# Metodologia de Pesquisa Jurimétrica

Ricardo Feliz Okamoto Julio Trecenti

11/01/2022

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Boas vindas!

Este livro foi produzido em *Quarto*. Todos os gráficos produzidos pela ABJ são reprodutíveis, ou seja, qualquer pessoa interessada em refazer qualquer gráfico ou revisar os códigos utilizados pode fazê-lo. O livro foi gerado com o software de publicação científica Quarto na versão 1.2. Os gráficos e análises foram gerados com o software estatístico R na versão 4.2.1.

# 1 Introdução

A Jurimetria é um ramo do Direito ainda em definição. Não se sabe ao certo a que métodos e técnicas se refere a Jurimetria, tampouco como a própria Jurimetria concebe o seu objeto de estudo, isto é, o Direito. Em outras palavras, a Jurimetria não é dotada de metodologia e epistemologia próprias, de forma que não é possível (ainda) distinguir completamente a Jurimetria da Análise Econômica do Direito, ou de outros estudos que tratem quantitativamente do Direito, tal como as Ciências Políticas podem fazer, ou como a própria Estatística o faz. Hoje em dia, a Jurimetria é definida pelas aplicações existentes.

Ainda assim é possível traçar algumas considerações sobre do que se trata a Jurimetria. A definição mais precisa que temos atualmente é a de Marcelo Guedes Nunes, que a define como uma (a) disciplina do conhecimento que (b) utiliza a metodologia estatística para (c) investigar o funcionamento de uma ordem jurídica[1]. Essa definição contém três elementos: (a) a taxonomia da jurimetria, (b) o seu método e (c) objeto.

Sobre a taxonomia, é importante frisar dessa definição a Jurimetria enquanto uma "disciplina do conhecimento", porque é muito comum encontrar uma redução dessa área do conhecimento a somente o seu método. Ou seja, é frequente a definição da Jurimetria como somente "a aplicação de métodos estatísticos ao Direito". Mas uma área do conhecimento envolve muito mais do que somente o seu método. Uma área do conhecimento implica uma forma de se conhecer o seu objeto de estudo própria. Então é disso que se trata a Jurimetria: sobre uma forma de conhecer e conceber o Direito.

Sobre o método, talvez esse seja o elemento mais claro da Jurimetria, ele é, de forma genérica, a aplicação de métodos estatísticos ao Direito. Isso significa que a Jurimetria incorpora os métodos e técnicas quantitativos no estudo do Direito. Entretanto é importante salientar que por "métodos e técnicas quantitativos" não estamos nos referindo a um único sistema homogêneo de métodos e de técnicas. Existem uma infinidade de discussões metodológicas e de divergências teóricas dentro da própria Estatística, como os debates sobre estatística frequentista e bayesiana. A Jurimetria não se posiciona de nenhum lado das discussões da Estatística; ela não designa nenhum método ou técnica em específico dentro da Estatística. A Jurimetria apenas significa a aplicação destes métodos, em suas mais variadas formas, em todas as suas vertentes.

Sobre este ponto, inclusive, é importante fazermos uma ressalva a respeito deste livro, porque, por mais que a Jurimetria enquanto área do conhecimento não se posicione em nenhuma discussão da Estatística, preferindo ou preterindo um ou outro método, não vamos abordar essas divergências, tampouco abordaremos todas as ramificações e abordagens possíveis dentro da Estatística. Em outras palavras, mesmo que a Jurimetria não se designe métodos e técnicas específicas, este livro irá abordar apenas algumas dessas técnicas, realizando, portanto, escolhas do que falar e do que não falar. Mas nada disso significa que os métodos e técnicas difundidos por este livro sejam exaustivos de todo o ramo da Jurimetria. Aqui, falaremos dos métodos e técnicas que a Associação Brasileira de Jurimetria mais comumente utiliza em suas pesquisas.

Por fim, a respeito do objeto, a Jurimetria se propõe a discutir "o funcionamento da ordem jurídica". São duas palavras importantes a respeito desse objeto: "funcionamento" e "ordem jurídica". A começar pela ordem jurídica, este conceito

[1] Nunes (2016)

é muito amplo e genérico. O que importa salientar a seu respeito é que por "ordem jurídica" nos referimos a um objeto maior do que simplesmente o conjunto de normas que compõem algum ordenamento jurídico. Além das normas, existem as suas interpretações, métodos de hermenêutica, as discussões doutrinárias e jurisprudenciais. Mas a Jurimetria não se delimita somente por estudar a "ordem jurídica". Mais do que isso, ela estuda o "funcionamento" dessa ordem. Isso significa que o que se estuda são, não as normas em abstrato, não a ordem jurídica no plano normativo, mas a ordem jurídica no seu plano concreto, factual, com especial atenção para a atuação dos tribunais, dos operadores do Direito e da Administração Pública.

Feita uma definição sobre a Jurimetria, é importante tratarmos de algumas notas epistemológicas sobre a forma (métodos quantitativos) como a Jurimetria conhece o seu objeto de estudo (o funcionamento da ordem jurídica). Tratar desses problemas é uma forma de alerta para os limites inerentes à Jurimetria, e também, possivelmente, uma forma de se preparar para as críticas a que estão sujeitos os trabalhos e estudos nessa área. Trataremos de dois problemas. O primeiro diz respeito a um problema geral de toda a pesquisa empírica do Direito (seja quantitativa, seja qualitativa); o segundo diz respeito a um problema geral de toda a pesquisa quantitativa (seja no Direito, ou fora dele). A Jurimetria é um ramo único pois é a única área do conhecimento que concentra esses dois problemas em um único lugar.

O primeiro problema epistemológico (o problema geral de pesquisa empírica no Direito) diz respeito à natureza das proposições que podemos fazer sobre o Direito. O Direito trata de afirmações *normativas* sobre o mundo, enquanto toda pesquisa empírica no Direito realiza afirmações *descritivas.* Dessa forma, surge uma questão que José Xavier coloca muito bem: as pesquisas empíricas cujo o objeto é o Direito são pesquisas **em** Direito, ou pesquisas **com** o Direito ou **sobre** o Direito?[2] Essa pergunta é importante, porque ela coloca uma questão epistemológica essencial: seriam as pesquisas empíricas no Direito pesquisas feitas a partir do ponto de vista interno ou externo do Direito? É possível uma pesquisa empírica em Direito, isto é, assumindo o seu ponto de vista interno? Essa pergunta indica que ainda está em debate se a Jurimetria pode realmente apresentar um ponto de vista interno ao Direito, ou se, por natureza, ela está condenada a apresentar sempre um ponto de vista externo?

Sobre o segundo problema (o problema geral de métodos quantitativos), é importante começarmos a delimitá-lo dizendo que todos os ramos das ciências sociais – e é importante aqui frisar que estamos falando das ciências sociais, e não das ciências naturais – buscam, de uma forma ou de outra, apreender a realidade social, dar sentido a ela, mas cada uma das diferentes áreas das ciências sociais possui um conceito de verdade distinto. Dessa forma, por mais que vários ramos das ciências humanas olhem para um mesmo objeto (como o Direito, por exemplo), elas vão olhar para ele a partir de lentes distintas e chegar a conclusões distintas. Essas lentes são essenciais porque em certo grau elas determinam o tipo de conclusão que se pode tirar a respeito desse objeto. Disso então decorre uma pergunta importante: que tipos de perguntas e de conclusões podemos tirar a respeito do Direito quando olhamos para ele a partir de dados, aplicando métodos quantitativos? E o que não podemos enxergar dessa forma?

Este livro não vai se debruçar sobre essas duas questões mais profundamente, mas são questões que devem ficar na cabeça de quem lê sempre que ela for aplicar as técnicas aprendidas aqui.

Estes eram os pontos iniciais importantes de serem tratados antes que nos iniciássemos no método e nas técnicas quantitativas. O que o livro se propõe a fazer de agora em diante é descrever as técnicas empregadas para fazer análises quantitativas.

[2] A citação é na verdade: "Se concebermos o direito como o mundo da doutrina, o mundo da elaboração teórica de categorias para a tomada de decisão, qual espaço resta para a pesquisa empírica **em** direito? Em outras palavras, para que a pesquisa empírica seja em direito, e não apenas **com** ou **sobre** o direito, é preciso ter uma concepção do direito que compreenda que o direito é aquilo que pode ser observado para além de construções doutrinárias e normas positivadas." Xavier (2015)

# 2 Planejamento de Pesquisa

A primeira tarefa a se realizar em qualquer pesquisa quantitativa é o planejamento da pesquisa. O planejamento da pesquisa, ou o *desenho de pesquisa*, é um plano no qual iremos expor de maneira clara e detalhada como pretendemos investigar o objeto que desejamos; e como iremos tirar conclusões a partir das análises que realizaremos.

No planejamento da pesquisa iremos falar dos métodos e técnicas que serão utilizadas (regressão, estudo de caso, etc), mas não podemos parar nisso. É fundamental que o planejamento de pesquisa deixe muito claro os pressupostos da pesquisa e que discorra sobre os detalhes dela. Não basta dizer, por exemplo, se o método utilizado será de regressão; é preciso dizer qual é a pergunta teórica que se está interessado em responder; e porque a regressão é um modelo adequado para responder a essa pergunta; e que variáveis serão utilizadas para representar os conceitos abstratos (operacionalização).

Para termos uma dimensão crítica do planejamento de pesquisa, precisamos tratar de alguns temas, a saber (i) qual é o papel da teoria em uma pesquisa empírica (*Teoria e Empiria*); (ii) o que difere uma pergunta jurídica, de uma pergunta jurimétrica (*Uma Pergunta de Pesquisa entre Dois Mundos*); (iii) como delimiltar o escopo da pesquisa (*Escopo do Estudo*); (iv) como operacionalizar conceitos abstratos em dados concretos (*Operacionalização de Conceitos*); (v) de onde extrair dados e o que isso implica (*Dados, Tribunais e Processos de Geração de Dados*); (vi) se o estudo será amostral ou populacional e como realizar amostragens, caso seja necessário (*Amostragem e População*).

## 2.1 Teoria e Empiria

Por mais que exista a distinção entre pesquisa teórica e pesquisa empírica, é importante começar a falar de pesquisa empírica deixando claro um ponto: *toda pesquisa empírica deverá se fundar, necessariamente, em uma teoria, pois não existe empiria sem teoria.* Um erro comum em pesquisas de dados é o de deixar os dados falarem por si só, deixar os dados guiarem sua pesquisa. Isso normalmente ocorre quando há uma falta de teoria para embasar a análise dos dados. Então não deixemos os dados liderarem a pesquisa, mas tomemos frente neste processo. A teoria é o ponto de partida em uma pesquisa empírica. É a partir dela que iremos pensar sobre o que queremos pesquisar; e é a partir dela que iremos formular nossa pergunta de pesquisa. Os dados nunca guiam a pesquisa. Para continuarmos essas reflexões, há algumas orientações gerais que devemos fazer sobre a teoria em uma pesquisa empírica.

### 2.1.1 O que a teoria pode nos dar em uma pesquisa empírica?

O primeiro grupo de orientações diz respeito ao que a teoria pode nos dar. A partir dos debates acadêmicos, iremos identificar perguntas relevantes *para o mundo e para os nossos pares.* As perguntas serão relevantes para o mundo na medida em que ela tentar responder a um problema real. Mas isso não basta. É preciso ter muita clareza do debate teórico e acadêmico dentro do qual estamos nos situando. Um mesmo problema pode ser abordado por diversas correntes acadêmicas de diferentes formas. A teoria, então, irá nos guiar a buscar perguntas relevantes e a nos colocar no debate.

Então, de alguma forma, a teoria irá nos ajudar a identificar problemas no mundo e a debater com nossos pares. Falando especificamente sobre teorias no Direito, como elas podem nos ajudar em pesquisas empíricas? Por "teorias no Direito" me refiro tanto a estudos sociojurídicos, como à doutrina como um todo. A doutrina é essencial para que alguém, interessado em realizar uma pesquisa jurimétrica, identifique questões relevantes a serem investigadas. Como saber de antemão que uma questão relevante em matéria de usucapião é a *res habilis*, isto é, qual bem pode ser usucapido? Como saber que é importante extrair o *tipo de parte* de quem ingressa com uma ação civil pública, senão por meio da doutrina informando e discutindo a legitimidade das partes para propor este tipo de ação?

Do outro lado, a sociologia jurídica pode fornecer outro tipo de substrato para embasar as pesquisas empíricas. Por exemplo, é a partir do teorema de Priest & Klein (1984) que podemos identificar o *sentido* da proporção de sentenças ser 50% favorável ao autor. Não cabe aqui explicar este teorema, mas basta indicar que ele abre portas analíticas importantes para futuras pesquisas em jurimetria.

Assim, o que podemos concluir é que a teoria sempre irá embasar e guiar um estudo empírico.

### 2.1.2 Como formular teorias a partir de dados empíricos?

O segundo grupo de orientações diz respeito a como a nossa pesquisa poderá contribuir com o conhecimento em geral, ou seja, a como podemos, a partir dos dados e da teoria que estimulou a pesquisa, criarmos uma teoria em retorno. Para tanto, é preciso ter em mente algumas noções.

Em primeiro lugar, é preciso que a teoria formulada seja *falseável*. Isso significa que a teoria deve ser formulada de uma forma que possa ser dita falsa. Uma "teoria" que nunca poderá ser falseada não é propriamente uma teoria, em termos científicos. E uma teoria só poderá ser falseada "se não estiver vazia a classe de falseadores potenciais"[1]

Partindo deste princípio da falseabilidade, como, então, devemos formular teorias? A falseabilidade da teoria advirá da quantidade de *implicações observáveis* que ela criar. Em quanto mais manifestações no mundo empírico a teoria for capaz de se exprimir, mais momentos ela terá para ser testada. E a quantos mais testes a teoria se mantiver de pé, mais robusta ela será. As implicações observáveis, portanto, são a forma de testar a teoria e, dessa forma, é por meio dessas implicações que conectamos a teoria com os dados[2].

Em segundo lugar, é preciso se atentar para um *trade off* entre a *generalidade da explicação* e a *parcimônia*. O que desejamos sempre é que a nossa teoria explique da melhor maneira possível a realidade. Entretanto, às vezes, a explicação mais precisa é uma explicação muito difícil de ser compreendida. Neste caso, talvez valha a pena sacrificar a generalidade da teoria, em prol de parcimônia. Por outro lado, teorias muito simples acabam sendo teorias não falseáveis, o que gera outro tipo de problema.

Esse tipo de *trade off* é muito presente em discussões sobre inteligência artificial. Alguns modelos preditivos conseguem, com muita acurácia, predizer certas situações. Entretanto, esses modelos precisam de aproximações para serem interpretados por humanos[3]. Em muitas decisões médicas, tem-se utilizado inteligências artificiais para diagnosticarem pacientes[4]. A acurácia dessas máquinas é muito maior do que a de um médico experiente. Entretanto, não há explicação clara para dar ao paciente. O que acontece nestes casos é que os modelos estatísticos que

[1] Popper (1934), p. 91
[2] Epstein e Martin (2014a)
[3] Para detalhes de como interpretar modelos black-box, ver Molnar (2022)
[4] Essas discussões todas sobre a interpretabilidade dos modelos de inteligência artificial estão presentes em uma reportagem de Laura Spinney para o The Guardian. Ver Spinney (2022)

constituem as redes neurais não foram feitos com o propósito de explicar o mundo; eles foram feitos para gerarem boas predições[5].

De alguma forma, as discussões sobre parcimônia e generalidade retratam esse mesmo tipo de problemática. Quando vamos elaborar um modelo matemático para explicar determinada relação no mundo, se perdermos de vista a parcimônia, podemos cair no mesmo problema da inteligência artificial e criamos modelos muito bons do ponto de vista explicativo, mas pouco práticos do ponto de vista de sua interpretação.

Assim, em linhas gerais, o que se recomenda é que a teoria seja falseável, genérica e parcimoniosa.

## 2.2 Uma Pergunta de Pesquisa entre Dois Mundos

A partir da teoria, criamos perguntas. Essas perguntas são de nosso interesse pessoal imediato, mas também contribuem para o conhecimento geral e discutem com os nossos pares. Uma vez formulada, a pergunta de pesquisa se torna o eixo fundador da pesquisa, pois ela irá guiar todo o projeto a seguir. Essa tarefa de elaborar uma pergunta de pesquisa, entretanto, se mostra difícil quando nos situamos entre duas áreas do conhecimento tão distintas. De um lado, há o Direito, cujas perguntas são elaboradas de forma normativa. Do outro, há a Estatística, que busca descrever o mundo por meio de dados. As perguntas de natureza normativa e descritiva são essencialmente diferentes. Daí a dificuldade de se elaborar perguntas em jurimetria.

No Direito, estamos acostumados com a elaboração de perguntas de natureza *normativa*. Queremos descobrir o conteúdo jurídico de normas, discutir a hermenêutica de leis, a *ratio decidendi* de tribunais ou o regime jurídico de determinado instituto. Todos esses temas acabam se voltando em perguntas cujo centro da investigação é a *norma*. Questões normativas se valem do léxico do dever ser, e mesmo quando se valem do "ser", elas o fazem descrevendo a norma, descrevendo o abstrato. Alguns exemplos são:

- Quem controla jurisdicionalmente a discricionariedade no Brasil? Qual é o objeto do controle? Como **deve ser** a discricionariedade controlada?[6]
- O que **é** o interesse público?[7]
- Como **devemos** tratar as demandas em direito à saúde, elas "**devem ser** consideradas como ações individuais e sujeitas às regras estabelecidas no Código de Processo Civil ou **deveriam ser** tratadas como coletivas e sujeitas às normas do Código de Defesa do Consumidor e da Lei da Ação Civil Pública? Ou deveriam receber um regime legal intermediário mais adequado?"[8]

A abordagem deve ser outra quando falamos de estudos jurimétricos. A questão central não é mais como as normas *devem ser*, não é mais interpretá-las em um plano abstrato; a questão agora passa a ser *descrever* a incidência dessas normas no mundo real. Aparecem, então, questões de muitas naturezas.

Um primeiro grupo de questões que aparecem quando olhamos de forma empírica para o direito é a discussão a respeito **do efeito e da eficácia** das normas. Alguns exemplos são:

- Quais os "impactos da MPV 1.040/2021 no tempo de abertura de empresas"?[9]

[5] Breiman (2001)
[6] Perez (2018)
[7] Lopes (2003)
[8] Grinover (2014)
[9] Trecenti e Nunes (2021)

- "Como os agentes tomadores de decisão reagiram à lei 11.343/2006? O padrão das apreensões mudou depois de 2006?"[10]

Ao lado destas questões, que buscam analisar especificamente o efeito de normas na realidade, aparece um segundo grupo de questões. Esse grupo traz questões que visam descrever a realidade subjacente à norma. Nesse sentido, aparecem pesquisas que buscam compreender:

- Qual é a amplitude dos *habeas corpus* nos tribunais superiores? E quais são as principais teses jurídicas que são levadas aos tribunais superiores por meio de *habeas corpus*?[11]
- Quem acessa o STF e quais questões são levadas à sua apreciação e deliberação? Como decide o STF?[12]
- Quais são as funções contemporâneas do mandado de injunção?[13]

A distinção entre as pesquisas dogmáticas e as pesquisas jurimétricas é evidente, quando olhamos para o tipo de perguntas de pesquisa que podem ser feitas. Dessa oposição entre as naturezas das perguntas, surge um questionamento comum: É possível dizer algo sobre o *dever* ser a partir do *ser*? Ou seja, é possível dizermos algo sobre o campo normativo a partir da descrição da realidade? E, de forma inversa, o que o mundo normativo pode nos informar sobre o mundo real? Essas são duas relações entre o direito e a estatística que devem ficar em mente quando alguém buscar realizar um estudo jurimétrico.

Essa indagação epistemológica (que busca relacionar o dever ser com o ser) não é exclusivamente da jurimetria, mas é de qualquer estudo empírico do direito, pois todos os estudos empíricos falam da realidade e não da norma. A única diferença da jurimetria para as demais ciências empíricas é que a forma de acessar a realidade, na jurimetria, é por meio da análise de dados quantitativos e não por outros métodos.

Não é o espaço deste livro discutir com mais afinco tais questões. Há, inclusive, uma lacuna muito grande no conhecimento para se discutir isso no momento. O importante é ter em mente que: a pessoa que pretender realizar uma pesquisa jurimétrica deve ter consciência de que as perguntas que ela pode responder por meio de métodos quantitativos talvez não respondam às suas indagações normativas. Essa consciência é importante para determinar se o desenho de pesquisa deverá envolver métodos empíricos quantitativos ou não. Para os fins deste livro, iremos trabalhar apenas com questões que necessariamente envolvem tais métodos, mas na prática do dia a dia, é possível que nem todos os questionamentos possam ser respondidos por métodos quantitativos. Essa limitação não é um problema, pois há muito o que pode ser respondido por esses métodos já.

## 2.3 Escopo do Estudo

Outra etapa do desenho de pesquisa é a delimitação do escopo do estudo. Há duas dimensões a que devemos nos atentar ao delimitar o escopo do estudo: a dimensão temporal e a dimensão geográfica. Temos que fazer essa delimitação porque, caso contrário, a pesquisa pode tornar-se inviável. Há duas dificuldades que podem inviabilizar a realização da pesquisa caso não haja um escopo bem definido. Primeiro, há um problema prático, de que é muito difícil obter dados para escopos muito amplos. E em segundo lugar, escopos muito largos, tanto geográfica, como temporalmente, podem tornar a variabilidade nos dados muito grandes, dificultando o seu controle na pesquisa.

[10] ABJ (2019)
[11] Bottino (2015)
[12] Sundfeld et al. (2011)
[13] Fulgêncio e Costa (2018)

### 2.3.1 Escopo temporal

Da dimensão temporal, há dois tipos de estudos possíveis. De um lado, temos estudos prospectivos; do outro, temos estudos retrospectivos. Estudo prospectivo é o estudo que acompanha o processo judicial desde a data de distribuição até o fim. O fim pode ser marcado pela data da sentença, acórdão, ou outro evento de interesse. Ou seja, os casos são indexados pela data de nascimento, e acompanhados até a data de sua morte. Em muitos casos, os processos ainda não atingiram o fim no momento da realização do estudo.

Estudo retrospectivo é o estudo que levanta processos que acabaram (por sentença ou por acórdão) e analisa suas características. Ou seja, os casos são indexados pela data de morte.

Estudos prospectivos são úteis quando o intuito é estudar o tempo das fases do processo. Já estudos retrospectivos são úteis para a análise do perfil de decisões. Estudos que analisam tempos em bases retrospectivas.

A Figura 2.1 mostra os diferentes escopos temporais de cada estudo.

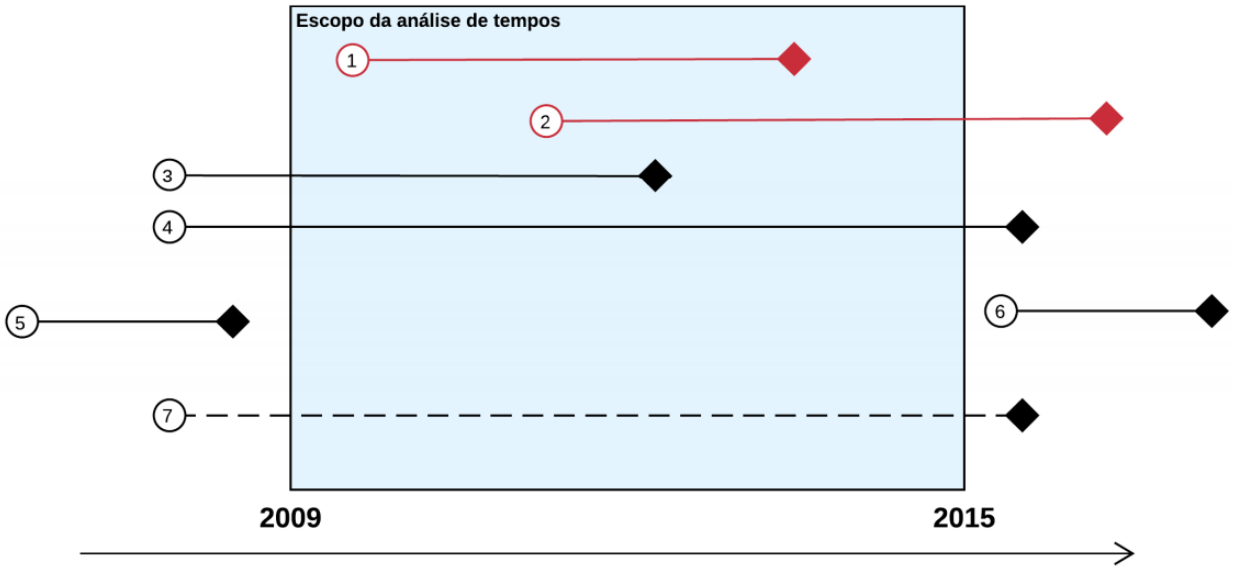



Figura 2.1: Diferentes escopos temporais de cada estudo

(1) Prospectivo e retrospectivo
(2) Apenas prospectivo
(3) Apenas retrospectivo
(4) Nenhum dos dois, mas poderia ser capturado por atividade no período
(5) fora do escopo
(6) fora do escopo
(7) Nenhum dos dois tipos e não poderia ser capturado (ficou inativo no período)

### 2.3.2 Escopo geográfico

Do lado geográfico, o que temos que determinar é, muitas vezes, a região que iremos estudar: será a Justiça Estadual ou Federal? Serão todas as Justiças Estaduais? Ou apenas uma única Justiça Estadual? Os Tribunais Superiores vão entrar na análise? Vou estudar Tribunais Administrativos também? De que lugares?

Todas essas questões dependem de uma série de discussões de Direito para serem respondidas. Existem ações que apenas os Tribunais Superiores têm competência originária, o que levaria a um interesse maior em se estudar o STJ e o STF. Pode também haver um interesse maior em se estudar crimes em determinada localidade, por ter nela alguma característica especial. Neste caso, deve-se determinar qual é o foro competente, pois será este o Tribunal de interesse.

O que acontece, porém, muitas vezes, é que não há um critério objetivo para se determinar o escopo geográfico da pesquisa. Neste caso, é muito frequente a utilização de um Tribunal de Justiça conveniente ao pesquisador.

Além da conveniência de se escolher um TJ com familiaridade, é muito importante também pensar na disponibilidade dos dados. Alguns lugares possuem dados mais estruturados, de mais fácil acesso; outros ainda, permitem a raspagem de dados. Mas há tribunais que dificultam a raspagem por meio de CAPTCHAs.

Todas essas questões devem ser registradas ao se escolher o escopo geográfico de pesquisa. Apesar de que escolhas de conveniência não sejam as mais recomendadas, na prática, são elas que guiam a delimitação do escopo geográfico. Desde que essa escolha seja fundamentada e reportada, não há problema nisso. Mas sempre que for possível utilizar outro critério para determinar o escopo, melhor não utilizar um critério de conveniência.

## 2.4 Operacionalização de Conceitos

Normalmente, as questões teóricas que nos interessam são formuladas em termos abstratos e conceituais, tais como: insegurança jurídica, independência do judiciário, pacificação de conflito social, eficiência, eficácia. A teoria se elabora em cima de conceitos, e conceitos não são visíveis, públicos, confrontáveis, observáveis. Assim, para se analisar empiricamente um conceito, uma grande questão aparece: como *representar* os conceitos de interesse a partir da realidade?

Preliminarmente, é necessário reconhecer que nenhuma representação de um conceito será precisa, absoluta, unânime e pacífica. Utilizamos representações específicas para conceitos específicos. Em um sentido, as representações são subjetivas, pois ligam-se ao sujeito que irá realizar a pesquisa. Por isso, é preciso *evidenciar* como essa representação está sendo feita e justificá-la teoricamente. Com isso, reforçamos o ponto de que *não existe estudo empírico sem teoria.* É uma ilusão acreditar que os dados poderão indicar algo sobre a realidade por si só. Ao lado de toda análise empírica, deve haver uma boa teoria. No caso da operacionalização, será a teoria que irá nos guiar a selecionar a melhor representação da realidade.

Tendo em vista a limitação das representações dos conceitos, podemos voltar a discutir a operacionalização. "Operacionalizar um conceito" significa selecionar alguma representação de um conceito. Se eu quero discutir se a independência judicial contribui para a liberdade econômica, eu devo pensar como eu vou representar os conceitos "independência judicial" e "liberdade econômica". Qualquer medida que eu queira fazer sobre esses conceitos necessariamente irá passar por uma transformação do conceito in abstrato para alguma representação concreta. Por causa dessa impossibilidade de se falar dos conceitos em abstrato nas pesquisas empíricas, a pergunta de interesse

(formulada em termos teóricos, abstratos e conceituais) acaba se transformando em algo muito distante da pergunta original.

Epstein e Martin trazem um exemplo interessante, a partir do estudo de La Porta et al (2014a). A pergunta teórica deste estudo é: "Do independent judiciaries promote economic freedom?". Entretanto, ao passar pelo processo de operacionalização, a questão que será testada se torna; "In 71 countries, do longer tenures for judges lead to fewer steps that a startup business must take in order to obtain legal status?". Ou seja, por mais que o estudo esteja interessado em discutir a relação entre independência judicial e liberdade econômica, o que se está sendo observado são outras coisas. Para representar a independência judicial, os autores usam a duração do tempo de estabilidade dos juízes. Quanto maior o tempo de estabilidade, maior a independência judicial nessa lógica. Do outro lado, para representar a liberdade econômica, os autores estão observando a burocracia para que uma startup consiga seu status jurídico. Quanto menos passos uma empresa tiver que tomar para tanto, maior será a liberdade econômica.

Podemos discutir, através desse exemplo, a questão principal por trás da operacionalização: como mensurar conceitos abstratos? Mensurações ruins distanciam de tal forma o mundo real observado da relação teórica buscada que elas podem minar a discussão por trás. Então em todas as pesquisas iremos passar por uma fase de operacionalização dos conceitos.

É na etapa da operacionalização que iremos buscar por indicadores para representar os conceitos. O mais importante é sempre reportar as escolhas feitas.

## 2.5 Dados, Tribunais e Processos de Geração de Dados

A operacionalização não pode ser pensada dissociada dos dados existentes no mundo. A forma como a teoria será investigada empiricamente está *constrangida* pelas limitações dos dados no mundo real. Alguns dados talvez não existam, ou sua qualidade por ser ruim, podendo haver muitas lacunas, informações faltantes ou informações inconsistentes. E todos esses problemas com os dados irão impactar a operacionalização dos conceitos. Por vezes, um conceito pode ser melhor representado por um determinado dado, mas a qualidade deste dado pode estar tão prejudicada, que isso inviabilize a pesquisa. Neste caso, talvez outra operacionalização deve tomar lugar.

Dessa forma é que se faz extremamente importante *conhecer os seus dados.* Por "conhecer os dados", queremos dizer que deve-se conhecer o *processo de geração dos dados* (*data generating process*). O processo de geração de dados é a forma, por meio da qual, determinada informação foi criada. É extremamente importante documentar e reportar todos esses processos, para que os pares acadêmicos possam validar a pesquisa e as escolhas metodológicas feitas nela.

### 2.5.1 Dados encontrados

A seguir, temos um exemplo de como os dados são gerados, em um contexto jurídico. Quando vamos analisar processos que tramitam no Judiciário, é muito frequente nos valermos das informações de capa do processo, a saber, o número do processo, nome e funções das partes, data de distribuição, valor da causa, classe e assunto processuais. Entretanto, ao utilizarmos essas informações, nem sempre temos clareza do que está por trás daquelas informações. O caso mais emblemático é aquele do assunto e classe processuais. O assunto e classe são informações padronizadas por meio da Resolução nº 46 do CNJ, que criou as Tabelas Processuais Unificadas. Em seu art. 3º, a Resolução disciplina que "todos os processos ajuizados (processos novos), *antes de distribuídos, deverão ser cadastrados de acordo com as tabelas unificadas de classes e assuntos processuais*", estabelecendo, pois, a obrigatoriedade de se seguir as TPUs.

Ao lado da obrigatoriedade de se cadastrar um processo com a sua classe e assunto processuais, é preciso saber também quem está encarregado de o fazer. Cada Tribunal possui uma Resolução própria regulamentando isso, mas a regra é sempre a mesma: o advogado, na hora de peticionar, é quem irá, obrigatoriamente, escolher a classe e o assunto processuais. No caso do TJSP, essa informação está disposta no art. 9º, inciso I, da Resolução nº 551/2011, que disciplina: "Art. 9º - A correta formação do processo eletrônico é responsabilidade do advogado ou procurador, que deverá: I - preencher os campos obrigatórios contidos no formulário eletrônico". Ao olharmos para os campos de preenchimento no ESAJ, encontramos "Classe" e "Assunto" com um asterisco, indicando a sua obrigatoriedade, conforme a Figura 2.2.

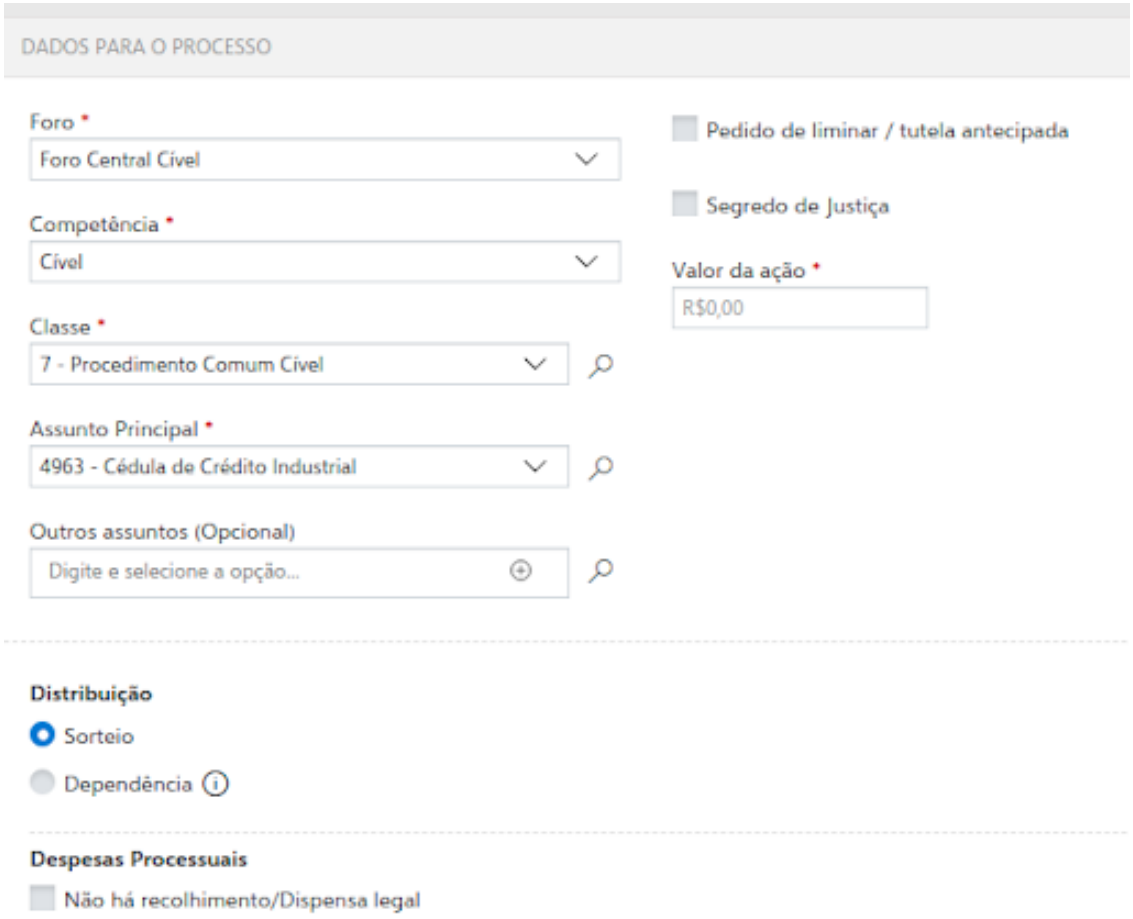



Figura 2.2: Tela de Peticionamento no TJSP.

Essa obrigatoriedade imposta aos advogados se repete em outros tribunais, seja no STF (art. 9°, Resolução n° 693/2020), seja no TJBA (art. 8º, I, Resolução nº 20/2013), no TJDFT (art. 14, Provimento 12/2017), entre outros. A questão é: uma vez que sabemos como o dado é gerado, o que isso implica para a nossa pesquisa?

O conhecimento do processo de geração de dados nos permite concluir sobre algumas incertezas relativas àquele dado. Como sabemos que a classe e o assunto são informações obrigatórias, podemos *presumir* que essa informação estará disponível para todos os processos; além disso, como sabemos que classe e assunto são informações padronizadas pelas TPUs, também conseguimos *presumir* que essa informação será padronizada entre processos e entre tribunais. Entretanto, sabendo que há um humano (no caso, um advogado), que está por trás da classificação dos processos,

devemos também esperar que a classificação possa estar errada. Há alguns tipos de erros possíveis: a classificação não condiz com o caso ou a classificação é mais genérica do que o possível. Por causa desses erros possíveis, devemos esperar que o dado de assunto e classe processuais contenham cifras ocultas[14]. A cifra oculta é a quantidade não observada de determinado dado, o que pode enviesar algumas análises. Se a quantidade não observada for aleatória, ela não irá gerar vieses. Caso contrário, ela será problemática para o estudo.

### 2.5.2 Dados criados

Sabendo da importância do processo de geração de dados, é que se recomenda sempre buscar ir atrás do processo de geração de dados. Entretanto, muitas vezes, não encontramos dados prontos no mundo; temos de criar dados. Fazemos isso, por exemplo, toda vez que lemos manualmente processos e tentamos classificar informações contidas nos autos. Se no primeiro caso, então, falávamos sobre investigar e reportar a forma como dados existentes foram gerados, neste segundo momento vamos tratar de reportar as modificações e transformações que realizamos nos dados. Dou aqui um exemplo.

É muito comum, ao recebermos uma base de um Tribunal de Justiça, buscarmos pela sentença de cada processo. Muitas bases possuem apenas as movimentações processuais de cada processo. Então, é a partir dessas movimentações, que devemos determinar se aquele processo teve ou não sentença. Essa pergunta pode se repetir para outras informações relevantes sobre o processo: teve ou não teve liminar? Teve ou não teve recurso? Teve ou não teve audiência de conciliação? Para todos estes casos, o procedimento é o mesmo: devemos olhar a descrição das movimentações processuais.

Assim como a Classe e o Assunto eram padronizados pelas Tabelas Processuais Unificadas, a descrição das movimentações também o é. Há um universo finito de movimentações possíveis. Entretanto, por mais que as movimentações sejam bem definidas, elas não trazem *em si* o seu sentido jurídico. Pode haver uma dúvida se uma determinada movimentação refere-se a uma sentença ou não. Neste momento, portanto, precisamos criar uma variável nova, por exemplo, a variável "teve_sentenca". O processo por meio do qual nós criamos essa variável é muito importante. É um processo que exige conhecimentos jurídicos e que, portanto, convida os juristas a participarem de sua discussão. Mas é um processo que pode gerar vieses nas análises. E se um pesquisador considerar decisões interlocutórias como sentenças? E se um pesquisador considerar acórdãos como sentenças? E se um pesquisador considerar decisões de conhecimento como sentenças? A depender da pesquisa, essas escolhas podem fazer sentido; mas via de regra, essas escolhas levarão a algum tipo de distorção dos dados, podendo superestimar ou subestimar a quantidade de processos que tiveram sentenças.

Dessa forma, a decisão mais acertada é simplesmente reportar o que foi e o que não foi considerado como sentença. Essa é uma forma simples, segura e, acima de tudo, reprodutível da pesquisa; é uma forma que permite a validação por pares. Veja, no exemplo acima, como o dado criado não existia de pronto no mundo. A variável "teve_sentenca" foi criada no meio da pesquisa. A equipe de pesquisa pôde exercer um grande nível de controle sobre a sua criação. E todas as decisões tomadas ao longo da criação dessa variável puderam ser reportadas e, eventualmente, poderão ser discutidas por outros pesquisadores.

[14] ABJ (2020)

### 2.5.3 Reporte os processos de geração de dados

Em resumo, há uma orientação geral a respeito dos dados que coletamos e criamos: devemos sempre reportar as incertezas. É comum omitirmos as incertezas e fraquezas dos nossos dados para dar mais credibilidade às nossas teorias; é comum também ficarmos desatentos às implicações dos dados que coletamos para as conclusões que tiramos e, por isso, não damos muita atenção para os processos de geração de dados. Entretanto, uma lição muito importante que devemos levar para a pesquisa quantitativa é que uma boa pesquisa reporta todos os processos de geração de dados, de cada um dos dados da pesquisa. Isso, além de deixar a pesquisa muito mais transparente para a comunidade, pode servir de base para que alguém continue a sua pesquisa, tentando cobrir os seus buracos, ou resolver as suas incertezas.

Além disso, como veremos mais para frente, o que desejamos futuramente realizar são *inferências*. A inferência é "conhecer o que não pode ser visto, a partir do que é visto". É o procedimento contrário à dedução – inferência é o mesmo que *indução*. Enquanto na dedução concluímos com certeza sobre algo, pois apenas derivamos um raciocínio de forma lógica, na inferência o que impera é a incerteza. Para cada conclusão sobre o "não visto" a partir do visto, carregamos muitas incertezas. Neste contexto, então, percebemos que é muita presunção não explicitar as incertezas, pois elas necessariamente existem, e todos sabem disso[15].

### 2.5.4 Uma nota sobre o processo de geração de dados e a inteligência artificial

A discussão a respeito do processo de geração de dados nos ajuda muito a compreender como funcionam inteligências artificiais. Os dados que existem no mundo foram gerados de alguma forma. Esse processo de geração do dado faz com que ele siga uma determinada distribuição no mundo. Este processo pode ser determinístico ou não determinístico. De qualquer forma, na maioria das vezes é, via de regra, inacessível para nós.

O que queremos com a inteligência artificial é conseguir, com os dados observados, reproduzir o processo de geração de dados de forma o mais precisa possível para gerar boas predições. A inteligência artificial, ao reproduzir um processo de geração de dados, pode ajudar a gerar peças processuais automaticamente, classificar informações em autos e até mesmo tentar predizer o resultado de um processo. Mas a inteligência artificial sem interferência humana não é capaz de compreender todos os problemas de viés, erros de preenchimento e conhecimento do *mecanismo de geração dos dados*.

Então o conhecimento sobre o processo de geração de dado pode nos ajudar a sermos mais críticos em relação às IAs, quando ele nos coloca a pensar se uma determinada inteligência artificial consegue ou não reproduzir tal processo. Sem a clareza do que está por trás de cada dado, não é possível fazer essa avaliação crítica dos robôs. Ao mesmo tempo, ao conhecer esses processos, podemos pensar cada vez mais, em como aprimorar as inteligências artificiais que vamos construir.

## 2.6 Amostragem e População

O último ponto de decisão importante no desenho de pesquisa é definir se o estudo será populacional ou amostral, e, caso seja amostral, como será feita a amostragem. Ao contrário do que acontece em muitas áreas do conhecimento, em jurimetria, muitos estudos são feitos de forma populacional. Mas a análise populacional impede a obtenção de alguns dados mais minuciosos, que precisariam ser coletados manualmente. Então, ainda que atualmente muitos estudos sejam

[15] King, Keohane, e Verba (1994)

feitos considerando a população como um todo, e não se valendo de técnicas de amostragem, é importante passar pelos princípios e técnicas específicas da amostragem.

O objetivo da amostragem é "fazer afirmações sobre uma população, baseando-se no resultado (informação) de uma amostra"[16]. Muitos dizem que para isso ser possível, é preciso que a amostra seja "representativa" da população. Entretanto, para sabermos se uma amostra "representa" uma população, isso exigiria que tivéssemos muitas informações a respeito da própria população, o que, em geral, é exatamente o oposto do que temos. Nós realizamos amostragens para se obter um conhecimento a respeito da população. Se, ao amostrar, eu busco descobrir informações ainda desconhecidas sobre uma população; e se, para eu obter uma boa amostragem, eu preciso que ela seja "representativa" da população; e se para uma amostra ser "representativa" da população, eu preciso saber previamente de informações da população; então, para aqueles que defendem que a amostra deve ser "representativa", há uma exigência de que para se conhecer algo sobre a população, eu devo conhecer muito sobre ela anteriormente, de modo que se torna inclusive desnecessária a coleta de amostra[17].

No lugar, então, de uma "amostra representativa", buscamos uma "amostra probabilística". A amostragem probabilística incorpora um elemento muito importante em seu método de coleta: a randomização. É a partir da randomização que nós conseguimos confiar que a amostragem possui uma distribuição similar à população sobre a qual nós desconhecemos suas informações.

Ao pensarmos na amostra, é muito importante pensarmos nas características da população. Essas características são normalmente desconhecidas para nós. Mas existe uma relação muito íntima entre a população e a amostra, pois a amostra deve ser simplesmente um reflexo da distribuição populacional. Se minha população for totalmente homogênea, então basta 1 única observação na minha amostra para que eu consiga refletir de forma adequada a distribuição da população. Conforme a variabilidade da população aumenta, é necessário aumentar o número de observações que serão amostradas, uma vez que é somente com mais observações que conseguiremos chegar a uma distribuição amostral próxima da distribuição populacional.

### 2.6.1 Por que randomizar?

Como acabamos de ver, a amostragem probabilística incorpora um elemento importantíssimo na sua coleta, a randomização. A questão que queremos confrontar a seguir é: por que randomizar? São várias respostas possíveis. Coletamos amostras de forma aleatória para garantir que os nossos resultados serão próximos aos que obteríamos caso medíssemos diretamente a população[18]; usamos a randomização também para evitar a introdução de vieses na amostra[19]; ou ainda. realizamos procedimentos de amostragem com elementos de randomização a fim de se evitar a introdução de *confounding effects*.

Todas as respostas estão apontando para a mesma questão: como a realização de procedimentos aleatórios de amostragem garante que possamos, ao analisar os dados, "fazer afirmações sobre uma população, baseando-se no resultado (informação) de uma amostra". Mas, apesar de haver todas essas definições, talvez a melhor forma de se compreender por que realizamos procedimentos aleatórios na amostragem seja por meio de exemplos. Vejamos dois exemplos a seguir.

[16] Bolfarine e Bussab (2005), p. 7
[17] Bolfarine e Bussab (2005), p. 14
[18] Shadish, Cook, e Campbell (s.d.)
[19] Winship&Mare1992

O primeiro exemplo foi extraído de uma revisão bibliográfica sobre viés de seleção (ou seja, a seleção da amostra realizada sem procedimentos de randomização), realizada por Winsihp e Mare (1992). Um dos estudos analisados trata da amostragem realizada dentre réus condenados no sistema penal. A questão é: o que representa amostrar pessoas que já foram condenadas pelo sistema de justiça? Se desejamos, por exemplo, estudar a probabilidade de alguém cometer um crime, a que resultados podemos chegar se a nossa amostra for feita somente dentre os réus já condenados? De todas as pessoas que existem, apenas algumas cometem crimes; de todos os crimes cometidos, apenas alguns são detectados pela polícia; dos crimes detectados, apenas alguns casos conseguem chegar em suspeitos; dos crimes que chegam a identificar suspeitos, apenas em alguns casos a polícia consegue prendê-los; das pessoas detidas, apenas algumas são processadas; e das pessoas processadas, apenas algumas são condenadas. Neste contexto, fica mais claro como a não utilização de um procedimento randômico pode gerar vieses.

Esse exemplo nos ajuda a elucidar também o viés em alguns estudos sobre violência doméstica. Se queremos responder a uma pergunta do tipo "quantas mulheres já sofreram violência doméstica em São Paulo?"; e, para responder a isto, utilizamos à base de dados da Secretaria de Segurança Pública de São Paulo (SSP/SP)[20], encontraremos um grande viés. Os dados da SSP/SP foram gerados a partir de boletins de ocorrência (BOs). O problema que se põe é: quantas mulheres nunca registraram violência doméstica por medos de agressão de seus parceiros, ou por descrença nas autoridades policiais e jurídicas para tratar de seus problemas?

Um segundo exemplo, nós extraímos do texto Randomization and Fair Judgment in Law and Science, realizado por Stern et al. (2020). Os autores tratam da randomização na distribuição de processos no Judiciário. Como cada juiz possui sua própria história, suas próprias opiniões e referências de mundo, então os autores colocam: > se a seleção dos juízes pudesse ser influenciada pelos litigantes ou por outra parte interessada, os ricos, aqueles mais bem informados, com melhores conexões, ou as partes mais fortes teriam maiores chances de obterem alguma vantagem em um processo direcionando o seu caso a um juiz simpático aos seus argumentos[21]

Tanto é um problema a seleção da jurisdição pelos litigantes que discute-se muito no Direito internacional privado a questão do *forum shopping*. O *forum shopping* é um efeito que acontece no Direito internacional em decorrência do conflito positivo de competências entre as jurisdições de mais de um país. Neste caso, não é que o processo não será distribuído aleatoriamente para um juízo; o problema é com a jurisdição. De qualquer forma, o efeito é o mesmo: as partes mais privilegiadas conseguem escolher com maior racionalidade a jurisdição, o que, novamente, mostra a importância da randomizaçã.

O terceiro e último exemplo se trata de um caso que ficou conhecido, relacionado ao treinamento de uma Inteligência Artificial para reconhecer rostos humanos. Por causa de uma base de dados composta, majoritariamente, por rostos de pessoas brancas, a IA criada para reconhecer faces humanas se mostrou incapaz de reconhecer alguns rostos de pessoas negras. O problema neste caso, não foi um comportamento imprevisível da IA, decorrente de sua natureza black-box; mas foi simplesmente um problema de viés de seleção na base de dados que a alimentou.

### 2.6.2 Amostragem aleatória simples

O procedimento mais básico da amostragem é a amostragem aleatória simples. A ideia é que, sem conhecer as informações a respeito da população, nós devemos escolher aleatoriamente quaisquer observações para compor a

[20] Disponíveis em: https://www.ssp.sp.gov.br/Estatistica/Pesquisa.aspx

[21] No original: "if the selection of judges could be influenced by the litigants or other interested parties, the richer, better informed, well connected, or otherwise more powerful parties would likely have an advantage in directing the case to a judge sympathetic to their arguments." Stern et al. (2020), p. 385

amostra. É uma escolha totalmente aleatória, sem seguir padrão algum.

Tem um exemplo interessante para pensarmos dentro da jurimetria. É muito comum conseguirmos obter todos os números de todos os processos em determinada vara a respeito de um mesmo tema. Basta pesquisarmos por esse tema nos campos da Consulta Pública, ou basta pesquisarmos pelo termo de interesse nos Diários Oficiais do Estado. Com os números dos processos em mãos, poderíamos muito bem iniciar um estudo populacional, em que analisamos toda a população de processos sobre determinado assunto. Entretanto, digamos que encontramos mais de 10 mil processos nessa coleta dos números processuais; e digamos também que precisamos realizar análises mais profundas desses processos que exigem uma etapa de leitura manual dos autos. Neste caso, não é muito recomendável uma análise populacional, pois isso envolveria a leitura manual de 10 mil processos. Nesta hora, é importante ter em mãos técnicas de amostragem. Dado que o conjunto de todos os números processuais coletados pela Consulta Pública de um TJ e pelos Diários Oficiais é a nossa população, então podemos fazer uma amostragem aleatória simples desses processos e escolher apenas um certo número de observações para analisarmos. No fim, iremos estudar apenas esses autos amostrados aleatoriamente (e não os 10 mil), mas iremos tirar conclusões sobre todos os 10 mil processos.

Para garantir que as conclusões sobre a amostra reflitam as conclusões sobre a população, temos de garantir que a amostra não está enviesada.

### 2.6.3 Viés e erro na amostragem

Devemos ter em mente que ao se realizar amostragens probabilísticas, sempre haverá um certo grau de erro, de descompasso entre a amostragem e a população. Entretanto, como a forma de amostragem foi aleatória, este erro também acabou sendo aleatório. Não há problema em ter erros, afinal, se uma amostra não tivesse erro algum em relação à população, seria porque a amostra é a população e, nesse caso, não haveria necessidade de amostrar.

O maior problema, portanto, não é o erro amostral (isso é natural, é normal, é esperado), mas é o *viés*. Haverá viés na amostra sempre que o erro não for aleatório, mas for sistemático. Por exemplo, imagine que temos um grupo de processos sobre determinado tema. Todos os processos sobre aquela tema representam a nossa população. E digamos que nós queremos amostrar este grupo de processos na população. Mas digamos também que essa população seja composta por 80% dos processos sendo julgados por juízes e apenas 20% por juízas. A partir deste cenário, mas sem que nós conheçamos de antemão a real distribuição de gênero entre os processos – como é de costume –, resolvemos estratificar a amostragem, de forma que metade da amostra seja composta por mulheres e a outra metade, por homens. Se essa população tinha uma distribuição de 80-20, mas amostra, uma distribuição de 50-50, houve um erro. Entretanto, esse erro não foi aleatório, pois fui introduzido por nós (os pesquisadores deste caso hipótetico). Neste caso, diz-se que houve um viés de seleção, pois **selecionamos** os casos que vão entrar na amostragem de forma enviesada. Para muitas análises, o gênero dos juízes não importa, pois o comportamento dos juízes e das juízes é o mesmo. Entretanto, pode ser que em algum caso, o julgamento feminino seja diferente do julgamento masculino. Neste caso, o viés de seleção de gênero introduzido na amostragem fará com que a análise final enviese a taxa de deferibilidade dos pedidos.

### 2.6.4 Tamanho da amostra

Quanto mais observações temos na nossa amostra, menores são os erros das nossas estimativas. Isso será melhor demonstrado no capítulo dedicado a estudar a Lei dos Grandes Números e o Teorema do Limite Central. Por hora, basta sabermos dessa relação entre o tamanho da amostra e o erro: é uma relação inversamente proporcional, de modo que quanto maior a amostra, menor o erro.

Dessa constatação, é possível concluir, de modo intuitivo e automático, que a nossa amostra deve ter sempre o maior tamanho possível. Se isso é verdade no plano teórico, no plano prático, essa afirmação não se sustenta. O que acontece é que no dia a dia da pesquisa, não dispomos de recurso e de tempo infinitos. Coletar dados é um procedimento caro, que demanda muitas horas para ser realizado. Assim, dadas as limitações do mundo real, discute-se qual é o tamanho ideal da amostra. Para cada tamanho de amostra, há uma precisão relacionada. O ideal seria calcular, a partir da precisão desejada (normalmente de 5% de erro), qual deveria ser o tamanho da amostra. Mas isso nem sempre é possível. Se não for possível determinar o tamanho da amostra, deve-se fazer o procedimento contrário de determinar qual é a precisão que uma amostra de determinado tamanho é capaz de gerar.

## 2.7 Viés de seleção em processos judiciais

### 2.7.1 Viés de seleção

A palavra **viés**, em estatística, pode ter múltiplos significados. Em geral, a palavra indica que existe diferença entre o que está sendo usado para fazer um estudo e o que se deseja estudar.

O tipo de viés mais conhecido é o *viés do estimador*. Ele é a *diferença entre a média de um estimador e o valor que está sendo estimado*. Por exemplo, ao estudar o valor médio de pedidos de indenização por dano moral em Alagoas (números fictícios), temos:

- Valor real: R$ 5.000,00
- Média do estimador: R$ 5.200,00
- Viés: R$ 5.200,00 - R$ 5.000,00 = R$ 200,00

**Viés de seleção** é um tipo de viés estatístico, indicando uma diferença entre os **indivíduos que estamos estudando na base de dados** de dados e **a população que desejamos estudar**. Coloquialmente, viés de seleção é uma forma de dizer que estamos **tirando conclusões sobre bananas ao estudar maçãs**.

Muitas vezes, no entanto, não é tão claro que estamos estudando maçãs. Para não concluir sobre bananas usando maçãs, é importante entender o **mecanismo de seleção** que gera os dados que observamos. O mecanismo de seleção é o conjunto de regras ou critérios que são usados para determinar quais indivíduos serão incluídos na base de dados e quais serão excluídos.

**Exemplo 1**. Queremos estudar a **letalidade** da Covid-19. Para isso, precisamos calcular a razão entre a quantidade de mortes e a quantidade de pessoas infectadas. Em um país fictício, o número de pessoas infectadas é calculado a partir dos relatórios de pacientes de hospitais. A estimativa de letalidade no país fictício ficou bem maior do que a letalidade real. Por quê?

Nesse caso, a letalidade estimada apresentará viés, porque as pessoas que vão ao hospital tendem a apresentar sintomas mais graves que a população em geral. O mecanismo de seleção, nesse caso, é a **entrada no hospital**: uma pessoa só entra na base de dados se for para o hospital.

Abaixo, temos também um exemplo computacional. Fizemos a simulação de 10 mil casos. A letalidade foi considerada como 1%, e a gravidade definida por um números entre zero e um, gerados aleatoriamente. O mecanismo de seleção então foi introduzido: a hospitalização como uma função da gravidade; quanto maior a gravidade, maior a probabilidade de

hospitalização. Em seguida, calculamos a proporção de pessoas que morreram, comparando com a proporção de pessoas que morreram nos hospitais (que é o que obervamos). Na simulação, a proporção observada é quase o dobro da real.

```
set.seed(3)

N <- 10000 # numero de pessoas infectadas
letalidade <- 0.01 # letalidade real
gravidade <- runif(N) # número aleatório indicando a gravidade da doença
morreu_real <- gravidade > 1 - letalidade # indica se a pessoa morreu

# hospitalizacao proporcional à gravidade
hospitalizacao <- as.logical(rbinom(N, 1, gravidade))

morreu_observados <- morreu_real[hospitalizacao]

mean(morreu_observados)
```

```
#> [1] 0.01917316
```

```
mean(morreu_real)
```

```
#> [1] 0.0097
```

**Exemplo 2**. Aviões foram enviados para a guerra, sendo que muitos caíram e os poucos que voltaram estavam com tiros nas pontas das asas. Os engenheiros, então colocaram mais proteções nas pontas das asas dos aviões, já que, estatisticamente, era a região mais atingidas. Os aviões reforçados continuaram caindo. Por quê?

Nesse caso, os aviões que caíram **foram atingidos em outros lugares**, diferentes da ponta das asas. Ou seja, os engenheiros protegeram a parte menos importante do avião! O mecanismo de seleção, nesse caso, é a **queda do avião**: o avião só entra na base de dados se sobreviver à guerra.

Existem vários tipos de viés de seleção. Alguns deles são:

- Amostragem (exemplos anteriores)
- Auto-seleção (viés do voluntário)
- Intervalo de tempo (vimos no módulo 1)

Quando passamos para processos judiciais, o mecanismo de seleção está relacionado **à escolha de litigar**. A Tabela 2.1 mostra algumas alternativas e ações de acordo com as opiniões do autor e do réu. Quando o autor acha que vai ganhar e o réu acha que vai perder, o esperado é que aconteça um acordo. Quando tanto autor quanto réu acham que vão ganhar, espera-se que o processo vá para o judiciário. Se o autor acredita que perderia o processo, em condições normais, o caso não iria a juízo. Ainda assim, se tanto autor quanto réu acreditam que perderiam o caso, dependendo das condições do problema, pode haver um litígio.

Tabela 2.1: Ação decorrente da opinião do autor e do réu sobre o possível resultado do processo.

| Expectativa autor | Expectativa réu | Ação |
|---|---|---|
| Vou ganhar | Vou perder | Acordo |
| Vou ganhar | Vou ganhar | Julgamento |
| Vou perder | Vou perder | Sem processo* |
| Vou perder | Vou ganhar | Sem processo |

A Tabela 2.1 leva a uma característica muito intrigante de processos judiciais: **o processo só acontece se as expectativas das partes forem diferentes.** Esse é o princípio por trás do mecanismo de seleção em processos judiciais.

Para modelar o problema, vamos criar algumas variáveis:

- $P_p$: Probabilidade do autor ganhar, segundo o autor (*plaintiff*)
- $P_d$: Probabilidade do autor ganhar, segundo o réu (*defendant*)
- $C_p$: Custo de litigar do réu
- $C_d$: Custo de litigar do autor
- $S_p$: Custo de acordo (*settlement*) do autor
- $S_d$: Custo de acordo do réu
- $J$: valor a ser pago pelo réu se o autor ganhar

Considere uma negociação hipotética entre o autor e o réu. O valor de acordo desejado pelo autor (*asking price*) é

$$A = P_p J - C_p + S_p$$

O componente $P_p J + C_d$ indica um cenário de litígio: o autor gastaria $C_p$ ganharia, em média, $P_p J$. O componente $S_p$ indica o custo de fazer o acordo (que pode ser zero).

Já o valor oferecido pelo réu (*bidding price*) é dado por

$$B = P_d J + C_d - S_d$$

O componente $P_d J + C_d$ indica um cenário de litígio: o réu gastaria, em média, $P_p J$ mais o custo do processo $C_p$. O componente $-S_d$ indica um cenário de acordo, onde o réu gastaria $S_d$.

O processo só acontece se $A > B$. Ou seja, o valor de acordo pedido é maior que o valor oferecido. Fazendo contas, é possível chegar na formulação alternativa abaixo:

$$P_p - P_d > \frac{C_p + C_d - S_p - S_d}{J}$$

Ou seja, o processo só ocorre se a diferença de expectativas for maior que a diferença entre os custos de processo e os custos de acordo, em relação ao valor da indenização. A equação mencionada acima é conhecida como *condição*

*de Landes-Posner-Gould* (LPG), atribuída aos autores que contribuíram, em momentos diferentes, para a construção da condição.

Neste link, temos um aplicativo que mostra a condição LPG em ação. O ponto em vermelho indica a relação entre as probabilidades subjetivas do autor ($P_p$) e do réu ($P_d$) sobre o evento do autor ganhar o processo. A linha é a condição LPG: se as probabilidades estiverem abaixo da linha, um litígio vai acontecer. Se estiver acima da linha, um acordo vai acontecer.

### 2.7.2 Priest & Klein

O artigo de Priest & Klein (1984), discute as implicações da condição LPG quando variamos o nível de informação das partes, criando um modelo para as quantidades $P_p$ e $P_d$.

O artigo "Selection of Disputes for Litigation" de George Priest e Benjamin Klein é uma análise econômica da escolha de disputas para litígios. O artigo argumenta que as partes envolvidas em uma disputa geralmente têm informações assimétricas sobre a probabilidade de sucesso e os custos de um litígio. Isso leva às partes a selecionarem disputas para litígios de forma diferente do que seria ótimo do ponto de vista social.

Os autores apresentam um modelo teórico que mostra como as partes selecionam disputas para litígios e como essa seleção afeta a eficiência econômica. Eles também discutem como as instituições legais, como a arbitragem e a regulamentação, podem afetar a seleção de disputas para litígios e, consequentemente, a eficiência econômica.

O artigo é amplamente citado e considerado como um dos principais trabalhos na área da teoria econômica da litigação. Ele fornece uma base para entender como as partes escolhem entre diferentes opções de resolução de disputas e como as instituições legais afetam a eficiência econômica do processo de resolução de disputas.

**O mérito como variável contínua**. O artigo assume que um conflito possui um **nível de mérito** $y$, desconhecido, que vem de uma distribuição de probabilidades também desconhecida. O juiz, analisando o caso, decide se $y > y*$, sendo $y*$ um valor limítrofe: se $y > y*$, a decisão é favorável ao autor. Caso contrário, é favorável ao réu.

**Princípio da simetria de informação**: Se as partes têm um nível de informação (medido pela variância das expectativas) parecido, os valores de $P_p$ e $P_d$ são, em média, iguais. Conforme o nível de informação aumenta, $P_p$ e $P_d$ vão ficando cada vez mais próximos, fazendo com que a ocorrência de um processo fique cada vez mais improvável.

O teorema de Priest e Klein (PK) mostra que, conforme o nível de informação aumenta, os casos que acabam indo ao tribunal acabam ficando próximos do nível de mérito limítrofe $y*$, de forma simétrica. Como efeito, a probabilidade de vitória do autor fica em torno de 50%.

Ou seja, o teorema mostra que, assumindo certas condições, como simetria de informação das partes e a validade da condição LPG, as proporções de vitória observadas nos tribunais serão de 50%.

O aplicativo deste link uma simulação dos resultados do teorema. O aplicativo parte das mesmas premissas da condição LPG, mas considerando uma amostra de vários processos. No gráfico de baixo, é apresentada a distribuição do mérito para todos os casos (em azul) e para os casos litigados (em verde). Conforme o nível de informação aumenta (e a variância cai), os casos abaixo da condição LPG vão ficando cada vez mais escassos e, nos poucos casos litigados, a distribuição do mérito fica próxima ao padrão de decisão, o que gera, no limite, proporções de decisão em torno de 50%.

O Teorema de Priest e Klein mostra que os casos observados no judiciário não são uma amostra aleatória dos conflitos que ocorrem na sociedade. Isso ocorre porque as partes envolvidas em uma disputa geralmente têm informações assimétricas

sobre a probabilidade de sucesso e os custos de um litígio. Devido a isso, as partes selecionam disputas para litígios de forma a otimizar o uso do litígio do ponto vista social.

O teorema também mostra que as taxas de vitória sempre ficarão em torno de 50%. Isso pode levar a uma (falsa) ideia de que não é possível fazer inferências sobre o judiciário a partir da análise de taxas de vitória.

No entanto, é importante notar que taxas muito altas ou muito baixas podem significar que os custos de litigar não estão bem ajustados ou que existe assimetria de informação entre as partes. Além disso, mudanças legislativas podem implicar em alterações temporárias nas taxas de vitória.

O Teorema de Priest e Klein tem fortes implicações na análise de processos judiciais. Sua existência deve ser levada em conta em estudos que envolvam análise de proporção de vitórias, para entender melhor as dinâmicas e os fatores que as influenciam. Isso pode ajudar a identificar problemas ou oportunidades para melhorar a eficiência e a eficácia do sistema judiciário.

# 3 Estatísticas

No capítulo anterior, vimos algumas questões essenciais por trás do método quantitativo. Neste capítulo vamos olhar para a estrutura dos dados e como resumir as informações a respeito desses dados. Para tanto, o capítulo está dividido em duas seções: (a) descrição das observações; (b) descrição do conjunto de observações.

Na seção sobre as observações, vamos ver como as bases de dados devem ser estruturadas e os tipos de dados possíveis. Na seção sobre a descrição do conjunto das observações, vamos ver as medidas que podem descrever conjuntos de observações, a saber, as medidas de posição, as medidas de variabilidade e as medidas de associação.

Os exemplos deste capítulo foram extraídos de duas bases de dados:

1. Base `consumo`: A base traz dados extraídos da jurisprudência do segundo grau no TJSP.
2. Base `leiloes`: A base leiloes traz dados que foram coletados no âmbito do Observatório da Insolvência - Fase 3 pela ABJ. Neste observatório, a ABJ coletou e analisou dados referentes a processos de falência distribuídos na Comarca de São Paulo, entre janeiro de 2010 e dezembro de 2020.

Você pode acessar essas bases de dados baixando, no R, o pacote `{abjData}`.

## 3.1 Olhando para as observações

O que queremos compreender nesta seção é a natureza das *observações* de uma base de dados. As observações correspondem às unidades amostrais, que podemos definir como "*cada uma das partes disjuntas em que uma população é exaustivamente decomposta, para* [que]*, do conjunto delas se façam extrações a fim de constituir uma amostra, ou estágio de uma amostra*" (Bolfarine e Bussab 2005, 263). **Em bases de dados, cada linha deve corresponder a uma unidade amostral; e cada coluna representa uma característica (também chamada de variável) dessa observação**.

Na Tabela 3.1, há um exemplo de uma base de dados da ABJ. São algumas linhas e colunas da base de leiloes do Observatório da Insolvência de São Paulo - Fase 2. Basicamente, esta base diz respeito aos bens que são levados a alienações nos processos de falência no Estado de São Paulo. Nessa base, cada linha representa um item levado a alienação de um processo, ou seja, a unidade amostral é um item levado a leilão; e cada coluna representa uma informação a respeito desse bem, a saber, quem é o leiloeiro responsável pela venda do bem, quando o bem foi levado a leilão, qual era o valor de avaliação inicial desse bem; e qual é o valor pelo qual ele foi arrematado (o que fica em branco, quando o bem nunca foi arrematado).

Tabela 3.1: Exemplo dos dados da base de leilões da ABJ

| id_processo | descricao | id_leiloeiro | data_edital | vendeu | av_inicial | arrematado |
|---|---|---|---|---|---|---|
| 00268835820128260100 | Lote 1C: Máquina ... | 5406 | 2015-10-05 | nao | 62000.0 | NA |

| id_processo | descricao | id_leiloeiro | data_edital | vendeu | av_inicial | arrematado |
|---|---|---|---|---|---|---|
| 10033914920148260100 | Monitores Dell | 466 | 2015-03-06 | nao | 400.0 | NA |
| 10018294820168260451 | Prensa Hidraucila... | 29062 | 2020-01-24 | nao | 540.0 | NA |
| 00468773820138260100 | Lote 143: Televis... | 7633 | 2016-08-17 | sim | 300.0 | 178.31 |
| 10012206520188260299 | BOMBA CENTRÍFUGA ... | 1748 | 2020-03-10 | nao | 4804.8 | NA |
| 00410345520108260114 | bancada de trabalho | 776 | 2014-02-24 | nao | 300.0 | NA |
| 10012206520188260299 | GUINCHO TIPO GIRA... | 1748 | 2020-03-10 | nao | 3225.6 | NA |
| 00268835820128260100 | Lote 159: Compres... | 5406 | 2015-10-05 | nao | 64800.0 | NA |
| 10488694620158260100 | Diversos itens de... | 5406 | 2018-04-03 | sim | 20.0 | 20.00 |
| 00268835820128260100 | Lote 1B: Geladeir... | 5406 | 2015-10-05 | nao | 11070.0 | NA |

Uma *observação*, então, como vimos, possui várias características. Todas essas características constituem variáveis a respeito da unidade amostral. O que precisamos ver, a seguir, são os tipos dessas características e as implicações de cada um dos tipos.

As variáveis podem pertencer a dois grupos: variáveis qualitativas (ou categóricas) ou variáveis quantitativas. As variáveis categóricas se subdividem ainda em nominais e ordinais; já as variáveis quantitativas podem ser do tipo discretas ou contínuas.

### 3.1.1 Variáveis qualitativas/categóricas

Todas as variáveis qualitativas representam algum tipo de categoria (por isso também chamamos essas variáveis de "categóricas"). Há dois tipos de variáveis categóricas, as nominais e as ordinais. As variáveis nominais são categorias de nomes, categorias sem ordenação possível. Já as variáveis ordinais representam categorias com algum tipo de ordenação universal, com algum *ranking* possível. O critério de distinção entre essas duas variáveis é a possibilidade de *ordenação universal* das respostas possíveis.

A seguir, temos alguns exemplos que discutem se determinadas variáveis categóricas são nominais ou ordinais.

1. **Unidade Federativa**: No Brasil, há 27 unidades federativas possíveis. É possível ordenar as unidades federativas, por exemplo, por ordem alfabética; ou até, se soubermos outras informações como PIB ou tamanho da população, podemos ordenar as UFs por algum critério outro. Apesar de essa variável ser *ordenável*, ela não pode ser ordenada a partir de um *critério universal*, ou seja, um critério intrínseco a ela mesma. Por isso, consideramos que UF é uma variável *nominal* (e não ordinal).

2. **Assunto processual**: Os assuntos processuais são dados pelas Tabelas Processuais Unificadas (TPUs)[1] do CNJ. Por mais haja uma numeração. Os assuntos também não possuem nenhuma ordenação universal. Portanto, esta é uma variável nominal.

[1]Para saber mais informações, ver Resolução nº 46 do CNJ, bem como o site dos assuntos do CNJ.

3. **Valor de bens no leilão judicial (categorizado)**: Uma possível variável que pode existir em processos judiciais é uma classificação para o valor dos bens em um leilão. Podemos classificar os valores, por exemplo, como "insignificante", "baixo", "médio", "alto", "extravagante". Neste caso, haveria uma ordenação intrínseca das categorias, sendo que "insignificante" é a categoria de menor valor e "extravagante", a de maior valor. Assim sendo, esta variável é ordinal.

4. **Resultado de uma sentença**: Uma sentença pode ter, de forma simplificada, três resultados possíveis: totalmente procedente, parcialmente procedente e improcedente. É um modelo simplificado, pois, para determinadas pesquisas, pode ser interessante diferenciar sentenças com julgamento de mérito de sentenças sem julgamento de mérito, ou de sentenças homologatórias. Por ora, vamos pensar apenas nessas três categorias. A questão que se põe é se há alguma ordenação universal entre essas três categorias ou não?

Este caso pode gerar algumas dúvidas, pois poderíamos ordenar a sentença de tal forma que a sentença "totalmente procedente" fosse a mais valiosa, em relação à sentença "improcedente". Entretanto, esse tipo de raciocínio pressupõe um valor intrínseco das sentenças, como se uma sentença "totalmente procedente" sempre fosse, de alguma forma, melhor do que uma sentença "improcedente". O problema desse raciocínio é que, a depender do *polo da parte*, o valor da sentença é exatamente o oposto: para o réu, a sentença "improcedente" é a de menor valor, enquanto, para o autor, a sentença "totalmente procedente" é a de maior valor.

Existem ainda mais ramificações dos tipos de variáveis qualitativas, por exemplo, variáveis intervalares, ou variáveis-razão. Essas demais ramificações não possuem muito uso prático no Direito, então não vamos nos aprofundar nelas. Há somente uma ramificação que ainda nos interessa que são as variáveis binárias, ou como são chamadas também, as variáveis *dummies*. As variáveis binárias só assumem dois valores possíveis, o valor de sucesso (representado numericamente pelo número 1, ou pela condição TRUE) e o valor de fracasso (representado numericamente pelo número 0, ou pela condição FALSE). Essas variáveis são importantes pois, como veremos no Capítulo 5 é possível representar qualquer variável categórica em um conjunto de variáveis binárias.

### 3.1.2 Variáveis quantitativas

A outra grande categoria de variáveis é a de variáveis quantitativas. Esse grupo se caracteriza por ter variáveis de valores numéricos. Há uma classificação dicotômica importante a respeito desses valores. Existem variáveis quantitativas discretas e contínuas.

As variáveis discretas são caracterizadas por valores numéricos que formam um conjunto finito ou enumerável de números. Usualmente, as variáveis desse tipo resultam de alguma *contagem*. Já as variáveis contínuas são valores numéricos que pertencem ao conjunto dos números reais. O critério de distinção entre essas duas categorias é a nossa capacidade de fazer uma correspondência dos números com o conjunto dos números naturais (0, 1, 2, ...).

A seguir, temos alguns exemplos se algumas variáveis são discretas ou contínuas.

- **Valor da causa**: Em um exemplo anterior (no caso do valor dos bens dos leilões), estávamos tratando de valores também, mas estávamos tratando de valores agrupados formando categorias. Agora vamos falar dos valores brutos, individualizados, e não das categorias a que eles pertencem. O valor da causa pode assumir incontáveis valores, não sendo, portanto, um valor enumerável. Assim sendo, ele é uma variável *contínua*.

- **Quantidade de partes em cada polo**: No caso de litisconsórcio ativo ou passivo, é possível contar quantas partes existem em cada polo. Essa informação pode ser relevante, por exemplo, ao se estudar direitos difusos e coletivos, pois pode ser importante saber quantas pessoas estão no polo ativo da demanda, para determinar se é uma demanda coletiva ou pseudo-coletiva[2]. A variável sobre a quantidade de partes em cada polo será do tipo *discreta*.

### 3.1.3 Considerações sobre os tipos de dados

Antes de prosseguir para as medidas desses dados, devemos fazer algumas considerações.

#### 3.1.3.1 Consideração 1: Cuidados ao se representar numericamente variáveis categóricas ordinais

O primeiro ponto a se destacar é sobre a representação numérica de variáveis categóricas ordinais. Vamos usar o exemplo de uma proposta do Center for *Court Innovation* de Nova Iorque[3]. Uma das iniciativas desse centro foi tornar os tribunais de Nova Iorque mais "amigáveis". Uma das técnicas propostas para tanto foram os Questionários de Satisfação sobre a prestação jurisdicional. A Tabela 3.2 resume algumas das perguntas do questionário elaborado pelo *Centro*.

Tabela 3.2: Perguntas selecionadas de um questionário de satisfação

| - | Concordo fortemente | Concordo | Neutro | Discordo | Discordo fortemente |
|---|---|---|---|---|---|
| O juiz compreendeu minha demanda | | | | | |
| O juiz levou a minha demanda a sério | | | | | |
| De forma geral, obtive o resultado esperado no tribunal | | | | | |
| Fui tratado com respeito pelo tribunal | | | | | |
| Fui tratado de forma justa pelo tribunal | | | | | |
| Eu pediria a ajuda ao tribunal no futuro, se necessário | | | | | |

Ao aplicar um questionário desses, estamos produzindo *dados*. A aplicação de vários questionários sucessivamente levaria à produção de uma base de dados em que cada linha (unidade amostral) seria um respondente, e cada coluna seria uma das perguntas. Todas essas perguntas recebem como resposta o nível de satisfação (concordo fortemente, concordo, neutro, discordo e discordo fortemente), sendo, portanto, variáveis de natureza *categórica ordinal*, pois há claramente uma ordem entre essas respostas.

O que queremos é discutir uma proposta de substituição dessas respostas para uma forma *numérica*. Como há uma ordem entre essas respostas, será que poderíamos olhar para elas de forma numérica? A transformação seria a seguinte:

- Concordo fortemente: 1
- Concordo: 2

[2]Usa-se aqui a distinção de Grinover (2014).
[3]CCI (2020)

- Neutro: 3
- Discordo: 4
- Discordo fortemente: 5

Essa transformação deve ser feita com cuidado. Por um lado, esse tipo de alteração não altera a ordenação dessas respostas. Entretanto, por outro lado, a representação numérica das categorias ordinais adiciona uma informação aos dados que não é verdadeira: a intensidade. O que estamos dizendo é que os números guardam, não só uma ordenação universal entre si, assim como as variáveis categóricas ordinais, mas eles guardam uma relação de intensidade entre si, algo que as variáveis categóricas ordinais não possuem. Assim, enquanto podemos dizer que o número 2 é o dobro do número 1, não podemos estabelecer essa relação entre as categorias "concordo" e "concordo totalmente".

#### 3.1.3.2 Consideração 2: As variáveis *dummies*

A segunda consideração que queremos fazer diz respeito à transformação das categorias nominais em variáveis *dummies*. Vamos tomar outro caso como exemplo para esta discussão.

No projeto que a ABJ realiza, em parceria com o CNJ, sobre adoção, tentamos auxiliar os pretendentes a escolherem o perfil da criança fornecendo a eles uma informação importante: o tempo que irá demorar adotar uma criança a depender do perfil escolhido para ela. Perfis mais restritivos em geral demoram mais tempo do que perfis mais permissivos. A questão é deixar claro que características importam para a mudança do tempo e o quanto cada característica importa para o tempo. Uma das variáveis é a variável de `tp_etnia`. Essa variável indica qual é a preferência de etnia preferida dos pretendentes em relação às crianças a serem adotadas. Há 6 respostas possíveis: A (de "amarelo"), B (de "branco"), I (de "indígena"), N (de "negro"), P (de "pardo") ou S (de "sem preferência"). Essas categorias não possuem um critério de ordenação universal, assim sendo, a variável `tp_etnia` é do tipo categórico nominal. Temos um exemplo dessa base na Tabela 3.3. Os valores usados são fictícios, por questões de sigilo da base.

Tabela 3.3: Variável categórica de etnia

| id_pretendente | tp_etnia |
|---|---|
| 1082 | I |
| 1083 | P |
| 1084 | B |
| 1085 | P |
| 1086 | S |
| 1087 | S |
| 1088 | N |
| 1089 | P |
| 1090 | N |
| 1091 | I |
| 1092 | A |

A questão de que queremos tratar é como transformar a `tp_etnia` em um formato dummy? Para realizar essa transformação, nós transformamos cada uma das etnias em uma variável que recebe apenas as respostas 0 ou 1. Mas temos que tomar um cuidado muito importante: **a quantidade de variáveis que criamos** é sempre o número de

categorias ($n$) menos 1, ficando $n-1$. No caso, como são 6 categorias possíveis, criamos n-1 variáveis, isto é, 5 variáveis. A base resultante está na Tabela 3.4.

Tabela 3.4: Transformação da variável categórica de etnia em dummy

| id_pretendente | tp_etnia | A | B | I | N | P |
|---|---|---|---|---|---|---|
| 1082 | I | 0 | 0 | 1 | 0 | 0 |
| 1083 | P | 0 | 0 | 0 | 0 | 1 |
| 1084 | B | 0 | 1 | 0 | 0 | 0 |
| 1085 | P | 0 | 0 | 0 | 0 | 1 |
| 1086 | S | 0 | 0 | 0 | 0 | 0 |
| 1087 | S | 0 | 0 | 0 | 0 | 0 |
| 1088 | N | 0 | 0 | 0 | 1 | 0 |
| 1089 | P | 0 | 0 | 0 | 0 | 1 |
| 1090 | N | 0 | 0 | 0 | 1 | 0 |
| 1091 | I | 0 | 0 | 1 | 0 | 0 |
| 1092 | A | 1 | 0 | 0 | 0 | 0 |

A questão importante dessa consideração era justamente chamar a atenção para o fato de que a quantidade de *dummies* criadas a partir da categorias é $n-1$. Isso não é uma escolha arbitrária, mas tem uma razão de ser. Mais para frente do livro, veremos que, se criássemos $n$ categorias, ao invés de $n-1$, teríamos um problema chamado *dependência linear*. Entretanto, por hora, de forma simplificada, podemos simplesmente afirmar que a categoria que foi deixada de fora, isto é, a categoria que não se transformou em *dummy* pode ser presumida.

Vejamos na Tabela 3.4 que a categoria deixada de lado foi "S" (ou "sem preferência"). Entretanto, conseguimos identificar um pretendente que não possui preferência por nenhuma etnia quando todas as dummies são iguais a 0. Assim, se A = 0; B = 0; I = 0; N = 0; e P = 0, então teríamos (caso essa categoria existisse) que S = 1. Dizer que S = 1 equivale a dizer que todas as outras categorias são iguais a 0. Justamente por haver essa fungibilidade entre a representação de S = 1 com tudo = 0 que não criamos exatamente $n$ dummies, mas $n-1$. O que deve ficar de lição é que a categoria deixada de lado está presumida pela resposta às demais dummies; ela estará presente sempre que todas as outras categorias forem 0.

## 3.2 Olhando para o conjunto das observações

Acima, estávamos discutindo as observações em si (unidades amostrais) e as informações que as caracterizam (variáveis). O que vamos falar a seguir é, não de uma única observação, mas do conjunto de várias observações. Vamos olhar, então, não para *uma* única linha, mas para um *conjunto* de linhas de uma base. A diferença da explicação anterior para a que se seguirá seria a de dizer, por exemplo (voltando ao exemplo da base de leilões do Observatório de Falências de São Paulo), que um bem específico foi vendido pelo dobro do seu preço de avaliação ou dizer que os *bens*, em média, são vendidos a 80% do seu preço de avaliação.

Uma primeira pergunta que podemos fazer para dar maior tecnicidade à explicação é o que esses conjuntos de dados representam? Se cada observação representava uma unidade amostral, então agora podemos dizer que o conjunto dessas unidades amostrais é a *amostra* em si. E para tirar informações e afirmar coisas sobre esse conjunto de observações, usamos *estatísticas de resumo*.

Estatísticas de resumo, como o próprio nome diz, resumem informações de grupos. Então, no lugar de olharmos para cada uma das observações individualmente, vamos olhar para uma única informação que irá dizer *algo* sobre as observações tomadas em conjunto. Esse "algo" que podemos falar sobre o conjunto das observações depende do tipo de variável para que estamos olhando. Se estamos olhando para variáveis categóricas (e, neste caso, pouco importa se são variáveis nominais ou ordinais), podemos resumir a sua frequência e proporção; já, se estamos olhando para variáveis contínuas, podemos olhar para as medidas de centro, de variabilidade e de posição do conjunto.

### 3.2.1 Medidas de resumo para variáveis categóricas

Não existe um repertório de medidas muito amplo para resumirmos variáveis categóricas. Isso se deve ao fato de que essas variáveis não possuem uma natureza numérica. Então, basicamente, para resumir estas variáveis precisamos transformá-las em números. A forma de fazer isto é uma: a contagem.

Por meio da *contagem* nós conseguimos *tabular* as respostas das variáveis categóricas, contando a sua frequência e, posteriormente, a proporção e porcentagem de cada categoria. A informação principal que temos é a frequência, pois é a partir dela que podemos pensar nas demais informações. Para além da frequência absoluta, temos também a frequência acumulada, a proporção, a proporção acumulada e a porcentagem.

A frequência é simplesmente a contagem de cada categoria. A frequência acumulada é a contagem de uma categoria somada com as categorias anteriormente contadas, de modo que a contagem da categoria final seja a soma total de todas as categorias. A proporção é a representação da contagem de uma categoria em relação a todas as observações. A proporção acumulada segue a mesma ideia da frequência acumulada, sendo a proporção de cada categoria, somada à proporção de todas as categorias anteriormente calculadas. Por fim, a porcentagem é simplesmente a proporção vezes 100.

Como exemplo, podemos olhar para a base de dados da ABJ sobre leilões nas falências em São Paulo, em que encontramos a Tabela 3.5.

Tabela 3.5: Tabela de frequências da modalidade do leilão.

| Modalidade | Frequência | Frequência acumulada | Proporção | Proporção acumulada | Porcentagem |
|---|---|---|---|---|---|
| leilao | 965 | 965 | 0.965 | 0.965 | 96.5% |
| pregao | 33 | 998 | 0.033 | 0.998 | 3.3% |
| proposta fechada | 2 | 1000 | 0.002 | 1.000 | 0.2% |
| total | 1000 | NA | 1.000 | NA | 100% |

A variável que está sendo resumida é a variável "modalidade", que indica se a falência foi ou não foi decretada. Há três respostas possíveis: "leilão", "pregão", ou "proposta fechada". Ao olharmos para todos os processos, podemos então realizar a contagem de cada uma dessas categorias. A partir da contagem, criamos a coluna de "frequência". A partir da frequência, nós criamos as próximas colunas, a saber, a frequência acumulada, a proporção, a proporção acumulada e a porcentagem. Note que as colunas "acumuladas" (isto é, as colunas de frequência acumulada e de proporção acumulada) não possuem um "total", pois o total já está expresso na última categoria.

### 3.2.2 Medidas de resumo para variáveis quantitativas

No caso das variáveis quantitativas, por elas já terem natureza numérica, há mais medidas de resumo possíveis. Vamos dividir a explicação em dois tipos de medidas: medidas de centro, medidas de variabilidade.

#### 3.2.2.1 Medidas de centro

O termo "medidas de centro" pode nos confundir. A nomenclatura pode nos induzir a pensar que essas medidas indicam o "meio" do gráfico, mas isso não é verdade. O "centro" a que se referem essas medidas é o ponto mais "típico" do conjunto em análise. São três medidas que estão nesta categoria: a média, a mediana e a moda.

##### 3.2.2.1.1 Média

A média é um conceito intuitivo, que designa basicamente a soma das observações dividida pelo total de observações. O símbolo da média é $\bar{x}$ (x-barra). O que queremos ver a seguir é a fórmula da média. Por mais que o conceito seja simples e intuitivo, precisamos formalizar um pouco mais o sentido da média. A fórmula é a que se segue:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

A fórmula principal é $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, mas precisamos compreender os elementos dessa fórmula. Vamos começar pelo símbolo $\sum$. Esse símbolo é a letra grega *sigma* (maiúscula), que na matemática usamos para representar *somatórios*. O somatório é uma notação que resume uma série de adições em sequência. Assim, dando um exemplo fácil, podemos representar a seguinte soma $4+8+12+16+20$ como $\sum_{k=1}^{5} 4k$. Vamos entender direito como que essa notação matemática representa a conta de adição.

O somatório tem três parâmetros: a notação que está embaixo dele (no caso, $k = 1$); a notação que está em cima dele (no caso, 5); e a notação que está na frente do sigma (no caso, $4k$).

A começar pelo que está na frente do sigma ($4k$), isso indica uma operação, que é 4 vezes $k$. Essa operação irá se repetir para 5 valores de $k$, sendo que o primeiro valor é $k = 1$ e o último valor é $k = 5$. De onde tiramos essas últimas informações? Das notações em cima e embaixo do $\sum$. Embaixo do $\sum$ encontramos o valor inicial de $k$; e em cima encontramos o valor final de $k$. Para cada valor que $k$ assumir, teremos uma expressão para somar. Então a primeira expressão é quando $k = 1$, ou seja, $4 \times 1 = 4$. Na segunda expressão, $k = 2$, ou seja, $4 \times 2 = 8$. A terceira expressão é $4 \times 3 = 12$; a quarta, $4 \times 4 = 16$; e a quinta e última, $4 \times 5 = 20$. Então nós pegamos o resultado das 5 expressões e as somamos, resultando em 4 (resultado da expressão quando $k = 1$) + 8 (resultado da expressão quando $k = 2$) + 12 (resultado da expressão quando $k = 3$) + 16 (resultado da expressão quando $k = 4$) + 20 (resultado da expressão quando $k = 5$), ou simplesmente $4 + 8 + 12 + 16 + 20$.

Então o $\sum$ (e todos os elementos que o acompanham) indica a soma de todas as observações. Na fórmula da média, há uma diferença importante: no lugar do $k$, temos o $x_i$. Acontece que o $i$ não é um número exatamente como funcionava com o $k$, ele é apenas um índice. Precisamos, então, verificar o que é $x_i$.

Voltemos à base de leilões, para pegar a variável *valor de avaliação*. Essa é uma variável quantitativa contínua. Pegando apenas as 10 primeiras observações dessa variável, temos os seguintes dados, resumidos na Tabela 3.6.

Tabela 3.6: Dados de leilões realizados

| Descrição | Valor de avaliação inicial |
|---|---|
| Volkswagem, modelo Santana, placa BPF-3434, cor azul renanavam 420.289.666 | 4934 |
| Vigas de Ferro para estruturas do barracão | 81000 |
| VIGA I DIM. 180 X 6.000 MM COM TALHA PNEUMÁTICA CAP. 1 TON | 1921 |
| VIBRADOR “QUIMIS” PARA PENEIRAS | 235 |
| ventilador de parede Delta Diâmetro 70 CM | 70 |
| Ventilador de parede | 234 |
| Veiculo ford Carrier | 2000 |
| vazo | 30 |
| Vasos - peso estimado de 90 a 130 kg | 7500 |
| vários fios eletricos | 0 |

A partir desse exemplo, podemos compreender o que significa $x_i$. Para cada $i$ diferente temos uma posição de $x$. Assim, por exemplo, $x_1$ será o valor de avaliação inicial do bem 1, ou seja, o valor do Volkswagen, de R$ 4.934,00. Podemos reescrever a tabela da seguinte forma:

Tabela 3.7: Dados de leilões realizados

| Índice | Descrição | Valor de avaliação inicial |
|---|---|---|
| $x_1$ | Volkswagem, modelo Santana, placa BPF-3434, cor azul renanavam 420.289.666 | 4934 |
| $x_2$ | Vigas de Ferro para estruturas do barracão | 81000 |
| $x_3$ | VIGA I DIM. 180 X 6.000 MM COM TALHA PNEUMÁTICA CAP. 1 TON | 1921 |
| $x_4$ | VIBRADOR “QUIMIS” PARA PENEIRAS | 235 |
| $x_5$ | ventilador de parede Delta Diâmetro 70 CM | 70 |
| $x_6$ | Ventilador de parede | 234 |
| $x_7$ | Veiculo ford Carrier | 2000 |
| $x_8$ | vazo | 30 |
| $x_9$ | Vasos - peso estimado de 90 a 130 kg | 7500 |
| $x_10$ | vários fios eletricos | 0 |

Compreendendo o que significam cada um parâmetros do somatório, bem como com interpretar o índice, podemos voltar para a fórmula da média.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Estamos, por enquanto, olhando para este elemento da fórmula $\sum_{i=1}^{n} x_i$. O que podemos concluir do significado disso? Vamos destrinchar com os elementos que vimos até agora. A somatória de $\sum x_i$ significa que vamos somar todos os

elementos $x_i$. Essa soma será feita a partir do primeiro elemento do conjunto, que é $x_1$ (por isso, embaixo do sigma tem $i = 1$), até o último elemento do conjunto, que é $x_n$. No caso, se temos 10 bens, então nosso $n = 10$.

Compreendido o somatório da fórmula, falta apenas o $\frac{1}{n}$. Basicamente, isso indica que devemos dividir o resultado do somatório pelo total de observações. O $n$ que encontramos nessa parte é o mesmo $n$ que está em cima do *sigma*. Então, no caso dos bens, nós somariamos o valor de avaliação dos 10 bens e depois dividiríamos o total resultante dessa soma por 10, finalizando assim, o cálculo da média.

Com isso, a fórmula da média está completa. Apesar de o conceito ser intuitivo, a sua representação matemática é um pouco mais complexa. Mas essa complexidade se mostra como uma ótima oportunidade para explicarmos um pouco sobre notação matemática. Compreender e operar fórmulas matemáticas ajuda a trabalhar com os conceitos matemáticos conforme eles ficam mais complexos.

#### 3.2.2.1.2 Mediana

Outra medida de centro importante é a **mediana**. A mediana é outro conceito intuitivo, mas menos conhecido. Literalmente, a mediana designa exatamente o número que está no meio do conjunto. Para encontrar este número, precisamos ordenar os números em ordem crescente. Uma vez encontrada esta ordem, há duas situações possíveis que podem acontecer. Se a quantidade de observações ($n$) for ímpar, o número exatamente ao meio será a mediana; se a quantidade de observações ($n$) for par, pegamos os dois números no centro e fazemos a média entre eles.

Em conjuntos ímpares, como encontramos o número que está exatamente ao centro? Ou seja, qual é a fórmula genérica para encontrarmos todos os números do meio em todos os conjuntos ímpares? Imaginemos um conjunto de 7 elementos. Intuitivamente conseguimos dizer que o elemento no meio será o 4º elemento, pois há 3 elementos à esquerda do 4º elemento e mais 3 elementos à direita dele. A partir disso, podemos pensar que 4 (isto é, a posisção do número central) é a metade de 8; e que 8 é um número acima de 7 (o número total de elementos do conjunto). Assim, a fórmula que temos é $\frac{n+1}{2}$. Então o índice do elemento central será $x_{\left(\frac{n+1}{2}\right)}$

$$\text{md}(\boldsymbol{x}) = x_{\left(\frac{n+1}{2}\right)}, \text{ se } n \text{ ímpar}$$

Em conjuntos pares, não existe nenhum número exatamente no centro. Se tivermos, ao invés de 7 observações, 8, não teremos nenhum elemento que, caso seja escolhido, divida o conjunto entre duas partes idênticas. Por exemplo, a metade de 8 é 4; então se pegarmos o 4o elemento, o que acontecerá? Haverá 3 elementos à esquerda deste elemento, e mais 4 elementos à sua direita. O 4º elemento, portanto, não pode ser a mediana. O mesmo problema acontece se escolhermos o 5º elemento, pois teremos 4 números à sua esquerda e mais 3 números à sua direita.

Como fazer então para encontrar a mediana neste caso? Basicamente, vamos, no exemplo, pegar o 4º e 5º elementos e fazer a média aritmética entre eles. Pensando de forma abstrata, o 4º elemento é $\frac{n}{2}$; e o 5º elemento é o próximo número da sequência em relação ao elemento do meio, ou seja, ele é $\frac{n}{2} + 1$. Para designar então esses dois elementos temos as seguintes notações: $x_{\left(\frac{n}{2}\right)}$ e $x_{\left(\frac{n}{2}+1\right)}$. A média aritmética entre esses dois elementos é simplesmente a soma entre eles dividido por dois. Assim, temos a seguinte fórmula para a mediana em conjuntos pares:

$$\text{md}(\boldsymbol{x}) = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$$

Vale fazermos um esclarecimento importante a respeito da diferença da média para a mediana. A mediana é preferível à média em muitas situações. Isso decorre de uma propriedade da mediana: ela é *robusta*. Robustez é uma palavra que indica a suscetibilidade de um valor aos seus extremos. Vamos dar dois exemplos para deixar isso claro.

Pensemos em um conjunto abstrato com os seguintes números:

```
10  11  12  13  14  15  16  17  18  19  20
```

Este conjunto contém $n = 11$ observações. Assim, a média é $(10+11+12+13+14+15+16+17+18+19+20)/11 = 15$. A mediana é exatamente o número central, pois o conjunto é ímpar, ou seja, a mediana é 15 também.

Agora olhemos para um segundo conjunto.

```
10  11  12  13  14  15  16  17  18  19  2000
```

A única diferença é o último elemento, ou seja, o número $x_{11} = 2000$. Temos 11 observações ainda. Neste caso a média é $(10+11+12+13+14+15+16+17+18+19+2000)/11 = 195$; mas a mediana continua sendo exatamente a mesma, ou seja, 15, porque 15 continua sendo o elemento do meio.

A partir deste exemplo podemos compreender o que significa dizer que uma medida é mais "robusta" do que outra. A mediana, neste caso, é mais robusta, pois ela se afeta menos com os valores extremos do que a média, ela é mais resistente a valores desviantes.

#### 3.2.2.1.3 Moda

Feita essa consideração, podemos falar da última medida de centro, a **moda**. A moda indica simplesmente o valor mais frequente do conjunto. Encontrar a moda pressupõe que saibamos todos os valores que aparecem no conjunto, bem como a sua contagem. O valor cuja contagem é maior será o valor da moda. Se o conjunto tiver apenas um valor com a maior contagem, ele será "unimodal"; se houver dois valores com a mesma contagem, então teremos um conjunto "bimodal"; se houver muitos valores, então será "multimodal". Se não houver nenhum valor que se destaque, então o conjunto será "uniforme". Por mais que este capítulo não se dedique a estudar gráficos ainda, neste caso, vale a pena demonstrar como cada tipo de conjunto (unimodal, bimodal, multimodal e uniforme) se comporta em termos gráficos. Vemos isso na Figura 3.1.

### 3.2.2.2 Medidas de dispersão

Para as variáveis quantitativas, além de falarmos dos valores mais "típicos" (ou "centrais", como convencionamos chamar), podemos falar também em como esses valores variam, ou como esses valores se dispersam ao longo do conjunto de observações. Com um exemplo simples, percebemos como a informação da medida central isoladamente não é capaz de contar a história inteira dos dados.

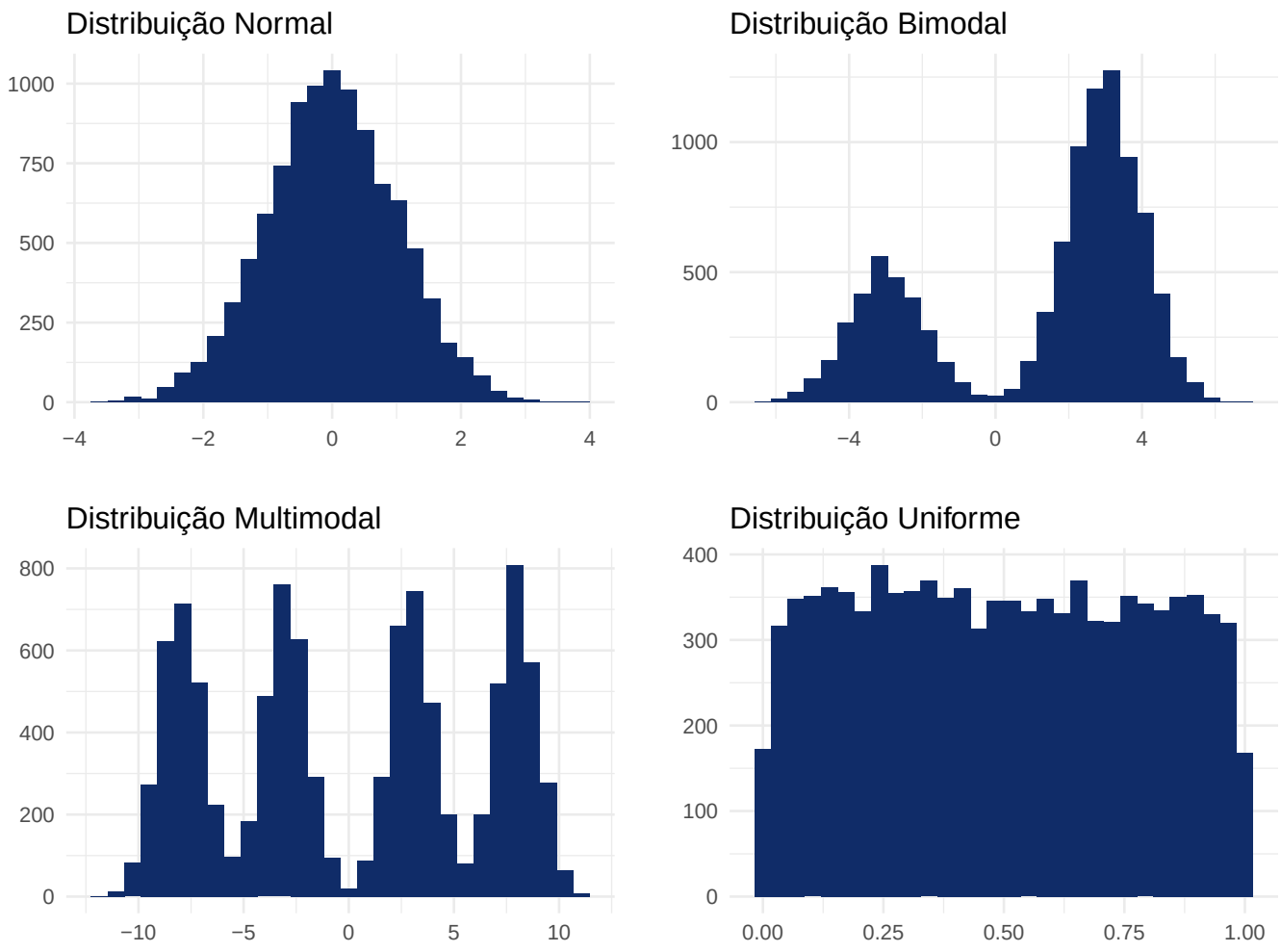



Figura 3.1: Distribuições

```
Grupo A (variável X):   3   4 5 6 7
Grupo B (variável Y):   1 2 5 7 9
Grupo C (variável Z):   5   5 5 5 5
Grupo D (variável W):   1   5 5 6 8
Grupo E (variável V):   3   5 5 6 6
```

Ao calcularmos a média desses 5 grupos, percebemos que todos possuem uma média 5,0. Entretanto, claramente os conjuntos são diferentes. A diferença que estes conjuntos apresentam não pode ser captada pelas medidas de centro. É por esta razão que utilizamos as medidas de dispersão. Há quatro medidas importantes aqui: amplitude, desvio médio, desvio padrão e intervalo inter-quartis, também chamado de IQR.

#### 3.2.2.2.1 Amplitude

A **amplitude** indica, basicamente, o espectro dentro do qual as observações variam. Para calculá-la basta identificar o valor máximo do conjunto, bem como o seu valor mínimo. A diferença entre os dois números é a amplitude.

$$A(\boldsymbol{x}) = \max(\boldsymbol{x}) - \min(\boldsymbol{x})$$

Podemos retornar ao exemplo da base de leilões. O menor valor de avaliação que encontramos nesta base é de R$ 0,00 (zero reais). Esse valor é até frequente na base, pois vários bens arrecadados não possuem valor algum. O maior valor da base é R$ 157.000.000,00, que representa o valor de um imóvel. Assim, a amplitude dessa variável é o valor mínimo subtraído do valor máximo, ou seja, $157.000.000,00 - 0$, ou seja, a amplitude dessa variável é de R$ 157.000.000,00. A interpretação deste valor é que todas as observações estão dadas dentro de um intervalo de cento e cinquenta e sete milhões de reais.

A amplitude é uma medida simples de ser calculada, entretanto, ela não consegue indicar a relação da dispersão com as medidas centrais. Em seu lugar, as medidas de dispersão mais frequentemente usadas são o desvio médio e o desvio padrão. Vamos, a seguir, realizar uma explicação sobre as duas medidas, pois elas estão muito próximas.

#### 3.2.2.2.2 Medidas de dispersão ao redor da média: desvio padrão e desvio médio

As duas medidas que estudaremos a seguir indicam a dispersão dos valores em torno da média. Ou seja, tomando a média (e não qualquer medida de centro) como referencial, essas duas medidas indicam como os dados variam ao redor do centro. Algumas perguntas importantes que essas medidas nos ajudam a compreender são: As observações estão concentradas ao redor da média? Ou elas estão dispersas e longe umas das outras? A Figura 3.2 mostra dois exemplo de conjuntos de dados com a mesma média, mas com dispersões das observações em torno da média totalmente diferentes. No primeiro gráfico, as observações estão concentradas; no segundo, elas estão dispersas. Observamos, ao compararmos estes dois gráficos, como que as medidas de centro não conseguem explicar as distribuições dos conjuntos muito bem, quando isoladas, mas quando combinadas com as medidas de dispersão, conseguimos descrever muito melhor os conjuntos.

Antes de entrarmos nas notações matemáticas, vamos pensar, intuitivamente como essas medidas de dispersão que se referem à média funcionam. Já deve ter ficado claro que a primeira informação essencial que temos de ter em mãos é a média. Sem o valor da média, não conseguimos calcular o desvio médio nem o desvio padrão. Uma vez que temos o valor médio, qual é o próximo passo?

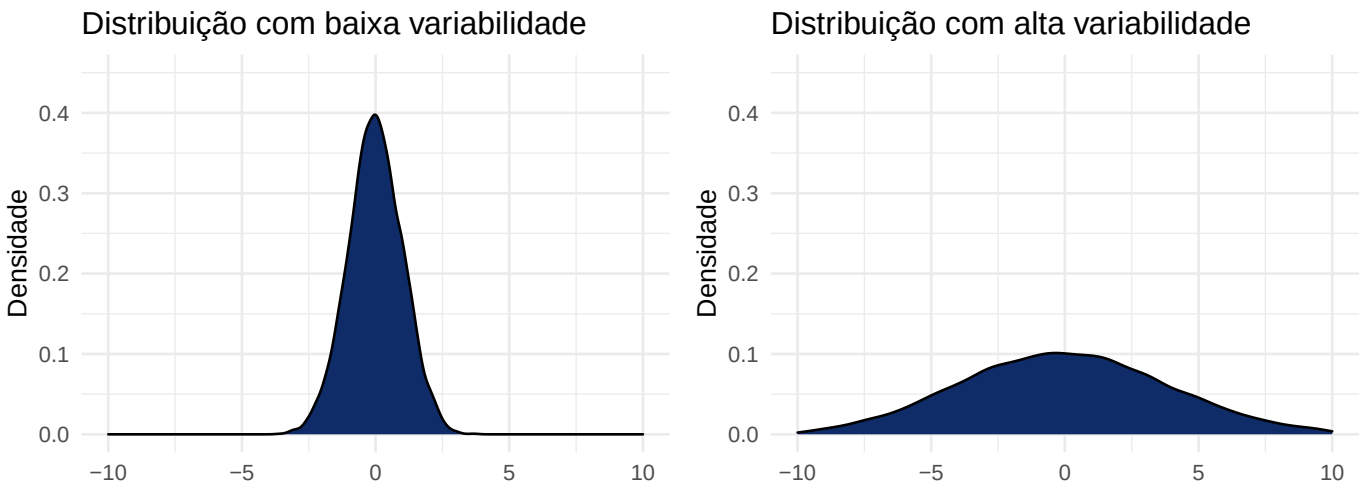



Figura 3.2: Duas distribuições com mesma média e variabilidades distintas.

Para responder a isso, imaginemos o seguinte conjunto de dados, cuja média é 5, conforme a Tabela 3.8.

Tabela 3.8: Dados para calcular desvios

| Índice | Valor |
|---|---|
| $x_1$ | 3 |
| $x_2$ | 4 |
| $x_3$ | 5 |
| $x_4$ | 6 |
| $x_5$ | 7 |

Uma pergunta que podemos fazer sobre esse conjunto, mas que ainda não fizemos, é: o quão distante da média está cada uma das observações? A partir disso, podemos discutir, por exemplo, que número está mais próximo da média 5, $x_4 = 6$ ou $x_5 = 7$? Para responder a essa pergunta, vamos olhar para a Figura 3.3:

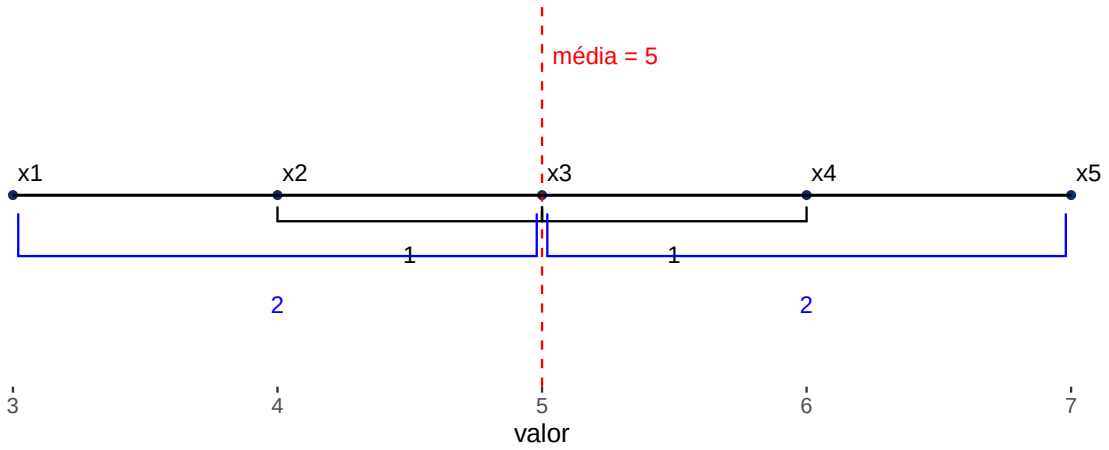



Figura 3.3: Distância

Pela figura, fica claro que a distância entre $x_4$ e a média é menor do que a distância entre $x_5$ e a média. São essas distâncias que queremos computar. Se calcularmos esse valor para todas as observações, então teremos uma informação a mais a respeito desse conjunto: teremos o desvio de cada observação em relação à média. Com isso, conseguimos completar o

| Índice | Valor | Desvio |
|---|---|---|
| $x_5$ | 7 | 2 |

Uma vez que sabemos como cada observação se desvia da média, falta sabermos uma última informação relevante. Lembrando que estamos falando de medidas sobre *conjuntos de informações*, é possível perceber que a informação da diferença é uma informação de *cada observação* e não do conjunto como um todo. Percebemos isso porque para cada índice ($x_1$, $x_2$, $x_3$, ...), temos uma única medida de desvio. A partir disso, o que podemos nos perguntar é: como eu trato das diferenças do *conjunto*, e não de cada observação? A essa pergunta a resposta é simples: basta fazer a média dos desvios, ficando $(2 + 1 + 0 + 1 + 2)/5 = 1,2$.

Dessa forma podemos dizer que o nosso conjunto tem média 5 e que ele tem um desvio médio de 1,2. Essa informação que encontramos (de 1,2 de desvio médio), ainda não explica a variância, mas ela dá uma noção de que tipo de informação a variância leva em conta, porque a lógica por trás da variância é a mesma por trás do desvio médio.

Até aqui vimos apenas uma noção intuitiva do que significam as medidas de "dispersão ao redor da média". Usamos como fio condutor para essa explicação a medida do desvio médio. Vamos, a seguir, formalizar o pensamento intuitivo, bem como desenvolver a outra medida restante o desvio padrão (e veremos que para chegar no desvio médio, precisamos explicar a variância antes).

A começar pela formalização matemática do desvio médio, é preciso notar uma operação matemática importante que fizemos ao longo da explicação. Quando comparamos a *distância* de cada observação em relação à média, usamos a Figura 3.3 para verificar essas medidas. Com isso, construímos a Tabela 3.9. O procedimento matemático que está por trás dessas etapas é o **módulo**. A notação do módulo de $x$ é dada por:

$$|x| = \begin{cases} x, & \text{se } x \geq 0, \\ -x, & \text{se } x < 0 \end{cases}$$

O módulo é usado para calcular distâncias até certos pontos. Essas distâncias nunca podem ser negativas. O que queremos são *valores absolutos*, ou seja, queremos apenas o número que indica a distância, sem considerar o sinal (positivo ou negativo).

Então o que fizemos quando construímos a Figura 3.3 foi representar graficamente o módulo das diferenças entre o valor e a média. Em seguida, computamos essas distâncias na Tabela 3.9. Em termos matemáticos, fizemos a seguinte operação:

$$desvio = |x - \bar{x}|$$

A partir dessa operação, podemos formalizar um pouco o pensamento. O que construímos até aqui foi a notação matemática do conceito de desvio. Essa notação representa o módulo da diferença entre a observação $x$ e a média do conjunto $\bar{x}$.

Devemos notar que calculamos os desvios **para cada** observação. A partir disso, somamos todos os desvios. Essas etapas (de calcular um desvio **para cada** observação, seguido de somar cada um desses desvios em um "desvio total), são representadas matematicamente por meio do somatório:

$$desvio\ total = \sum_{i=1}^{n} |x_i - \bar{x}|$$

Lembrando da notação do somatório, ele significa que para cada valor de $x_i$, em que $i$ pode assumir os valores de 1 a $n$, iremos fazer o módulo da diferença entre o valor e a média, somando todos os resultados.

Por fim, o procedimento final que realizamos na explicação intuitiva foi calcular a média do desvio total, isto é, pegamos o desvio total e dividimos pelo total de observações. De modo mais formalizado, o que fizemos foi a seguinte conta:

$$dm(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

Esta é a fórmula do desvio médio de $x$, ou simplesmente, $dm(x)$.

Para seguirmos adiante com a explicação do desvio médio (e da variância), tem uma questão importante que devemos nos fazer: porque devemos somar o módulo das diferenças e não apenas as diferenças, ou seja, porque a conta para o desvio médio não foi simplesmente $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$?

A resposta para essa pergunta não é trivial. Se a gente calculasse o desvio total pela soma das diferenças (e não pela soma do módulos das diferenças), depois esbarraríamos em um problema ao calcular a média desse desvio. O problema é que a média das diferenças é sempre zero. A prova matemática disso é:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} \bar{x} = \bar{x} - \frac{n}{n}\bar{x} = \bar{x} - \bar{x} = 0$$

Não há necessidade de se compreender toda a demonstração de que a média das diferenças é sempre igual a 0 (mas vale a pena tentar!). O que realmente precisamos entender é simplesmente que se calculássemos cada um dos desvios, sem levar em consideração o módulo, teríamos uma série de números negativos na nossa conta e esses números negativos iriam "anular" os números positivos, resultando em zero sempre. Podemos ver o resultado disso na Tabela 3.10.

Tabela 3.10: Comparação dos desvios com as diferenças

| Índice | Valor | Desvio | Diferença |
|---|---|---|---|
| $x_1$ | 3 | 2 | -2 |
| $x_2$ | 4 | 1 | -1 |
| $x_3$ | 5 | 0 | 0 |
| $x_4$ | 6 | 1 | 1 |
| $x_5$ | 7 | 2 | 2 |
| total | NA | 6 | 0 |

Então, como estávamos falando, para seguirmos adiante na explicação do desvio médio (e da variância), precisávamos entender o problema de não considerar o valor absoluto (ou seja, o valor em módulo) dos números. *O problema são os valores negativos.* Se, para calcular o desvio médio o problema dos números negativos é resolvido pelo módulo, veremos

na variância e no desvio padrão que a forma de resovler este problema é outra. No lugar do módulo, calculamos a diferença ao quadrado. Apenas trocando o módulo pelo quadrado do fórmula do desvio médio, encontramos a fórmula da variância, conforme observamos na seguinte equação:

$$var(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Quando elevamos uma diferença ao quadrado, por mais que a diferença seja negativa, o quadrado sempre será positivo, de modo que podemos observar na Tabela 3.11 que na coluna *Diferença ao quadrado* não há nenhum número negativo.

Tabela 3.11: Comparação dos desvios com as diferenças e com as diferenças ao quadrado

| Índice | Valor | Desvio | Diferença | Diferença ao quadrado |
|---|---|---|---|---|
| $x_1$ | 3 | 2 | -2 | 4 |
| $x_2$ | 4 | 1 | -1 | 1 |
| $x_3$ | 5 | 0 | 0 | 0 |
| $x_4$ | 6 | 1 | 1 | 1 |
| $x_5$ | 7 | 2 | 2 | 4 |

A diferença da variância para o desvio médio é simplesmente a forma como eles lidam com o problema dos valores negativos. Dessa forma, a mesma intuição que desenvolvemos para o desvio médio pode ser usada para compreender a variância, com a diferença que a variância apresenta os resultados distorcidos por estarem ao quadrado. Entretanto, o significado da variância ainda representa algo muito similar ao do desvio médio: a variância apresenta a dispersão dos valores ao redor da média.

Por causa dessa distorção que a variância gera ao elevar os números ao quadrado, pode parecer que o desvio médio é preferível em relação a ela. Acontece que o uso do módulo no cálculo da medida gera alguns problemas matemáticos em algumas operações. Assim, não usamos nem o desvio médio (por usar o módulo) e nem a variância (por estar um pouco distorcida pelos números ao quadrado). No lugar dessas medidas, fazemos uma "correção" da variância. Já que a variância altera a unidade de medida para a unidade ao quadrado, podemos voltar à dimensão anterior simplesmente fazendo a raiz quadrada da variância. Afinal de contas, a raiz quadrada é a operação matemática inversa de elevar ao quadrado. Quando realizamos este procedimento final, de tirar a raiz quadrada da variância, obtemos o chamado *desvio padrão* (e note a sutileza nas palavras, pois estamos falando de desvio *padrão* e não mais de desvio *médio*). A fórmula do desvio padrão é representada da seguinte maneira:

$$dp(\boldsymbol{x}) = \sqrt{var(\boldsymbol{x})} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x - \bar{x})^2}$$

Fica bem claro pela fórmula do desvio padrão que ele é simplesmente a raíz quadrada da variância. Exatamente por guardar essa relação intrínseca com a variância, é que precisávamos explicá-la antes de explicar o desvio padrão.

Com esta última fórmula, conseguimos compreender o significado e as notações matemáticas das duas medidas de dispersão ao redor da média que nos propusemos a discutir: o desvio médio e o desvio padrão (definindo, no meio do

caminho, a variância). Dessas medidas, as mais mais utilizadas são a variância e o desvio padrão. O desvio padrão é mais popular do que o desvio médio por ser mais sensível a valores atípicos (o que geralmente é desejável aqui) e por ter algumas propriedades matemáticas de interesse na modelagem estatística. Não trataremos dessas propriedades por enquanto.

Vamos antes de finalizar essa explicação das medidas que indicam a dispersão dos dados em relação à média, dar um exemplo para dar concretude dessas medidas, pois os conceitos apresentados são difíceis. O exemplo é o mesmo que vem acompanhando a gente nesta seção: o valor de avaliação dos bens alienados em processos de falência. A base usada é a base `leiloes`. Dessa vez, não vamos olhar apenas para 10 observações, mas para 483 observações da base (de um total de 1000 observações). Realizamos uma filtragem dessa base com fins didáticos apenas, para ficar mais fácil de calcular. A base se encontra na Tabela 3.12.

Tabela 3.12: Amostra de 10 bens aleatórios da base

| descricao | valor_avaliacao_inicial |
|---|---|
| bem 1 | 126 |
| bem 2 | 286 |
| bem 3 | 420 |
| bem 4 | 432 |
| bem 5 | 750 |
| bem 6 | 1660 |
| bem 7 | 1900 |
| bem 8 | 1949 |
| bem 9 | 2500 |
| bem 10 | 3700 |

Recomendamos fortemente que você tente calcular as duas medidas de dispersão desse conjunto. A média do valor de avaliação é R$ 1.372,30. Em primeiro lugar, tente calcular as diferenças de cada um desses bens em relação à média. O resultado esperado está resumido na Tabela 3.13

Tabela 3.13: Amostra de 10 bens aleatórios da base com o valor das diferenças

| descricao | valor_avaliacao_inicial | diferenca |
|---|---|---|
| bem 1 | 126 | -1246.3 |
| bem 2 | 286 | -1086.3 |
| bem 3 | 420 | -952.3 |
| bem 4 | 432 | -940.3 |
| bem 5 | 750 | -622.3 |
| bem 6 | 1660 | 287.7 |
| bem 7 | 1900 | 527.7 |
| bem 8 | 1949 | 576.7 |
| bem 9 | 2500 | 1127.7 |
| bem 10 | 3700 | 2327.7 |

A partir destes valores, podemos calcular o desvio médio, a variância e o desvio padrão. Tente calcular por sua conta o valor do desvio médio, da variância e do desvio padrão. Os resultados devem ser:

Tabela 3.14: Resumo das medidas de dispersão ao redor da média do exemplo de leilões

| Medida de dispersão | Valor |
|---|---|
| desvio médio | 969.50 |
| variância | 1229532.41 |
| desvio padrão | 1108.84 |

Como interpretamos esses resultados? Primeiro, devemos nos lembrar de que a média de avaliação de cada bem é R$ 1.372,30, pois essas três medidas dizem respeito à distribuição dos resultados ao redor da média. Uma vez que temos a média, então podemos dizer que:

- Pelo desvio médio, podemos dizer que os valores variam em média R$ 969,50 ao redor da média. Isso significa que os valores estão pouco concentrados e muito dispersos. Percebemos isso porque o valor do desvio médio é muito próximo ao valor da média.
- Pela variância, podemos dizer que os valores ao quadrado variam em média 1.229.532,41 reais ao quadrado ao redor da média. Essa interpretação é difícil de ser feita, por isso olhamos para o desvio padrão.
- Pelo desvio padrão, podemos dizer que os valores variam em média R$ 1.108,84 ao redor da média. Vemos também como o desvio padrão acaba se assemelhando ao valor do desvio médio. Como comentado anteriormente, o desvio padrão tende a considerar mais os valores atípicos e por isso é um pouco maior que o desvio médio, mas os valores são próximos. Aqui, novamente, percebemos que os valores estão dispersos, jáq ue o desvio padrão e a média são próximos.

Vale notar que, na interpretação desses resultados, essas medidas sempre indicam dispersão ao redor da média.

#### 3.2.2.2.3 Interquartile range (IQR) e os quantis empíricos

Há uma última medida de variabilidade que é o *interquartile range* (IQR). Podemos dizer que os IQRs estão para o desvio padrão assim como a mediana está para a média. Se bem lembrarmos, a média padece de um problema: ela não é robusta, ou seja, ela é influenciável pelos valores extremos. Em seu lugar, podemos usar a mediana para indicar o centro da distribuição, pois essa medida é mais robusta. Da mesma forma, como a variância e o desvio padrão dependem da média para serem calculadas, elas também são medidas pouco robustas. O IQR, então, pode ser visto como uma medida de dispersão robusta. **Mas para compreendermos o IQR, devemos compreender o que são quantis empíricos**.

Um **quantil empírico** é um valor que divide um conjunto de dados em determinada proporção. Por exemplo, a mediana é o quantil empírico que indica o valor que divide o conjunto de dados em 50%. Existem outros quantis empíricos, como o quantil de 25%, que é um valor que divide o conjunto de observações entre 25-75, ou seja, 25% das observações estão antes deste valor e 75% das observações estão depois deste valor.

Existe um quantil empírico para cada porcentagem que se queira. Representamos, matematicamente, cada percentil como $q(p)$, em que $p$ é a proporção desejada. Por exemplo, o quantil de 25% é o $q(0,25)$; o quantil de 78% é o $q(0,78)$; o quantil de 1% é o $q(0,01)$, e assim por diante.

Alguns quantis possuem nomes específicos. Se dividirmos os quantis em 100, cada quantil será um "percentil". E se dividirmos o conjunto em 10 partes iguais, cada quantil será um "decil". Ainda, se dividirmos os quantis em 4 partes iguais, cada quantil será um "quartil". Esta última divisão é uma das divisões mais importantes. Apesar de dividirmos o conjunto em 4 partes iguais, e apesar de o nome desse quantil ser "quartil" (o que, novamente, remete à ideia de 4 partes), há apenas 3 quartis:

- quartil inferior, $q(0,25)$: Divide o conjunto no 25%, ou seja, um quarto dos dados estão abaixo deste quartil.
- mediana, $q(0,50)$: Divide o conjunto no 50%. É a mediana.
- quartil superior, $q(0,75)$: Divide o conjunto no 75%, ou seja, apenas um quarto dos dados estão acima deste quartil.

Chamamos essas três medidas de quartis, não porque são 4 valores, mas porque conseguimos dividir o conjunto dos dados em 4 partes iguais a partir destes valores. A primeira é calculado como a amplitude entre o ponto mínimo, $q(0,00)$, e o primeiro quartil, $q(0,25)$; a segunda, entre o segundo quartil, $q(0,25)$ e a mediana, $q(0,50)$; a terceira, entre a mediana, $q(0,50)$ e o terceiro quartil, $q(0,75)$; e a quarta, entre o terceiro quartil, $q(0,75)$ e o valor máximo, $q(1,00)$. Assim temos:

- Conjunto 1: $q(0,25) - q(0,00)$
- Conjunto 2: $q(0,50) - q(0,25)$
- Conjunto 3: $q(0,75) - q(0,50)$
- Conjunto 4: $q(1,00) - q(0,75)$

A partir dos quartis (isto é, os quantis que dividem o conjunto em 4 partes iguais), podemos construir a medida do IQR. Enquanto os quartis indicam pontos no conjunto de observações, o IQR indica uma amplitude. Entretanto, ao contrário da amplitude geral que vimos acima, que é construída a partir do valor mínimo, $q(0,00)$, e do valor máximo, $q(1,00)$, o IQR é construído pela amplitude entre o quartil inferior, $q(0,25)$, e o quartil superior, $q(0,75)$.

Justamente por excluir os valores extremos (que estão no conjunto 1 e no conjunto 4, ou seja, entre o valor mínimo e o quartil inferior e entre o quartil superior e o valor máximo), dizemos que o IQR é uma medida mais robusta do que o desvio padrão para indicar a dispersão do conjunto.

Como exemplo final, podemos voltar à base de leilões. Dessa vez, vamos olhar para a base completa (as 1000 observações), ao invés de alguma amostra dela. Temos os seguintes quantis empíricos da variável sobre o valor de avaliação

Tabela 3.15: Resumo dos quantis empíricos do valor de avaliação

| Quantil | Valor |
|---|---|
| min(0%) | 0 |
| quartil inferior (25%) | 104 |
| mediana (50%) | 500 |
| quartil superior (75%) | 3755 |
| máximo (100%) | 157000000 |

Para calcular o IQR, basta subtrair o quartil inferior do quartil superior, ou seja 3755-104, que resulta em 3651. A interpretação desse resultado é os valores centrais da amostra se distribuem dentro de uma amplitude de R$ 3.651,00. Ou seja, 50% dos valores estão distribuídos dentro de um intervalo de R$ 3.651,00. É interessante comparar o IQR com a amplitude. A amplitude, como vimos, é igual a R$ 157.000.000,00, ou seja, todos os valores da base estão dentro

de um intervalo de R$ 157.000.000,00, mas desse intervalo inteiro, 50% dos valores (e não quaisquer 50% dos valores, mas a metade dos valores centrais) estão apenas em um intervalo de amplitude R$ 3.651,00. Isso indica como que a população está muito dispersa, pois há uma concentração de metade dos valores em um pequeno intervalo, e uma outra metade distribuída em um grande intervalo. Por mais que ainda não tenhamos estudado gráficos, é útil neste momento compreender para o que estamos falando de forma visual (Figura 3.4).

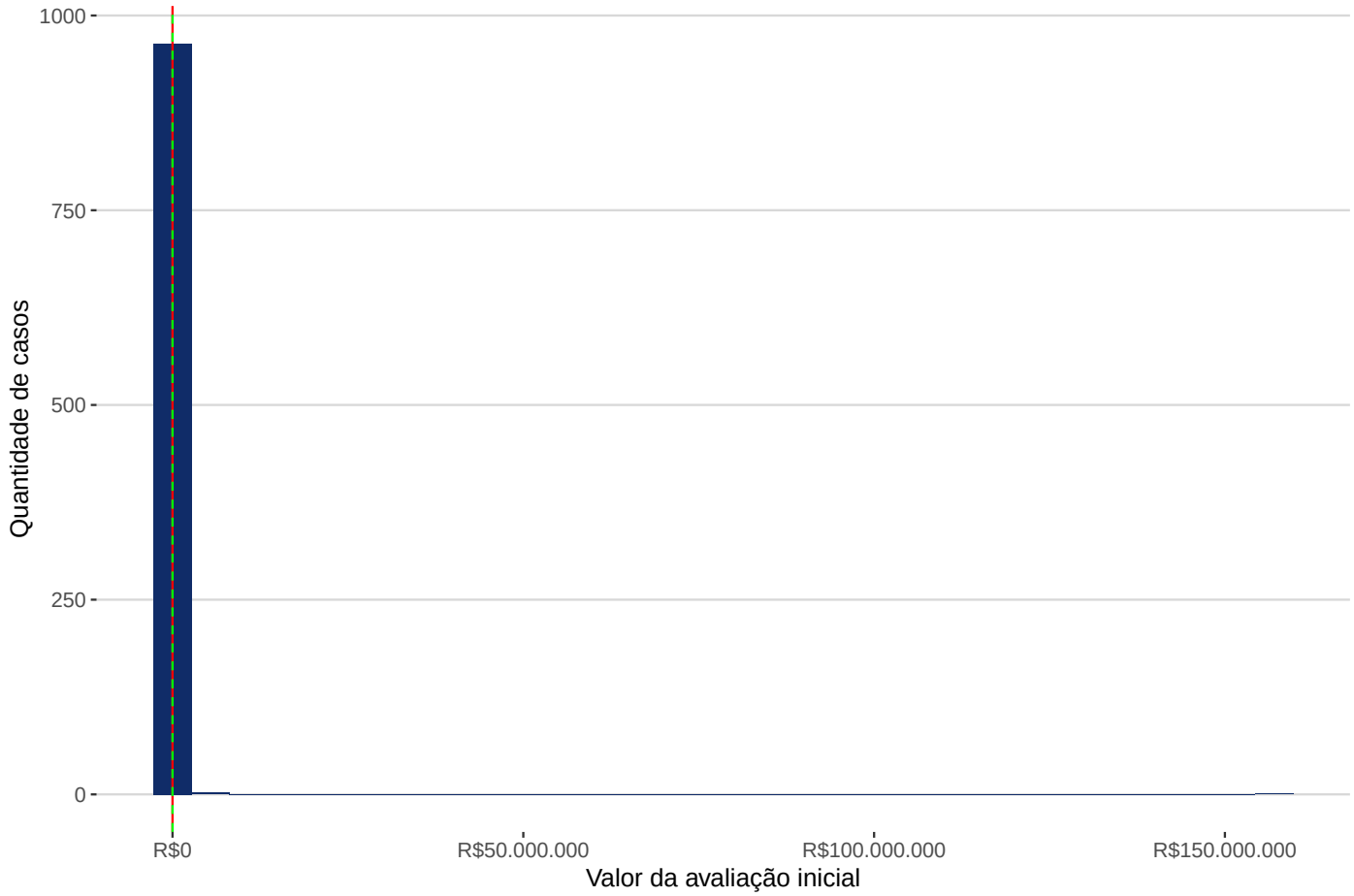



Figura 3.4: Dados de valores de avaliação com alta variabilidade

Observamos neste gráfico uma altíssima concentração em valores, que, por causa da escala, estão representados no 0. Não é que, de fato, os valores sejam 0, mas é que, dentro de uma escala que vai até 150 milhões, a faixa de valores em que se concentram os dados é muito pequena (50% estão na faixa entre R$ 104,00 e R$ 3.755,00).

# 4 Visualização

No Capítulo 2, nós vimos questões sobre o método quantitativo, já no Capítulo 3, passamos do método às medidas. Nesta ocasião, vimos as medidas de duas formas diferentes: as medidas das observações consideradas individualmente e as medidas de resumo do conjunto de observações. Na parte das medidas individuais de cada observação, vimos que as características das observações podem ter duas naturezas distintas, ou elas são categóricas (pois indicam uma categoria de resposta), ou elas são quantitativas (pois indicam uma resposta numérica). Essas duas naturezas possíveis possuem suas subdivisões, mas o mais importante é considerarmos essas duas grandes classes. Em seguida, na parte de medidas resumo do conjunto das observações, tivemos que separar as medidas resumo das variáveis categóricas das medidas resumo das variáveis quantitativas.

Neste capítulo, ainda vamos nos valer desta distinção – entre variáveis categóricas e numéricas –, mas não vamos falar de medidas de resumo. O foco deste capítulo são as visualizações relacionadas a cada tipo de dado. Inicialmente, precisamos discutir para que servem as visualizações suas vantagens e desvantagens. Em seguida, vamos falar das visualizações específicas de cada tipo de variável (categórica ou quantitativa), começando pelas visualizações das variáveis categóricas para seguir então para as visualizações das variáveis quantitativas. Depois de olhar como representar individualmente as variáveis, vamos olhar para gráficos bivariados.

## 4.1 Para que servem visualizações?

Hoje em dia, nós vemos gráficos em vários contextos da vida cotidiana. Vemos gráficos em jornais, apresentações, livros, artigos de revista e em um monte de outros lugares. Mas nem sempre foi assim. Tratar de *dados* por meio de *visualizações gráficas* é apenas um jeito de contar a história. No capítulo anterior, vimos como contar a história, não por meio de gráficos, mas por meio de medidas de resumo. Se os gráficos nem sempre foram utilizados, cabe nos perguntarmos por que eles ganharam essa relevância nos dias atuais? O que eles trazem que outras formas de comunicar as informações (como em tabelas) não trazem?

Vamos começar essa discussão com uma citação importante, retirada do livro The *Elements of Graphing Data* [Os Elementos da Representação Gráfica dos Dados] (1985) do William S. Cleveland. Na sessão *the power of graphical display* [o poder da representação gráfica], Cleveland diz:

> Representações gráficas são uma ferramenta excepcional para a análise de dados. A razão disso está bem resumida em uma sentença de uma carta de 1982 escrita pelo senhor W. Edwards Deming a mim: "Os métodos gráficos podem reter a informação nos dados". Procedimentos numéricos de análise de dados – tais como a média, o desvio padrão, coeficientes de correlação e teste-t – são essencialmente técnicas de redução dos dados. Os métodos gráficos complementam essas técnicas. Os métodos gráficos tendem a mostrar conjuntos de dados como um todo, permitindo-nos resumir o comportamento geral e estudar o detalhe. Isso nos leva a uma análise de dados mais minuciosa. Uma razão para as representações gráficas

> conseguirem reter as informações dos dados é que uma grande quantidade de informação quantitativa pode ser exibida e absorvida[1].

Para dar um exemplo do que significa conseguir "exibir e absorver" informações quantitativas, e para explicar também o que significa "técnicas de redução de dados", vamos olhar para um exemplo, no qual nós comparamos três formas de apresentar os mesmos dados. Os dados expostos dizem respeito aos valores das causas de ações de consumo.

Tabela 4.1: Tabela com dados de valores

| valor | | | | |
|---|---|---|---|---|
| 3373.20 | 29940.00 | 10000.00 | 12500.00 | 20000.00 |
| 5000.00 | 2546.50 | 1200.00 | 15000.00 | 52250.00 |
| 43873.64 | 15665.01 | 16294.08 | 11794.29 | 5843.75 |
| 4224.50 | 10000.00 | 463.31 | 15000.00 | 10000.00 |
| 52900.00 | 6125.96 | 9449.46 | 6299.77 | 54880.14 |
| 4241.39 | 17293.00 | 18710.00 | 26664.00 | 8299.46 |
| 8594.40 | 5158.70 | 10000.00 | 23390.00 | 11067.60 |
| 17407.35 | 15000.00 | 21650.34 | 4408.35 | 3612.00 |
| 9980.00 | 10450.00 | 30000.00 | 10925.93 | 1000.00 |
| 95735.92 | 31236.32 | 16050.00 | 39073.68 | 10000.00 |
| 13179.28 | 23082.44 | 13180.95 | 5000.00 | 17681.82 |
| 3677.88 | 1000.00 | 19224.00 | 52582.70 | 12105.00 |
| 15000.00 | 23289.31 | 1180.00 | 20000.00 | 26802.24 |
| 9980.00 | 6028.00 | 134311.00 | 2003.40 | 9730.00 |
| 20000.00 | 1617.07 | 40000.00 | 6810.80 | 31462.73 |
| 15144.08 | 16297.08 | 21236.52 | 2763.90 | 5450.00 |
| 10000.00 | 4036.91 | 20000.00 | 49964.81 | 12343.62 |
| 15000.00 | 13454.08 | 2944.42 | 990.51 | 2581.76 |
| 22267.55 | 13778.04 | 1000.00 | 6533.40 | 20531.12 |
| 8784.00 | 17284.38 | 84523.57 | 10000.00 | 15000.00 |

Tabela 4.2: Medidas de resumo

| medidas | valores |
|---|---|
| média | 7654.10 |
| desvio padrão | 174.82 |
| mínimo | 63.31 |
| quartil inferior | 981.94 |

[1] Cleveland (1985), pp. 9-10, tradução livre, grifos no original No original: "Graphs are exceptionally powerful tools for data analysis. The reason is nicely encapsulated in a sentence from a 1982 letter written to me by W. Edwards Deming:"Graphical methods can retain the information in the data." Numerical data analytic procedures – such as means, standard deviations, correlation coefficients, and t-tests – are essentially data reduction techniques. Graphical methods complement such numerical techniques. Graphical methods tend to show data sets as a whole, allowing us to summarize the general behavior and to study detail. This leads to much more thorough data analysis. One reason why graphical displays can retain the information in the data is that a large amount of quantitative information can be displayed and absorbed."

| medidas | valores |
|---|---|
| mediana | 2421.81 |
| quartil superior | 0.00 |
| máximo | 34311.00 |

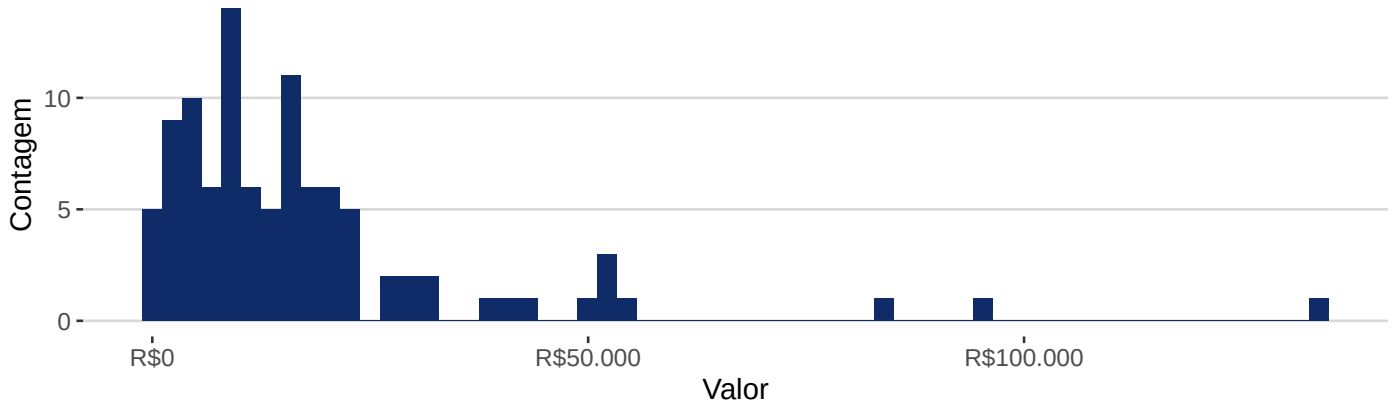



Figura 4.1: Exemplo de visualização

As três apresentações de dados que fizemos foram, respectivamente, uma tabela de observações, apresentando 100 valores de causa em ações de consumo; seguido das medidas resumo dessas observações; finalizando com um histograma dessas observações. O que queremos discutir é *o que a apresentação gráfica nos mostra que as demais formas não nos mostram*?

A primeira apresentação dos dados nos diz muito pouco. Por mais que tenhamos uma visão de *todos* os dados, não conseguimos tirar conclusão alguma sobre eles. As informações ali estão muito cruas. É difícil de ordenar as informações e de observar tendências gerais.

Na segunda forma de apresentar os dados, todo aquele conjunto imenso de dados foi reduzido a algumas poucas estatísticas de resumo. No caso, escolhemos mostrar a média, o desvio padrão e algumas medidas de posição importantes (o mínimo, o máximo, os quartis inferior e superior e a mediana). Nesse caso, perdemos a noção do *todo* (pois não vemos mais cada uma das observações), para olharmos para tendências gerais. Essas medidas nos permitem olhar para algumas tendências importantes, por exemplo, conseguimos perceber que, como a média está mais à direita do que a mediana, pois R$ 7654.1 (média) é maior do que R$ 2421.81 (mediana), que a distribuição não é simétrica, mas que existem pontos à direita que estão "puxando" a média para a direito, enquanto a mediana, que é robusta como vimos, não está recebendo efeito dessas observações à direita. Confirmamos essa mesma tendência pela observação de que o desvio padrão é muito maior do que a média.

Todas essas conclusões que tiramos a partir das medidas resumo nos levam a perceber que a distribuição dos valores de causa de ações de consumo possui uma distribuição assimétrica para a direita (ou seja, muitos dados concentrados próximo de zero, e poucos dados muito distantes do zero). Acontece que o processo para chegarmos a essa conclusão não foi intuitivo, ele necessitou de algumas análises anteriores.

É com esse problema em mente que podemos olhar para o gráfico que está apresentando os mesmos dados. Duas características são importantes da representação gráfica pelo histograma: (a) o gráfico não "reduz" os dados a algumas poucas medidas, mas ele exibe todas as observações, aproximando-se, dessa forma, da primeira apresentação dos dados (isto é, a tabela de observações); (b) ao mesmo tempo em que o gráfico mostra *todas* as observações, ele ainda consegue indicar tendências gerais da distribuição dos dados, nesse sentido, aproximando-se da segunda apresentação dos dados (isto é, a tabela de medidas resumo). Aquela conclusão que tiramos a partir das medidas resumo (de que a distribuição é assimétrica para a direita, pois ela possui muitos dados concentrados próximo de zero e poucos dados distantes do zero) é reforçada no gráfico. A diferença é que conseguimos chegar a essa conclusão de uma forma muito mais intuitiva. É por essa razão que Cleveland diz que "uma grande quantidade de informação quantitativa pode ser exibida e **absorvida**", com destaque para a palavra "absorvida".

A partir desse exemplo, conseguimos compreender algumas características importantes das visualizações gráficas: elas conseguem manter um olhar para *cada uma das observações* (sem reduzir os dados a algumas poucas medidas), e, ao mesmo tempo, conseguem apresentar tendências gerais da distribuição dos dados de uma forma de fácil absorção e compreensão por aquele que visualiza.

Com essas características em mente, então, podemos prosseguir à próxima pergunta: para que servem visualizações?. Há duas respostas importantes: as visualizações gráficas servem para investigar os dados, bem como servem para comunicar resultados. Vamos olhar para as duas funções

### 4.1.1 Detetive de dados - Análise exploratória de dados

Uma forma de encarar as visualizações gráficas é criar visualizações a fim de se obter evidências e pistas para analisar outros fenômenos. Isso é o que Tukey (1977) chama de "quantitative detective work" [trabalho de detetive quantitativo]. Nesse sentido, a visualização de dados não é um fim em si mesmo, mas é uma etapa para um processo de compreensão maior. Nesse sentido, usamos visualizações gráficas para descobrir como um determinado dado se distribui; e a partir dessa informação, podemos perceber que talvez seja adequado fazer uma transformação de variável. Por exemplo, ao analisar o valor da causa, podemos nos deparar com a seguinte distribuição, conforme a Figura 4.2.

Não queremos colocar um gráfico desses em um relatório, ou em um artigo a ser publicado. Mas esse gráfico é importante para percebemos que o eixo x (valor) precisa passar por um ajuste. O ajuste que vamos fazer é colocar o valor em escala logarítmica. Essa é uma transformação muito comum. Mais para frente veremos por que, quando e como fazer essa transformação, por hora, basta sabermos que o primeiro gráfico foi um indicativo de que precisávamos mudar a escala da variável `valor`. Esse gráfico conseguiu nos indicar isso pois ele nos revelou a distribuição das observações sobre essa variável. E ao saber a distribuição, conseguimos padronizar melhor os dados. Na Figura 4.3, vemos as mesmas informações dispostas na Figura 4.2, ajustadas com log.

A partir da Figura 4.3, conseguimos observar uma distribuição totalmente diferente daquela expressa na Figura 4.2. Essa nova figura tem alguns problemas de interpretação, por exemplo, o que significa que o valor da causa ser log 8? A informação não está equivocada, mas ela só está de difícil interpretação. O que queríamos mostrar é que a Figura 4.2 foi usada, não para apresentar os dados que ela continha, mas como um meio para chegarmos à Figura 4.3.

São usos deste tipo que aqui estamos chamando de uma análise exploratória de dados. Nesse sentido (de exploratório), os gráficos servem de instrumento de investigação de relações. Veremos ao longo do livro várias formas de usar gráficos nesse sentido, tais como testes de normalidade, verificar pressupostos de modelos, descobrir outliers, entre muitos outros.

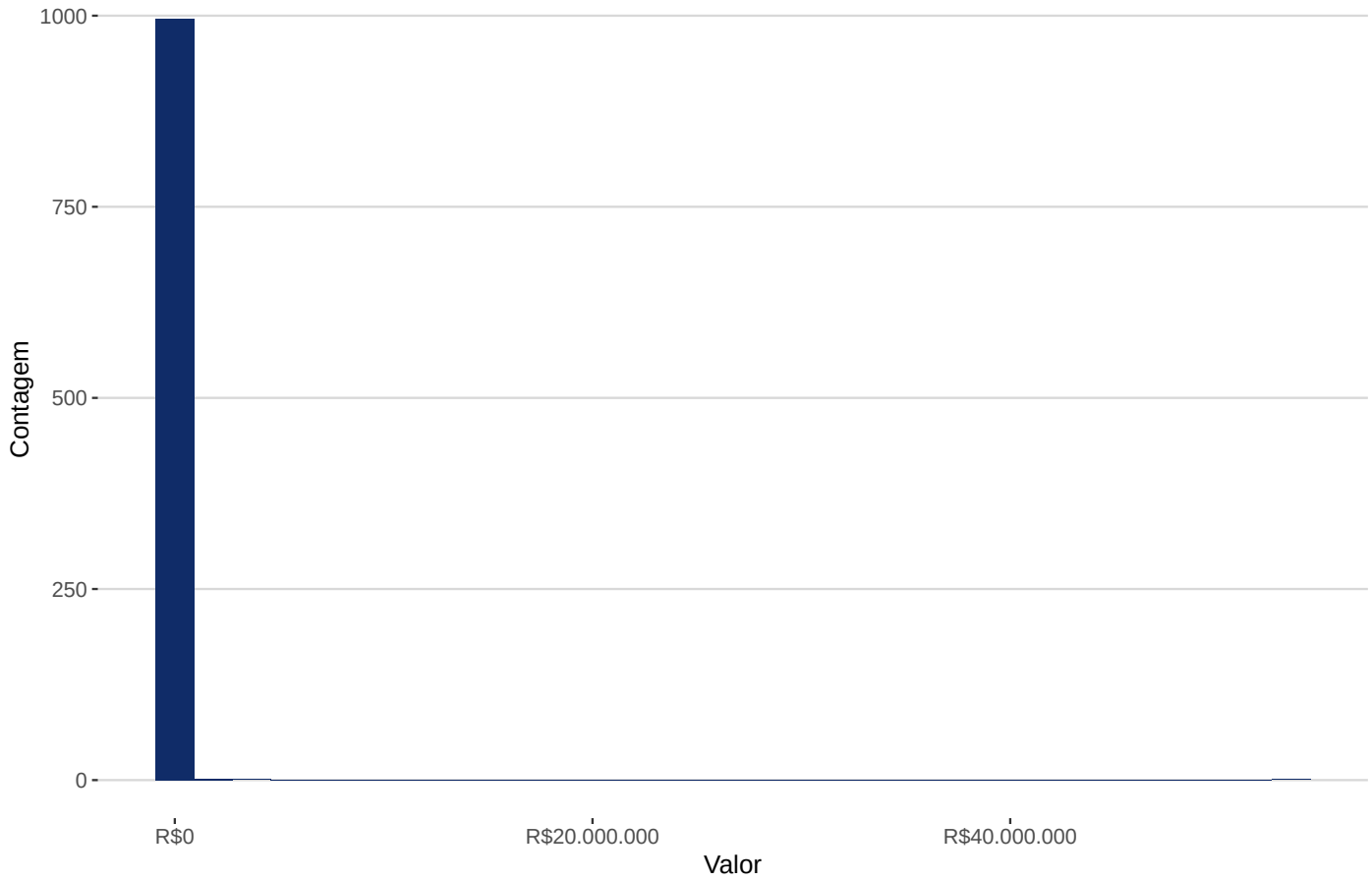



Figura 4.2: Distribuição do valor da causa

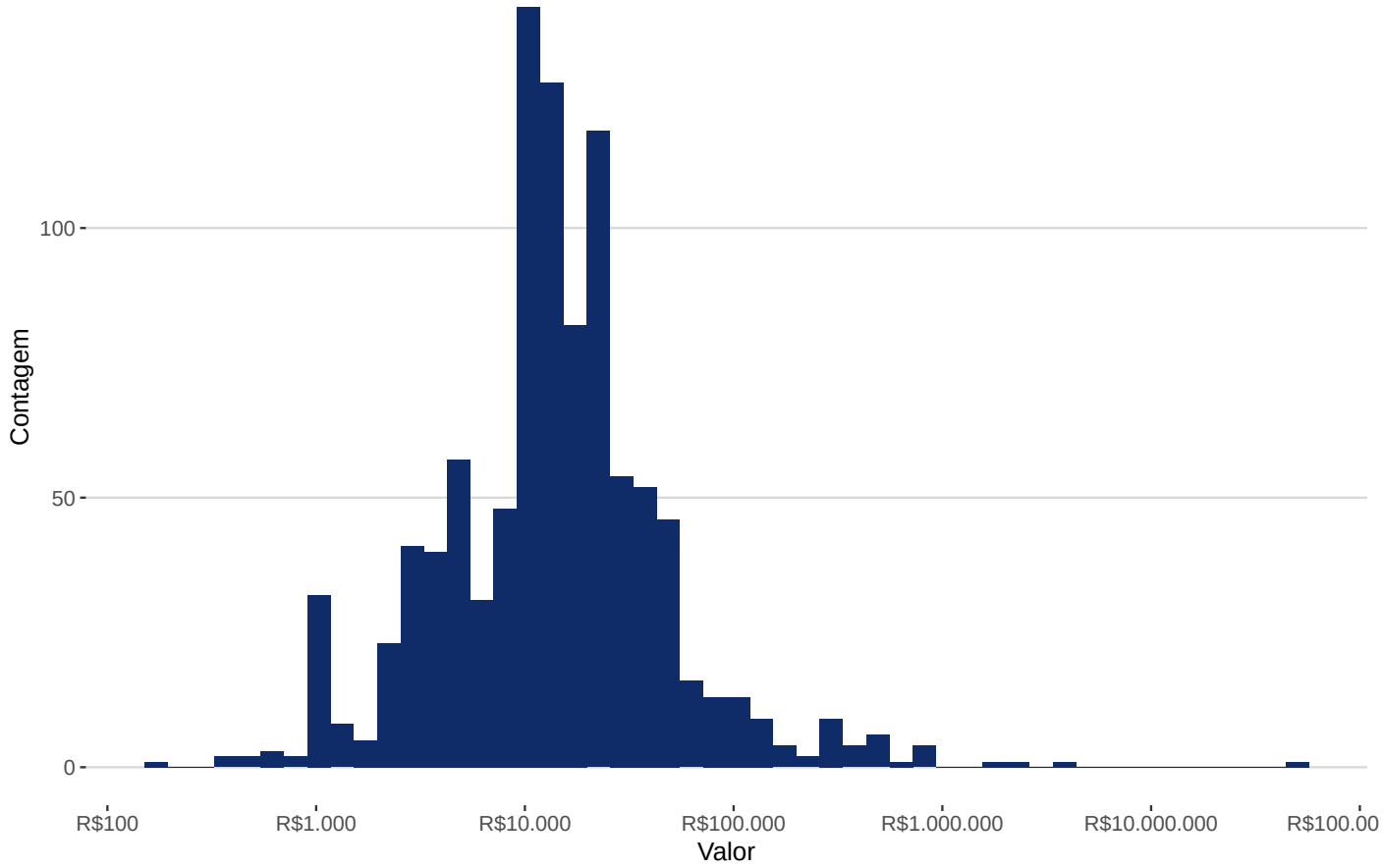



Figura 4.3: Distribuição do valor de causa (com log);

## 4.2 Comunicadora de dados - Apresentação de dados

Outra função que podemos dar à visualização de dados é a comunicação dos resultados. Bons gráficos devem ser gráficos que consigam ser facilmente compreendidos; são gráficos claros e intuitivos. Existe uma série de estudos sobre como construir gráficos para comunicação, assim como existem muitos exemplos, principalmente no jornalismo, de gráficos que contam boas histórias.

> **Nota**
>
> Recomendamos aqui a leitura de Cleveland, William S. The Elements of Graphing Design. California: Wadsworth Advanced Book Program. 1985; ou para um resumo e aplicação prática de Cleveland, recomendamos Kozak, Marcin. Basic principles of graphing data. Sci. Agric. (Piracicaba, Braz.), v.67, n.4, p.483-494, July/August 2010. Neste artigo de Kozak, ele resume os princípios elencados por Cleveland em The Elements of Graphing Data, bem como fornece valiosos exemplos de como esses princípios podem melhorar a apresentação gráfica.

> **Nota**
>
> Aqui recomendamos os gráficos do Nexo Jornal(https://www.nexojornal.com.br/grafico/); a sessão Igualdades da Revista Piaui; os gráficos do New York Times; as reportagens do Núcleo de Jornalismo; e, por fim, esta reportagem do Estadão sobre adoção de crianças. Todos os sites foram acessados em 28/04/2022.

Temos que ter alguns cuidados nessa parte de comunicação dos resultados. A comunicação é uma etapa que, se feita de forma equivocada, pode não conseguir passar a interpretação correta dos resultados, ou ainda, se for feita de forma maliciosa, pode passar informações falsas.

> **Nota**
>
> Ver o exemplo discutido por Cleveland, 1985, sobre o gráfico que o Carl Sagan apresentou em seu livro Os Dragões do Éden a respeito da relação entre a proporção da massa do cérebro em relação à massa do corpo e a inteligência de diversas espécies. O exemplo se encontra em na sessão 1.3. The Challenge of Graphical Display do livro Cleveland, William S. The Elements of Graphing Design. California: Wadsworth Advanced Book Program. 1985.

> **Nota**
>
> Ver o artigo de Caio Lente no blog da Curso-R, em que ele descreve como os dados podem mentir, ao se analisar uma imagem sobre o desenvolvimento econômico da Argentina após a Reforma Constitucional que ensejou o início da Justiça Social. O artigo está disponível neste link, acessado em 28/04/2022.

## 4.3 Visualizações em espécie

Tendo em mente as características e funções das visualizações, podemos prosseguir ao detalhamento de algumas espécies de visualizações. Existem muitos tipos de gráficos. Você pode ver uma tentativa de documentação completa desses gráficos em From data to Viz. Não vamos falar de todos esses gráficos, mas apenas dos tipos mais frequentes.

Da mesma forma como nós dividimos a explicação das medidas de resumo entre as explicações das medidas para variáveis categóricas e as medidas para variáveis numéricas, vamos, novamente, seguir este padrão, de modo que apresentaremos:

1. Visualizações para variáveis categóricas
    1. Gráficos univariados
    2. Gráficos bivariados
        1. Com explicativa categórica
        2. Com explicativa numérica (só faz sentido regressão)
2. Visualizações para variáveis numéricas
    1. Gráficos univariados
    2. Gráficos bivariados
        1. Com explicativa categórica
        2. Com explicativa numérica

Para cada gráfico, iremos ver (i) como desenhar o gráfico e (ii) como interpretá-lo. A explicação do ponto (i) já diz muito sobre o ponto (ii), mas iremos tentar discriminar cada uma das análises especificamente.

Antes de continuar as explicações, precisamos explicar o que significa ser um gráfico univariado ou bivariado?.

Um gráfico univariado representa apenas uma única variável nele. Isso **não significa que** o gráfico tenha apenas um eixo. Na verdade, os gráficos univariados normalmente têm dois eixos. Então a quantidade de variáveis não diz exatamente respeito à quantidade de eixos.

Já o gráfico bivariado representa duas variáveis nele, em que uma variável é a variável resposta e a outra variável é a variável explicativa. A variável resposta é aquela que nós queremos compreender; e a variável explicativa é a variável que queremos usar para explicar a variável resposta.

Vamos dar um exemplo para compreender o que são as variáveis explicativa e resposta. Ao analisar instância recursal, podemos nos interssar por ver se a reforma da decisão de primeiro grau é afetada pelo valor da causa. Então surge a seguinte pergunta de psquisa: será que causas de valores maiores tendem a ser menos reformadas? Veja que a relação que queremos avaliar tem uma ordem certa: importa dizer que é o valor da causa que explica a reforma, e não o contrário, de que é a reforma da sentença que explica o valor da causa. Essas duas relações são possíveis de serem analisadas (apesar de que faz pouco sentido teórico explicar o valor da causa pela reforma). O importante é notar que essas duas relações são distintas, e aquilo em que elas se distinguem é justamente o que cada uma delas considera como variável resposta e como variável explicativa. Na pergunta de se a reforma da sentença é afetada pelo valor da causa, a variável "reforma" é a variável resposta (ou seja, a variável que queremos compreender); e a variável "valor da causa" é a variável explicativa (ou seja, a variável que queremos que explique a variável resposta). Em termos matemáticos, a variável resposta é a variável Y e a variável explicativa é a variável X.

### 4.3.1 Visualizações de variáveis categóricas

#### 4.3.1.1 Gráficos univariados

Os gráficos univariados de variáveis categóricas são marcados por conterem sempre duas informações: as categorias analisadas e a contagem ou proporção dessas categorias. Por mais que os gráficos tenham duas dimensões, eles não contém duas "variáveis". A única variável são as categorias; a contagem/proporção é simplesmente um atributo dessa variável. Vamos ver duas visualizações mais difundidas desse tipo de variável. Para mais exemplos, ver From data to Viz.

##### 4.3.1.1.1 Gráfico de barras

O gráfico de barras é o principal meio de visualizar variáveis categóricas. Ele contém dois eixos: a variável categórica e a contagem/proporção. Não importa o eixo em que cada uma dessas informações está. A variável pode ser apresentada tanto no eixo x (horizontal), como no eixo y (vertical, também chamado de gráfico de colunas), conforme a Figura 4.4.

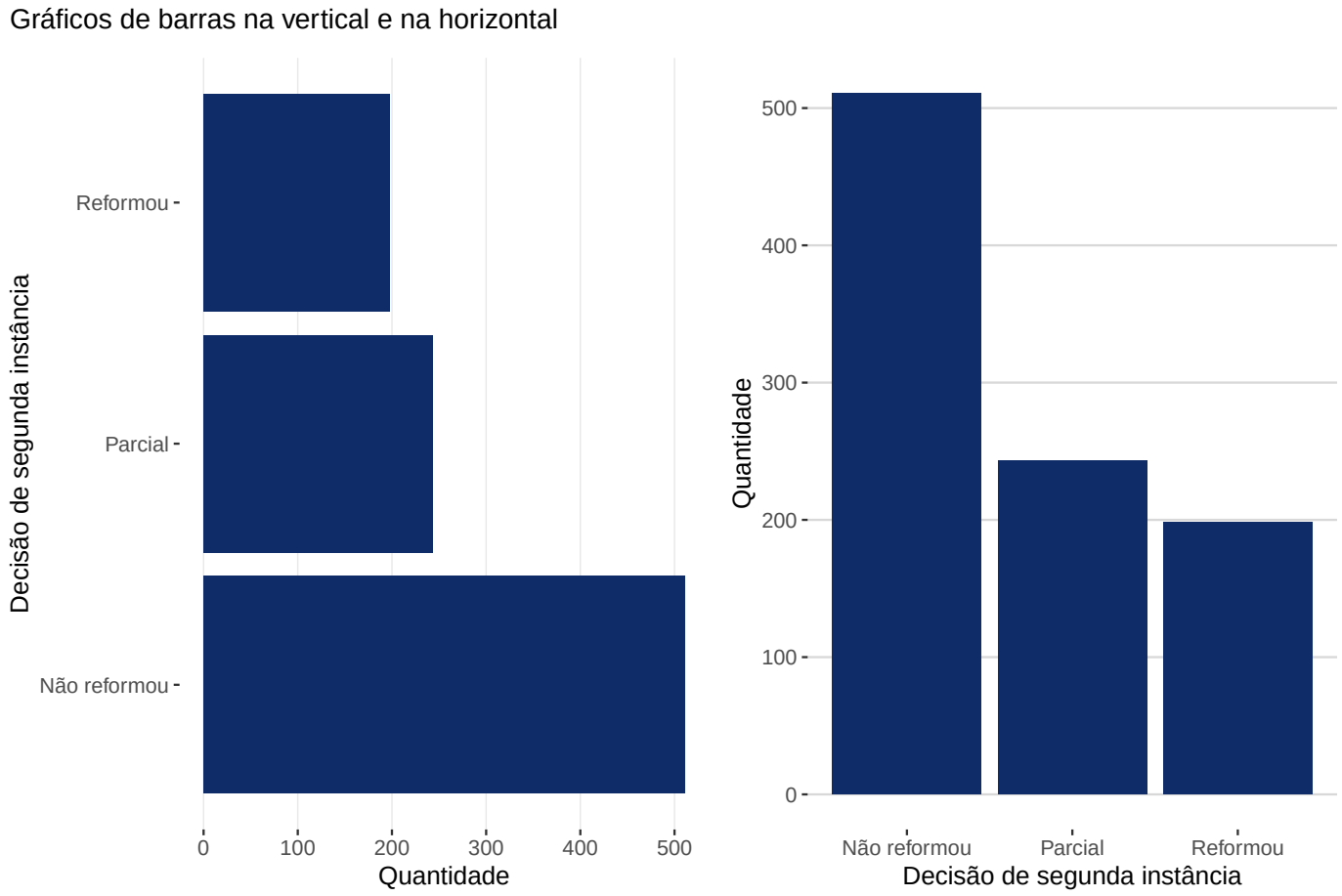



Figura 4.4: Gráficos de barras

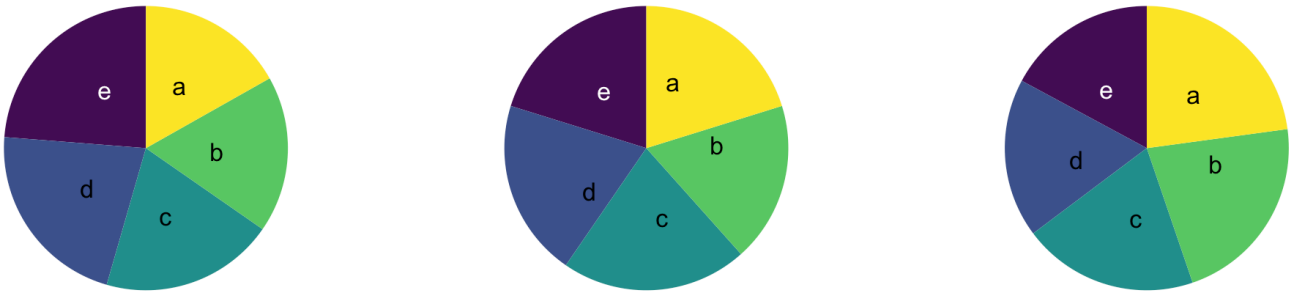
O gráfico é simples, bem como a sua interpretação. Basta verificar a contagem de cada categoria. Usualmente, tira-se desses gráficos a conclusão de quais grupos são mais relevantes, qual é a tendência de resposta de determinada categoria. No caso em tela, observamos que os processos de consumo, em segunda instância, tendem a não ser reformados.

#### 4.3.1.1.2 Gráfico de setores (pizza)

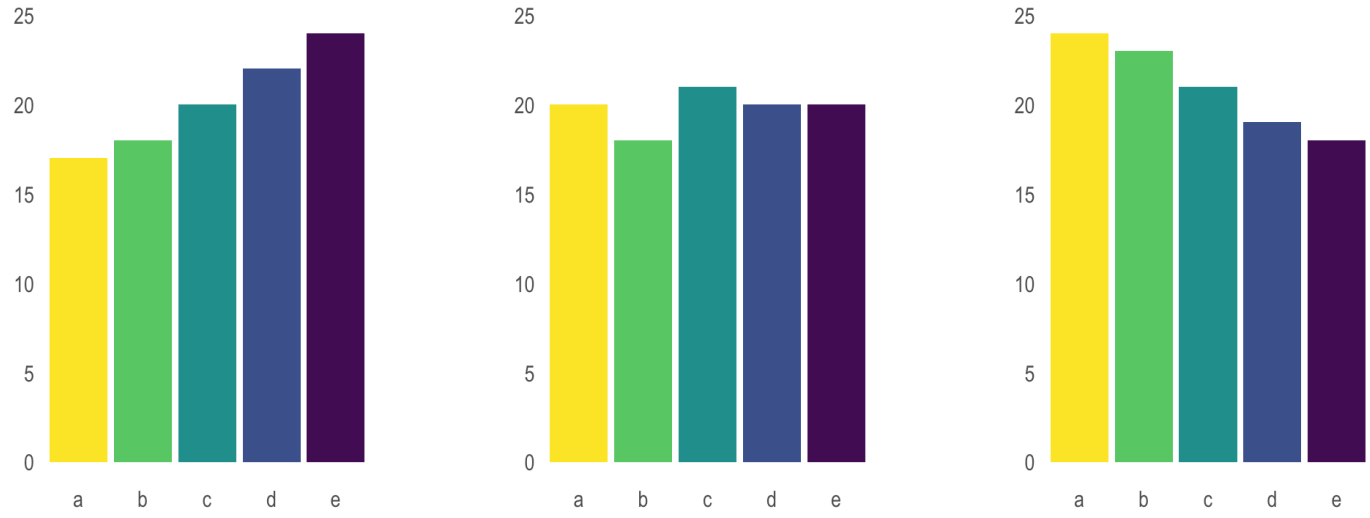
O segundo tipo de visualização categórica são os gráficos de setores, ou gráficos de pizza (ou *pie charts*, em inglês). Esta visualização é uma forma muito frequente de apresentar os dados. A diferença dos gráficos de pizza para os gráficos de barras é que eles apresentam informações dispostas, não lado a lado, mas dentro de um círculo. O círculo inteiro compreende todas as observações possíveis, enquanto as suas ramificações (ou pedaços de pizza) representam a parcela do todo que diz respeito a uma determinada categoria.

Esse gráfico, apesar de ser amplamente utilizado, ele traz um problema de visualização. Por mais que o gráfico funcione bem para representar duas ou três categorias, com mais categorias do que isso ele passa a não ser muito funcional.

Aqui vamos replicar o exemplo elaborado pelo Data to Viz em The issue with pie chart. O exemplo que eles dão parte da seguinte pergunta: Tente descobrir nos três grupos abaixo qual é o grupo com mais observaçẽs.




A resposta deveria ser, para a primeira pergunta, que no primeiro gráfico, o maior grupo é o E; no segundo, o C; e no terceiro, o A. Acontece que, por causa da forma como os dados estão dispostos, é difícil chegar a essa conclusão, enquanto, se repetíssemos a pergunta para os mesmos dados dispostos em barras, essa pergunta seria trivial.




Então, a questão com os gráficos de pizza é que eles possuem um iminente problema de comunicação e interpretação. A assimilação desses gráficos não é intuitiva, muitas vezes, sequer possível a olho nu. Dessa forma, deve-se preferir, na maioria das vezes, a utilização de gráficos de barras no lugar de gráficos de pizza. Recomenda-se que se utilize gráficos de pizza para representar 2 ou, no máximo, 3 categorias. Mais do que isso, a visualização fica prejudicada.

#### 4.3.1.2 Gráficos bivariados (com explicativa categórica)

##### 4.3.1.2.1 Gráfico de barras

O mesmo gráfico de barras que pode representar uma única variável, pode ser usado para representar duas variáveis categóricas. Neste caso, não mudamos os eixos. O que fazemos, no lugar, é "quebrar" cada uma das categorias do eixo categórico em algum subgrupo. Por exemplo, na Figura 4.5, pegamos as categorias referentes a "decisões de segunda instância" e quebramos pelo tipo de litígio. Há 4 tipos de litígio possíveis: uma pessoa física (PF) no polo ativo contra alguma não-pessoa física (nPF), tal como uma empresa, um espólio ou o Poder Público (PF-nPF); uma pessoa física contra outra pessoa física (PF-PF); uma pessoa não física no polo ativo, contra uma pessoa física (nPF-PF); ou uma disputa entre duas pessoas não físicas (nPF-nPF).

Essa "quebra" pode ser feita de diversas maneiras. A escolha entre cada uma delas dependerá dos propósitos do gráfico, ou até mesmo do estilo do próprio pesquisador. A Figura 4.6 traz outra forma de realizar essa quebra.

O que precisamos compreender é a complexidade que essa segunda variável adiciona à interpretação do gráfico. No gráfico de barras univariado, podíamos comparar apenas o tamanho de cada um dos desfechos da sentença. Na Figura 4.7, damos um exemplo de um tipo de comparação possível a partir do gráfico univariado, em que comparamos os os grupos "Reformou" com "Não reformou".

Com os gráficos de barras bivariados, outras comparações são possíveis. No caso, podemos realizar 2 novas comparações. A primeira comparação possível é aquela em que fixamos a variável de interesse para compararmos a variável explicativa. No nosso caso, isso significa comparar, dentro de um único desfecho da sentença, alguns tipos de litígio. Vemos um exemplo desta comparação na Figura 4.8.

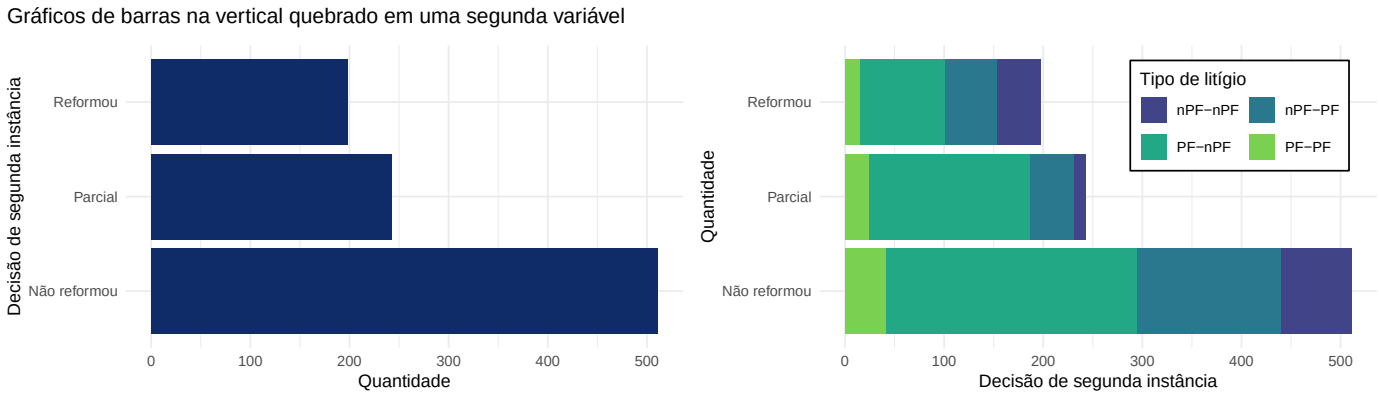



Figura 4.5: Gráficos de barras com duas variáveis, empilhado.

A segunda nova comparação possível é aquela em que fixamos a variável explicativa, ou seja, a variável de quebra, e a analisamos através de todas as categorias de interesse. Vemos um exemplo disto na Figura 4.9, em que olhamos para o tipo de litígio "PF-nPF" para todos os três tipos de desfechos possíveis das decisões de segunda instância.

Notemos, então, que adicionar uma nova variável ao gráfico de barras, para torná-lo bivariado, aumenta a sua complexidade de análise. Isso pode ser bom, pois nos permite visualizar novas relações, mas também pode ser ruim, na medida em que dificulta a interpretação do gráfico.

#### 4.3.1.3 Gráficos bivariados (com explicativa numérica)

Nos gráficos de barras bivariados que mostramos acima, a variável explicativa era categórica (no caso, "Tipo de litígio"). A partir dessa variável explicativa categórica, nós pudemos "quebrar" as barras em categorias menores. Vimos diversas formas de realizar essa quebra, bem como as novas interpretações que isso permitia realizar.

Poderíamos nos perguntar, então, o que aconteceria se, no lugar de "Tipo de litígio", colocássemos alguma variável numérica contínua, tal como "valor da ação"? Como o gráfico ficaria?

Temos de ter cautela com essa pergunta, pois, por mais intuitivo que seja se fazer uma pergunta dessas, esse tipo de visualização só faz sentido em um contexto bem específico: o de regressão. Veremos regressões apenas mais para frente do livro. Quando chegarmos lá, poderemos completar essa explicação.

### 4.3.2 Visualizações de variáveis quantitativas

#### 4.3.2.1 Gráficos univariados

No caso das variáveis quantitativas, também temos formas de representá-las de forma univariada. São duas as visualizações mais frequentes: o histograma e o boxplot.

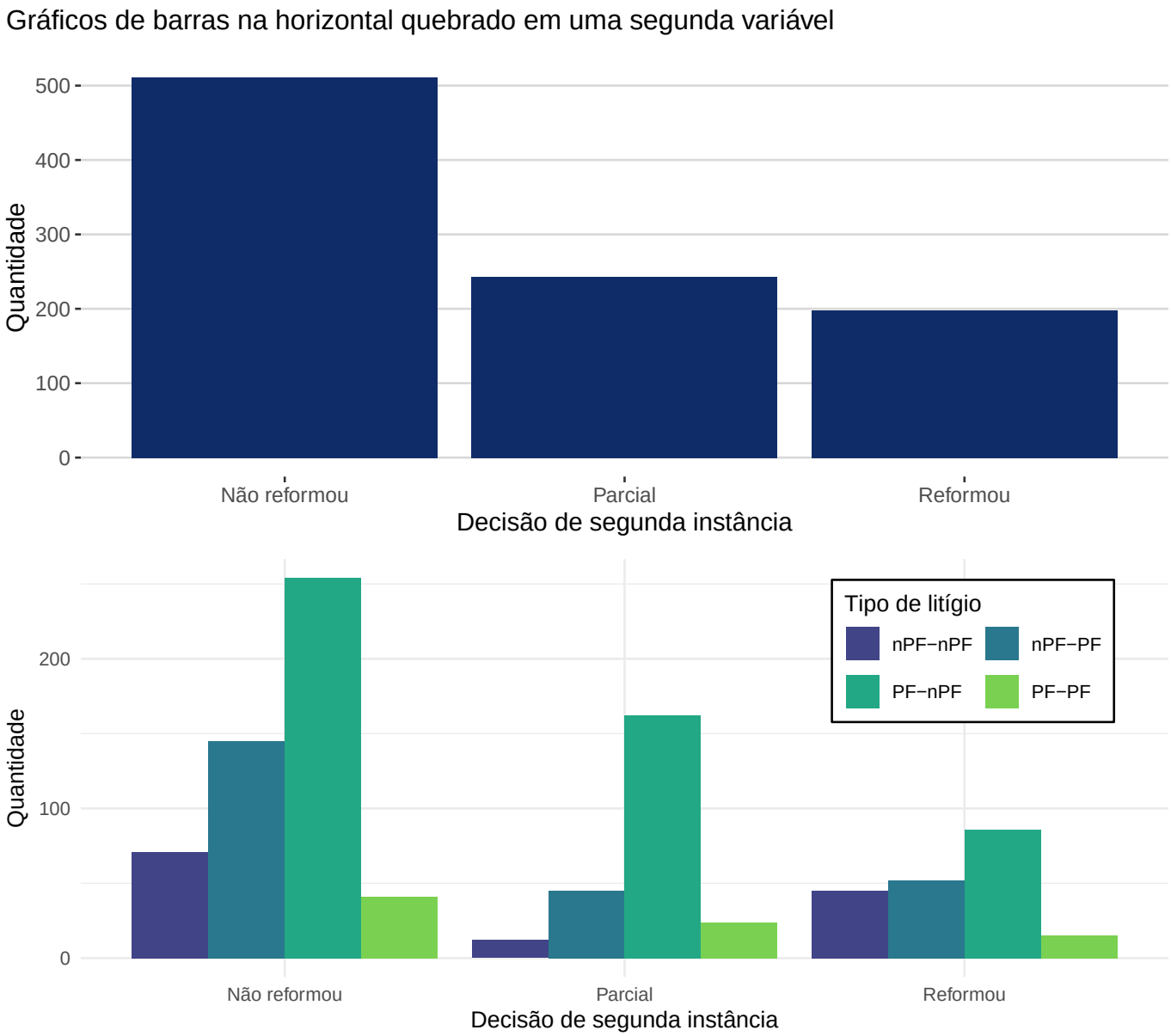



Figura 4.6: Gráficos de barras com duas variáveis, lado a lado.

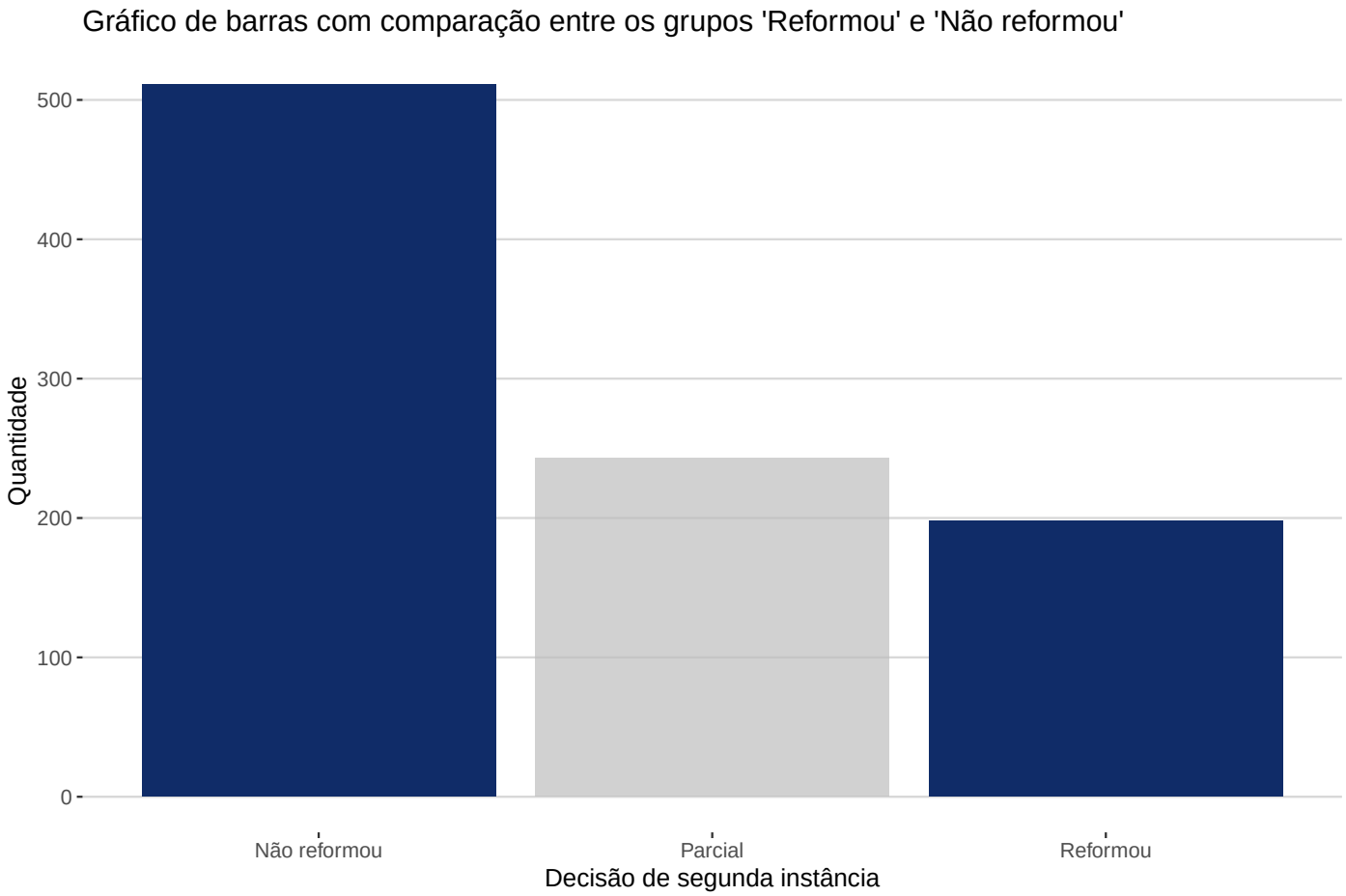



Figura 4.7: Gráfico de barras com comparação entre os grupos ‘Reformou’ e ‘Não reformou’

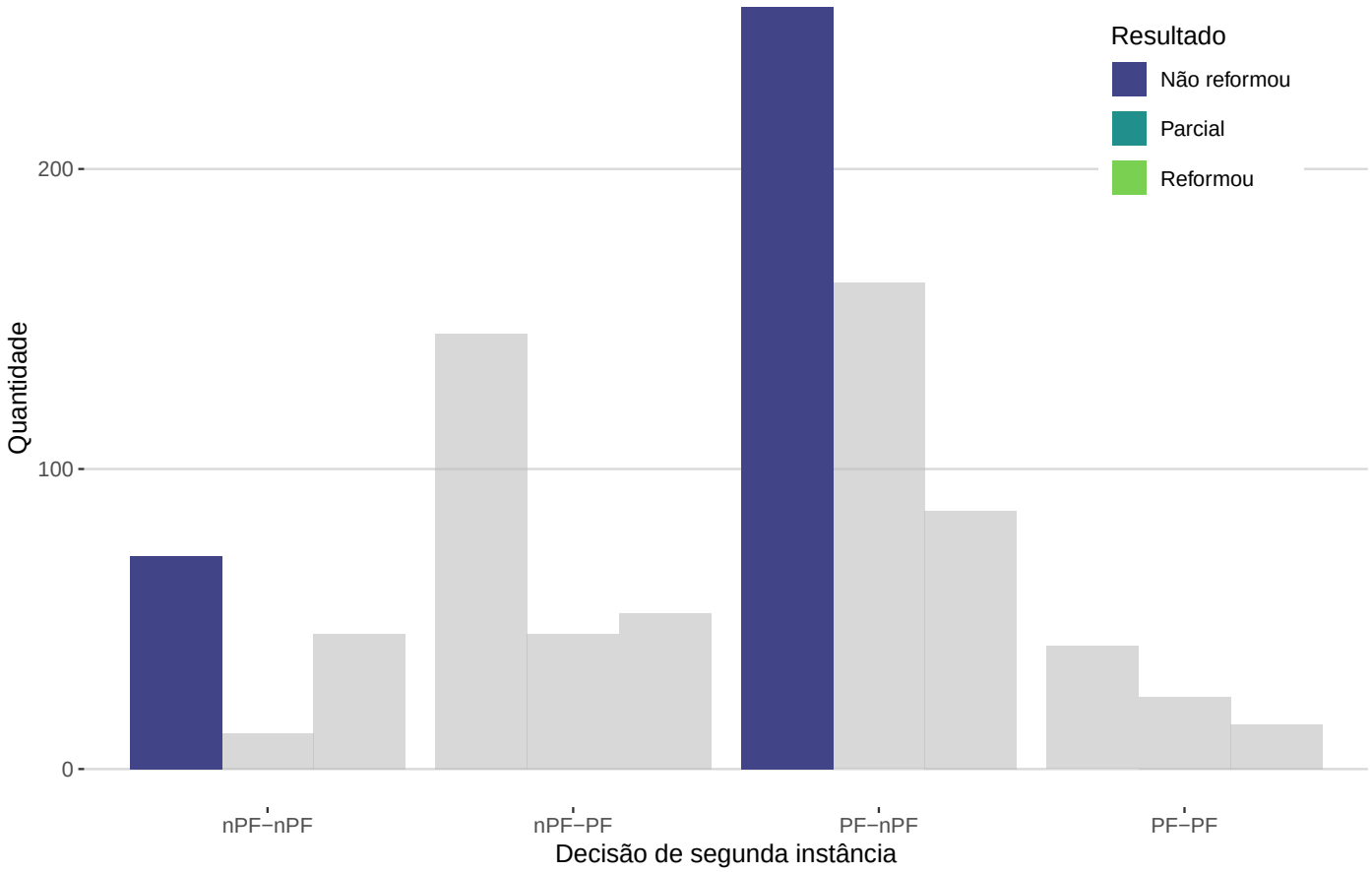



Figura 4.8: Contagens de não reforma e configuração das partes.

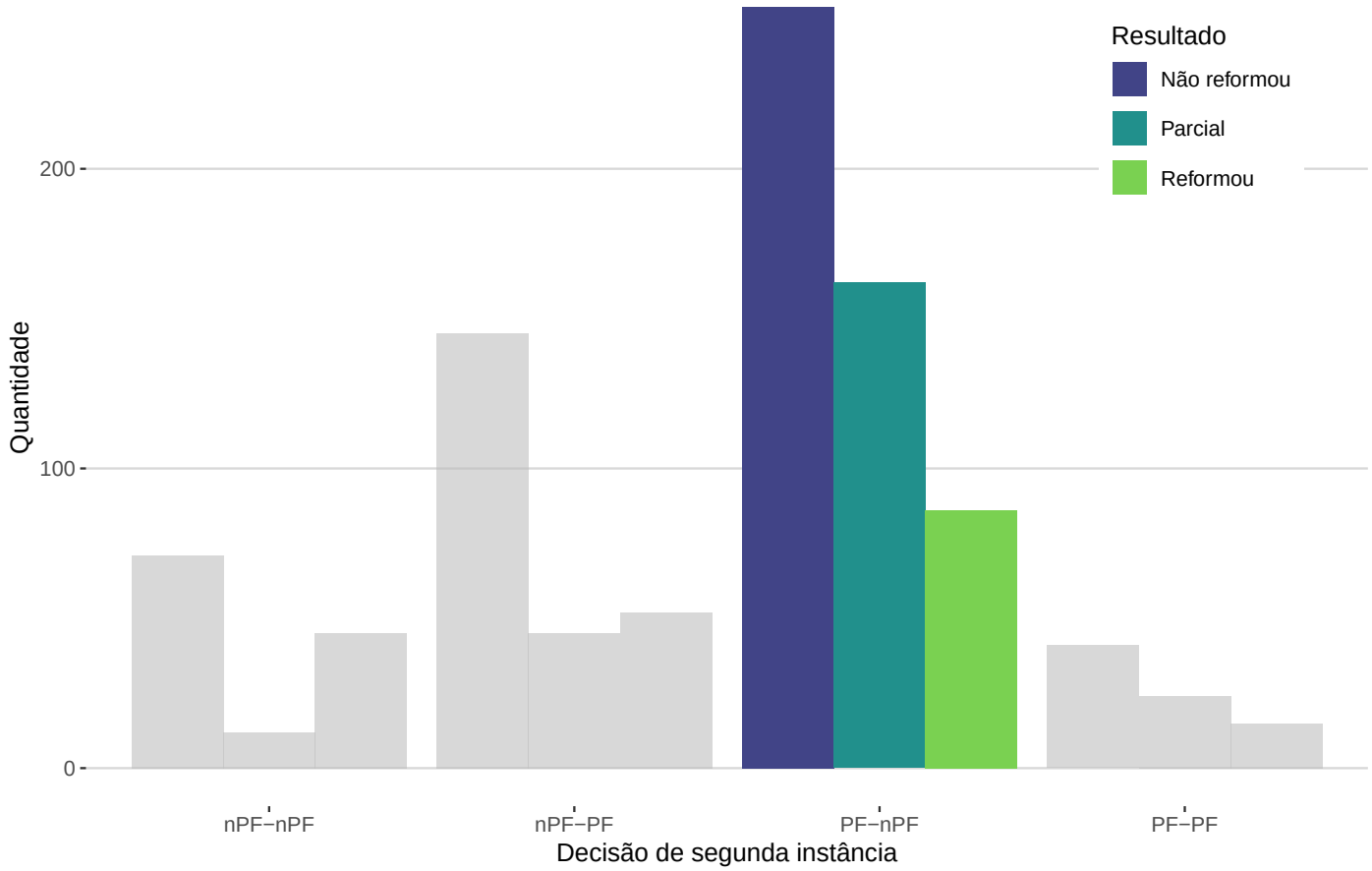



Figura 4.9: Contagens de reforma/não reforma e configuração das partes.

#### 4.3.2.1.1 Histograma

O histograma assemelha-se muito ao gráfico de barras. Eles não devem, entretanto, ser confundidos. O histograma representa variáveis quantitativas contínuas, enquanto o gráfico de barras representa categorias. A diferença essencial é que, enquanto no gráfico de barras nos interesse o tamanho de cada barra, no histograma nos interessa mais a distribuição geral dos dados ao longo do eixo. A Figura 4.10 traz um exemplo de um histograma com o valor de tempo de cada processo em dias.

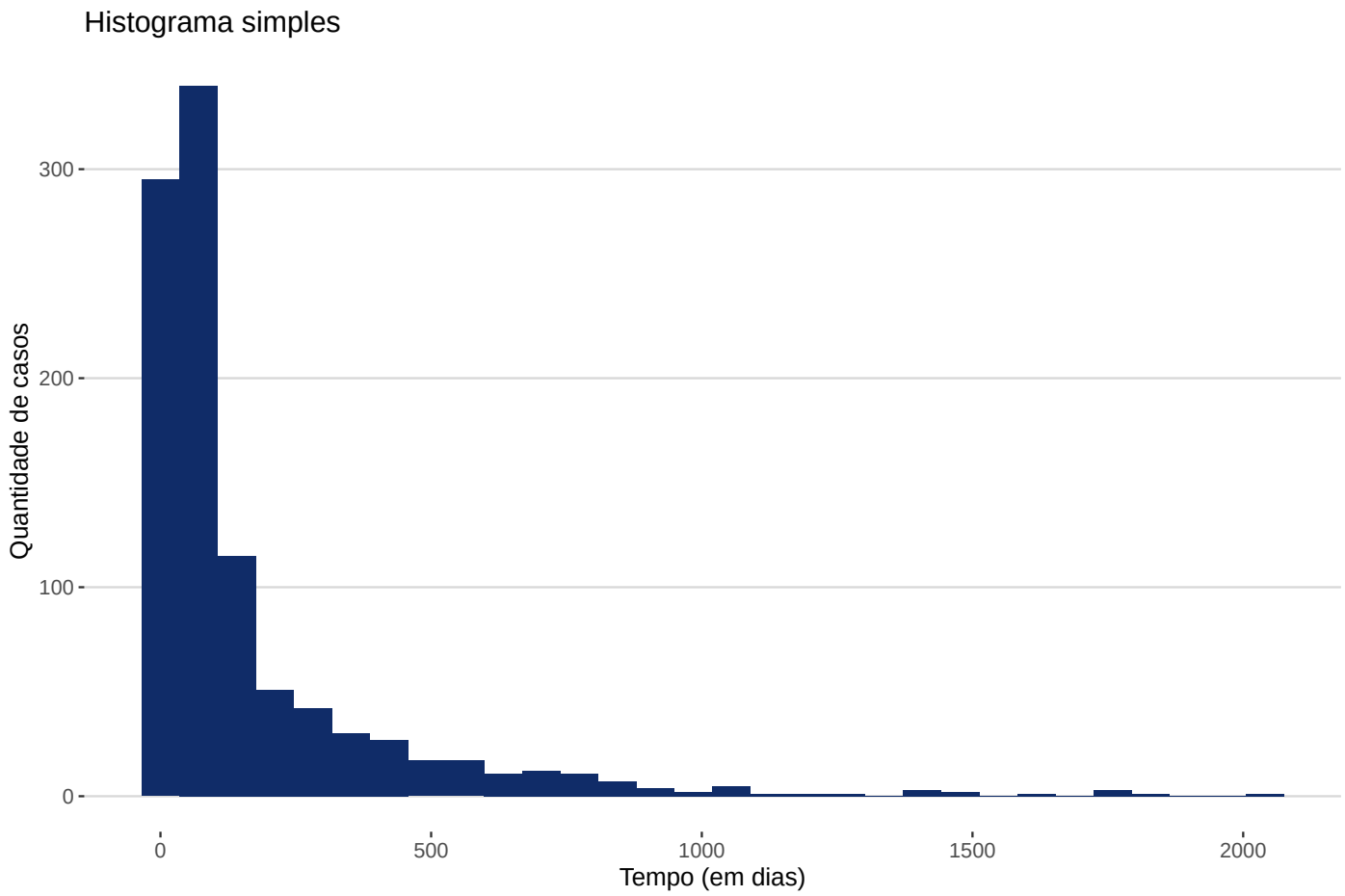



Figura 4.10: Histograma simples

Trabalhando em cima deste exemplo, vamos compreender (a) como se formam as barras do histograma; (b) como interpretar este gráfico; (c) propriedades do histograma; (d) problemas de visualização e transformação dos dados.

A começar pela formação das barras do histograma, precisamos voltar um pouco na natureza das variáveis quantitativas. Existem dois tipos de variáveis quantitativas: as discretas e as contínuas. As discretas representam apenas a contagem de certo número, por exemplo, quando se deseja saber a quantidade de juízes que determinado tribunal possui. Essa quantidade, que expressa apenas uma contagem, será uma variável discreta. As características dessa variável são que ela não pode assumir nem valores negativos, nem valores fracionados, mas apenas valores inteiros. As variáveis contínuas, por outro lado, seriam variáveis que expressam números reais, podendo por natureza envolver números negativos ou

fracionados. Acontecem alguns casos em que os números negativos não fazem sentido no mundo real, por exemplo, quando temos uma variável contínua sobre tempos. Por exemplo, se tivermos uma variável sobre o tempo que demora para um processo morrer, não faz sentido que essa variável indique "-100 dias".

Essa recapitulação dos tipos de variáveis quantitativas é importante porque os histogramas representam apenas variáveis quantitativas *contínuas*, e isso traz um problema para a representação gráfica. O problema é, se eu for fazer uma barra para cada valor possível, tratando eles como categorias próprias, eu teria, teoricamente, infinitas barras. Por exemplo, eu poderia ter, em uma variável de valor, uma barra para todos os processos cujo valor da ação é de R$ 1,00; e depois outra barra para os valores de R$ 1,10. Mas entre uma barra eu poderia ter R$ 1,05; ou ainda 1,025, etc.

A fim de eliminar esse problemas das barras infinitas, cada barra do histograma acaba representando um intervalo, de modo que nenhuma barra represente um único valor, mas sim um intervalo de valores. Assim, no lugar de uma barra para o valor de R$ 1,00, temos uma barra para os valores de R$ 1,00 a R$ 2,00, por exemplo. O tamanho desse intervalo é variável, não é fixo; ele não está predeterminado. O que precisamos saber deste intervalo é que ele é fechado no início e aberto no final, ou seja, se tivermos o intervalo de R$ 1,00 a R$ 2,00 e outro intervalo de R$ 2,00 a R$ 3,00, isso significa que um processo cujo valor da causa seja R$ 2,00, ele estará dentro do segundo intervalo e não do primeiro, mas o valor da causa for de R$ 1,99, ele estará dentro do primeiro intervalo mesmo.

A representação dos intervalos segue uma notação específica: para os intervalos fechados, representamos um colchete voltado para o número; para os intervalos abertos, um colchete "de costas" para o número. A seguir, damos alguns exemplos para compreender melhor essa notação.

- [1,10] significa que o intervalo é fechado no 1 (porque usamos um colchete) e fechado no 10 (porque usamos um colchete);
- [1, 10) significa que o intervalo é fechado no 1 (porque usamos um colchete) e aberto no 10 (porque usamos um parêntese);
- (1, 10] significa que o intervalo é aberto no 1 (porque usamos um parêntese) e fechado no 10 (porque usamos um colchete);
- (1, 10[ significa que o intervalo é aberto no 1 (porque usamos um parêntese) e aberto no 10 (porque usamos um parêntese).

A partir da criação dos intervalos, cada uma das observações da base será colocada dentro de uma das categorias especificadas. Assim, cada um dos intervalos terá uma contagem (igual ao caso do gráfico de barras). Todas as categorias devem ter exatamente o mesmo tamanho. Assim, se foi criada uma categoria que vai do 1 ao 10, então a próxima categoria *tem que ser* do 10 ao 19, pois elas devem ter exatamente o mesmo tamanho.

Mexer no tamanho das categorias pode gerar histogramas diferentes. Veja o mesmo histograma do exemplo acima com três variações do tamanho dos intervalos de cada barra na Figura 4.11.

Uma vez que descobrimos como que as barras do histograma são criadas, podemos passar para o tópico seguinte: como interpretar? No caso do histograma, não nos interessa muito saber a contagem exata de cada "categoria" (até porque as barras não representam categorias, mas apenas intervalos arbitrários). A contagem específica de cada categoria era uma informação importante quando olhávamos para o gráfico de barras, mas no histograma, não é essa informação por que estamos buscando. O essencial do histograma é ver a *distribuição* dos dados como um todo, e não cada barra.

Voltando ao nosso exemplo, conseguimos ver que existe uma barra cuja contagem de casos está acima dos 300 casos. Não nos importa saber a contagem exata desta barra. Ao que devemos nos atentar é que a distribuição do tempo dos processos se dá de forma concentrada na esquerda. Esse tipo de distribuição chamamos de distribuição assimétrica para

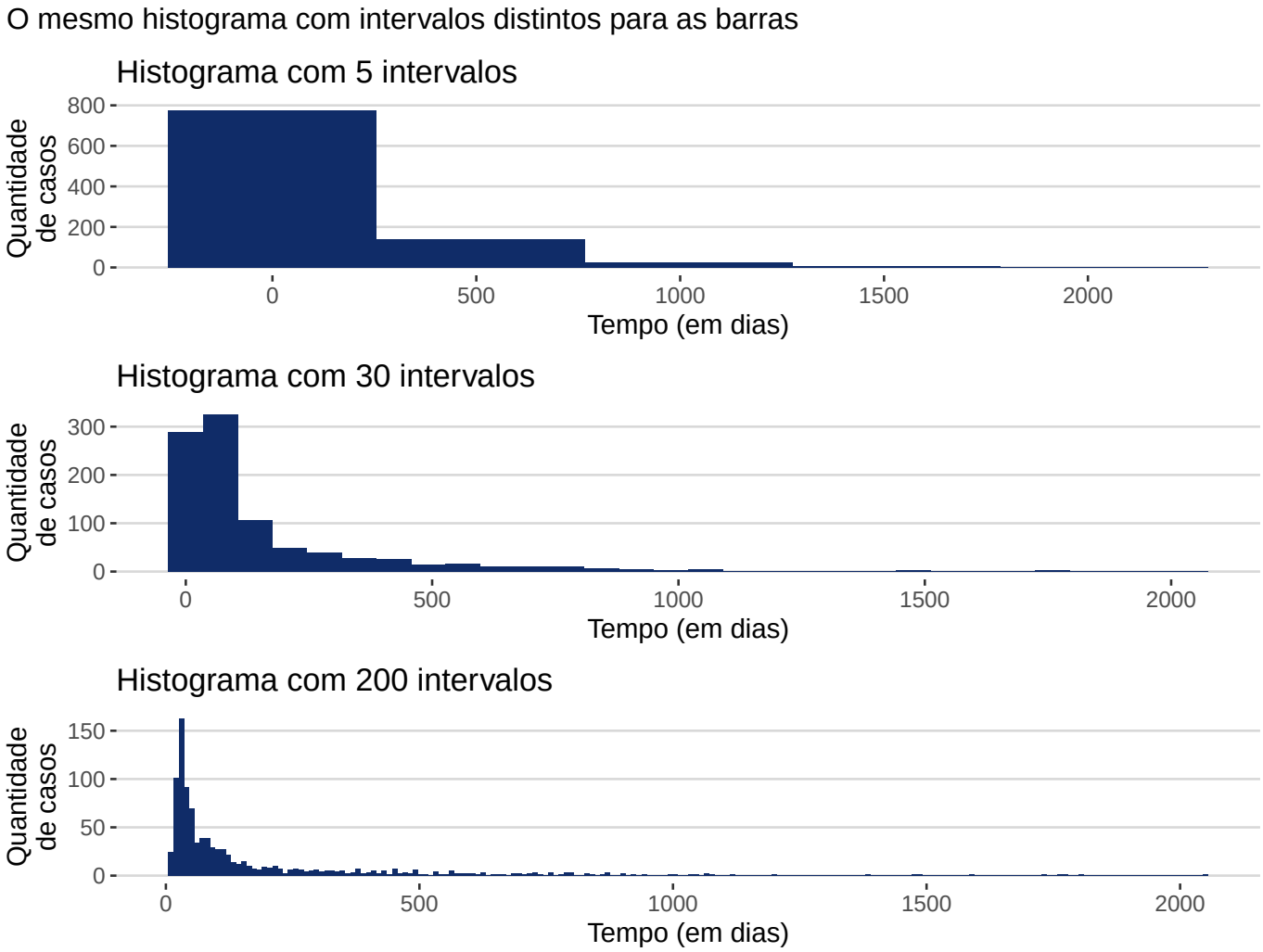



Figura 4.11: O mesmo histograma com intervalos distintos para as barras

a direita. A indicação da assimetria diz respeito à localização da "cauda" do gráfico. No nosso caso, temos uma série de valores concentrados no início, em tempos mais baixos, e alguns poucos casos esparramados no fim, em tempos mais longos. Então, quando dizemos que a distribuição é assimétrica "para a direita" estamos nos referindo à assimetria que estes casos da cauda do gráfico criaram.

Como estamos falando, o mais importante dos histogramas é verificar a distribuição dos dados, e não a contagem específica de cada barra. O que precisamos falar ainda é: quais são as distribuições possíveis? Existem muitas distribuições, mas destacamos 3 principais: a distribuição simétrica, a distribuição assimétrica para a direita e a distribuição assimétrica para a esquerda. Essas distribuições estão resumidas na Figura 4.12.

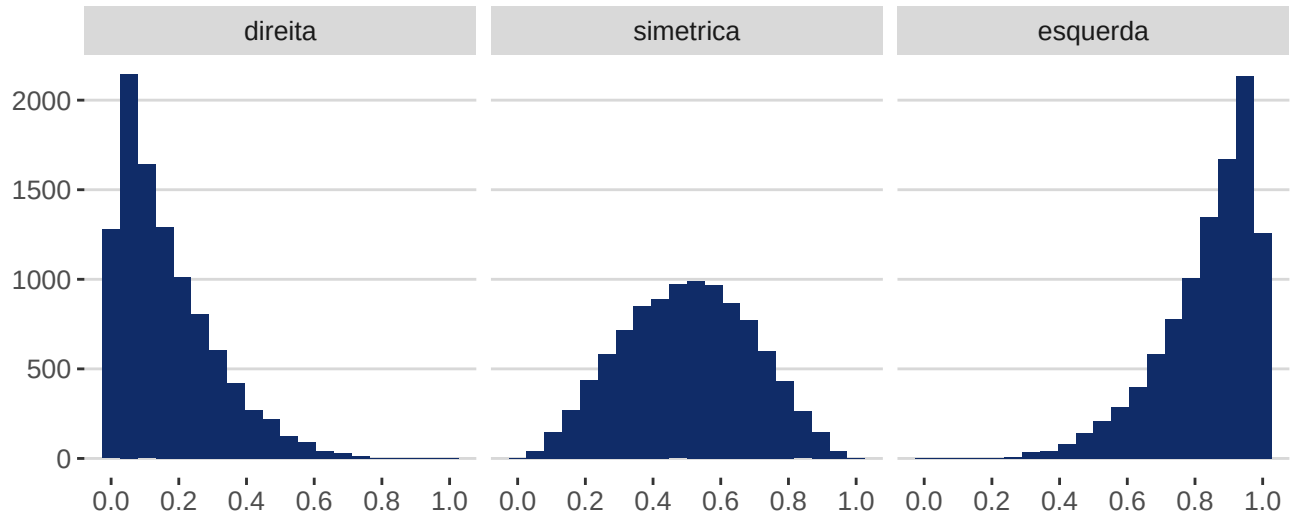



Figura 4.12: Distribuições

A propriedade importante do histograma diz respeito à posição relativa entre média e mediana. Queremos sempre determinar onde os dados estão concentrados. Para tanto, as medidas de resumo que indicam tendência central são excelentes. Lembrando, temos três medidas: a média, a mediana e a moda. A média e a mediana se diferenciam porque a mediana é considerada robusta (isto é, ela não era influenciada pelos valores extremos), enquanto a média é considerada uma medida não robusta. Assim, quando estamos diante de distribuições assimétricas – sejam elas para a esquerda ou para a direita –, a média será afetada pelos valores que estão na cauda, enquanto a mediana não. Essa diferença entre as duas medidas cria um importante atributo para interpretarmos o histograma. O que acontece é que, em distribuições simétricas, a média e a medina irão se sobrepor, assumindo o mesmo valor, ou valores muito próximos, enquanto, no caso das distribuições assimétricas, os valores da média e da mediana irão ficar descompassados. Assim, podemos visualizar, a partir do histograma, como que a média e a mediana se comportam, conforme a Figura 4.13.

A partir desses três histogramas, podemos tirar as seguintes propriedades:

- Distribuição assimétrica para esquerda: Média < Mediana
- Distribuição simétrica: Média = Mediana
- Distribuição assimétrica para direita: Média > Mediana

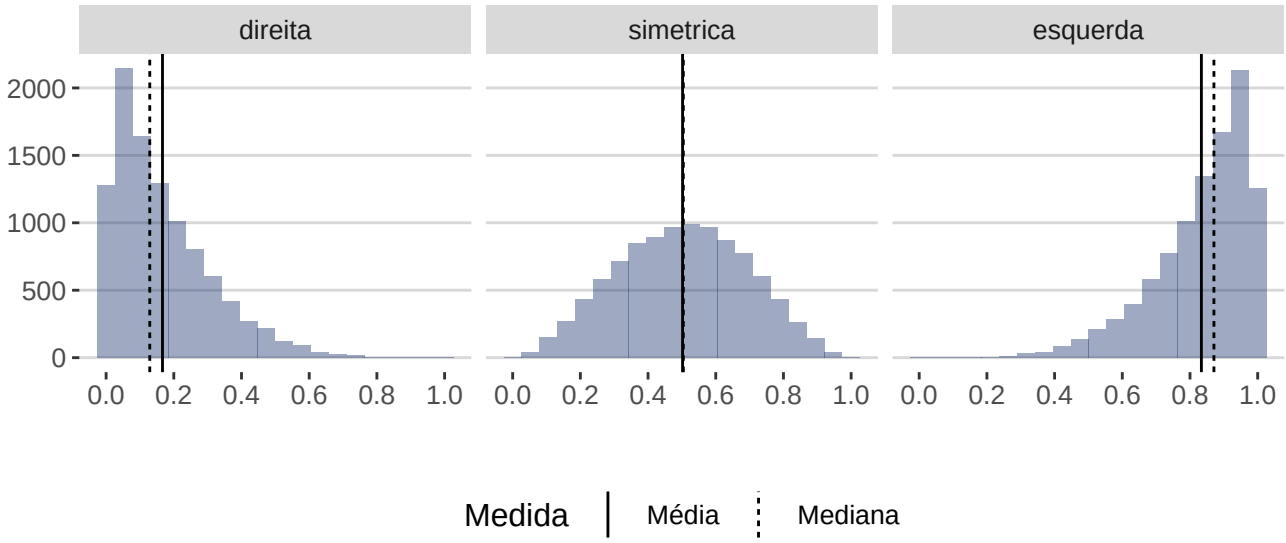



Figura 4.13: Distribuições

Por fim, podemos nos atentar à última questão com os histogramas: problemas de visualização e a transformação dos dados. O que acontece é que muitas variáveis possuem uma distribuição muito assimétrica, de forma que existem muitos valores concentrados em um pico e pouquíssimos valores dispersos em uma grande amplitude. Isso é extremamente frequente em todos os dados envolvendo valores. Até agora, estávamos usando como exemplo apenas gráficos com tempos, justamente porque as visualizações com valores são muito difíceis. Na Figura 4.14, temos um exemplo do que acontece com os valores na base de dados de consumo.

O que acontece é que existem pouquíssimos valores *muito altos* (maiores do que R$ 40 milhões) e uma massa de casos com valores *muito baixos*. Uma assimetria deste tamanho faz parecer que todos os casos possuem um valor de causa igual a zero. Entretanto, isso não é verdade. Se, simplesmente retirarmos da base 10% dos processos, envolvendo os maiores valores, obtemos o histograma da Figura 4.15.

A partir dessa figura conseguimos perceber que são somente os processos acima do quantil(0.9) que estão prejudicando a visualização do histograma. Diante desse problema, o que devemos fazer?

Uma solução já foi dada, que foi justamente excluir da base os processos que estão atravancando a análise e depois analisá-los separadamente. Temos um bom exemplo disto na seguinte reportagem do Nexo Jornal O saldo do cinema brasileiro. Essa solução pode servir para alguns casos, principalmente quando estamos falando da função comunicativa das visualizações. Entretanto, para a função exploratória, essa função nem sempre é adequada, pois, como ela fragmenta a base, ela perde de vista a distribuição do todo.

Uma solução muito frequente para este problema, então, é ao invés de excluir os dados, transformá-los. A transformação mais comum é a transformação em log10. No nosso exemplo, iríamos então transformar os valores exatos de cada causa, para os valores em log de base 10. Assim, aquele histograma da Figura 4.15 ficaria com esta aparência depois da transformação em log, conforme a Figura 4.16.

Por mais que a visualização das barras fique melhorada, a interpretação deste gráfico fica prejudicada, pois agora temos que interpretar o valor em log. O que significa que existe uma concentração de processos com valor em log igual a 4?

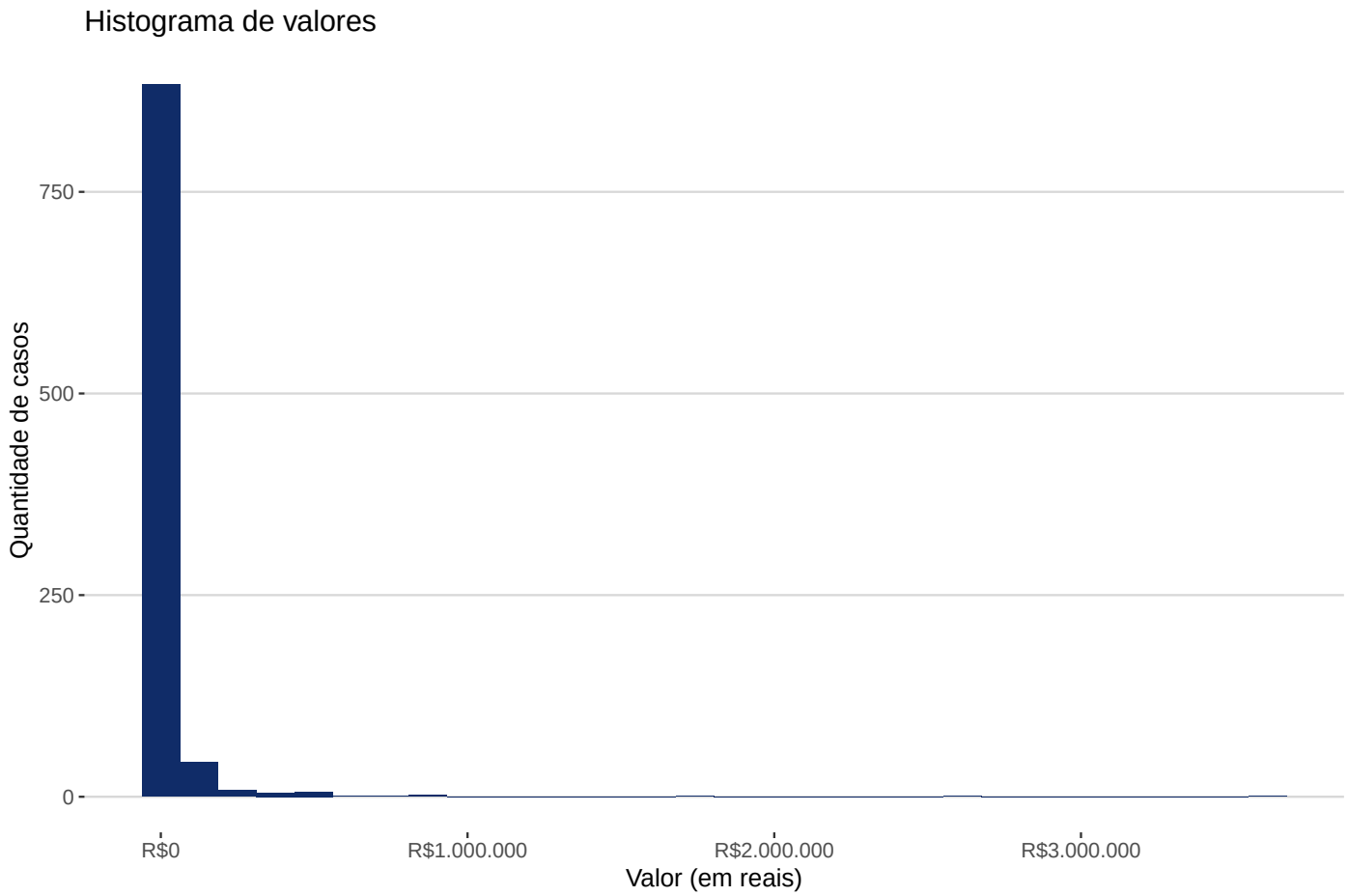



Figura 4.14: Histograma de valores

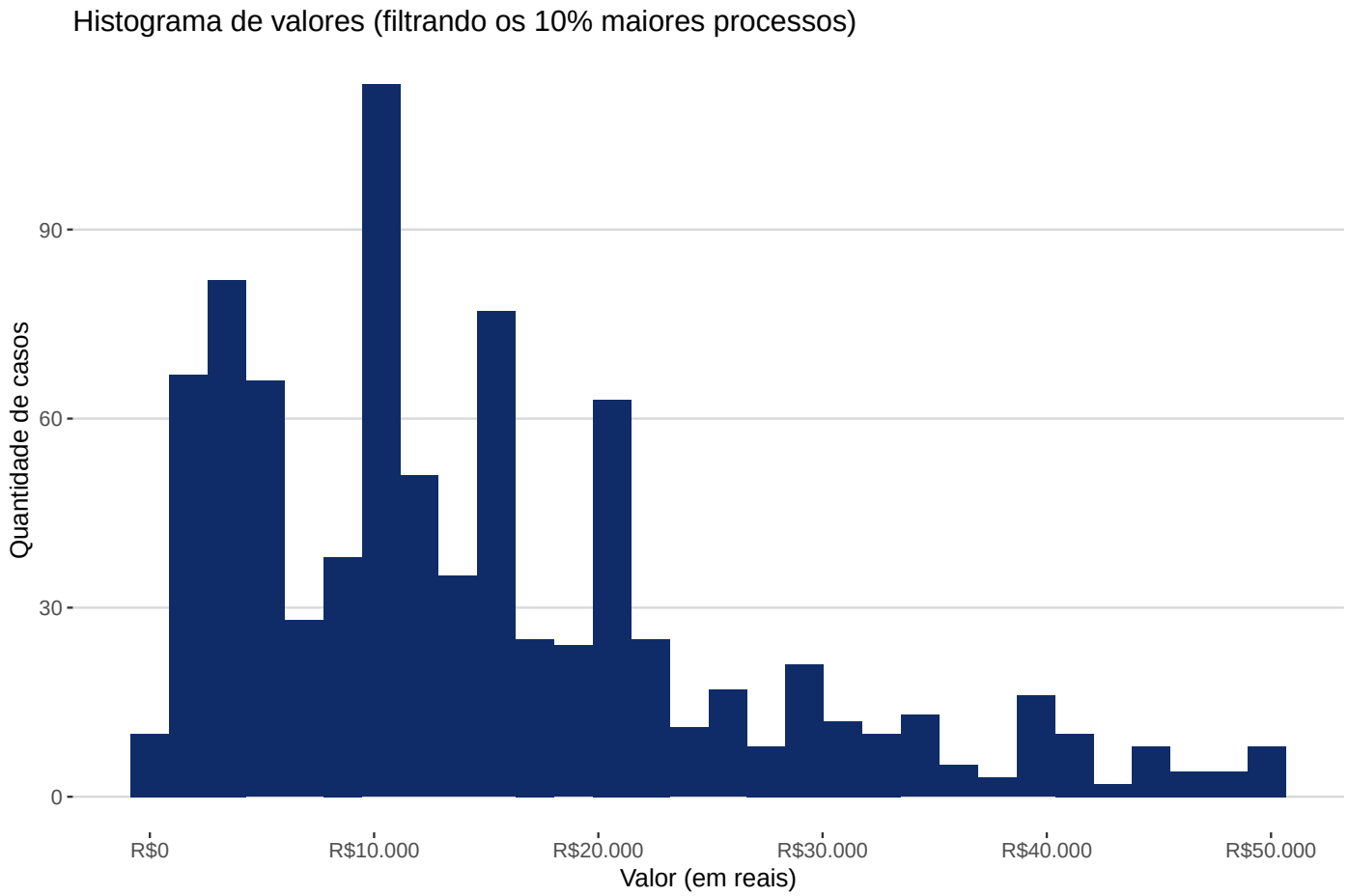



Figura 4.15: Histograma de valores (filtrando os 10% maiores processos)

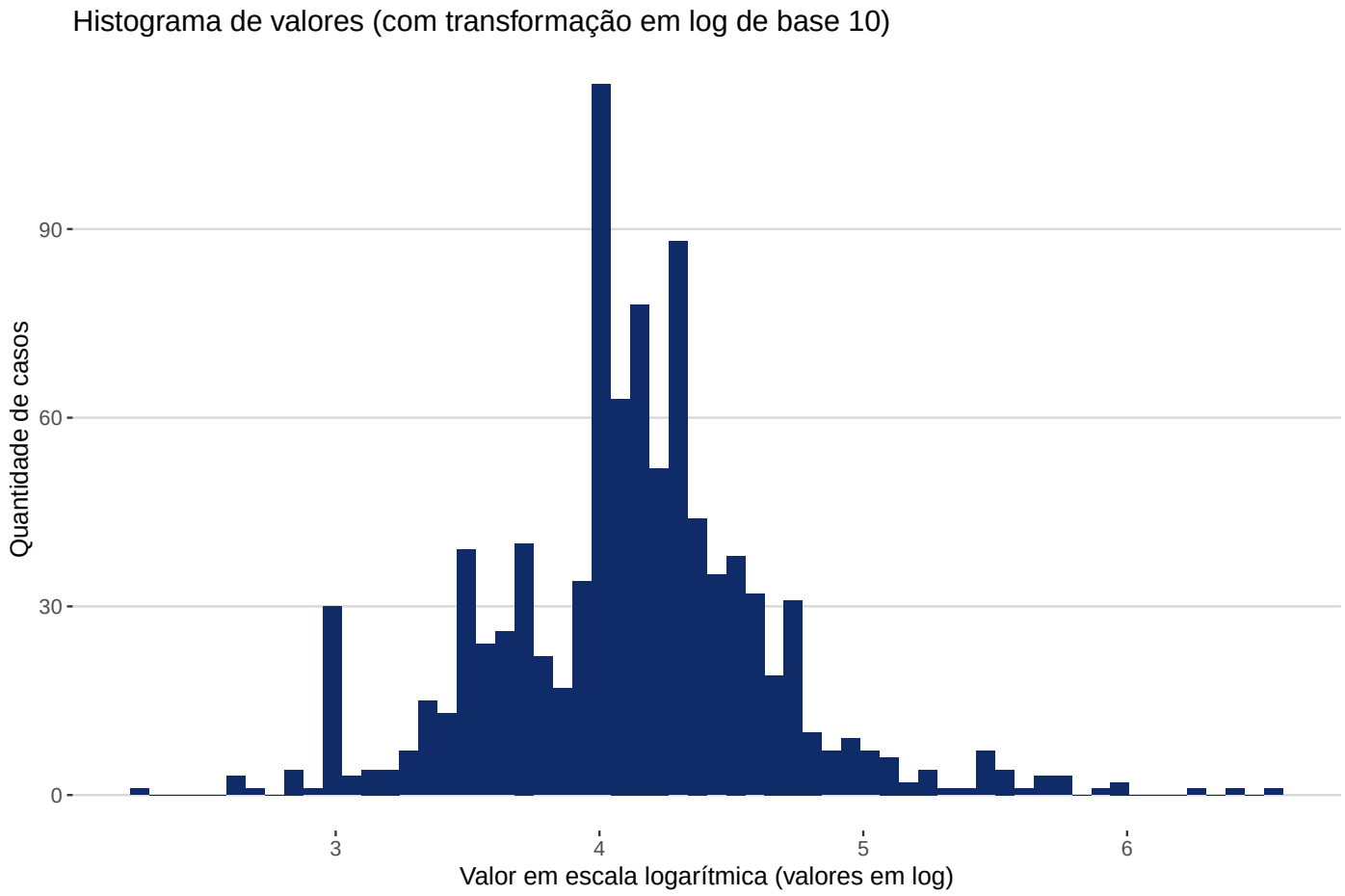



Figura 4.16: Histograma de valores (com transformação em log de base 10)

Isso significa que existe uma concentração de processos em torno de $10^4 = 10.000$, ou seja, em torno de R$ 10.000,00. Por mais que a interpretação fique dificultada, a visualização decerto fica melhor.

A melhor alternativa para se apresentar um histograma em escala logarítmica, talvez seja a de substituir os valores em log pelos valores reais. A única atenção que temos de ter neste caso é para não se deixar confundir com a escala do gráfico (Figura 4.17: a escala continua sendo logarítmica, apesar de as marcações não indicarem isso.

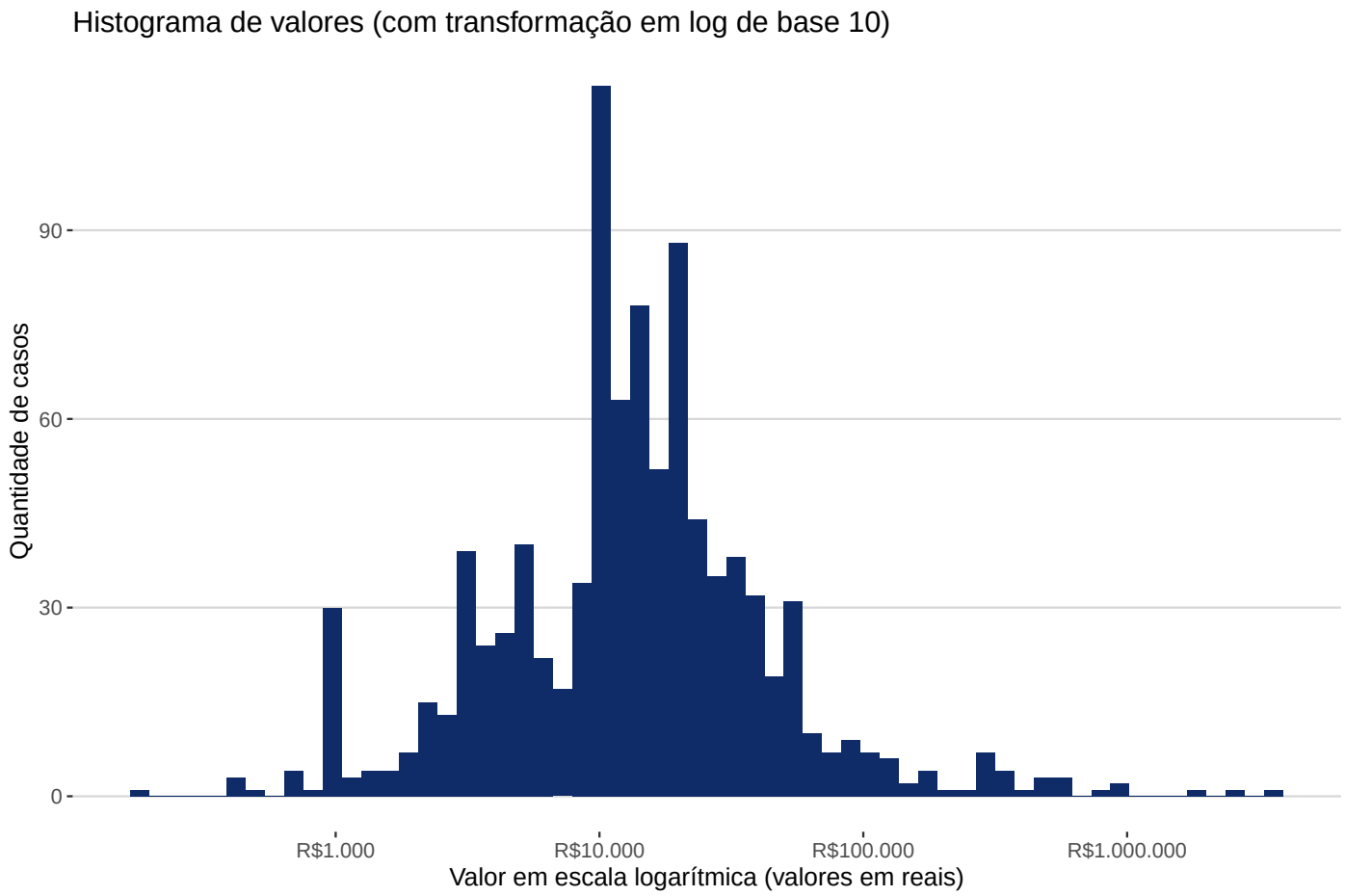



Figura 4.17: Histograma de valores (com transformação em log de base 10)

#### 4.3.2.1.2 Boxplot

A próxima visualização importante para variáveis numéricas é o boxplot. O boxplot é um gráfico que apresenta as estatísticas resumo de uma determinada variável, ao invés de apresentar as unidades amostrais, tal como ocorre no histograma. Para explicar o boxplot, vamos falar de 3 pontos: (a) como se forma o boxplot; (b) como interpretar este gráfico; e (c) problemas relacionados a ele.

Para tratar da formação do boxplot, é importante começar visualizando-o, para que possamos compreender as diferentes partes que o compõem.

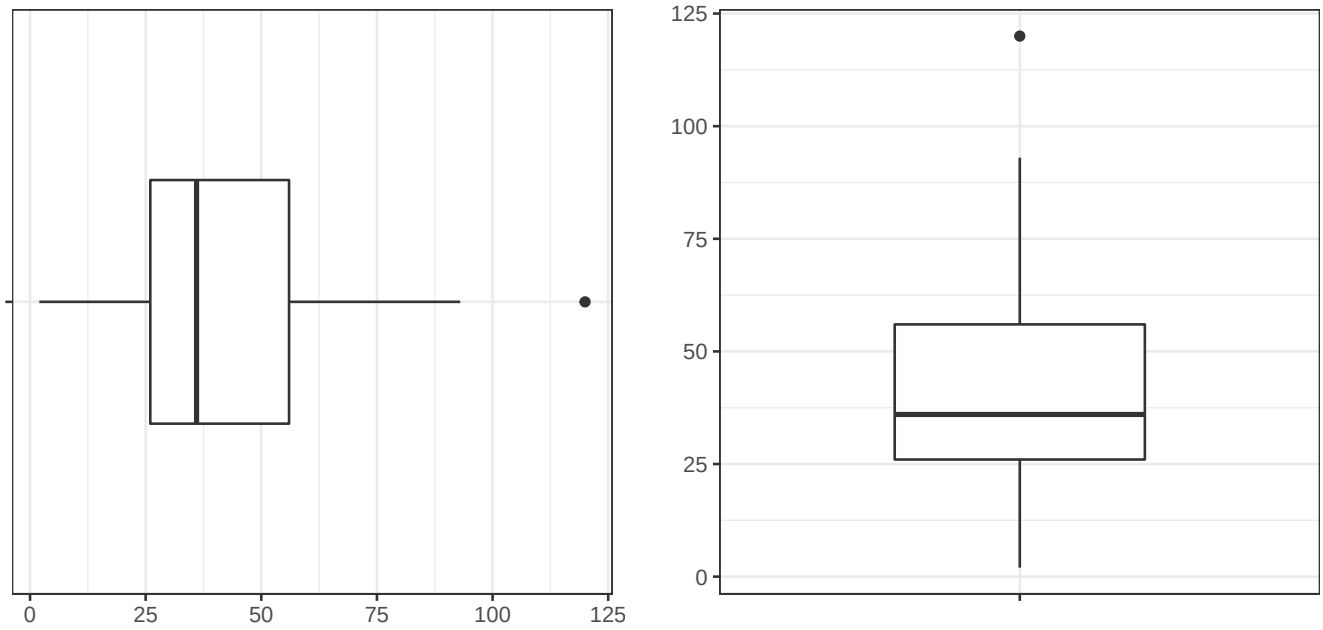



Figura 4.18: Exemplos de boxplot, um em pé e outro deitado

Com a imagem do boxplot em mente, podemos dissecá-lo em três partes, para fins didáticos: o centro, os bigodes e os pontos.

O centro é o retângulo que fica no meio do boxplot. Esse retângulo é definido por três parâmetros. Para entendê-los, precisamos nos lembrar dos quantis empíricos e, mais especificamente, dos quartis. Essa matéria foi vista no capítulo anterior, mas ela é importante neste contexto porque o centro do boxplot é definido justamente pelos três quartis: o quartil superior, a mediana e o quartil inferior. Basicamente, o perímetro externo do centro do boxplot é delimitado pelos quartis superior e inferior e a linha que corta o retângulo boxplot é a mediana. Não necessariamente a mediana está ao centro do retângulo. Mais para frente veremos o que significa a posição da linha mediana dentro do retângulo. O que importa, neste momento, é saber que o centro do boxplot é definido pelos três quartis empíricos, sendo que a mediana representa a linha que corta o retângulo central.

A seguir, a segunda parte do boxplot são os bigodes. Existem dois bigodes, um para cima e outro para baixo. Ambos se formam da mesma maneira: eles se estendem do perímetro externo do retângulo central até 3/2 do IQR. Vale lembrar que, conforme vimos no capítulo anterior, IQR (Interquartile Range) representa a amplitude dos quartis superior e inferior, ou seja, ele é calculado pela diferença do quartil superior (75%) com o quartil inferior (25%). Dessa forma, o IQR diz qual é a amplitude em que se encontram os 50% centrais dos dados. Lembrando desta informação, podemos voltar aos bigodes. Os bigodes são linhas que se estendem desde a área externa do retângulo central até um ponto que representa 3/2 do IQR. O bigode inferior, se inicia do quartil inferior e vai até -3/2 IQR; enquanto o bigode superior se inicia do quartil superior e vai até 3/2 IQR.

Por que usamos o valor de 3/2 do IQR? Isso acontece porque, em distribuições normais, 3/2 do IQR representa 99% de todos os dados. Essa afirmação faz pouco sentido agora, entretanto, mais para frente do livro, daremos sentido a ele,

definindo melhor o que significa uma distribuição normal e como podemos saber que 3/2 do IQR representa 99% dos dados. Essa última afirmação está relacionada a algumas propriedades da distribuição normal.

É preciso saber também, sobre os bigodes, que nem sempre o bigode, de fato, vai até os 3/2 do IQR. Isso se dá porque, no boxplot, o bigode pode ir até os 3/2 do IQR, entretanto, ele só irá até a última observação antes de esse ponto chegar. Isso significa que mesmo se existirem pontos além dos 3/2 do IQR, se o último ponto antes desse limite estiver, por exemplo, a $\frac{1}{2}$ do IQR, então o bigode vai andar somente até esse ponto.

E se existir algum ponto além dos 3/2 do IQR? Então o bigode irá representar as observações **antes** de se passar desse limite; e os valores que passarem deste limite serão representados com pontos. Daí a explicação da terceira parte do boxplot, os pontos. Os pontos, nada mais são do que os dados que estão além dos 3/2 de IQR, seja para cima ou para baixo. Se não houver dados acima desses 3/2 de IQR, então não haverá pontos. Essa é a única parte do gráfico que pode ou não aparecer, todas as demais são partes essenciais do gráfico.

Antes de prosseguirmos a explicação para outras questões relacionadas ao boxplot, vamos montar parte a parte de um boxplot a partir de um conjunto de dados. O conjunto de dados que utilizaremos é a base consumo. Essa base contém 1000 observações, cuja unidade amostral são processos de consumo. Vamos, para os fins dessa explicação, amostrar somente 30 processos.

O primeiro passo para montar o boxplot é determinar os eixos do boxplot. É preciso ter clareza de que o boxplot que estamos vendo agora é univariado, ou seja, ele possui apenas um eixo. Este eixo é o eixo que contém os valores da variável numérica. No nosso caso, será o eixo de valor da causa. Não importa se esse eixo for o x ou o y, pois o boxplot pode ser apresentado tanto na horizontal como na vertical.

O segundo passo é realizar os cálculos das estatísticas de resumo do nosso conjunto de dados. São 5 informações de que precisamos (máximo, mínimo, mediana, quartil superior e inferior), e outras 7 medidas decorrentes dessas 5 (IQR, o limite superior do bigode, o limite inferior do bigode, o ponto superior, o ponto inferior e os potenciais outliers nos limites inferior e superior). A Tabela 4.3 resume essas informações.

Tabela 4.3: Medidas resumo que compõem o boxplot.

| medidas | parte do boxplot | valores |
|---|---|---|
| valor máximo | bigodes | R$ 82.279,86 |
| quartil superior | centro | R$ 22.370,00 |
| mediana | centro | R$ 14.066,00 |
| quartil inferior | centro | R$ 10.003,23 |
| valor mínimo | bigodes | R$ 395,26 |
| IQR | bigodes | R$ 12.366,77 |
| quartil superior + 3/2 IQR | bigodes | R$ 40.920,16 |
| quartil inferior - 3/2 IQR | bigodes | R$ -8.546,94 |
| último ponto superior | bigodes | R$ 40.400,74 |
| último ponto inferior | bigodes | R$ 395,26 |
| quantidade de pontos superiores | pontos | 3 pontos |
| quantidade de pontos inferiores | pontos | 0 pontos |

A começar pelo centro do boxplot, ele está contido dentro dos limites dos quartis superior e inferior e, portanto, ele começa, no ponto R$ 10.003,23 e acaba no ponto R$ 22.370,00. A mediana, que corta o retângulo central, está no ponto

R$ 14.066,00. Percebemos já, com isso, que a mediana não está no centro. O retângulo central está representado na Figura 4.19.

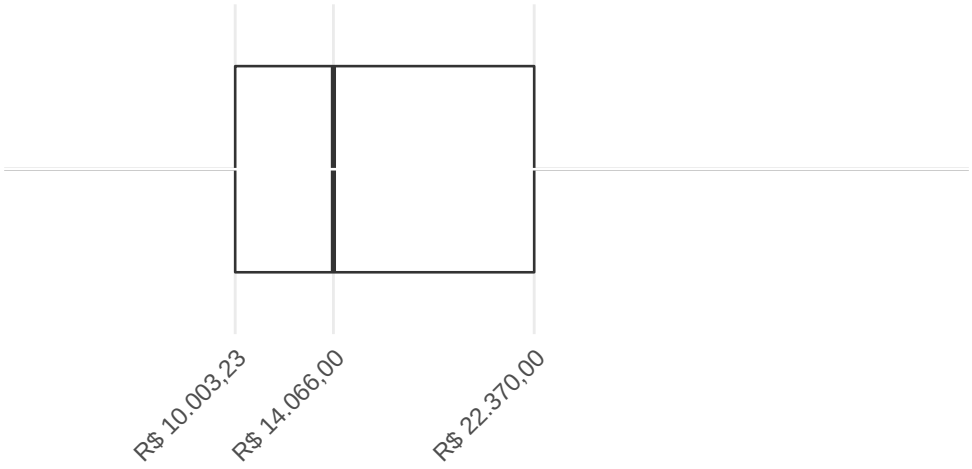



Figura 4.19: Boxplot sem bigodes

A seguir, para construir os bigodes, temos que calcular o IQR. O IQR é simplesmente a diferença do quartil superior com o quartil inferior, dando R$ 12.366,77. Então 3/2 do IQR representa R$ 18.550,15 (3/2 * 12.366,77). Com base nisso, podemos calcular os bigodes inferior e superior.

O bigode superior pode andar 18.550,15 unidades acima do quartil superior, podendo assumir, portanto, no máximo o valor de R$ 40.920,16. Quanto ao bigode inferior, precisamos subtrair 3/2 do IQR do quartil inferior, resultando em R$ -8.546,94. Como não é possível que o valor da causa seja negativo, então esse número, na verdade, equivale simplesmente a R$ 0,00, ou seja, o menor valor que o bigode inferior pode assumir é zero. O zero está contido dentro do limite do bigode inferior, então não haverá nenhum valor além do limite inferior para o bigode e, portanto, não haverá nenhum ponto na região inferior. Os limites dos bigodes estão representados na Figura 4.20.

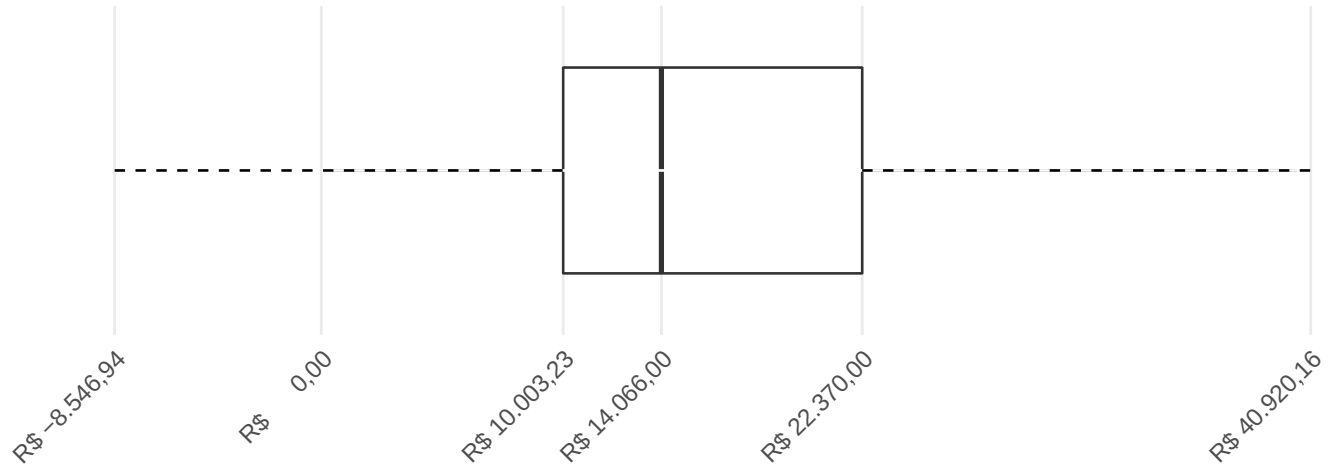



Figura 4.20: Boxplot com bigodes estendidos

Mesmo sabendo os pontos máximo que os bigodes poderão assumir, isso ainda não nos diz onde os bigodes irão parar, uma vez que não necessariamente (e via de regra não é assim) o maior valor dentro da região dos bigodes é de fato o maior valor possível. No nosso caso, a observação de maior valor que está dentro do bigode inferior é o valor mínimo, de R$ 395,26; e a observação de maior valor dentro do limite do bigode superior é o ponto R$ 40.400,74. Os bigodes estão representados na Figura 4.21.

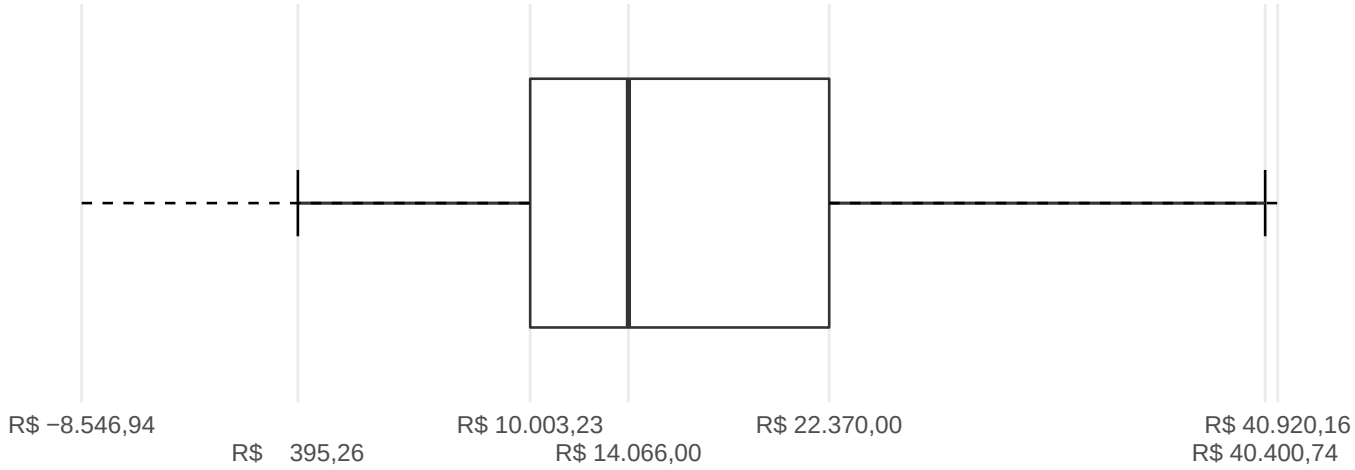



Figura 4.21: Boxplot com bigodes cortados

Por fim, uma vez que os bigodes foram construídos, falta apenas a última parte dos boxplots, que são os pontos. Na parte inferior do gráfico, como vimos, como o limite inferior vai até um número negativo, mas os valores reais podem ir, no máximo, até R$ 0,00, então não há pontos na parte inferior. Na parte superior do gráfico, há três pontos que extrapolam o limite do gráfico. Estes pontos estão representados na Figura 4.22.

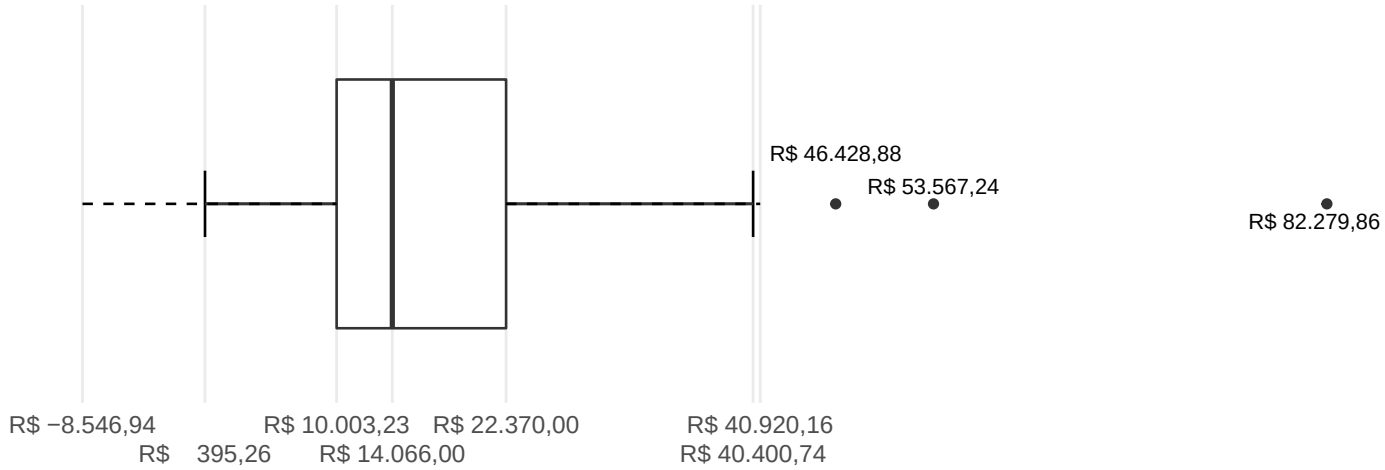



Figura 4.22: Boxplot com bigodes cortados e valores atípicos

Com a explicação de como construir o boxplot, seguida de um exemplo, podemos prosseguir ao segundo ponto importante a respeito dos boxplots: como interpretar os boxplots? Os boxplots, assim como os histogramas, indicam

distribuições. Então, assim como vimos alguns tipos de distribuições nos histogramas, podemos verificar como essas distribuições reverberam na construção dos boxplots. Na Figura 4.23, vemos a relação do boxplot com as diferentes distribuições existentes.

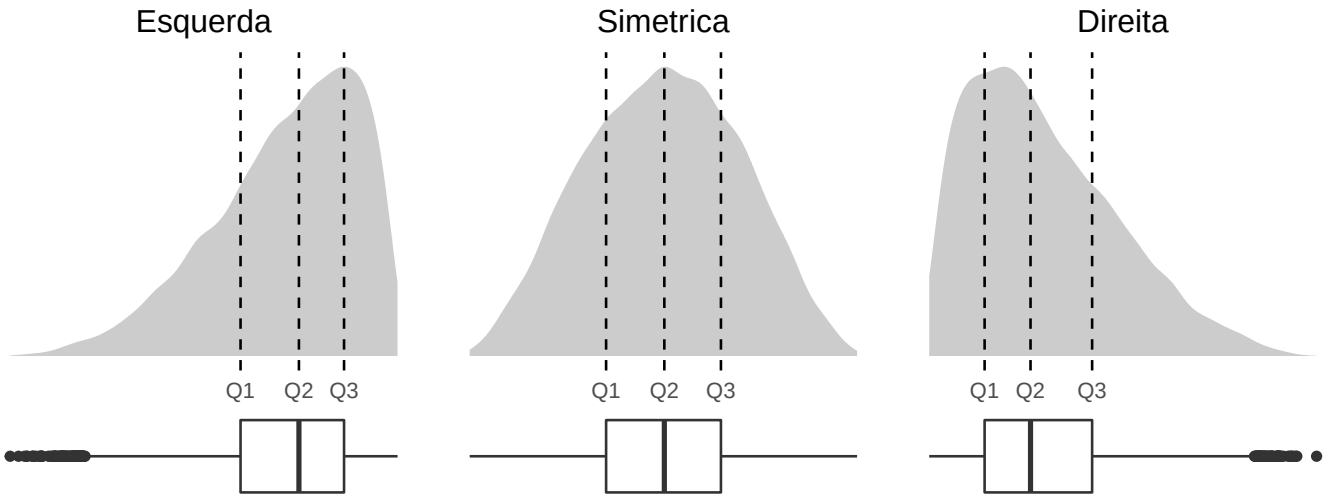



Figura 4.23: Distribuições e boxplots

O que podemos notar é que a posição da linha da mediana *dentro* do retângulo, indica o tipo de distribuição. Se a linha mediana está bem ao centro do retângulo central, então podemos concluir que a distribuição é simétrica; se a linha mediana está deslocada, então temos distribuições assimétricas. Com isso, percebemos a que elemento do retângulo central temos de nos atentar quando analisamos um boxplot.

A representação das distribuições por meio do boxplot é uma informação que já está presente no histograma e, nesse sentido, o boxplot não acrescenta muito à análise. Entretanto, o boxplot adiciona alguns elementos à visualização que não estão presentes no histograma. Há dois elementos importantes para a análise.

Primeiro, o boxplot consegue destacar os dois quartis centrais (o segundo e o terceiro quartis), que, juntos, representam exatamente o meio dos dados. Esse destaque nos permite inferir conclusões mais *robustas* a respeito do comportamento central do nosso conjunto. Se nos lembrarmos da discussão a respeito das medidas no capítulo anterior, uma medida *robusta* é uma medida menos suscetível aos valores extremos. Em muitos conjuntos de dados, os valores extremos distorcem os valores centrais. Então o boxplot consegue representar muito bem os valores centrais, dando destaque a esses valores. A Figura Y deixa em realce a porção dos dados que estão sendo representadas pelo retângulo central do boxplot.

O último ponto que merece destaque da interpretação dos boxplots são os pontos. A interpretação dos pontos é um ponto importante dos boxplots. Esses pontos indicam potenciais *outliers*. Esta é a primeira vez que falamos de outliers neste livro. Outliers são pontos que se diferem drasticamente do resto dos dados. Essa definição é muito genérica, pois ela é uma definição guarda-chuva. Mas ela não é suficiente. O que significa dizer que algum ponto "difere drasticamente" do resto dos dados? Como que eu meço o que são "o resto dos dados" para comparar com os pontos que "diferem drasticamente"?

Essas perguntas são essenciais, porque, na verdade, são justamente elas que dão contornos estatísticos para os outliers. Para cada resposta distinta que damos a essas perguntas, teremos uma definição distinta de outlier e existem muitas

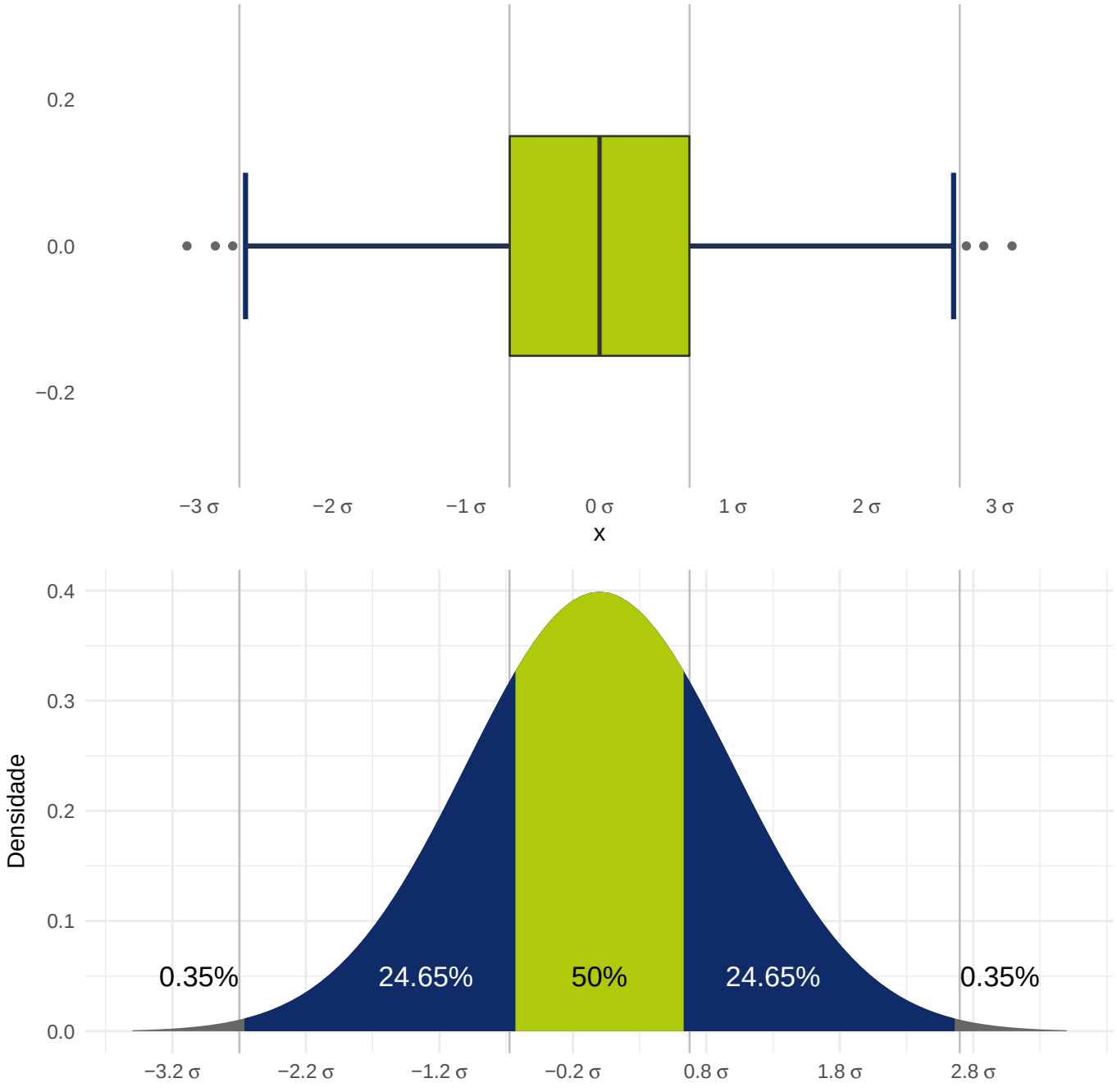



Figura 4.24: Boxplot comparado com distribuição normal.

definições possíveis. A questão que une todos os outliers é que, em todas as suas acepções, os outliers distorcem os valores dos dados e são pontos cuja incidência na população é muito rara.

Há duas dificuldades ao lidar com outliers: a identificação e a solução. Sobre a identificação, o problema central é definir o ponto de corte dos dados: a partir de que valores poderemos dizer que existem outliers? A segunda questão é saber o que fazer depois de identificarmos os outliers? Retiramos eles da base? Olharmos separadamente para os outliers e os pontos centrais? Essas perguntas (de identificação e de solução) irão acompanhar o livro em muitos momentos, pois os outliers geram problemas que se manifestam de diversas maneiras na estatística. Voltaremos várias vezes nisso, portanto.

Neste momento introdutório, em que estamos vendo os outliers pela primeira vez, basta sabermos que eles são pontos discrepantes e que existe uma dificuldade tremenda em se determinar um ponto de corte para bater o martelo e dizer: "qualquer valor que ultrapasse essa fronteira, será considerado um outlier!". No caso do boxplot, por causa da forma como ele é construído, todos os pontos representados no gráfico estão a 3/2 do IQR, contado a partir dos quartis inferior ou superior. Então, se assumirmos que todos os pontos que aparecem no boxplot são outliers, estaremos definido que "diferir drasticamente" significa "distanciar-se 3/2 do IQR contados a partir do quartil superior ou inferior". Esse ponto de corte (quartil + 3/2 do IQR) é apenas uma definição dentre muitas possíveis para os outliers. Mas podemos usar critérios de corte mais brandos ou mais severos. Por isso dizemos que os pontos do boxplot indicam apenas "potenciais outliers".

É importante pontuar aqui que o critério que usaremos para decidir o ponto de corte do outlier não deve ser um critério automático (como o é no caso dos pontos do boxplot). Para cada caso, para cada dado, para cada propósito, teremos um critério distinto para definir outliers. Cada critério resolve problemas distintos, mas gera vieses distintos também. Veremos com mais detalhes as formas de mensurar outliers ao longo do livro.

Dada toda essa discussão a respeito dos outliers, podemos voltar à interpretação dos boxplots. Em primeiro lugar, vimos que havia uma semelhança dos boxplots com os histogramas: ambos representavam distribuições dos dados, mas de formas distintos. Em seguida, passamos a ver duas características específicas do boxplot, que não estavam presentes no histograma: o foco no conjunto central e os outliers. Sobre este último ponto, vimos que os pontos que aparecem no boxplot representam apenas "potenciais outliers", justamente porque eles assumem um critério automático de representação. Em alguns contextos, fará sentido assumir que os outliers são esses pontos mesmos; em outros, entretanto, não será assim.

Explicada a interpretação e a formação dos boxplots, resta somente abordar seus principais problemas. Há dois problemas: distribuições assimétricas e dados escassos.

A começar pelo problema das distribuições assimétricas, este é o mesmo problema que acontecia com o histograma, de que a visualização ficava impossibilitada por causa de distribuições muito assimétricas. Basta nos lembrarmos da Figura 4.14. No caso do boxplot, aqueles mesmos dados ficam representados conforme a Figura 4.25.

Vemos que o retângulo central está tão achatado, próximo do 0, que ele se torna uma linha. Perdemos, com isso, toda a noção das medidas centrais. Entretanto, se olharmos para a tabela de medidas para compôr este boxplot, veremos os seguintes valores, conforme a Tabela 4.4.

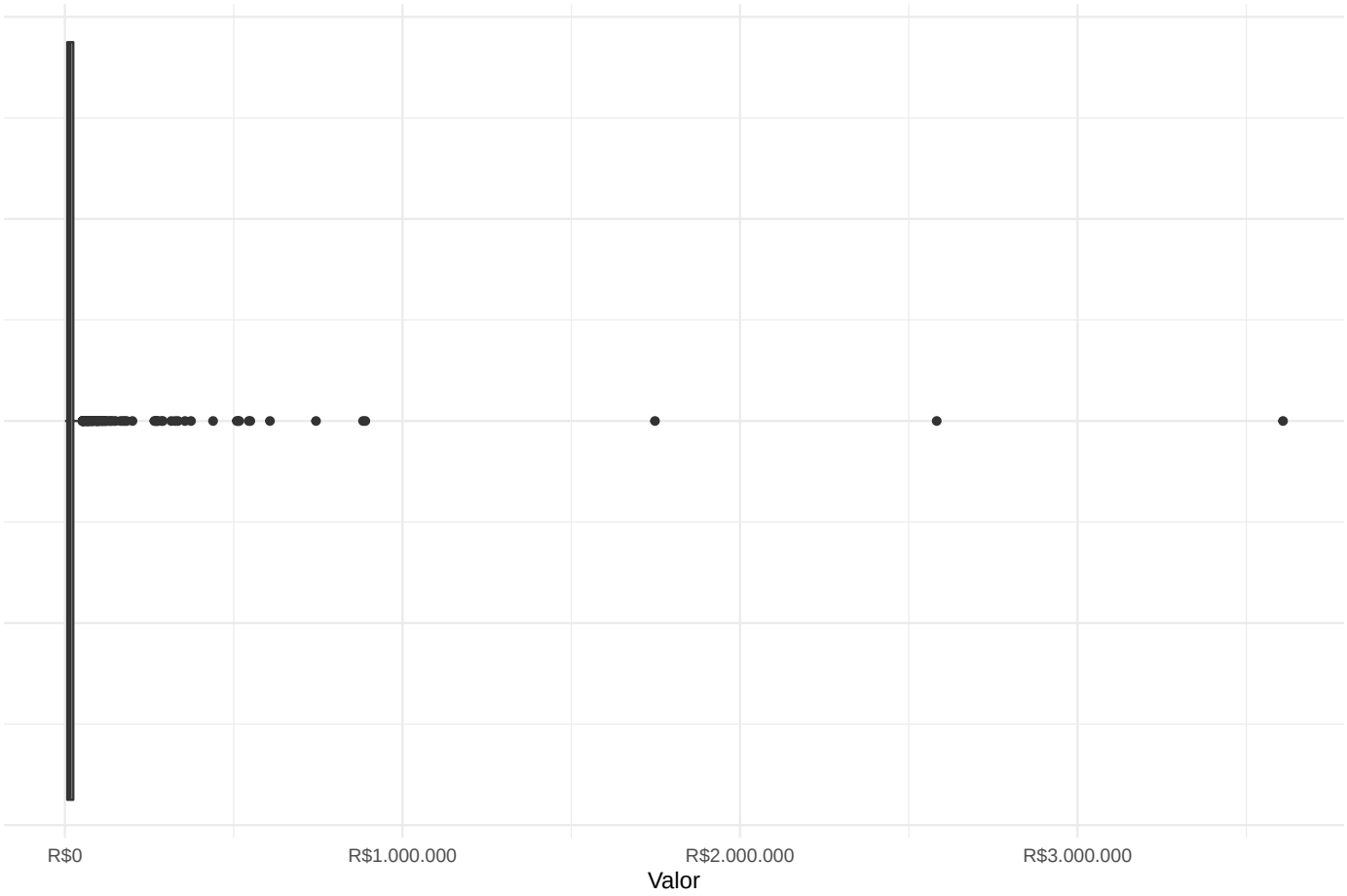



Figura 4.25: Boxplot da variável consumo

Tabela 4.4: Tabela de estatísticas-resumo para construção do boxplot

| medidas | parte do boxplot | valores |
|---|---|---|
| valor máximo | bigodes | R$ 3.608.780,16 |
| quartil superior | centro | R$ 24.799,48 |
| mediana | centro | R$ 13.590,47 |
| quartil inferior | centro | R$ 6.854,77 |
| valor mínimo | bigodes | R$ 181,84 |
| IQR | bigodes | R$ 17.944,71 |
| quartil superior + 3/2 IQR | bigodes | R$ 51.716,54 |
| quartil inferior - 3/2 IQR | bigodes | R$ -20.062,29 |
| último ponto superior | bigodes | R$ 51.587,91 |
| último ponto inferior | bigodes | R$ 181,84 |
| quantidade de pontos superiores | pontos | 87 pontos |
| quantidade de pontos inferiores | pontos | 0 pontos |

Por esta tabela, vemos que o valor máximo que estará contido nos bigodes é de aproximadamente R$ 51 mil, entretanto, o valor máximo é de mais de R$ 53 milhões. Ou seja, o valor máximo que deveríamos visualizar não é nada perto do último ponto que observamos, por isso o centro do boxplot e os seus bigodes ficam achatados, parecendo uma linha no zero.

Quanto aos pontos, vemos que há 99 pontos acima do limite superior do boxplot. Entretanto, o que efetivamente conseguimos observar são apenas 4 pontos. Os demais 95 pontos estão todos concentrados, de forma que eles ficam indiscriminados entre si.

Então o primeiro problema diz respeito à visualização de dados muito dispersos e assimétricos. Como já vimos, isso pode ser resolvido colocando o gráfico em escala logarítmica. Na Figura 4.26, o gráfico está em log de base 10.

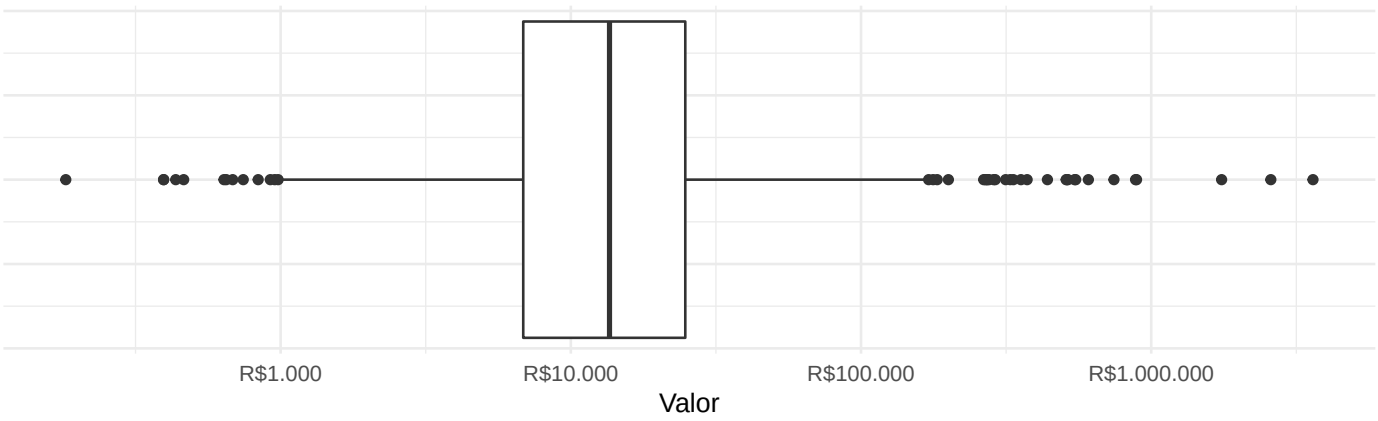



Figura 4.26: Boxplot dos valores em log na base 10

Talvez o problema dessa transformação de variável seja que os parâmetros do gráfico se alteram. Vemos algo que não víamos nos outros boxplots representando esta mesma variável: pontos abaixo do bigode inferior! Isso acontece porque,

com os valores em log, a variabilidade dos dados é muito diferente.

O outro problema dos boxplots são as informações que eles escondem. Como os boxplots apresentam somente as medidas de resumo, então eles escondem o tamanho da amostra. Veremos mais para frente que precisamos de grandes números de dados para realizar algumas inferências. Para um exemplo, ver a discussão do Data to Viz em The Boxplot and its pitfalls.

#### 4.3.2.2 Gráficos bivariados (com explicativa categórica)

Se antes estávamos vendo os gráficos univariados para variáveis numéricas, agora vamos ver como esses gráficos podem representar variáveis explicativas. Vamos começar com as variáveis explicativas categóricas, ou seja, gráficos que misturam uma informação numérica com outra categórica. Usualmente, usamos as categorias para explicar o comportamento da distribuição numérica.

A combinação de uma variável numérica com uma variável categórica não gera nenhum gráfico novo, mas apenas adiciona novas camadas de densidade para os dois gráficos de variáveis numéricas que já vimos: o histograma e o boxplot. Os exemplos serão breves, portanto.

##### 4.3.2.2.1 Histograma

Para adicionar uma informação categórica nos histogramas, basta "quebrarmos" os dados em várias categorias, indicando como que cada categoria contribui para a distribuição geral. Na Figura 4.27 observamos o mesmo histograma de tempos que já vimos anteriormente, mas dividido pelo resultado final da decisão. Cruzando o tempo de decisão com o resultado da decisão poderíamos, por exemplo, investigar se os recursos protelatórios (cujo resultado é sempre negativo, ou, "Não reformou").

##### 4.3.2.2.2 Boxplot

Para o boxplot, podemos realizar o mesmo tipo de visualização, em que quebramos o boxplot principal em grupos menores. Isso nos permite comparar medianas e IQRs com maior facilidade.

#### 4.3.2.3 Gráficos bivariados (com explicativa numérica)

Ao contrário dos gráficos de variável numérica bivariados com uma explicativa categórica, que não formavam nenhum tipo de gráfico novo, quando cruzamos duas informações numéricas, aparece um novo gráfico, muito comum: o gráfico de dispersão (em inglês, scatter plot). Veremos esta visualização em mais detalhes a seguir.

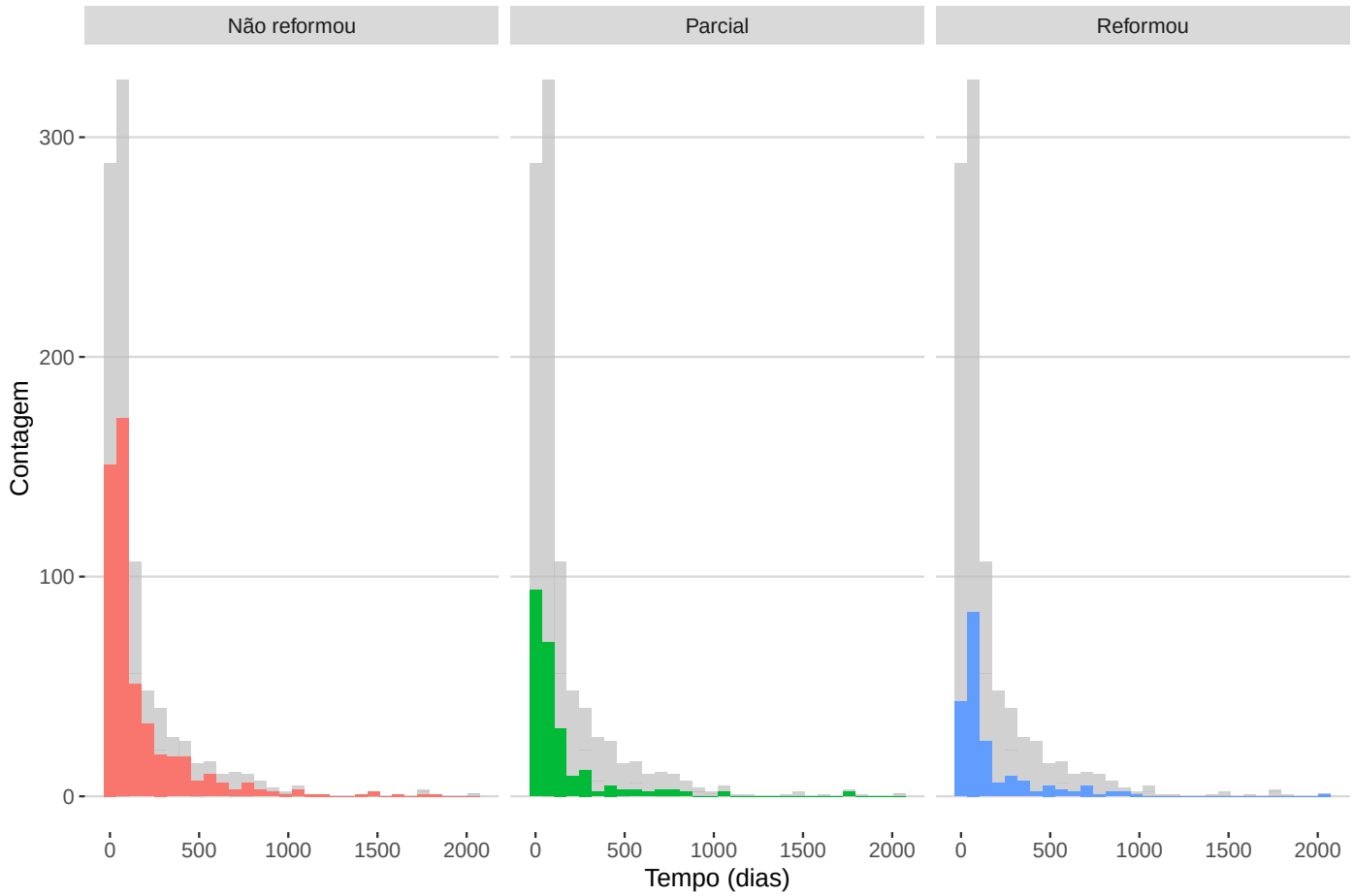



Figura 4.27: Histograma para cada tipo de decisão

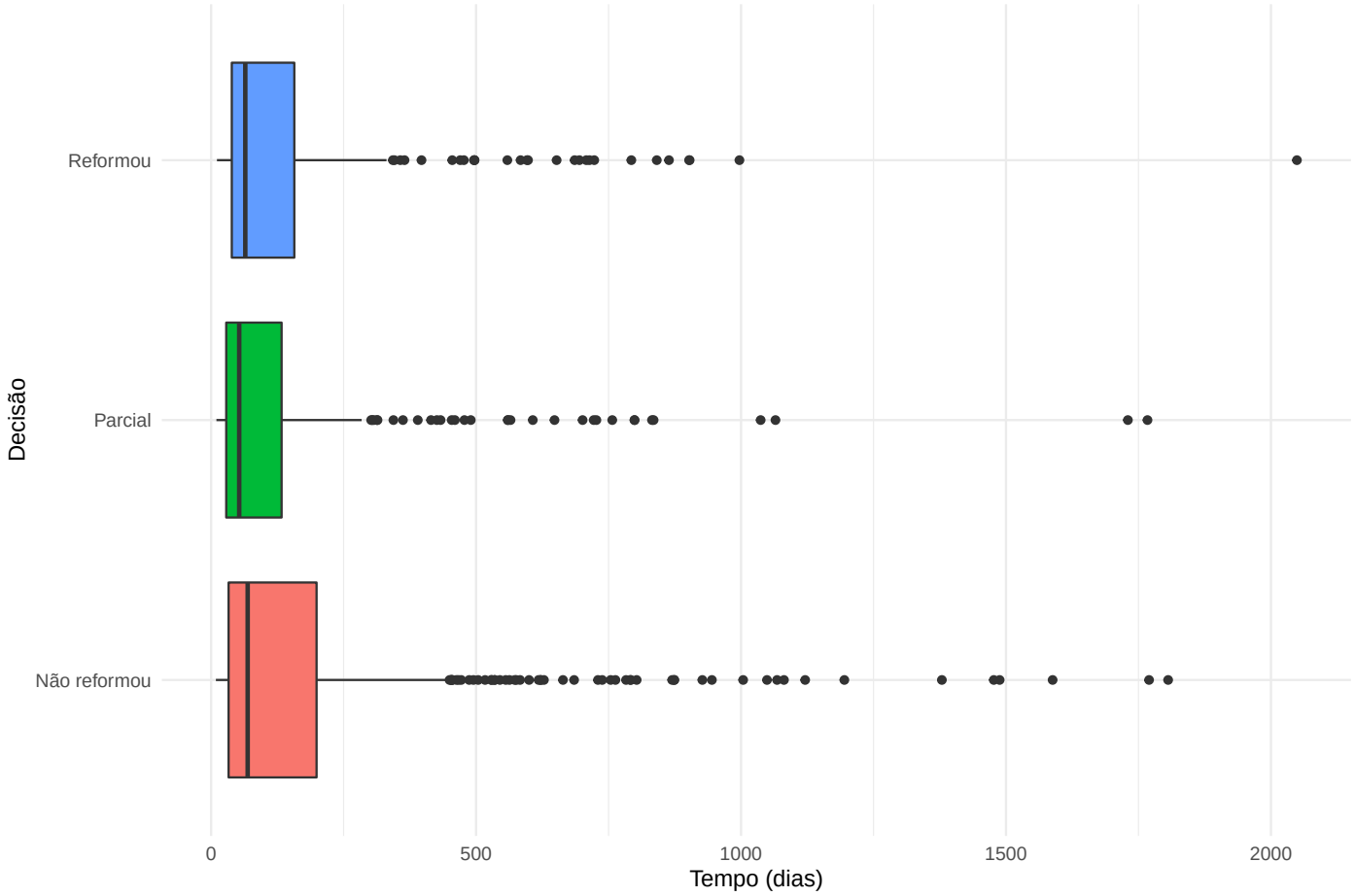



Figura 4.28: Histograma para cada tipo de decisão

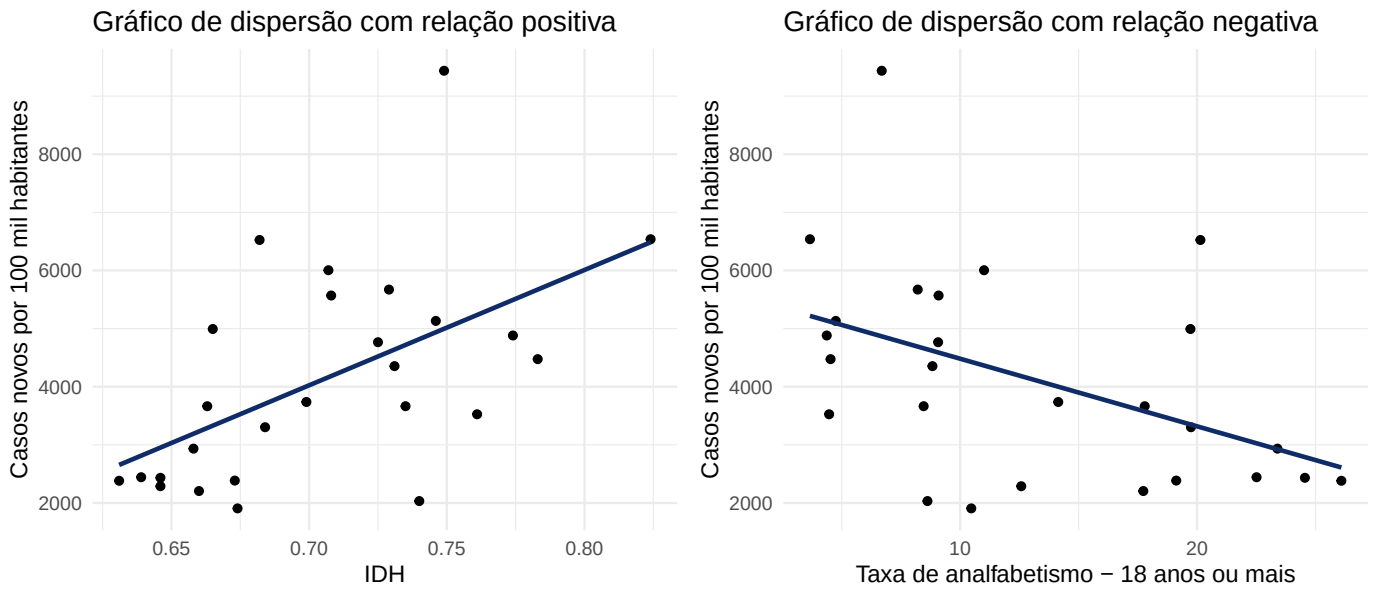



Figura 4.29: Gráficos de dispersão sobre litigiosidade

#### 4.3.2.3.1 Gráficos de dispersão

O gráfico de dispersão, como já dito, é usado para representar duas variáveis numéricas. Esse tipo de representação nos indica como que duas variáveis numéricas se relacionam. Há duas relações possíveis: uma relação positiva (isto é, quanto maior uma variável, maior a outra também), ou uma relação negativa (isto é, quanto mais de uma variável, menos eu tenho da outra). A Figura 4.29 traz dois gráficos de dispersão, demonstrando as relações possíveis.

A linha que demonstra a relação não será tratada neste capítulo. Ela foi meramente ilustrativa para indicar o tipo de relação que está sendo representada entre as variáveis. O que importa é que, ao colocar, nos dois eixos, variáveis numéricas, conseguimos observar como essas variáveis se associam entre si.

Por mais que os gráficos de dispersão sejam usualmente utilizados para relacionar duas variáveis contínuas, podemos dar outras funções a eles. Em um caso exemplar, do Observatório da Insolvência da ABJ sobre Recuperações Judiciais no Estado de São Paulo, analisamos a relação entre a dívida da recuperação com a remuneração do administrador judicial. Ao apresentar a associação entre as variáveis, construímos um gráfico de dispersão. Uma forma de se encarar este gráfico é justamente pela indagação de se quanto maior a dívida total, maior a remuneração do administrador judicial. Mas, para além disso, podemos ter um outro olhar para o gráfico de dispersão. O que sabemos é que o art. 24, § 1º da lei 11.101/2005 estabelece como limite à remuneração do AJ que "o total pago ao administrador judicial não excederá 5% (cinco por cento) do valor devido aos credores submetidos à recuperação judicial ou do valor de venda dos bens na falência." Neste caso, então, ao colocarmos uma linha de 5% em relação ao total da dívida, conseguimos observar a distância da remuneração dos AJs em relação ao máximo permitido por lei. Quando adicionamos esta camada de interpretação ao gráfico, não estamos olhando para a relação entre as variáveis (remuneração do AJ e total da dívida), mas estamos olhando para a legalidade dessa remuneração.

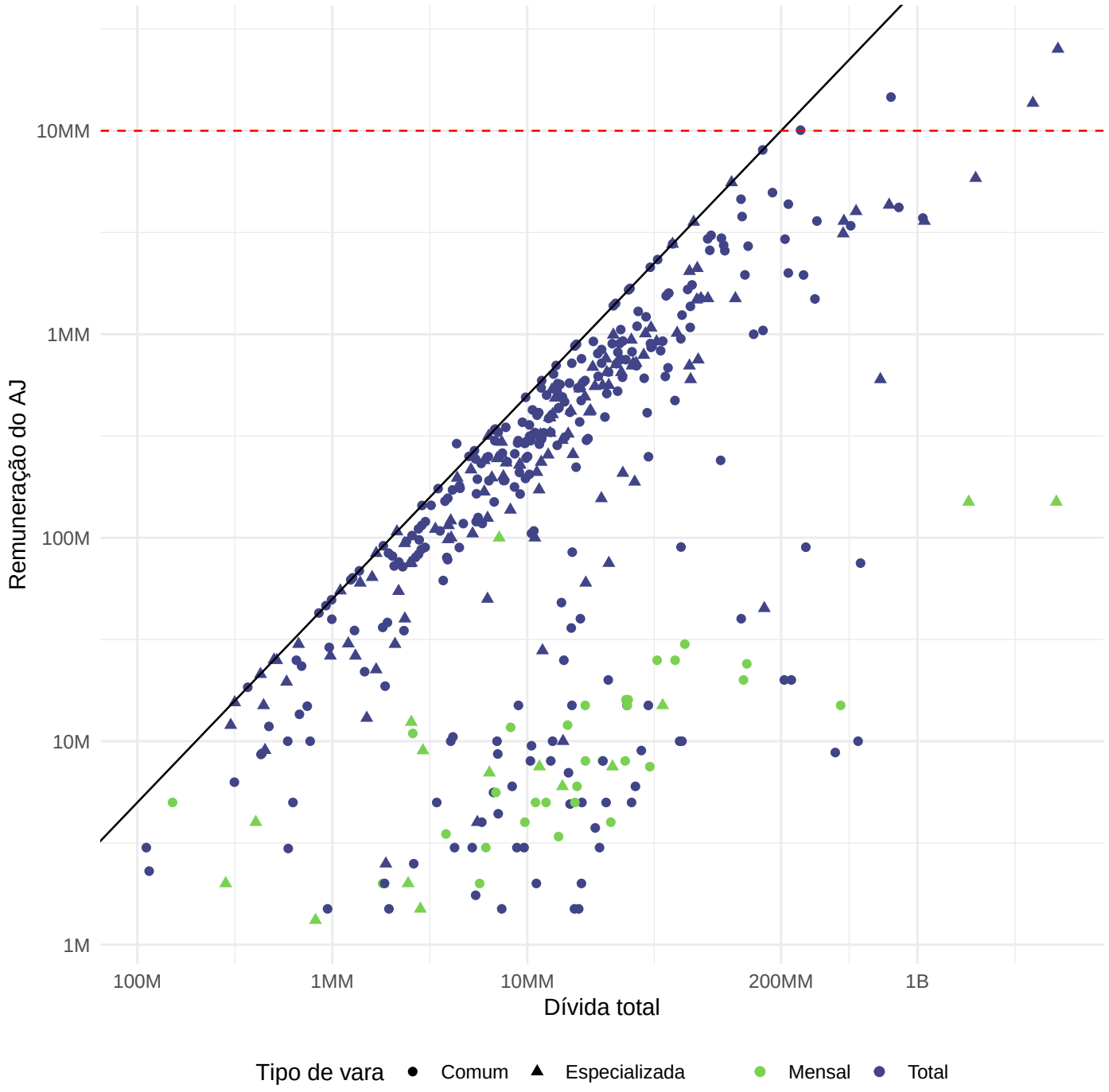



Figura 4.30: Relação entre dívida e remuneração de Administradores Judiciais: Observatório da Insolvência

#### 4.3.2.4 Gráficos bivariados (no tempo)

Chegamos, então, ao fim dos gráficos bivariados com uma variável numérica como variável de interesse. O último gráfico que nos resta observar é o caso das variáveis explicativas temporais. Esse caso é especial porque as variáveis temporais não possuem uma natureza muito definida.

Por exemplo, se pegarmos a variável categórica de resultado da decisão de segunda instância, ela pode assumir três valores: reformou, não reformou ou reformou parcialmente. Faria sentido, diante dessas três informações possíveis, que a gente substraísse "reformou" de "não reformou"? Obviamente não, pois não há sentido algum neste cálculo. Entretanto, se tivermos duas datas, a subtração dessas duas informações irá sim fazer sentido: ela irá nos indicar um intervalo de tempo. Com isso, já vemos que as variáveis temporais não se comportam exatamente como variáveis categóricas.

Agora, pensando nas variáveis numéricas contínuas, normalmente, entre um valor e outro, existem infinitos valores possíveis. Por exemplo, entre 1 e 2, existe o número 1,1 e também o 1,2. Mas entre esses valores existe o 1,11 e o 1,12. Mas entre eles, existe o 1,111 e o 1,112. E assim por diante. Essa propriedade das variáveis numéricas contínuas não está presente nas variáveis temporais. Se a variável temporal representa um ano, por exemplo, não há nada entre os anos de 2016 e 2017. Ou ainda, se a variável temporal representa datas, não há nenhum outro número entre um dia e outro.

Por fim, então, comparando as variáveis temporais com variáveis numéricas discretas, elas também apresentam diferenças entre si. Uma variável numérica normalmente expressa a contagem de alguma informação, por exemplo, a contagem de juízes por comarca. Esse valor é sempre positivo, inteiro e maior ou igual a zero. Se tentarmos pensar dessa forma com as variáveis temporais, o que seria uma data maior ou igual a zero? A variável igual a zero seria o dia 0/0/0000? Não faz sentido isso. Entretanto, por mais que haja algumas diferenças, a variável numérica discreta é a que mais se aproxima das variáveis temporais.

Por causa dessa natureza especial das variáveis temporais é que faz sentido pensarmos nos gráficos bivariados no tempo como um caso especial.

Antes de entrarmos na visualização gráfica dos gráficos bivariados no tempo, é importante diferenciarmos dois tipos de variáveis relacionadas a tempos. Uma coisa é uma variável que pode receber valores como "27/08/2021"; mas existe outra informação relacionada a tempo que diz respeito ao tempo transcorrido entre duas datas, por exemplo, o tempo entre a distribuição de um processo e a sua sentença. No primeiro caso, a natureza da variável é temporal (uma data); no segundo caso, a natureza da variável é uma variável numérica comum. O tipo de variável especial de que vamos tratar nesta sessão diz respeito somente ao primeiro tipo, e não ao segundo, tanto que já fizemos um gráfico com a variável "tempo".

##### 4.3.2.4.1 Séries temporais

O nome do gráfico que representa uma variável numérica em relação ao tempo se chama série temporal. O gráfico de uma série temporal se apresenta sempre com o tempo no eixo x e a variável numérica no eixo y.

Este gráfico é utilizado para "descrever apenas o comportamaneto da série", o que vai envolver "a verificação da existência de tendências, ciclos e variações sazonais".[2]

Na Figura 4.31, vemos um exemplo desse tipo de relação.

[2]MORETTIN, Pedro A; TOLOI, Clélia M. C. Análise de Séries Temporais, 2a ed. São Paulo: Editora Blucher. 2006, p. 3.

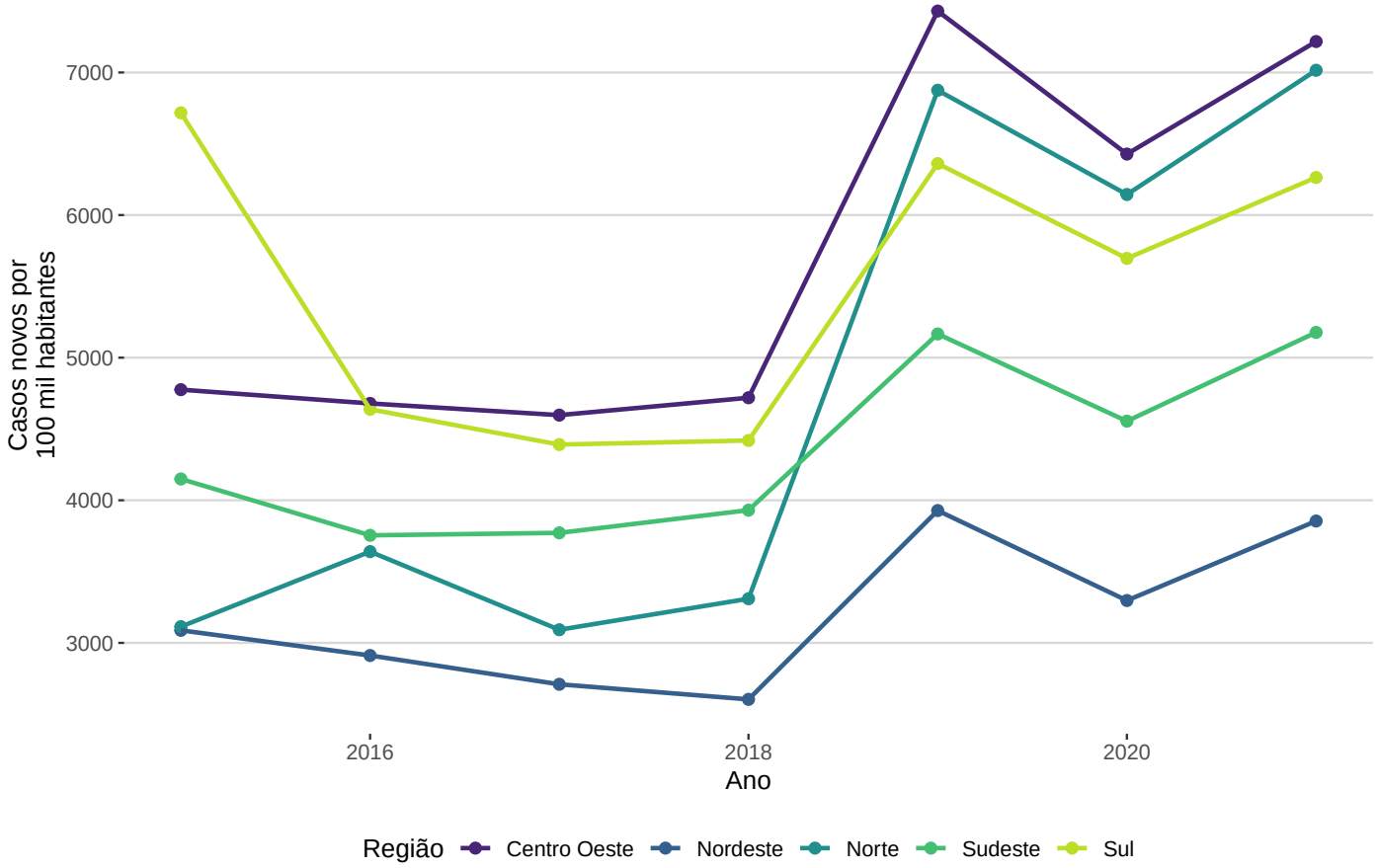



Figura 4.31: Série de tempo dos dados de litigiosidade.

A figura nos mostra uma tendência se aumentar os casos em todas as regiões do Brasil a partir de 2018. Além disso, vemos que, na região Sul, entre 2015 e 2016, houve uma clara queda, no total de casos por 100 mi habitantes, em comparação com outras regiões, que ou demonstraram aumentos, ou demonstraram quedas não tão intensas.

O que vemos é que não estamos fazendo afirmações sobre a relação entre a variável contínua (número de casos novos por 100 mil habitantes) e a variável temporal (ano) do tipo "quanto mais tempo, mais casos", porque não é esse tipo de relação que séries temporais nos indicam. Esse tipo de visualização está nos apontando para tendências temporais, sejam elas cíclicas, sazonais ou episódios pontuais.

# 5 Modelagem

(em construção)

# Bibliografia


ABJ. 2019. ”Avaliação do Impacto de Critérios Objetivos na Distinção Entre Posse para Uso e Posse para Tráfico: um estudo jurimétrico”. https://abj.org.br/cases/drogas-stf/ .

———. 2020. ”O problema da cifra oculta nos tribunais brasileiros”. https://lab.abj.org.br/posts/2020-12-07-cifra-oculta/.

Agresti, Alan, e Barbara Finlay. 2009. *Statistical Methods for the Social Sciences*. 4.ª ed. London: Pearson.

Bolfarine, Heleno, e Wilton O. Bussab. 2005. *Elementos de Amostragem*. São Paulo: Blucher.

Bottino, Thiago. 2015. ”Panaceia universal ou remédio constitucional? Habeas corpus nos Tribunais Superiores”. Brasília: Ipea. http://pensando.mj.gov.br/wp-content/uploads/2015/06/thiago_55_finalizada_web.pdf.

Breiman, Leo. 2001. ”Statistical Modeling: The Two Cultures”. *Statistical Science* 16 (3): 199–215.

Bussab, Wilton O., e Pedro A. Morettin. 2017. *Estatística Básica*. 9.ª ed. São Paulo: Saraiva.

Camargo, Solano de. 2015. ”Forum shopping: modo lícito de escolha de jurisdição?” Dissertação de mestrado, Universidade de São Paulo. https://doi.org/10.11606/D.2.2016.tde-21122015-193317.

CCI, CENTER FOR COURT INNOVATION. 2020. ”Can Courts Be More User-Friendly? How Satisfaction Surveys Can Promote Trust and Access to Justice”. https://www.courtinnovation.org/sites/default/files/media/document/2020/CCI_FactSheet_SatisfactionSurveys_04202020.pdf.

Cleveland, William S. 1985. *The Elements of Graphing Data*. California: Wadsworth Advanced Book Program.

Epstein, Lee, e Andrew D. Martin. 2014b. *An Introduction to Empirical Legal Research*. United Kingdom: Oxford University Press.

———. 2014a. *An Introduction to Empirical Legal Research*. United Kingdom: Oxford University Press.

Fulgêncio, Henrique Augusto Figueiredo, e Alexandre Araújo Costa. 2018. ”As funções contemporâneas do mandado de injunção: análise empírica sobre o perfil das ações ajuizadas perante o Supremo Tribunal Federal”. *Revista da Faculdade de Direito do Sul de Minas* 34 (2): 451–88. https://revista.fdsm.edu.br/index.php/revistafdsm/article/view/202/211.

Grinover, Ada Pellegrini, ed. 2014. ”Avaliação da Prestação Jurisdicional Coletiva e Individual a partir da Judicialização da Saúde”. CEBEPEJ.

King, Gary, Robert O Keohane, e Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

Kozak, Marcin. 2010. ”Basic principles of graphing data”. *Sci. Agric.* 67 (4): 483–94.

Lambert, Paul C. 2007. ”Modeling of the cure fraction in survival studies”. *The Stata Journal* 7 (3): 351–75.

Lopes, José Reinaldo de Lima. 2003. ”A definição de interesse público”. Em *Processo Civil e Interesse Público: o processo como instrumento de defesa social*, editado por Carlos Alberto de Salles. São Paulo: Revista dos Tribunais.

Molnar, Christoph. 2022. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2.ª ed. https://christophm.github.io/interpretable-ml-book.

Nunes, Marcelo Guedes. 2016. *Jurimetria: Como a Estatística Pode Reinventar o Direito*. São Paulo: Revista dos Tribunais.

Perez, Marco Augusto. 2018. ”O Controle Jurisdicional da Discricionariedade Administrativa: métodos para uma jurisdição ampla das decisões administrativas”. Tese de livre docência, Universidade de São Paulo.

Popper, Karl. 1934. *A Lógica da Pesquisa Científica*. São Paulo: Cultrix.

Priest, George L, e Benjamin Klein. 1984. "The selection of disputes for litigation". *The Journal of Legal Studies* 13 (1): 1–55.

Shadish, William R., Thomas D. Cook, e Donald T. Campbell. s.d. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Cengage Learning.

Spinney, Laura. 2022. "Are we witnessing the dawn of post-theory science?" The Guardian. 2022. https://www.theguardian.com/technology/2022/jan/09/are-we-witnessing-the-dawn-of-post-theory-science.

Stern, Julio Michael, Marcos Antonio Simplicio, Marcos Vinicius M. Silva, e Roberto A. Castellanos Pfeiffer. 2020. "Randomization and Fair Judgment in Law and Science". arXiv. https://doi.org/10.48550/ARXIV.2008.06709.

Sundfeld, Carlos Ari, Ester Gammardella Rizzi, Evorah Lusci Costa Cardoso, Flávio Beicker, Francisco Carvalho de Brito Cruz, Gabriele Estábile Bezerra, Gustavo Cesar Mazutti, et al. 2011. *Controle de constitucionalidade e judicialização: o STF frente à sociedade e aos Poderes*. Belo Horizonte: Faculdade de Filosofia e Ciências Humanas; SBPD. https://sbdp.org.br/wp/wp-content/uploads/2018/01/05-controle_de_constitucionalidade_e_judicializacao.pdf.

Trecenti, Julio, e Marcelo Guedes Nunes. 2021. "Impactos da MPV 1.040/2021 no tempo de abertura de empresas". https://lab.abj.org.br/posts/2021-06-11-analise-1040/.

Tukey, John W. 1977. *Exploratory Data Analysis*. Addison-Wesleyl Publishing Company.

Whitten, Paul M. Kellstedt Guy D. 2015. *Fundamentos da pesquisa em ciência política*. Sao Paulo: Blucher.

Winsihp, Christopher, e Robert D. Mare. 1992. "Models for sample selection bias". *Annual review of sociology* 18 (1): 327–50.

Xavier, José Roberto Franco. 2015. "Algumas notas teóricas sobre a pesquisa empírica em direito". *São Paulo Law School of Fundação Getúlio Vargas – FGV DIREITO SP, Research Paper Series – Legal Studies*, n. 122.

Xie, Yihui. 2015. *Dynamic Documents with R and knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. http://yihui.name/knitr/.