos testes de hipótese. Vamos definir uma hipótese nula do tipo:

$$H_0: X \sim N(\mu, \sigma)$$

Esta hipótese significa que a distribuição da amostra se aproxima da distribuição normal.

Em alternativa, definimos uma hipótese do tipo:

$$H_1: X \neq N(\mu, \sigma)$$

Aqui, H_1 representa a hipótese da distribuição da amostra ser diferente da distribuição normal.

Assim, se se verificar H_0 podemos usar testes paramétricos. Se, pelo contrário, se verificar H_1 teremos de usar testes não paramétricos. A rejeição ou aceitação da hipótese determina-se pelo apuramento do valor crítico da distribuição da estatística de K-S (Pestana & Gageiro, 2005). Assim, rejeita-se H_0 se $D \geq D_{tabela}(\alpha)$. Para tal, temos de calcular o *p-value* da hipótese, isto é, o menor valor de α a partir do qual $D \leq D_{tabela}(\alpha)$. O *c* produzido é calculado usando a aproximação analítica estatística de testes de Lilliefors proposta por Dallal e Wilkinson (Maroco, 2007). Portanto, a regra é rejeitar H_0 se o *p-value* $\leq \alpha$.

4.2. Estatística Descritiva

Neste ponto passaremos muito rapidamente as principais medidas de tendência central e de dispersão.

Enquanto medidas de tendência central temos as tradicionais formas de média, mediana e moda.

A média aritmética é provavelmente um dos indicadores mais usados e conhecidos em qualquer estudo ou análise realizados. Obtêm-se somando todas as amostras e dividindo-as pelo número de observações.

$$(\chi): \chi = 1/n \; \Sigma ni{=}1 \; \chi j$$

Em termos gráficos, a média é uma recta que minimiza a distância entre todas as observações.

Exemplo 7

Assumindo casuisticamente valores para três variáveis, realizamos o respectivo teste de normalidade.

Teste de Normalidade K-S

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Variável A	0,146	18	0,2	0,963	18	0,659
Variável B	0,112	12	0,2	0,962	12	0,807
Variável C	0,119	12	0,2	0,951	12	0,652

Assim, segundo a tabela, podemos verificar que em todos os casos o *p–value* (0,2) > $\alpha = 0{,}05$. Portanto, não se rejeita H_0 e podemos concluir, com uma probabilidade de erro de 5%, que a distribuição das variáveis é aproximada à Normal.

Outra condição para que possamos verificar a Normalidade da distribuição é verificar se as variâncias populacionais são homogéneas. O teste de *Levene* é um meio bastante robusto para verificar desvios de normalidade (Maroco, 2007). Desta forma pretendemos testar:

$$H_0: \sigma^2_1 = \sigma^2_2 = \cdots = \sigma^2_k$$
$$H_1: \exists, i, j: \sigma^2_i \neq \sigma^2_j \ (i \neq j; i, j = 1, \ldots, k)$$

Ou seja, a hipótese nula que atesta a homogeneidade das variáveis; e pelo contrário, a hipótese inversa visa verificar se existe pelo menos um i e um j tal que $\sigma_i \neq \sigma_j$.

Teste de Homogeneidade da Variância

	Levene Statistic	df1	df2	Sig.
Baseada na média	0,247	2	39,00	0,782
Baseada na mediana	0,162	2	39,00	0,851
Baseada na mediana corrigida (amostra)	0,162	2	33,72	0,851
Baseada na média aparada	0,245	2	39,00	0,784

Da tabela, podemos concluir que as variâncias populacionais, estimadas a partir das amostras, são homogéneas já que os diferentes *p–value* > $\alpha = 0{,}05$ e como tal não rejeitamos H_0.

A mediana é o valor que divide em partes iguais as observações da variável. Ou seja, a mediana é o valor que determina que 50% das observações estejam acima desse valor e 50% abaixo.

$$\chi = \begin{cases} \dfrac{X_{n/2} + X_{(n+2)/2}}{2} & \text{se } n \text{ é par} \\ X_{(n+1)/2} & \text{se } n \text{ é ímpar} \end{cases}$$

A moda define-se como o valor mais frequente da variável usada na investigação.

Quanto às medidas de dispersão, os valores mais usados são os da variância e desvio padrão. A variância mede a dispersão dos valores em torno da média:

$$\sigma^2 = 1/n \ \Sigma_n^{i=1} (Xi - \mu)^2$$

O desvio-padrão, sendo a raiz quadrada da variância, reduz a dispersão à mesma dimensão da variável em estudo. Continua a medir a dispersão média que as observações têm em torno da média das observações. Portanto, quanto maior for o desvio-padrão menor será a consistência do valor apresentado pela média.

Exemplo 8

A título de exemplo seleccionámos 6 variáveis disponíveis no INE relativas à caracterização demográfica, social e económica dos concelhos em Portugal Continental: *Rendimento per capita*, *Área*, *Área do concelho destinada a uso urbano*, *Área do concelho destinada a uso industrial*.

	Rendimento per capita	Área	Solo Urbano	Solo Industrial	Independência Financeira	IDS
Observações	278	278	278	278	278	278
Média	718,812	269,432	1730,51	270,3273	0,3394450	0,8948058
Mediana	686,12	211,50	1165,613	119,7940	0,3012909	0,8965000
Moda	595,98	95,00	92,33	—	0,0253800	0,9000000
Desvio Padrão	133,273	193,162	1721,78	408,97991	0,19864520	0,0258957
Variância	17761,70	37311,75	2964522,57	167264,564	0,039	0,001

De uma rápida leitura facilmente verificámos, por exemplo, que a consistência da média aritmética na variável *rendimento per capita* é maior do que na variável *solo urbano*. Vejamos, neste último caso o valor do desvio padrão (1721,78) quase ultrapassa o valor da média (1730,51). Isto significa que a variabilidade das observações é de tal ordem que, em muitos casos, os valores registados ficam muito aquém ou vão muito mais além, do que o valor médio deixa transparecer.

4.3. Estatística de Associação

Há diferentes testes que podem ser feitos para comprovar a relação existente entre duas variáveis, sejam elas dicotómicas, ordinais, nominais, etc.. Portanto é nossa intenção, apresentar os diferentes métodos, em função da natureza das variáveis, de que o investigador dispõe para testar/medir o nível de associação entre as variáveis que compõem o seu estudo. No entanto, convém referir que as medidas de associação quantificam unicamente a intensidade e a direcção de associação das variáveis (Maroco, 2007). Desta associação, não se deve retirar qualquer tipo de conclusão acerca da relação de causalidade que possa existir entre ambas (Bryman, 2008).

As medidas de associação que apresentaremos pretendem portanto mostrar uma relação (in)directamente proporcional entre as variáveis, conforme assumam valores próximos de 1 ou –1. No caso do coeficiente de correlação assumir o valor zero, poderemos concluir pela total ausência de correlação entre as variáveis em análise.

De seguida, apresentamos então alguns dos mais frequentes coeficientes de correlação: *Pearsons'r, Cramér's V, Chi-Quadrado, Spearman's rho, phi.*

O coeficiente de correlação de *Pearson's r* é o instrumento adequado para medir a relação, isto é, a intensidade e a direcção da associação, de tipo linear (Maroco, 2007), entre duas variáveis quantitativas, isto é, duas variáveis de tipo intervalares/razão. O resultado do teste *Pearson's r*, também designado por «ρ» varia entre a unidade positiva e negativa ($-1 \leq \rho \leq 1$). Conforme já dissemos, o sinal do coeficiente indica a direcção da relação, enquanto a sua grandeza indica a força da relação.

O coeficiente de correlação de *Spearman's rho* é um tipo de teste estatístico não paramétrico, ou seja, pode ser usado em amostras que não tenham distribuição normal. É uma medida de associação normalmente associada a variáveis ordinais ou quando uma é ordinal e as outras são intervalo/rácio ou dicotómicas. Tal como o teste *Pearson's r*, o coeficiente obtido varia entre a unidade positiva e negativa ($-1 \leq \rho \leq 1$) atestando a relação entre variáveis bem como a sua direcção e intensidade.

O coeficiente de correlação *V de Cramer* também é, à imagem do teste *Spearman's rho*, um tipo de teste estatístico não paramétrico. É usado para medir a intensidade da associação de variáveis nominais. Ao contrário dos restantes testes desta secção, o seu valor só assume valores positivos ($0 \leq V \leq 1$). Como tal, nada nos diz sobre a direcção da relação. Só nos dá indicação relativamente à intensidade da sua relação. O coeficiente de correlação *phi* é usado para medir a intensidade da associação de variáveis dicotómicas. A sua interpretação faz-se muito à imagem das correlações *Pearson's r* e *Spearman's rho*. Assim, o coeficiente varia entre a unidade assumindo valor quer positivo ou negativo ($-1 \leq \Phi \leq 1$).

Exemplo 9

Imagine que queremos testar a correlação entre o salário auferido por um trabalhador e o montante de horas extra que, em média, faz por mês. Assim, teríamos duas variáveis quantitativas estruturadas em três classes:

	Rendimento Mensal
Classe 1	$450 \leq x < 650$
Classe 2	$450 \leq x < 800$
Classe 3	+ de 800

	Horas Extra
Classe 1	$1 \leq x < 2$
Classe 2	$2 \leq x \leq 3$
Classe 3	+ de 3

As hipóteses do teste são:

H_0: As variáveis Rendimento Mensal e Horas Extra são independentes

H_1: As variáveis Rendimento Mensal e Horas Extra não são independentes

Recordamos que a regra é rejeitar H_0 se o *p-value* $\leq \alpha$.

		Rendimento	Horas Extra
Rendimento	Pearson Corr	1	0,606**
	Sig. (2-tailed)		0,000
	N	100	100
Horas Extra	Pearson Corr	0,606**	1
	Sig. (2-tailed)	0,000	
	N	100	100

** nível de significância de 99%

Como podemos verificar, como o *p-value* $\leq \alpha$. ($0{,}00 \leq 0{,}05$), existe evidência estatística para, com 99% de certeza, rejeitar H_0, e concluir pela dependência das variáveis em causa. Mais, interpretando o valor do coeficiente *pearson*, podemos concluir que se trata de uma correlação positiva. Ou seja, quanto mais horas extra são realizadas por um indivíduo, maior o seu rendimento.

Exemplo 10

Imagine agora que desejamos verificar se existe uma relação entre o salário auferido por um trabalhador (mesma variável de intervalo usado no caso anterior) e as habilitações literárias (variável ordinal). Assim, teríamos duas variáveis quantitativas estruturadas em três classes:

	Rendimento Mensal
Classe 1	450 ≤ x < 650
Classe 2	450 ≤ x < 800
Classe 3	+ de 800

	Habilitação Literária
1	Ensino Obrigatório
2	12º ano
3	Licenciatura
4	Pós-Graduação

As hipóteses do teste são:

H_0: As variáveis Rendimento Mensal e Habilitação Literária são independentes

H_1: As variáveis Rendimento Mensal e Horas Extra não são independentes

Recordamos que a regra é rejeitar H_0 se o *p-value* ≤ α.

			Rendimento	Hab. Literárias
Spearman's rho	Rendimento	Correlation Coefficient	1	0,796**
		Sig. (2-tailed)	.	0
		N	100	100
	Habilitações Literárias	Correlation Coefficient	0,796**	1
		Sig. (2-tailed)	0	.
		N	100	100

** nível de significância de 99%

Como podemos verificar, como o *p-value* ≤ α. (0,00 ≤ 0,05), existe evidência estatística para, com 99% de certeza, rejeitar H_0, e concluir pela dependência das variáveis em causa. Mais, interpretando o valor do coeficiente *Spearman's rho* podemos concluir que se trata de uma correlação positiva. Ou seja, indivíduos com nível de habilitações literárias mais elevadas, auferem maior rendimento.

Exemplo 11

Vamos agora usar duas variáveis nominais: natureza do curso superior, tipo de gostos musicais. Assim, teríamos duas variáveis quantitativas estruturadas em três classes:

	Natureza do Curso
1	Gestão
2	Arquitectura
3	Medicina

	Tipos de Música
1	Rock
2	Opera
3	Pop
4	Rn'B

As hipóteses do teste são:

H_0: As variáveis Natureza do Curso e Tipos de Música são independentes

H_1: As variáveis Natureza do Curso e Tipos de não são independentes

Recordamos que a regra é rejeitar H_0 se o *p-value* ≤ α.

Symmetric Measures

		Value	Asymp. Std.	Approx. Tb	Approx. Sig.
Nominal by Nominal	Phi	0,287			0,41
	Cramer's V	0,203			0,41
Interval by Interval	Pearson's R	–0,025	0,099	–0,252	,801c
Ordinal by Ordinal	Spearman Correlation	–0,024	0,1	–0,237	,813c
N of Valid Cases		100			

** nível de significância de 99%

Como podemos verificar, uma vez que o *p-value* ≤ α. (0,41 ≥ 0,05), não existe evidência estatística para, com 99% de certeza, rejeitar H_0, e concluir pela independência das variáveis em causa.

Portanto, não há qualquer tipo de conclusão relativamente à correlação entre as variáveis em causa. Atenção, o facto de não podermos tirar qualquer conclusão não significa que não haja qualquer correlação. Esta seria a conclusão se o teste fosse significativo e a estatística de *Cramer's V* fosse 0.

Exemplo 12

Vamos usar duas variáveis dicotómicas: Localização e Ensino Universitário. Imagine que desejamos verificar se o Ensino Universitário se localiza tendencialmente no litoral. Construímos uma variável dicotómica para classificar a localização dos concelhos em Portugal. A outra variável, também ela dicotómica, destina-se a verificar a existência de estabelecimentos de ensino superior público no concelho.

Valor	Localização	Ensino Universitário
1	Litoral	Sim
0	Interior	Não

As hipóteses do teste são:

H_0: As variáveis Localização e Ensino Universitário são independentes

H_1: As variáveis Localização e Ensino Universitário não são independentes

Recordamos que a regra é rejeitar H_0 se o *p-value* $\leq \alpha$.

		Value	Approx. Sig.
Nominal by Nominal	Phi	–0,098	0,104

Como podemos verificar, como o *p-value* $\geq \alpha$. (0,104 $\geq$ 0,05) não existe evidência estatística para, com 99% de certeza, rejeitar H_0, e concluir pela dependência das variáveis em causa. Portanto, aqui não podemos retirar qualquer tipo de conclusões senão a de que a localização das instituições de ensino superior público não são determinadas pelo facto ser do litoral ou não.

Na tabela 9.1 fica um resumo dos tipos de testes possíveis de serem realizados, em função da natureza das variáveis

Tabela 9.1. O Tratamento e análise de dados

	Nominal	Ordinal	Intervalo/Rácio	Dicotómica
Nominal	Cramér's V	Cramér's V		
Ordinal		Spearman's rho	Spearman's rho	Spearman's rho
Intervalo/Rácio			Pearsons'r	
Dicotómica			Spearman's rho	phi

Fonte: Adaptado de Bryman (2008)

4.4. Estudo das Proporções

Este ponto está sobretudo direccionado e centrado no tratamento de variáveis qualitativas (Maroco, 2007). Tem por objectivo analisar a distribuição por cada uma das categorias ou classes da variável em Estudo.

O teste Chi-Quadrado (χ^2), é um dos mais populares e usados testes paramétricos, usados para testar se dois, ou mais, grupos independentes diferem relativamente a uma mesma característica. Ou seja, este teste serve sobretudo quando, tendo por base uma variável comum, queremos analisar o comportamento/desempenho de dois grupos/populações independentes. Não se deve confundir este tipo de testes com aqueles que acabamos de apresentar. Na secção anterior procurávamos verificar a relação entre duas variáveis, agora, procuramos analisar o comportamento de dois grupos independentes em função de uma variável comum.

Exemplo 13

Para exemplificarmos o teste Chi-Quadrado (χ^2), vamos simular um inquérito à população de um concelho. Nesse inquérito é pedido ao cidadão que classifique a sua satisfação em relação a 50 serviços municipais (usando uma escala de Likert para medir a satisfação) Estes 50 serviços são fornecidos à população, quer por serviços directos do município, quer por serviços contratados ao exterior. A nossa intenção, enquanto investigadores, é a de verificar se há uma diferença na avaliação dos serviços em função da entidade que o presta (município ou agente externo).

Valor	Entidade Responsável
1	Município
2	Externalização

Nível de Satisfação				
1	2	3	4	5
Insatisfeito	Pouco Satisfeiro	Satisfeito	Bastante Satisfeito	Muito Satisfeito

Tendo em consideração a variável «satisfação dos cidadãos», as hipóteses do teste são:

H_0: A satisfação dos cidadãos é independente da natureza do agente que presta o serviço

H_1: A satisfação dos cidadãos é dependente da natureza do agente que presta o serviço

Recordámos que a regra é rejeitar H_0 se o *p-value* $\leq \alpha$.

Exemplo 13 (cont.)

Chi-Quadrado

	Value	df	Approx. Sig. (2-sided)
Pearson Chi-Square	13,510[a]	4	0,009
Likelihood Ratio	14,747	4	0,005
Linear-by-Linear Association	8,233	1	0004
N of Valid Cases	50		

Como podemos verificar, como o *p-value* $\leq \alpha$. (0,009 $\leq$ 0,05), existe evidência estatística para, com 99% de certeza, rejeitar H_0, e concluir pela dependência das variáveis em causa. Portanto, podemos afirmar que a natureza do agente é relevante no grau de satisfação do cidadão. No entanto, este teste nada nos diz acerca do agente que tem maior nível de satisfação. Para tal temos de avaliar e compara os valores obtidos com os esperados.

Crosstabulation Measures

			Natureza		
			Publico	Externalizado	Total
Satisfação	Insatisfeito	Count	1	1	2
		Expected Count	1,1	0,9	2
	Pouco Satisfeito	Count	9	1	10
		Expected Count	5,4	4,6	10
	Satisfeito	Count	8	2	10
		Expected Count	5,4	4,6	10
	Bastante Satisfeito	Count	3	8	11
		Expected Count	5,9	5,1	11
	Muito Satisfeito	Count	6	11	17
		Expected Count	9,2	7,8	17
Total		Count	27	23	50
		Expected Count	27	23	50

Daqui podemos verificar que o serviço prestado por um agente externo tem uma melhor avaliação de satisfação, por parte do cidadão.

4.5. Comparação de Populações

Por vezes torna-se necessário fazer comparações das médias de duas ou várias populações. Segundo Maroco (2007), o melhor mecanismo para atingir este objectivo é usar a Análise de Variância, mais conhecido por ANOVA (**AN**alysis **O**f **VA**riance). Existindo uma variável dependente, trata-se de comparar a sua variância com a da variável independente (*one-way ANOVA*) ou variáveis dependentes (ANOVA factorial). Em função de existirem diferentes variáveis dependentes, poderão existir factores associados a cada uma (Tabela 9.2).

Tabela 9.2. O Tratamento e análise de dados

	Factores fixos	Factores não fixados
Uma Variável Independente	ANOVA Tipo I	ANOVA Tipo II
Mais de duas V.Independentes		ANOVA Tipo III

Assim, a ANOVA permite, tendo uma variável dependente, verificar como os factores da variável independente exercem a sua influência. A lógica da análise está na comparação entre a variância registada dentro das amostras e as registadas entre amostras. Se a variância residual for significativamente inferior à variância entre as amostras, então as médias populacionais estimadas são significativamente diferentes. Trata-se de um teste paramétrico que permite a comparação simultânea de várias categorias da variável independente, sendo que esta pode ser de origem qualitativa ou quantitativa, no entanto, a variável dependente terá de ser contínua.

Exemplo 14

Voltamos a usar a variável rendimento *per capita* por concelho juntamente com a variável «Localização». Dividimos esta em três categorias: Pequeno, Médio e Grande Centro Urbano. Com a ajuda da tabela ANOVA vamos verificar se a localização é um factor determinante no rendimento *per capita*.

Começamos por analisar a tabela ANOVA:

H_0: O rendimento *per capita* é independente das categorias de localização.

H_1: O rendimento *per capita* é dependente das categorias de localização.

Exemplo 14 (cont.)

ANOVA

Rendimento

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1549202,92	2	774601,46	63,195	0
Within Groups	3370788,528	275	12257,413		
Total	4919991,448	277			

Aqui interessa verificar o resultado do teste F. Como podemos verificar, uma vez o *p–value* $\leq \alpha$. (0,00 ≤ 0,05), existe evidência estatística para, com 99% de certeza, rejeitar H_0, e concluir pela dependência das variáveis em causa.

Agora, trata-se de saber qual das três categorias exercer influência na variável dependente. Para cada categoria definidas as hipóteses:
H_0: O rendimento *per capita* é independente da categoria de localização em evidência.
H_1: O rendimento *per capita* é dependente da categoria de localização em evidência.

(I) LocalDese	(J) LocalDese	Mean Difference (I–J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Pequeno	Semi Centro	–88,20481*	14,87846	0,000	–123,265	–53,1441
Centro	Grande Centro	–262,86739*	25,04391	0,000	–321,882	–203,852
Semi	Pequeno Centro	88,20481*	14,87846	0,000	53,1441	123,2655
Centro	Grande Centro	–174,66258*	26,61731	0,000	–237,385	–111,9396
Grande	Pequeno Centro	262,86739*	25,04391	0,000	203,852	321,8827
Centro	Semi Centro	174,66258*	26,61731	0,000	111,9396	237,3856

Desta tabela podemos verificar que, em todas as situações, temos níveis de significância (0,00 ≤ 0,05) que nos permitem concluir que todas as categorias exercem uma influência na variável dependente.

		Subset for alpha = 0.05		
Localização	N	1	2	3
Pequeno Centro	175	672,31 €		
Semi Centro	81		760,51 €	
Grande Centro	22			935,18 €
Sig.		1	1	1

Aqui verificamos que o facto de se viver num maior centro urbano determina o nível de rendimento *per capita*.

4.6. Estatística de Causalidade

Nesta secção do capítulo de análise e tratamento de dados vamos tratar o caso das regressões, lineares ou categoriais, como forma de estimar causalidade entre variáveis. Assim, uma regressão é uma técnica de modelar relações entre variáveis de maneira a obtermos um meio capaz de determinar o comportamento de uma variável, dita dependente, em função de outras variáveis, ditas independentes.

Há inúmeros modelos de regressão que podem ser usados para analisar a relação entre variáveis. O tipo de regressão a ser utilizada depende do tipo de variável dependente em causa, bem como ao comportamento da mesma. Nesta secção, e a título exemplificativo, serão apresentados casos de regressões lineares e categoriais Se esta variável dependente for contínua, podem ser aplicados métodos de regressão linear, se, pelo contrário, a variável dependente for dicotómica, ou assumir valores em classes, então teremos de usar regressões do tipo categorial. As variáveis de contagem de eventos são tratadas através de regressões de poisson.

4.6.1.Regressão Linear

A regressão linear é uma ferramenta estatística usada quando se pressupõe existir uma associação linear entre uma variável endógena *Y*, e uma ou mais variáveis exógenas *X*'s (Pestana e Gegeiro, 2005). O modelo de regressão é do tipo:

$$Yj = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_p X_{pj} + \varepsilon_j \quad (j = 1, \ldots, n)$$

Y_j — representa a variável dependente;
β_0 — representa a constante do modelo;
β_j — representam os coeficientes de cada variável independente do modelo;
ε_j — representa os resíduos do modelo.

Nas regressões deste tipo, segundo Maroco (2007), é necessário que se verifiquem algumas condições importantes

1. Apenas a variável *Y* é afectada pelos erros de medição do modelo;
2. Os erros do modelo devem ser aleatórios, independentes e com distribuição normal. A falta desta condição poderá indiciar a existência de uma variável explicativa, ignorada no modelo definido;
3. As variáveis independentes não podem estar correlacionadas.

Os coeficientes β_j do modelo são apurados utilizando o método dos mínimos quadrados. Isto é, na impossibilidade de recolher informação, relativa à variável dependente de toda a população, usam-se mecanismos de maneira a minorar os erros entre as observações reais e os coeficientes estimados. Construída a regressão

com base nos dados amostrais, torna-se necessário verificar a sua inferência ao resto da população. Neste âmbito, a significância do modelo é verificada com a estatística F (com distribuição *F-Snedecor*). De seguida, é necessário avaliar a significância individual de cada dos estimadores usando teste *t*-Student. Finalmente, é necessário prestar a devida atenção ao coeficiente de determinação (representado por R^2). Este coeficiente varia entre zero e um ($0 \leq R^2 \leq 1$) e representa de que forma a variabilidade na variável dependente Y é função das variáveis independentes, X_i, introduzidas no modelo.

Exemplo 15

Usando a variável rendimento *per capita* por concelho como variável dependente, definimos um conjunto de variáveis (*Uso de Solo Urbano*, *Uso de Solo Industrial*, *Localização* e *Índice de Desenvolvimento Social* «IDS») que, em nosso entender, poderão ter capacidade explicativa da dispersão dos valores do rendimento, pelas diferentes zonas do país.

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,623a	0,389	0,38	104,97526	1,716

a. Predictors: (Constant), IDS, SoloIndustrial, SoloUrbano, LocalDese
b. Dependent Variable: rendto

Na tabela resumo podemos verificar que o R^2, que define o nível de explicação que as variáveis independentes têm para com o variável dependente, não vai além dos 38%. Como tal, trata-se de um registo algo baixo para ciências sociais. Independentemente deste facto, continuamos a análise ao nosso exemplo.

Passamos para a tabela ANOVA:

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1911584,605	4	477896,151	43,367	0,000a
	Residual	3008406,843	273	11019,805		
	Total	4919991,448	277			

a. Predictors: (Constant), IDS, SoloIndustrial, SoloUrbano, LocalDese
b. Dependent Variable: rendto

Exemplo 15 (cont.)

Como dissemos anteriormente, neste quadro é relevante analisar o resultados do teste F. Como podemos verificar, uma vez o *p-value* $\leq \alpha$. (0,00 ≤ 0,05), existe evidência estatística para, com 99% de certeza, rejeitar H_0, e concluir que o modelo é significativo.

Passando agora para o quadro dos coeficientes da regressão[3]:

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	–681,723	263,874		–2,584	0,01
	SoloUrbano	–0,014	0,005	–0,177	–2,948	0,003
	SoloIndustrial	0,072	0,018	0,22	4,094	0,000
	LocalDese	82,667	13,027	0,396	6,346	0,000
	IDS	1436,038	305,121	0,279	4,706	0,000

a. Dependent Variable: rendto

Desta tabela podemos verificar que todas as variáveis enumeradas são, individualmente, significativas na explicação da variável dependente (em todos os casos *p-value* $\leq \alpha$). Assim, o nosso modelo ajustado (depois da normalização das variáveis) fica:

$$\text{Rendimento} = -681{,}72 - 0{,}177(\text{SoloUrbano}) + 0{,}22(\text{SoloIndustrial}) + 0{,}396(\text{Localização}) + 0{,}279(\text{IDS})$$

Resta agora verificar se os pressupostos do modelo são cumpridos.

Desta forma, passamos a análise dos resíduos de maneira a comprovar a homocedasticidade da distribuição dos erros. Ou seja, procuramos verificar se existem erros de heteroscedasticidade que possam comprometer o nosso modelo.

No primeiro quadro temos calculado a estatística Durbin-Watson. Sabemos que:

- Se Dw < dL, então há evidência estatística que os erros estão positivamente autocorrelacionados;
- Se Dw > dU, então há evidência estatística que os erros estejam negativamente autocorrelacionados;
- Se dL < Dw < dU então o teste é inconclusivo;

Das tabelas Savin e Withe: dL: 1,57 e dU:1,78, logo o teste é inconclusivo.

Exemplo 15 (cont.)

Para proceder à verificação deste pressuposto tentamos então analisar a distribuição dos erros do modelo. Como foi enunciado, se o erros do modelo não seguirem uma distribuição normal, isso indicia a existência de variáveis explicativas não incluídas no modelo. Para tal, decidimos realizar um teste de normalidade à distribuição dos erros.

One-Sample Kolmogorov-Smirnov Test

		Studentized Deleted Residual
N		278
Normal Parameters a,b	Mean	0,003737
	Std. Deviation	1,03073593
Most Extreme Differences	Absolute	0,114
	Positive	0,114
	Negative	–0,076
Kolmogorov-Smirnov Z		1,901
Asymp. Sig. (2-tailed)		0,001

a. Test distribution is Normal.
b. Calculated from data.

Assim, segundo a Tabela, podemos verificar que o *p-value* (0,001) < $\alpha = 0,05$. Portanto rejeita-se H_0 e podemos concluir, com uma probabilidade de erro de 5%, que a distribuição não segue uma distribuição aproximada à normal. Logo, o modelo não cumpre o pressuposto da homocedasticidade.

Resta ainda verificar o problema de multicolinearidade. Esta é uma situação onde pelo menos duas variáveis estão correlacionadas e representam a mesma dimensão explicativa.

Coefficients(a)

	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
	B	Std. Error	Beta	t	Sig.	Tolerance	VIF
(Constant)	–681,723	263,874		–2,584	0,01		
SoloUrbano	–0,014	0,005	–0,177	–2,948	0,003	0,622	1,608
SoloIndustrial	0,072	0,018	0,22	4,094	0,00	0,773	1,294
LocalDese	82,667	13,027	0,396	6,346	0,00	0,576	1,737
IDS	1436,038	305,121	0,279	4,706	0,00	0,637	1,569

Do quadro vemos que os valores de colinearidade são baixos, anulando assim, os riscos de multicolinearidade.

4.6.2. Regressão Categorial

Como já referimos, as regressões são formas de se estabelecerem relações de causalidade entre um conjunto de variáveis independentes e uma variável dependente. No ponto anterior, analisámos o caso das regressões lineares usadas para tratar situações onde temos uma variável dependente contínua. No caso das regressões categoriais, a nossa variável dependente deixa de ser contínua para assumir uma das categorias que a compõem. Isto é, as regressões lineares são adequadas para variáveis dependentes quantitativas enquanto as regressões categoriais são próprias para variáveis dependentes qualitativas ou categóricas (sendo que as variáveis independentes podem ser do tipo qualitativo e quantitativo). Assim, o rendimento *per capita* pode ser entendido como uma variável contínua, já a variável *Destinos Turísticos* (pode ser categorizada, entre outros, em turismo de cultura, de saúde, de desporto, natureza) assume um dos tipos teoricamente definidos.

O modelo funciona assumindo uma das classes como base de referência e estabelece comparações, em termos probabilísticos, de pertencer às restantes categorias (Menard, 2002; Borooah, 2002; Aldrich & Nelson, 1984; Liao, 1994). Este modelo econométrico usa o método da máxima verosimilhança para a estimação dos coeficientes da regressão. Isto é, estima os coeficientes que maximizam a probabilidade de encontrar as realizações da variável dependente (Maroco, 2007). Este método é usado nos modelos categoriais porque os erros não seguem nem uma distribuição normal nem apresentam uma variância constante. Assim, dada a impossibilidade de estimar os coeficientes com base no método dos mínimos quadrados usado nas regressões lineares, o método de máxima verosimilhança apresenta-se como o mais adequado.

Temos três tipos de regressões categoriais: Probit, Logit, Multinominal Logístico. As regressões do tipo Probit e Logit são adequadas para tratar de variáveis dependentes dicotómicas. Os resultados das estimações são muito similares, uma vez que os modelos seguem o mesmo método de estimação, distinguindo-se unicamente na sua função de distribuição associada. Enquanto o modelo Probit segue uma distribuição mais linear, entre os limites da variável, o modelo Logit fá-lo de forma menos abrupta. O modelo Multinominal Logístico é uma expansão do modelo de regressão logística onde a variável dependente deixa de ter um comportamento dicotómico, entre duas classes, para poder assumir mais de duas classes mutuamente exclusivas. Surge assim como a opção mais apropriada no tratamento de variáveis dependentes que assumam diferentes categorias da variável sob a qual não podemos impor uma ordem ou sequência natural. (Menard, 2002; Borooah, 2002; Aldrich & Nelson, 1984; Liao, 1994). A literatura especializada indica que, no caso de tratar-

mos variáveis dependentes sem ordem ou sequência natural, este será o modelo mais adequado (Menard, 2002; Borooah, 2002; Aldrich & Nelson, 1984; Liao, 1994).

A leitura dos coeficientes da regressão não se consegue fazer de uma forma tão linear e simples como acontecia com a regressão linear. Nestes casos, os coeficientes são expressos em rácios de hipóteses (*Odds ratio*), que nos indicam o impacto na probabilidade de um acontecimento, em relação à categoria alternativa. Ao contrário do modelo lógico binário, as probabilidades, no modelo multinominal logístico, não são calculadas em função da única categoria alternativa, mas sim em função da categoria assumida como referência para todas as comparações entre categorias (Liao, 1994).

Finalmente, resta referir a estatística pseudo R^2. Esta medida não pode ser interpretada nem directamente nem extrapolada do R^2 da regressão linear (Zimmermann, 1996). Nas regressões dos mínimos quadrados afere-se a dimensão do efeito linear das variáveis independentes sobre a variável dependente. Nas regressões logísticas não é possível calcular o R^2, uma vez que a variância da variável dependente, depende da probabilidade em que ocorrem os seus valores (Maroco, 2007). Segundo Dhrymes (1986) a estatística do R^2 tem três propriedades que seriam desejáveis numa estatística do pseudo R^2:

1. Tem uma relação directa com a estatística F para testar a significância dos coeficientes das variáveis explicativas;
2. É uma medida da redução da variabilidade da variável dependente, através da significância das variáveis explicativas;
3. É o quadrado do coeficiente de correlação simples entre os valores previstos e reais da variável dependente dentro da amostra.

Ora, segundo Zimmermann (1996) e Dhryme (1986) a única similitude relevante encontra-se na primeira propriedade para modelos probit binários, onde, em vez de existir uma relação com a estatística F, a relação é estabelecida com o rácio de verosimilhança. Assim, não havendo um equivalente para o R^2 para as regressões logísticas, foi criada uma ampla variedade de pseudo R^2. Segundo Zimmermann (1996) as comparações entre o R^2 e o pseudo R^2 deverão ser feitas com muitas reservas e em condições específicas. Assim, a comparabilidade é possível para modelos tobit, binários e modelos probit e logit ordinários, sendo de excluir modelos multinomiais logit e probit.

Exemplo 16

Imaginemos que queremos analisar se os factores socioeconómicos condicionam as escolhas que as Autarquias Locais fazem, em termos organizacionais, para a prestação de bens e serviços públicos. Para tal, identificámos sete variáveis independentes (*IDS, Rendimento per capital, Crescimento Populacional, Densidade Populacional, Área, Uso de Solo Urbano, Uso de Solo Industrial*). A nossa variável dependente é a forma organizacional assumida pelas Autarquias Locais para prestar bens e serviços públicos. Esta variável foi dividida em três categorias: 1. Soluções de Hierarquias; 2. Soluções de Mercado; 3. Soluções de Parcerias.

A primeira estatística a ter em atenção, após a construção da regressão categorial logística, é aquela que visa testar a significância do modelo. A regra é recorrente, concluímos pela significância do modelo se o *p-value* $\leq \alpha$. Ora, como podemos verificar na tabela seguinte, o 0,0171 $\leq$ 0,05, logo existe evidência estatística para, com 99% de certeza, rejeitar H_0, e concluir que o modelo é significativo.

	Mlogit	
	RRR	P > |z|
Mercados de Governação	**Mercado**	
IDS	3,039442	0,802
Rendimento *Per Capita*	1,001123	0,047*
Crescimento Populacional	0,740261	0,814
Densidade Populacional	0,9999096	0,370
Área	1,000164	0,527
Solo Urbano	1,000067	0,070
Solo Industrial	0,995523	0,010*
Mercados de Governação	**Parcerias**	
IDS	1,536465	0,002*
Rendimento *Per Capita*	0,9999568	0,941
Crescimento Populacional	0,3084034	0,347
Densidade Populacional	0,9997957	0,062
Área	0,999986	0,960
Solo Urbano	0,9999997	0,992
Solo Industrial	0,9998761	0,425
Mecanismo de Referência	**Hierarquia**	
Observações	20814	
Log pseudolikelihood	–1925,1344	
Prob > chi2	0,0171	
Pseudo R^2	0,0071	

* p < 10; ** p < .05; *** p < .01; two.tailed tests. Robust Standard errors.

Exemplo 16 (cont.)

Depois trata-se de analisarmos, para cada categoria, quais as variáveis independentes significativas (em que o *p–value* $\leq \alpha$). Na categoria Mercado, as variáveis *Rendimento Per Capita* e *Solo Industrial* são significativas. Na categoria Parcerias somente, a variável *IDS* tem capacidade explicativa.

Recordamos que os coeficientes da regressão são rácios de hipóteses e devem ser lidos como tal. Em todas as situações, as probabilidades são assumidas tendo por base a categoria de referência, neste caso a hierarquia.

Assim, verificamos que o aumento de uma unidade no *Rendimento per capita* provoca um aumento que proporcional na probabilidade relativa do município recorrer a mecanismos de mercado (1,001123). Também nas soluções de mercado, vemos que a probabilidades das Autarquias Locais recorrerem a esta alternativa, diminui (0,995523) sempre que há um aumento de uma unidade na variável *Solo Urbano*.

Olhando agora para as soluções de Parcerias, verificámos que estas vêem a sua probabilidade aumentada (1,536465), de ser usadas como alternativa ao método tradicional da hierarquia, sempre que aumenta uma unidade o IDS.

4.6.3.Regressão de Poisson

É um tipo de regressão usada fazendo uso de variáveis discretas e é adequada quando se pretende registar a ocorrência de fenómenos observáveis. Ou seja, é uma ferramenta de análise que se torna útil num tipo de investigação onde a variável dependente não é uma realidade contínua ou uma variável qualitativa mas sim o resultado de uma contagem relativa ao número de ocorrências discretas, num determinado período de tempo. O modelo de regressão é de tipo:

$$f(k;\lambda)\frac{e^{-\lambda}\lambda^{k}}{k!}$$

Onde:

K — factorial de k

Λ — Valor esperado de ocorrências num determinado espaço de tempo.

Exemplo 17

Imaginemos que queremos analisar os factores que determinam a criação de estabelecimentos de saúde em cada concelho. A variável dependente seria, neste caso, o número de estabelecimentos de saúde por concelho, e como variáveis independentes teríamos, por exemplo: *LogPopulação*, *LogRendimento*, *Índice de Dependência do Idosos* e a *Taxa de Natalidade*. A dependente apresenta a seguinte distribuição:

Número de Estab.	Freq.	Perc.	Ac.
0	1	0,36	0,36
1	242	87,05	87,41
2	6	2,16	89,57
3	8	2,88	92,45
4	1	0,36	92,81
5	1	0,36	93,17
6	4	1,44	94,60
7	1	0,36	94,96
8	3	1,08	96,04
9	3	1,08	97,12
11	1	0,36	97,48
13	1	0,36	97,84
14	1	0,36	98,20
15	3	1,08	99,28
16	1	0,36	99,64
19	1	0,36	100

A distribuição das observações segue a distribuição de *poisson*, uma vez que apresenta um enviesamento, neste caso à esquerda, não estando assim reunidas as condições para usarmos o tipo de regressão linear (ver secção anterior). Usando a regressão de *poisson* devemos, ainda que se trate de uma contagem de um evento, ter o cuidado de testar a sua adequabilidade à variável dependente. Assim, definimos o teste de hipótese relativamente à adequabilidade da utilização de uma regressão de *poisson*. As hipóteses do teste são:

H_0: A distribuição da variável não segue uma distribuição de *poisson*.

H_1: A distribuição da variável segue uma distribuição de *poisson*.

Exemplo 17 (cont.)

Recordamos que a regra é rejeitar H_0 se o *p-value* $\leq \alpha$.

	Value	Prob > χ^2
Goodness-of-fit	2927.078	0,000

Uma vez que 0,000 ≤ 0,01, existe evidência estatística para, com 99% de certeza, rejeitar H_0, e concluir que o modelo segue uma distribuição de *poisson*. Se o resultado não permitisse rejeitar H_0 poderíamos, e tratando-se de uma variável de contagem, usar uma regressão binomial negativa.

Assim temos:

PseudoR2	Wald chi2(4)	Prob > chi2	Log pseudolikelihood
0,2626	293,1	0,000	–416,23882

Podemos verificar que o resultado do teste qui-quadrado (0,000 ≤ 0,01) permite concluir que existe uma evidência estatística, com 99% de certeza, para concluir que o modelo é significativo. Olhando agora para os coeficientes do mesmo modelo, temos:

Var. Dep	Número				
	Estabelecimentos Saúde	Coef.	Std. Error	Z	P>|z|
	(Constant)	–16,30967	1,950216	–8,36	0,000
	LogPopulação	0,4546561	0,0939675	4,84	0,000
	LogRendimento	1,835922	0,4167628	4,41	0,000
	Tx. Natalidade	–0,0203378	0,320452	–0,63	0,526
	Ind.Dep.Idosos	0,0078596	1,950216	1,36	0,175

Dos resultados obtidos verificamos que há evidência estatística para aceitar as variáveis *LogPopulação* e *LogRendimento* como explicativas, já que o resultado de cada uma obedece à condição *p-value* $\leq \alpha$ (em ambos casos *p-value* (0,000) < α = 0,01). Assim, podemos concluir que o número de estabelecimentos de saúde criados num concelho está positivamente relacionado com o tamanho da sua população bem como com o nível de rendimento da mesma.

5. Tratamento Qualitativo

O tipo de análise qualitativa é manifestamente diferente dos casos até agora apresentados por se mostrarem mais adequados a uma abordagem quantitativa. Tal está obviamente relacionado com o facto destas duas abordagens assentarem em estratégias de investigação alternativas: a abordagem quantitativa tem os seus fundamentos, em termos epistemológicos, numa posição positivista, enquanto que a abordagem qualitativa está relacionada com a corrente interpretativista. Estas estratégias são alternativas já que advogam diferentes designs de investigação e protagonizam duas posições dissemelhantes do mesmo *continuum* em termos da metodologia empregue. De um lado, a abordagem positivista que procura, através de ferramentas quantitativas, validar as hipóteses previamente definidas e que decorrem de uma teoria de suporte. Do outro lado, verificamos que a postura interpretativista procura, através de uma colecção massiva de dados, encontrar ligações entre categorias e conceitos de maneira a construir pressupostos teóricos suficientemente válidos que permitam a sua generalização.

A abordagem qualitativa começa por definir uma pergunta de investigação primária e embrionária. De seguida, são seleccionados e identificados os indivíduos/grupos alvo da recolha de dados. Ao invés da abordagem quantitativa, a teoria não surge de seguida, como elemento de suporte e sustentação das hipóteses de trabalho. Ela não desempenha aqui o papel de fornecer as hipóteses ou os conceitos e os indicadores a serem testados. A abordagem qualitativa, após a definição da pergunta de investigação que coloca assim como da identificação do seu grupo de análise, enceta um longo processo de recolha de dados. Este processo tem por objectivo permitir a constante verificação do fenómeno. É ao longo deste processo de interpretação e recolha de dados adicionais que se constituí um longo processo contínuo de redefinição e aperfeiçoamento da teoria que está a ser construída. Já na posição interpretativista, a teoria está constantemente a ser desenvolvida, testada e criada (Bryman, 2004). As ferramentas mais utilizadas no processo qualitativo para a recolha de dados, passam pela utilização de entrevistas, observação participante, análise de conteúdos, grupos de estudo, entrevistas semi-estruturadas, etc.

Dada a estratégia epistemológica assumida, o tratamento de dados qualitativo não é tão linear e sistematizado como os outros que foram apresentados nos pontos anteriores. No entanto, gostaríamos de partilhar algumas ferramentas utilizadas no tratamento de dados qualitativos.

a) A inductive analysis

Em primeiro lugar temos a *inductive analysis*. Trata-se de uma ferramenta que se adequa com uma estratégia de constante construção e validação teórica. O acumular de dados permite ao investigador proceder a uma constante interpretação e verificação da sua conformidade relativamente à pergunta de investigação. Desta forma, logo que se encontrem dados que comprometam a validade dos conceitos e das relações propostas, toda a teoria é adequada de maneira a compreender esta nova constatação. Segue-se um novo processo de recolha de dados por forma a tornar a teoria suficientemente consistente e generalizável. Este processo é difícil e extremamente trabalhoso já que obriga a uma constante adaptação do trabalho do investigador aos dados recolhidos. Pode, num extremo, a investigação nunca se dar por terminada já que a teoria só é considerada concluída quando não existam dados que a contrariem.

b) A análise de conteúdo

Em segundo lugar e embora a maior parte da pesquisa social se baseie em inquéritos ou entrevistas, os textos também são sobre os pensamentos das pessoas, sentimentos, memórias, planos e discussões e poderão ser utilizados como fontes para o desenvolvimento de investigação social. A análise que aqui se aplica denomina-se Análise de Conteúdo e poderá revestir uma forma qualitativa ou quantitativa.

Desde há décadas que se reconhece a importância de analisar textos recolhidos em primeira mão ou já existentes, importância acentuada hoje com a existência de novas tecnologias que possibilitam o acesso a uma quantidade enorme de informação em plataformas *on-line*, como os arquivos de jornais, programas de rádio e TV e Internet (Gunter, 2000).

Verifica-se então actualmente um interesse renovado em Análise de Conteúdo, em particular nas suas técnicas assistidas por computador (por exemplo MAXQDA para a análise qualitativa e SPSS para a análise quantitativa). No entanto, esta técnica só fará sentido se a investigação assim o demandar.

Exemplos de perguntas de partida que requeiram Análise de Conteúdo:

- Como é que os jornais de referência e os populares/sensacionalistas diferem na sua forma de noticiar a ciência e a tecnologia?
- A televisão comercial dirige-se às suas audiências de forma diferente da televisão pública — como?
- Quando e como é que o tema do ‹sucesso› aparece representado nos livros de crianças?
- Como é que a informação de memorandos internos de uma organização comunica?

A análise de conteúdo é então uma técnica muito utilizada em várias áreas do saber, nomeadamente em presença de fontes secundárias como textos provenientes de meios de comunicação social, como notícias, imagens, discursos ou guiões de séries ou filmes ou mesmo de fontes primárias como entrevistas.

No entanto, e embora tivéssemos optado por situar aqui esta análise, a nível do tratamento de informação, os mesmos dados poderão ser submetidos a estratégias de tratamento quantitativas ou qualitativas. A Análise de Conteúdo quantitativa contrasta com abordagens qualitativas como a Semiótica — o estudo/ciência dos signos. Esta é uma abordagem à análise de documentos que enfatiza a importância de procurar o significado profundo destes fenómenos. Preocupa-se por exemplo com o descobrir do processo de construção de significado e em saber como os signos são desenhados para terem um efeito nos seus consumidores. Outra abordagem qualitativa é designada como análise de conteúdo etnográfica — uma abordagem a documentos que enfatiza o papel do investigador na construção de significado dos e nos textos.

De acordo com Gunter (2000) existem vários tipos de análise qualitativa ao conteúdo:

1. Análise estruturalista semiótica/semiológica: para saber os significados profundos da mensagem;
2. Análise de discurso: também vista como uma forma de linguística crítica, presta atenção especial à componente linguística usada na linguagem dos media e avalia a prática ideológica da representação através da linguagem;
3. Análise retórica: analisa como a mensagem é apresentada visualmente ou textualmente (análise ‹estilística›); o núcleo da análise é a organização e a apresentação da mensagem e as escolhas do comunicador; foca-se em características distintivas como a composição, a forma, o uso de metáforas e a estrutura da argumentação;
4. Análise narrativa: neste tipo de análise o mais importante são as personagens e os seus actos, dificuldades, escolhas e desenvolvimentos e não tanto as características do texto; os textos são considerados histórias — a mensagem é tomada como uma versão editada de uma sequência de eventos, cujos elementos são descritos e caracterizados segundo a sua estrutura;
5. Análise interpretativa: nesta análise são usadas questões de pesquisa descritivas dirigidas a descobrir e formar teoria; os procedimentos de análise são cumulativos e comparativos; a relação entre os dados e os conceitos é fundamentalmente aberta.

Estas formas de análise são referidas como análise ao conteúdo qualitativa, com ênfase em permitir que as categorias emerjam dos dados e em reconhecer a importância do contexto.

Por outro lado, aquilo que comummente se designa *content analysis* tem um carácter marcadamente quantitativo e tem sido objecto de várias definições segundo diferentes autores. Segundo Bryman (2004), podemos defini-la como uma abordagem à análise de documentos e textos que procura quantificar conteúdo em termos de categorias pré-determinadas e de uma forma sistemática e replicável. Trata-se pois de análise de conteúdo de carácter quantitativo que não pode ser confundida com a primeira.

Segundo Krippendorf (1980), um dos mais reconhecidos autores sobre esta técnica, as estratégias de pesquisa com Análise de Conteúdo poderão servir para:

1. Construir um corpus como um sistema aberto para detectar tendências e padrões em mudança (ex: monitorização dos media — uma amostra é regularmente codificada para detectar mudanças e núcleos num conjunto de temas);
2. Fazer comparações que revelam diferenças que podem ser observadas entre a cobertura de diferentes jornais (por fontes), em discursos de um político a diferentes constituintes (por audiências) e entre artigos científicos e versões mais populares (*input-output*);
3. Construir índices como sinais de mudança social (ex: quantidade de cobertura científica em jornais pode ser uma medida da posição da ciência e tecnologia na sociedade);
4. Reconstruir ‹mapas de conhecimento› — as pessoas usam a linguagem para representar o mundo e neste caso a Análise de Conteúdo deve classificar os elementos não só pelas suas unidades como pelas inter-relações.

Os principais procedimentos na análise de conteúdo na vertente quantitativa compreendem uma sequência lógica da qual se identificam os principais passos:

1. Selecção de certos textos, sugerida pela teoria e circunstâncias;
2. Amostragem de textos, se forem demasiados para analisar completamente;
3. Construção de uma grelha de codificação que sirva tanto as considerações teóricas como os materiais;
4. Experimentar e rever a grelha de codificação e explicitamente definir as regras da codificação;
5. Testar a validade de todos os códigos e — quando é o caso — sensibilizar os codificadores para as ambiguidades;

6. Codificar todos os materiais da amostra e estabelecer a validade total do processo;
7. Construir um ficheiro para análise estatística — poderá ser analisado em SPSS recorrendo às principais técnicas descritas neste manual, de acordo com as necessidades do investigador;
8. Escrever um livro de códigos.

Em relação ao que contar, irá naturalmente depender da natureza da investigação (Bryman, 2004), mas podem referir-se como exemplos:

1. Actores significativos — principais intervenientes:
   a) Repórter generalista ou especialista;
   b) Foco do artigo — político, perito, representante de uma organização, etc.;
   c) Vozes alternativas — político, perito, representante de uma organização, etc.;
   d) Contexto para o item — entrevista, saída de relatório, acontecimento;
2. Palavras:
   a) Presença de certas palavras;
   b) Relação entre palavras (ex: estudo de Hansen de 1995 sobre a BSE mostrou associação entre ‹carne›, ‹comida›, ‹Governo›, ‹banir› e ‹louca›, ‹vaca› e ‹doença›);
3. Sujeitos e temas — deve-se procurar conteúdo manifesto mas também latente;
4. Disposições — ex: se o editorial tem comentários positivos, negativos ou apenas descritivos ou através de ideologias, crenças ou princípios.

c) A grounded theory

O terceiro processo de análise e tratamento de dados qualitativos é proposto por Strauss (1987) e Glaser (1992): a *grounded theory*. Trata-se de um processo mais estruturado do que o anterior, assenta numa técnica de codificação dos dados de maneira a permitir a criação de conceitos, a organização de categorias e o estabelecimento de relações entre conceitos para a criação de uma teoria. Por exemplo, numa entrevista é necessário codificar várias expressões utilizadas. Este processo de codificação é extremamente importante, já que permite a indexação do conteúdo da entrevista com categorias definidas como base da teoria que se pretende desenvolver. Portanto, a codificação é um processo de rotulação, separação e organização da informação. Depois de codificadas as expressões, constituem-se conceitos organizados que posteriormente serão analisados de modo a solidificar as relações entre os mesmos e assim sustentarem as assunções teóricas do investigador. Este processo

repete-se até conseguirmos um elevado nível de saturação teórica. Isto é, até termos dados suficientes para completarmos toda a amplitude do conceito e permitir a generalização da teoria.

d) A análise normativa

Finalmente, temos a análise normativa sugerida por Bryman (2004), como forma de tratamento de dados qualitativos. É bastante similar à *inductive analysis*, no entanto, neste caso, o investigador pretende obter toda a informação possível usando meios mais indirectos. Numa situação social, por exemplo, o investigador procura obter informação relativa à mesma situação, usando diferentes fontes. Não procura directamente saber o que aconteceu, mas procura entender a forma como cada indivíduo viveu aquela situação. Esta estratégia é adequada para entender as motivações ou as estruturas sociais dos indivíduos já que procuramos absorver a forma como cada indivíduo descodifica a sua própria realidade.

6. Considerações finais

Ao longo deste capítulo procurámos dar, ao investigador, uma imagem das várias formas alternativas de análise e tratamento de dados. Se a forma como o investigador vai tratar os dados, do seu trabalho, estiver definida logo à partida, o restante trabalho de investigação poderá ganhar uma orientação mais clara. Ou seja, conhecendo o tipo de tratamento a fazer conseguimos agilizar a revisão de literatura com os conceitos a serem tratados; melhoramos o nosso modelo de análise; aperfeiçoamos as hipóteses de trabalho. Esta clareza e linearidade de pensamento reflectir-se-á, indubitavelmente, na qualidade do trabalho. Desta forma, tenta-se evitar o inconveniente para o investigador de, após a construção do modelo e a enumeração de hipótese, ter dificuldades em encontrar os melhores mecanismos para concluir a parte empírica do seu trabalho.

Começámos este capítulo com uma breve discussão acerca das estratégias alternativas de investigação: Qualitativa *vs.* Quantitativa. Depois, com base nesta diferenciação, apresentámos diferentes tipos de investigação e maneiras alternativas da recolha e tratamento de dados. Continuamos que um erro a este nível, e neste momento da investigação, tenderá a comprometer toda a análise de dados que possa vir a ser feita.

Do tratamento e análise de dados, foram apresentadas as principais técnicas de tratamento e análise de dados, através de métodos estatísticos de descrição, associação, comparação de populações, estudos de proporções e estudos de causalidade.

Com a sua apresentação e análise esperamos ter dotado o investigador de competências para, em função da natureza das suas variáveis, poder associar variáveis, testar correlações, criar relações de causalidade, etc. No entanto, gostaríamos de chamar a atenção para o facto de, por vezes, existir um excesso de confiança nas estatísticas obtidas. Em caso algum podemos esquecer que, tratando com ciências sociais, devem procurar a explicação inerente à realidade latente através dos números. Isto é, devemos ir para além dos resultados dos testes e procurar compreender a sua total amplitude e profundidade. Como tal, apresentamos também algumas ferramentas de análise e tratamento de dados qualitativos.

Notas

1. Em termos econométricos é possível verificar se existem variáveis explicativas excluídas do modelo. Para tal, testa-se a heteroscedasticidade dos erros, permitindo, caso exista, concluir pela existência de uma, ou mais, variáveis explicativas, inadvertidamente, excluídas da análise.
2. Em termos econométricos é possível realizarem-se testes de autocorrelação. Isto é, verificar se, entre as variáveis dependentes, existe uma independente que afecte o comportamento de outras variáveis independentes, perdendo assim as suas características explicativas.
3. A regressão foi criada usando o método enter que utiliza todas as variáveis independentes. Há outras alternativas, como o método *stepwise*, no qual as variáveis independentes são acrescentadas gradualmente e mantidas em função da significância que acrescentam ao modelo. Ainda existem os métodos *remove, backward, forward.*
4. Todas as análises estatísticas deste capítulo foram realizadas com valores fictícios e com recurso aos programas estatísticos SPSS v.17 e o STATA v.10.
5. No que diz respeito à *grounded theory*, as abordagens de Strauss e de Glaser são, na sua especialidade diferentes, no entanto, para o objectivo deste capítulo, foram tratadas como iguais.

Referências Bibliográficas

Aldrich, J., F. Nelson, 1984. *Linear Probability, Logit, and Probit Models.* Newbury Park: Sage Publication.

Blaikie, Norman, 2003. *Analyzing Quantitative Data: From Description to Explanation.* London: Sage Publications.

Blalikie, Norman, 2008. *Designing Social Research.* Cambridge: Polity Press.

Borooah, V. K. 2002. *Logit and Probit: Ordered and Multinomial Models.* SAGE.

Bryman, Alan, 1998a. *Doing Research in Organizations.* London: Routledge.

Bryman, Alan, 1998b. «Introduction», *in* Bryman, Alan (eds.), *Doing Research in Organizations*, London: Routledge, pp.: 1–20.

Bryman, Alan, 2004. *Social Research Methods* — 2nd Edition. Oxford: University Press.

Dhrymes, P., 1986. «Limited Dependent Variables». *In* Griliches, Intriligator eds., *Handbook of Econometrics Amesterdam*: North-Holland, pp.: 1567–1631.

Gunter, B., 2000. *Media Research Methods — Measuring Audiences, Reactions and Impact.* London: Sage Publications.

Glaser, B., 1992. *Basics of grounded theory analysis.* Mill Valley, CA: Sociology Press.

Krippendorff, K., 1980. *Content Analysis — An introduction to its methodology.* California: Sage Publications.

Liao, T., 1994. *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models.* Thousand Oaks: Sage Publication.

Maroco, J., 2007. *Análise Estatística.* Lisboa: Sílabo.

Menard, S., 2002. *Applied Logistic Regression analysis.* Thousand Oaks, CA: SAGE.

Pestana, M., Gageiro, J., 2005. *Análise de Dados para Ciências Sociais.* Lisboa: Sílabo.

Strauss, A., 1987. *Qualitative analysis for social scientists.* Cambridge, England: Cambridge University Press.

Zimmermann, V., 1996. «Pseudo-R2 Measures for Some Common Limited Dependent Variable Models.» *Sonderforschungsbereich 386*, paper 18.