

Manipulações estatísticas e anomalias visuais: design de visualização de dados e reconhecimento de vieses estatísticos

Statistical manipulations and visual anomalies: data visualization design and statistical bias recognition

Jaqueline Vasconcelos Braga,
Tiago Barros Pontes e Silva & Virgínia Tiradentes Souto

viés estatístico, gráficos,
visualização de dados,
anomalias visuais

O mundo contemporâneo é caracterizado por um amplo volume de informações produzidas. Contudo, proceder a seleção e leitura dessas informações por meio de relatos de pesquisa ou de notícias ainda é um desafio. Entre os obstáculos presentes se destacam os vieses da informação, originados por tratamentos de jornalistas ou pesquisadores, ou mesmo provocados intencionalmente para subverter a representação da realidade a partir dos dados obtidos. Assim, o presente estudo visa discutir a interpretação de informações visuais em representações gráficas de cálculos estatísticos de modo a contextualizar alguns dos principais recursos visuais de enviesamento de pesquisa. Para tanto, aborda os principais modos de enviesamento em pesquisas a partir das representações da estatística e da visualização de dados e identifica alguns passos nos quais o enviesamento se traduz em informações visuais. A partir do levantamento realizado, sugere-se que a compreensão visual dos recursos de visualização de dados pode ao menos instigar a indagação do leitor acerca do possível viés.

*statistical bias, graphs,
data visualization, visual
anomalies*

The contemporary world is characterized by a large volume of produced information. However, selecting and reading this information through research reports or news is still a challenge. Among the present obstacles stand out the information bias, originated by treatments of journalists or researchers, or even intentionally provoked to subvert the representation of reality from the obtained data. Thus, the present study aims to discuss the interpretation of visual information in graphical representations of statistical calculations in order to contextualize some of the main visual bias features of research. To this end, it addresses the main modes of search bias from statistical representations and data visualization and identifies some steps in which bias translates into visual information. From the study, it is suggested that the visual understanding of data visualization resources may at least instigate the reader's question about the possible bias.

1 Introdução

Em meio à massiva produção e difusão de dados no contexto contemporâneo, é imprescindível categorizar e interpretar as informações disponíveis. Elas são empregadas como ferramentas contextualizadoras e, em diversos aspectos, conduzem e orientam o processo decisório. Ainda que os dados não permitam inicialmente uma significação ou apropriação direta (Cunha & Cavalcanti, 2008), cabe àquele que lhe provê tratamento interpretá-lo e dotá-lo de significados para os outros leitores. Tanto no momento da análise de dados, quanto da difusão da informação, o tratamento visual dos dados exerce um papel chave: ele traduz um referencial visual agregado que pode conectar evidências e alinhar fatores para que o significado seja expresso. A difusão de informações depende, por sua vez, dos resultados obtidos; ou seja, algo que não possua relativa significância tende a ser deixado de lado em nome da difusão de conteúdos de maior relevância. Assim, a difusão de estudos, em especial aqueles que partem de dados quantitativos, podem ser alvo de vieses estatísticos ou de representação, reduzindo sua confiabilidade.

Lipton (1991), com a “inferência da melhor explicação” (IME), aponta para a possibilidade de se obter a resposta mais ajustada nas diversas circunstâncias a partir de inferências. Tal relação de causalidade, mesmo que não possa ser explicitada por meios estatísticos, é apontada como um vetor de inferenciação de uma realidade suportada por dados e robustamente testada e falseada. Inferir causalidade em visualização de dados é, inclusive, uma premissa apontada por Tufte (2013) como fundamental para contextualizar seu bom funcionamento. Enquanto a interpretação estatística se configura como uma leitura precisa, trata-se de uma habilidade que requer tanto um aprendizado particular, quanto experiência na prática analítica. Assim, a leitura de uma visualização de dados assume, como apontado por Cairo (2016), um perfil de apoio cognitivo; um recurso capaz de tornar visível e inteligível nuances e informações que estão codificadas a partir de dados. Embora a natureza das visualizações de dados não seja “simplificadora” para este, é possível admitir que, em função de seu uso enquanto amparo cognitivo, sob algumas instâncias, eles se transformem em um traço agregador no qual traduzem-se as informações para o sujeito que lhe faz a leitura. Meirelles (2013) traça uma rede de propósitos da utilização de visualizações de informação mais abrangente, destacando seu uso enquanto ferramenta de registro da informação, enquanto agregador de significados, facilitador de pesquisas, amparo à memória, facilitador de descobertas, ampliação da detecção e reconhecimento de elementos, apoio a processos inferenciais, provedor de modelos hipotéticos e realistas e também como ferramenta de manuseio/manipulação de dados. Em todos estes casos, os autores promovem relações nas quais a leitura das informações nos recursos visuais

subsídia a ampliação do processamento e inferenciação enquanto ferramenta a priori e, posteriormente, enquanto um registro fatural.

Ao passo que clareza, objetividade e, sobretudo, a capacidade de permitir inferências lógicas em meio aos dados representam um aspecto da visualização de dados, a fidedignidade ao banco de dados e à realidade à qual este representa possui uma expressa relação de valor com o recurso visual. Os dados, contudo, não podem falar por si e demandam interlocutores que possam articular-lhes as variáveis em um discurso do qual espera-se que possa surgir a inferência esperada de seu leitor. Este trabalho, ora atribuído ao jornalista de dados, ora atribuído ao pesquisador, condensa, aglutina, recorta ou ajusta os dados de forma a permitir sua análise e difusão. A neutralidade destes agentes enquanto autores da informação, assim como daqueles que previamente delinearão os modos de aquisição dos dados e variáveis computáveis, é passível de análise tanto quanto aquilo que estes agentes informam. Deste modo, tão amplamente quanto se discute a possibilidade de promover a inteligibilidade da informação quantitativa por meio de recursos cognitivos, é fundamental que se discuta a amplitude dos vieses que podem ser agregados nesta mesma fonte de informação e também os aspectos que tangenciam sua confiabilidade. Segundo Mayr et al. (2019), a confiabilidade em visualização de informações consiste em “uma tendência implícita ou explícita de se ancorar na representação dos dados para compor a informação exibida”.

No âmbito da apresentação de vieses, o presente artigo realiza um compêndio bibliográfico de referenciais para fomentar a interpretação de informações visuais em representações gráficas de cálculos estatísticos de modo a contextualizar alguns dos principais recursos visuais de enviesamento de pesquisa. Em um primeiro momento esse processo é descrito sob a ótica da estatística e, em seguida, sob a ótica do design de visualização de dados, sobretudo no aspecto da literacia visual. Seu objetivo é, portanto, permitir que sejam reconhecidas algumas estruturas visuais e contextuais das visualizações da informação de modo a promover o reconhecimento de possíveis fontes de vieses e despertar novas dimensões de leituras críticas em visualizações de dados.

2 Representação gráfica sob aspectos estatísticos

Reilly (2017) traz luz à questão da publicização de informações por meio do jornalismo de dados, sustentando que a inserção destas representações visuais gráficas confere maior dinamismo face às informações textuais. A autora traz ainda que os responsáveis por produzir estes conteúdos comumente não tem formação em estatística ou programação, o que dificulta a produção destes conteúdos. Assim como Reilly (2017), acreditamos que o caminho para o melhor desenvolvimento de visualizações de informações

passa pela colaboração entre profissionais de áreas distintas. Contudo, defendemos também que o conhecimento estatístico elementar gráfico componha o referencial daqueles que trabalham com a composição visual propriamente dita. Neste sentido, trataremos a seguir de alguns elementos estatísticos com natureza visual, mais diretamente relacionados ao design da informação.

O campo da estatística subdivide-se na indução estatística e na descrição estatística; sendo a primeira direcionada às análises e a segunda, direcionada às representações gráficas (Akamine & Yamamoto, 2013). A representação gráfica dos dados estatísticos, por sua vez, pode também ser caracterizada por assumir dois perfis distintos. O primeiro, referente aos processos de análise e interpretação dos dados durante a análise exploratória; o segundo, voltado a apresentar os resultados obtidos pela análise ao leitor não habituado aos conceitos e critérios analíticos da estatística. Enquanto recurso de análise exploratória dos dados, sua função específica direciona-se a propiciar ao analista uma melhor compreensão e uma visão ampliada do banco de dados em trabalho; requerendo assim requisitos específicos de representação distintos daqueles voltados à legibilidade da análise concluída. Os recursos visuais utilizados nas análises, em geral, não são apresentados aos leitores. Contudo, a utilização destes elementos visuais enquanto recurso retórico pode enriquecer o processo de inferência das informações apontadas na análise, contribuindo com a persuasão do leitor quanto ao resultado obtido (Pandey et al., 2014).

Dado o seu uso conjunto com as estatísticas descritivas, indicam-se como recursos visuais gráficos principalmente: as curvas de distribuição, os gráficos de dispersão, linhas e barras de erro, e também os histogramas (NIST/SEMATECH, 2013). Todos estes recursos são descritos e interpretados a seguir de modo a sistematizar a percepção de elementos discrepantes em cada tipo de recurso visual ao longo da análise estatística descritiva. Ainda, em conjunto a esta apresentação dos recursos visuais, são discutidos os modelos de análises estatísticas que possam auxiliar na compreensão das relações destes elementos com uma possível composição de vieses nesta etapa de produção em estatística.

2.1 Curva de distribuição: médias e outliers

No contexto de análise exploratória, algumas premissas devem ser observadas nos dados de modo a possibilitar as análises posteriores. Primeiramente, espera-se que o grupo amostral seja, de fato, representativo da população de referência para que, neste caso, os efeitos observados na amostra possam refletir um percentual amplo da população. Este ponto se reflete na tendência à distribuição normal dos efeitos, uma curva equilibrada em altura e extensão, de modo a evidenciar que a maior frequência dos efeitos observados

correspondem à maior incidência na população, fazendo com que, em situações ideais, todas as medidas centrais - médias, mediana e moda - se encontrem ao centro da distribuição. A normalidade na distribuição é, contudo, uma abstração suscetível ao tamanho do grupo amostral na mesma medida em que depende da ótima obtenção dos integrantes da amostra (Fox & Levin, 2004). Embora a normalidade seja esperada a um determinado nível, a perfeita normalidade é um critério utópico, embora saiba-se que, para os fenômenos que demonstram essa distribuição, quanto maior a amostra, mais próxima ela estará da distribuição normal. A zona central da curva de distribuição normal representa 95% da população retratada pela amostra, e suas extremidades não representam mais que 0,5% destes (Figura 1).

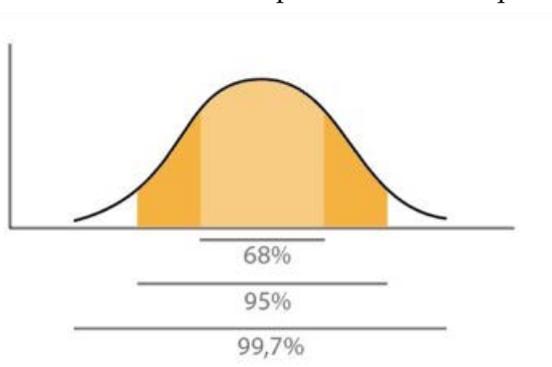


Figura 1 Curva de distribuição normal com projeção de amplitude amostral compreendida; o intervalo de confiança dado pelo p-valor ($p < 0,05$) corresponde à área central da curva. Fonte: imagem dos autores.

Ao tratar uma distribuição anormal como normal, aumenta-se significativamente a possibilidade de que os resultados obtidos na análise não retratem a realidade da população. Os dados que fogem da normalidade requerem tratamentos específicos e análises não paramétricas. Entretanto, recursos de tratamento de dados como o bootstrapping permitem aproximar a amostra da normalidade e viabilizar a execução de testes mais robustos. A capacidade de reconhecer desvios na curva de normalidade, assim como identificar ajustes nas distribuições, permite identificar a pertinência do grupo amostral à população que ele representa e é, deste modo, um elemento necessário tanto ao analista quanto ao leitor de informações estatísticas.

Huff (1954) já elucidou a importância de uma média bem selecionada para justificar um determinado ponto em específico. Em um de seus exemplos, Huff ilustra a diferença entre salários dos trabalhadores de uma fábrica: enquanto a média aritmética aponta o valor de USD\$ 5.700, a mediana atingiu um patamar inferior, USD\$ 3.000, sendo que a moda entre os valores alcançava apenas USD\$ 2.000. Deste modo, ainda segundo Huff (1954), a escolha da média representava um recurso de enviesamento, tornando a medida de tendência central incapaz de representar o grupo.

As discrepâncias nas medidas de variabilidade são, em sua maioria, causadas pela presença de dados discrepantes, chamados outliers, que forçam tanto as medidas de tendência central quanto a curva de normalidade para pontos extremados, reduzindo a fidedignidade da amostra à população (Fox & Levin, 2004). É possível tratar os dados de modo a minimizar efeitos de presença de outliers. Contudo, excluí-los de modo completo do grupo amostral torna-se um recurso que igualmente envia a percepção da população: assim como o outlier se apresenta na amostra, ele pode ser encontrado na população, e sua presença não deve ser descartada de modo a compreender os efeitos buscados ao falsear as hipóteses. A Figura 2A (esquerda) ilustra um exemplo de Boxplot que evidencia dois outliers, e a Figura 2B (direita) apresenta uma curva de distribuição anormal com skewness, uma irregularidade vertical.

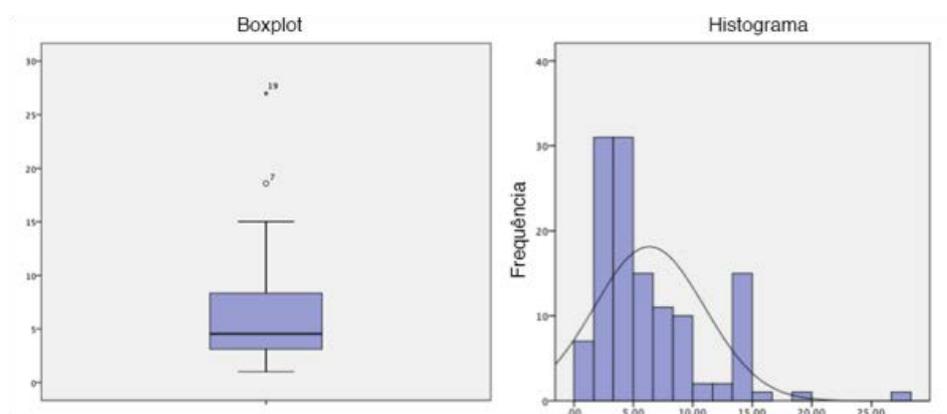


Figura 2 (A) À esquerda, o Boxplot evidencia dois outliers; a barra de erros ampliada acima da caixa e a linha central com a média dos valores obtidos nos dados na porção inferior da caixa evidenciam os efeitos da presença dos casos discrepantes; (B) à direita, a curva de distribuição anormal com skewness - irregularidade vertical na centralidade das médias - aponta efeitos da presença dos outliers, os casos extremos puxam a cauda direita da curva de distribuição, acentuando a irregularidade da curva. Fonte: imagem dos autores elaborada utilizando a versão de avaliação do software IBM SPSS.

2.2 Medidas de efeito, p-valor e p-curva

Além das características referentes ao grupo amostral, é possível identificar questões relativas ao poder estatístico dos testes performados por meio da análise gráfica. Um ponto frequentemente relatado é referente aos testes de significância, que indica o pertencimento do efeito no grupo amostral a um volume igual ou superior de 95% da amostra; ou seja, abertura para aceite de um erro equivalente a 5% da amostra. A definição de aceite do p-valor

menor ou igual a 0,05 tem sido frequentemente a métrica que define a relevância do estudo.

Alguns autores, como Simonsohn et al (2004), ao diagnosticar a incidência de um volume cada vez mais amplo de estudos nos limites do p-valor ou em valores “marginalmente significativos”, desenvolveram um modelo em que é possível prever a presença de alterações amostrais para modificar os p-valores obtidos, um efeito conhecido estatisticamente como p-hacking. A p-curva (Simonsohn et al, 2004) permite estimar o tamanho do poder estatístico que o ensaio alcança, assim como permite estimar o quanto o p-valor sob efeito de hacking inflaciona o tamanho do efeito observável na população.

A Figura 3 ilustra um exemplo de p-curva gerada sem indicativa de manipulação tipo p-hacking. Ao se adquirir uma base de dados, é recomendável que sejam verificados não apenas os intervalos de confiança dos números obtidos, mas também o poder estatístico e a relevância do objeto, mesmo quando os valores obtidos são marginalmente significativos, sempre considerando o tamanho do efeito em relação ao modelo proposto e os benefícios ou traços que visa-se obter por meio do falseamento das hipóteses testadas. O objetivo é verificar, por meio da relevância e da estimativa do poder obtido pelo efeito, se alguma estrutura enviesante acometeu o estudo no momento da análise estatística.

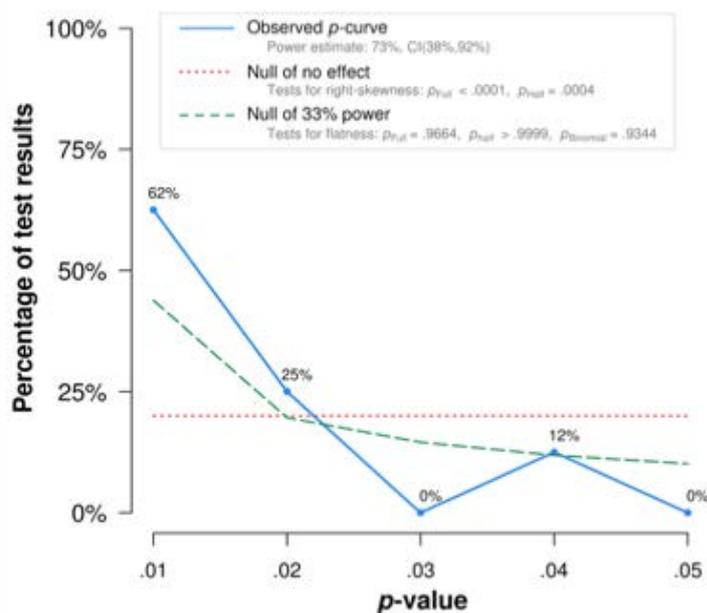


Figura 3 p-curva gerada sem indicativa de manipulação tipo p-hacking - sua assimetria à direita é um indício de que os valores marginais do p-valor não foram tratados como significativos. De acordo com os dados inseridos neste caso, embora os efeitos observados possam ser mais modestos, conforme observado pela linha tracejada em verde, o valor observado que é expresso pela linha azul indica uma tendência semelhante. Fonte: imagem dos autores elaborada com dados de simulação via aplicação web de cálculo da p-curva em <<http://www.p-curve.com/app4/pcurve4.php>> em 18/06/2019.

2.3 Histogramas, linhas e barras de erros

Na representação dos grupos, os histogramas potencializam características de comparação entre os dados retratados. Embora possa conter uma representação direta do número de elementos componentes de uma informação, como nos boxplots, eles podem também ser usados como representante de uma medida central, abrindo margem para que sejam inseridas variações em sua extensão. Esta variação expressa a estimativa de erro da medida tomada para a elaboração do histograma; o intervalo de confiança da amostra (Fox & Levin, 2004). A inserção de barras e linhas de erro na composição do histograma permite que seja analisada, de modo mais verossímil, a dimensão retratada pelas barras dentro de suas variações positivas e negativas.

A adoção de histogramas ilustrando dados de medidas centrais sem a devida linha de erro pode criar a ilusão de que um elemento seja mais representativo que outro enquanto eles podem ser iguais. Ou ainda, que a relação de grandeza expressa pelas barras seja de fato oposta na população além do recorte amostral. As diferenças possíveis de interpretação destes valores podem inverter o resultado esperado em cada circunstância e, desconsiderá-los, consiste em abrir margem para que um viés se configure neste tipo de representação. A Figura 4 apresenta um exemplo de gráfico de barras com linha de erro.

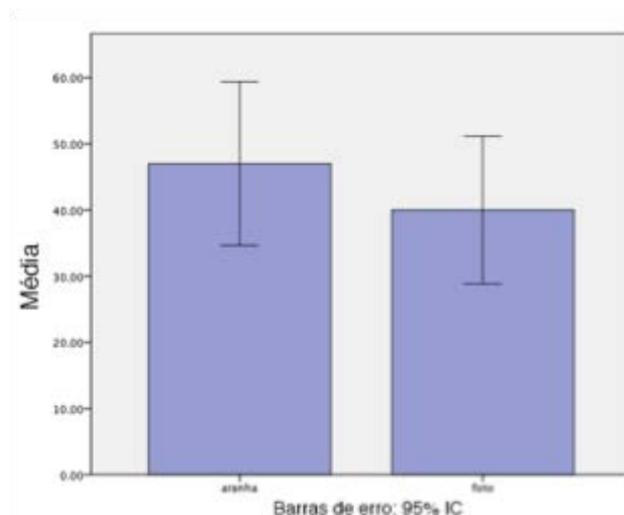


Figura 4 As barras, quando acrescidas das linhas de erro, consideradas pelo intervalo de confiança da amostra de 95%, permitem que, em casos extremos de amostragem, a menor barra - em seu maior valor esperado - possa representar uma frequência superior à maior barra - em seu menor valor esperado dentro dos limites do intervalo de confiança. Fonte: imagem dos autores elaborada utilizando a versão de avaliação do software IBM SPSS.

2.4 Gráficos de dispersão e regressão

Na representação de dados de variáveis contínuas, gráficos de dispersão permitem observar as entidades conforme sua proximidade de comportamento. Enquanto é marcada a incidência de respostas por meio de pontos em um plano cartesiano, um agrupamento de pontos indica então uma tendência a se observar refletida na população. Pontos muito distantes dos demais, por sua vez, demonstram respostas díspares das outras respostas obtidas, os outliers. Estes casos não ameaçam a validade da pesquisa e devem ser mantidos como reflexo da amostra (Bakker & Wicherts, 2014).

Gráficos de dispersão retratam também a relação entre duas variáveis distintas que se inter-relacionam. A disposição dos pontos, neste caso, insere o conceito de que exista uma tendência que acompanhe ambos os elementos em sua evolução. As linhas e curvas de regressão são traçadas com base nestes pontos para indicar uma tendência positiva, neutra ou negativa, dada uma alteração na variável descrita pelo eixo ortogonalmente oposto.

Enquanto indicadores diretos da amostra, eles são também uma referência do agrupamento e volume de casos sob um aspecto em particular: ao exibir um gráfico de dispersão para o leitor, é possível inculcar-lhe sensos de comunidade e marginalidade oriundos da presença ao centro e dos outliers. Ainda e, mais diretamente passível de enviesamentos, estão as assimilações relativas à regressão: o recorte ou ajuste dos dados expressos no gráfico de dispersão podem afetar a percepção da relação direta entre os elementos, ou criar a ilusão de que um determinado recorte é representativo da realidade. O aspecto preditivo deste tipo de recurso, que indica o comportamento esperado de uma variável em função de outra, assume posturas manipulativas caso a consideração de outliers ou recortes seja ocultada do leitor. A Figura 5 ilustra um exemplo do uso da linha de regressão em um gráfico de dispersão.

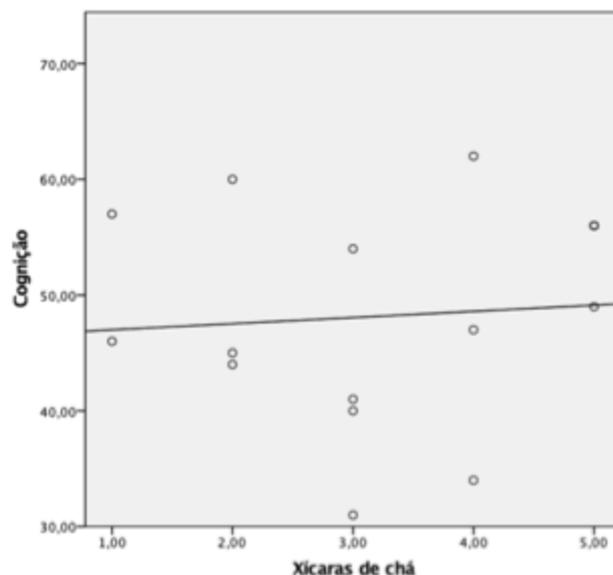


Figura 5 Simulação hipotética em que se relacionam o volume de xícaras de chá ingerido enquanto auxílio cognitivo. As circunferências dispersas na caixa indicam os valores de cada elemento amostral; a linha exibe a relação de regressão entre as variáveis; neste caso, moderadamente positivo. Embora diversos casos ilustram que há desempenhos baixos com 3 e 4 xícaras de chá, a linha de regressão indica que quanto maior o volume ingerido melhor será o desempenho cognitivo médio. A presença de outliers afeta a angulação da regressão, mostrando uma relação mais ou menos proporcionalmente relacionada entre as variáveis. Fonte: imagem dos autores elaborada utilizando a versão de avaliação do software IBM SPSS.

3 Representação gráfica sob a perspectiva de visualização de dados

Paradoxalmente, ainda que os aspectos referentes à capacidade de interpretar gráficos estatísticos componham uma disciplina isolada - a literacia visual gráfica - e relacione-se inclusive com os conteúdos fundamentais do ensino de matemática desde a educação básica, estas habilidades raramente são trabalhadas nas escolas, tanto brasileiras quanto no exterior (Börner et al., 2019; Curi & Nascimento, 2016; Deahl, 2014; Vivacqua et al., 2019). Deste modo, uma visão semelhante à perspectiva do leitor pode ser encontrada junto à interpretação de fundamentos do design visual, replicados enquanto componentes da cultura visual vigente. Aqui agregamos os conceitos de Lupton e Phillips (2008) com relação aos critérios de configuração visual de um produto gráfico. As autoras ressaltam as relações entre ponto, linhas e planos; ritmo e equilíbrio; escala; textura; cor; figura/fundo; enquadramento; hierarquia; camadas; transparência; modularidade; grid e, finalmente, padronagem como elementos básicos do design. Tomita (2015), por sua vez, enumera princípios da composição visual gráfica voltada ao melhor processo de leitura: equilíbrio, unidade,

proximidade, contraste, ênfase e alinhamento. Ainda que a utilização destes conceitos pelo leitor do material de visualização de informações seja subjetiva, é sua simplicidade que fomenta a literacia visual gráfica em indivíduos distantes do uso de dados (Bhargava & Ignazio, 2015). Utilizaremos os elementos de Lupton e Phillips (2008) e Tomita (2015) em consonância com os princípios de Bresciani e Eppler (2008, 2015) para interpretar os recursos de visualização de dados quanto a modelagens de viés.

3.1 Pontos, linhas e planos

A relação visual entre pontos, linhas e planos, dentro das concepções dos princípios do design, conduz à percepção de unidades, continuidade, união e áreas. É imprescindível que elementos intrincados aos demais sejam compreendidos considerando sua relação com o todo. Contudo, ressalta-se aqui os princípios de contraste e alinhamento.

Um ponto em contraste aos demais remarca um destaque. Interpretar um recurso de visualização de dados requer compreender o valor dos destaques apontados, pois o foco pode indicar um recorte. Linhas desalinhadas, e sua relação com a verticalidade e horizontalidade, podem consolidar um viés no raciocínio relacional, aumentando ou diminuindo a percepção de contiguidade e intensidade entre os pontos que conectam. Por sua vez, o plano pode tanto afetar a compreensão de pontos quanto a direção tomada pela linha. Em configurações enviesadas, tais planos podem ocultar elementos fundamentais à leitura dos dados, mascarando efeitos e recortes adotados pelos autores do recurso. Este desajuste quanto ao foco dos elementos, a inconsistência quanto a planos e linhas, assim como a falta de clareza entre semelhantes e discrepantes são elementos que compõem o que Bresciani e Eppler (2015) definem como uma ameaça cognitiva no âmbito da codificação do recurso visual.

3.2 Escala, ritmo e equilíbrio

A compreensão escalar, de ritmo e de equilíbrio da composição caminham juntas: ao passo que promovem uma sensação de ordenação fluida, a compreensão da escala pode atenuar ou acentuar diferenças significativas para a leitura da informação. A Figura 6 ilustra dois casos em que a escala foi utilizada para alterar a percepção dos dados: no primeiro, a percepção de crescimento dentre pontos é mais forte; no segundo, a escala considerando os graus de liberdade na amostra apresenta uma relação ínfima de progressão. A Figura 7, por sua vez, mostra como uma alteração no ritmo, condensando ou estendendo as linhas em suas respectivas escalas, também afeta a percepção da evolução do efeito.

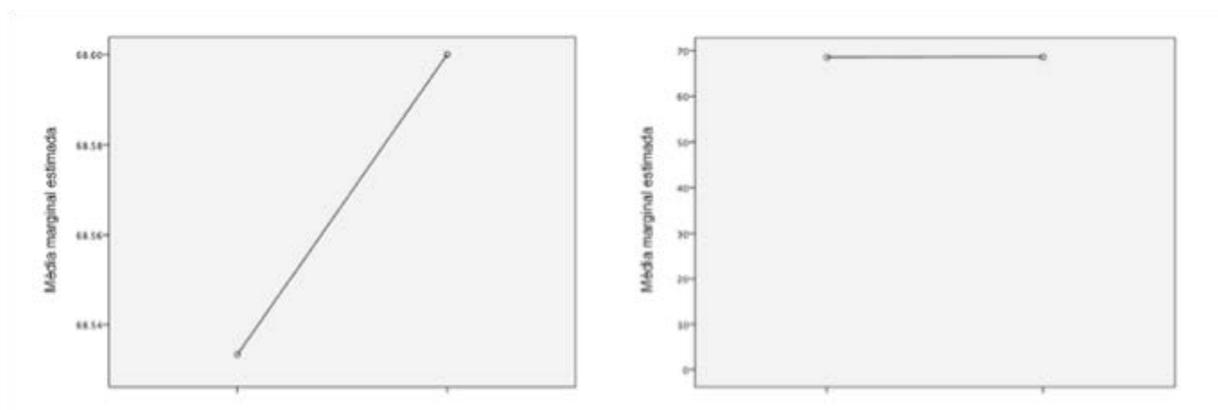


Figura 6 Alterações escalares em linhas de progressão de médias via software de análise estatística: (A) gráfico à esquerda, em que a escala é voltada para variações decimais, e (B) gráfico à direita, em que a escala é indicada pelo número de graus de liberdade da amostra. Fonte: imagem dos autores elaborada utilizando a versão de avaliação do software IBM SPSS.

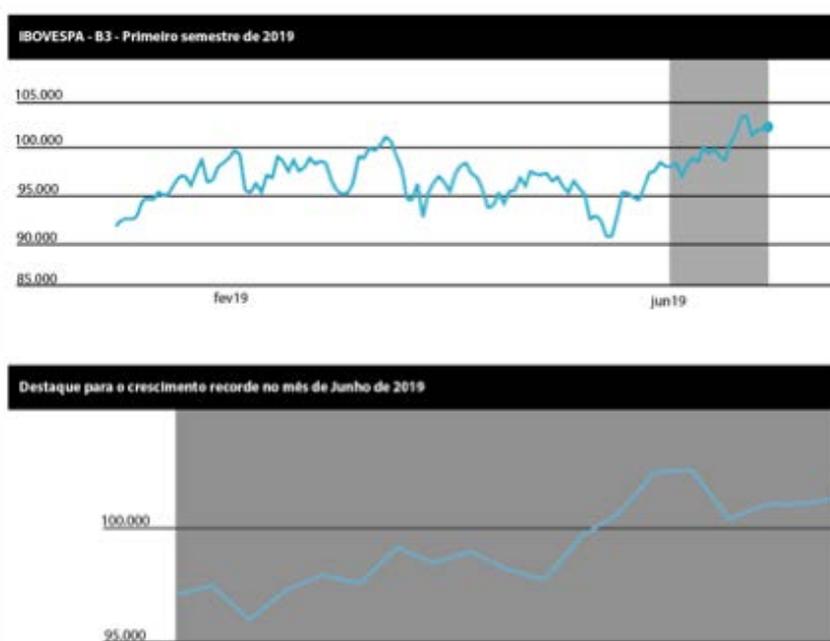


Figura 7 Evolução do índice B3 ao longo dos seis primeiros meses do ano de 2019, exibindo o crescimento da Bolsa de Valores. O destaque dado ao recorte do histórico do mês de junho – no qual o índice superou seu desempenho desde a abertura do mercado de ações contido – exibe resultados menos impressionantes em função da orientação escalar. Fonte: imagem dos autores elaborada sobre planilha de evolução do índice Bovespa oferecido pelo B3.

Alterações quanto à percepção escalar conduzem à interpretações errôneas no ritmo do desenvolvimento reportado a partir da linha ou histograma. Ainda, o foco de destaque, assim como potenciais sobrecargas visuais, tende a aumentar a relevância informacional de elementos de baixo impacto e diminuir, por conseguinte, a relevância de elementos de maior importância na composição da informação. Estes aspectos são de igual modo ameaças à apreensão cognitiva da visualização de informações. Entretanto, sua manifestação pode ser percebida no momento de decodificação da informação pelo leitor (Bresciani e Eppler, 2015).

3.3 Cor, textura, figura/fundo, camadas e transparência

Rogers (2007) ressalta que a cor é um dos elementos que mais rapidamente podem atrair a atenção do leitor. As relações de cores, texturas, camadas e transparências, em conjunto com a compreensão das relações de figura/fundo causadas pela interação entre os elementos, podem de igual maneira conduzir a uma percepção em que há a atenuação ou acentuação dos efeitos observados. Enquanto Tufte (2016) recomenda buscar a “menor diferença efetivamente perceptível” (p. 73) no que tange os contrastes, a noção voltada ao espectro cromático requer parcimônia ao se adotar progressões sutis, tanto quanto ao adotar contrastes marcados: opticamente a percepção destes elementos é modificada quando percebida em conjunto (Figura 8).

O uso de transparência como forma de apresentar camadas contrastantes de informação fomenta a comparação entre os elementos sobrepostos. Contudo, as interferências entre os elementos passam a figurar também de modo mais evidente. O ruído na informação ganha peso e apelo de elemento complementar, e deve ser observado como um potencial enviesamento, ainda que não intencional. O uso de cores e contrastes de modo errado, pouco atraentes ao leitor, repetitivos ou mesmo perturbadores, são entendidos como uma ameaça que afeta os aspectos emocionais da codificação de visualizações de informações para Bresciani e Eppler (2015), apontando para uma referência na qual a estética da visualização afeta a motivação do leitor junto ao conteúdo lido.



Figura 8 Representação circular de percentuais com cores repetidas e sem identificação adequada nos sub-níveis. A variação tonal favorece o agrupamento das categorias, mas pode influenciar a percepção das suas proporções. Fonte: imagem dos autores elaborada sobre gráfico de software de planilhas eletrônicas.

3.4 Hierarquia, enquadramento, modularidade, padronagem e grid

Relações de layout, posicionamento e enfoque constituem um sutil, porém importante, elemento na elaboração de visualizações de dados. Estas relações permitem que, ainda que a representação exata dos dados seja exibida, sua visibilidade seja alvo de maior ou menor foco. Sob este aspecto, a orientação vinda do suporte pode transportar uma relação de assimilação, direção ou mesmo temporalidade para o produto visual disponível. Considerando-se o processo virtualizador da leitura conforme Lévy (1996), é possível inferir que a adoção de estruturas conscientes deste elemento induz o leitor a um percurso no qual a informação é adquirida. Este percurso pode conferir maior ou menor importância a um determinado elemento em detrimento aos demais e, assim, configurar uma lógica de enviesamento. A disposição hierarquizada dos elementos direciona não apenas o olhar, mas, sobretudo, a zona de foco e interesse do leitor. Assim, elementos que fazem referência ao inadequado uso da percepção gestaltica, de convenções comuns, usam de ambiguidades e inconsistências, afetam igualmente a percepto-cognição no âmbito da codificação da visualização (Bresciani & Eppler, 2015).

4 A percepção visual de vieses

Se interferências de representação visual podem contribuir para a obtenção de um produto com viés inerente ainda na finalização do trabalho estatístico, o fluxo de produção do material dirigido ao leitor comum requer semelhante atenção para não propagar uma informação incongruente. Além destes pontos, e focando diretamente no recurso de visualização de dados, Tufte (2013) concebeu princípios para nortear sua elaboração que continuam relevantes atualmente. Os princípios propostos por Tufte (2013) são: comparabilidade,

causalidade, multi-variabilidade, integração de evidências e documentação. Quando relacionados com estes recursos visuais, eles podem fomentar uma maior criticidade à percepção de cada elemento gráfico de visualização de dados.

Neste texto exemplificamos caminhos nos quais a informação traduzida em gráficos pode se apresentar de modo questionável, assim como relacionamos elementos visuais que corroboram com este desalinho. Escolhas nocivas à percepção da informação, como a incapacidade de comparar medidas em função da adoção de uma escala específica, ou um conjunto cromático em que os contrastes não se fazem evidentes, podem de fato contribuir com a leitura enviesada. Da mesma maneira, a falha no reconhecimento de anormalidades, ou mesmo a falta de tratamento a dados outliers, podem corromper a análise estatística.

A capacidade de compreender as informações estatísticas a despeito dos possíveis vieses de suas representações é uma habilidade que foi ressaltada por Huff (1954) e que hoje ainda configura como imprescindível, visto o crescente volume de informações aos quais somos cotidianamente submetidos. Informes dispostos entre hoaxes, boatos, fake news, notícias e pesquisas podem ser desenvolvidas ou manipuladas de modo a ancorar a percepção de um ponto enviesado de um evento, efeito ou fato. Se, ao final, o leitor é incapaz de avaliar a robustez estatística da informação direcionada a ele, a compreensão visual dos recursos de visualização de dados pode ao menos instigar a indagação acerca do possível viés.

O olhar do leitor aos elementos que suportam a inferência ou causalidade nos gráficos, assim como as sutis presenças de elementos que não auxiliam a percepção, o que Tufte (2013) nomeia como *chart junk*, são também fontes de informação, e cabe ao leitor atentar-se a estes elementos para inferir e compreender de modo mais profundo aquilo que seus dados explicam.

5 Considerações

Destacamos a percepção da literacia visual de gráficos como uma ferramenta fundamental à percepção e análise de dados. Seu uso enquanto facilitador do reconhecimento de estruturas propícias a vieses pode ser enfatizado e, sob tal aspecto, é primordial que suas características e métodos sejam difundidos. Ainda, é de suma importância que, em conjunto às análises, todos os recursos de informação sejam contextualizados quanto a seus autores e quanto a possíveis alinhamentos entre os objetivos dos autores e aquilo que é sublinhado pelos dados. No contexto contemporâneo, entende-se que a facilitação do uso de recursos interativos pode ser empregada para favorecer a manipulação e tratamento das visualizações para que permitam um maior entendimento de seus vieses.

Embora nem sempre seja possível ter acesso à análise estatística ou à análise exploratória, é possível atentar-se às evidências de que o resultado obtido ao longo destas análises seja passível de enviesamento. Para isto, a experiência do leitor ao interpretar informações torna-se fundamental, e estes elementos de reconhecimento visual de traços de vieses faz-se útil. Contudo, é necessário sustentar que há diversas maneiras de corromper por enviesamento uma pesquisa e sua divulgação; desde as motivações e delineamentos da pesquisa, até estágios de coleta, análise e distribuição dos resultados. Estes pontos modificam a estrutura do processo inferencial e a interpretação da narrativa dos dados e gráficos que, sob diversos aspectos, contribuem para a de desinformação. A abordagem visual narrativa complementa o aspecto de representação visual tratado neste artigo e, em conjunto, buscam fomentar a melhora nos processos de aquisição de informação baseada em dados.

Finalmente, ressalta-se que a prática analítica na leitura de representações visuais de dados pode amparar a compreensão de informações além de suas fontes de vieses, e recomenda-se, neste sentido, que os parâmetros descritos sejam cotidianamente observados e falseados por leitores e analistas de dados.

6 Agradecimento

Agradecemos aos avaliadores do periódico pelo cuidado na leitura do texto e pelo detalhamento das sugestões encaminhadas. A leitura crítica do artigo contribuiu não apenas para a sua revisão, mas também para a construção do pensamento que acompanha as pesquisas subsequentes, ainda em fase de realização.

Referências

- Akamine, C. T., Yamamoto, R. K. (2013). *Estudo Dirigido de Estatística Descritiva* - 3. ed. rev. São Paulo: Érica.
- Bakker, M., Wicherts, J. M. (2014). Outlier Removal, Sum Scores, and the Inflation of the Type I Error Rate in Independent Samples t Tests: The Power of Alternatives and Recommendations. *Psychological Methods*, Vol 19(3), pp-409-427. American Psychological Association: Washington-DC.
- Bhargava, R., & Ignazio, C. D. (2015). Designing Tools and Activities for Data Literacy Learners. *Workshop on Data Literacy*, 1–5.
- Börner, K., Bueckle, A., & Ginda, M. (2019). Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences of the United States of America*, 116(6).
- Bresciani, S., & Eppler, M. J. (2008). The Risks of Visualization A Classification of Disadvantages Associated with Graphic Representations of Information.
- Bresciani, S., & Eppler, M. J. (2015). The pitfalls of visual representations: A review and classification of common errors made while designing and interpreting visualizations. *SAGE Open*, 5(4). <https://doi.org/10.1177/2158244015611451>
- Cairo, A. (2012). *The Functional Art*. Kindle Edition - São Francisco: New Riders.
- Cairo, A. (2015). *Graphics Lies, Misleading Visuals*. New Challenges for Data Design - Cap. 5, pp. 103-116. Londres: Springer.
- Cairo, A. (2016). *The Truthful Art*. Kindle Edition - São Francisco: New Riders.
- Cunha, M. B.; Cavalcanti, C. R. de O. (2008). *Dicionário de biblioteconomia e arquivologia*. Brasília: Briquet de Lemos. xvi, 451 p.
- Curi, E., & Nascimento, J. de C. P. do. (2016). O ensino de gráficos e tabelas nos anos iniciais do ensino fundamental: resultados de pesquisa nas várias instâncias curriculares. *Revista de Educação Matemática e Tecnológica Iberoamericana*, 7(1).
- Deahl, E. S. (2014). Better the Data You Know: Developing Youth Data Literacy in Schools and Informal Learning Environments. *SSRN Electronic Journal*.
- Fox, J. A., Levin, J. (2004). *Estatística para Ciências Humanas*. São Paulo: Prentice Hall.
- Lévy, P. (1996). *O Que é o Virtual?* São Paulo: Editora 34.
- Lipton, P. (1991). *Inference to the Best Explanation*. London: Routledge.
- Lupton, E., Phillips, J. C. (2013). *Novos Fundamentos do Design*. São Paulo: Cosac Naify.
- Mayr, E., Hynek, N., Salisu, S., & Windhager, F. (2019). Trust in Information Visualization. Em Kosara, R., Lawonn, K., Linsen, L., Smit, N. (Editors), *EuroVis Workshop on Trustworthy Visualization (TrustVis)*. The Eurographics Association.
- Meirelles, I. (2013). *Design for Information: an introduction to the histories, theories, and best practices behind effective information visualizations*. Beverly: Rockport.
- NIST/SEMATECH. (2013). *e-Handbook of Statistical Methods*, disponível em: <<http://www.itl.nist.gov/div898/handbook/>>, acesso em 23/06/2019. Departamento de Comércio EUA. Washington.

- Pandey, A. V., Manivannan, A., Nov, O., Satterthwaite, M., Bertini, E. (2014). The persuasive power of data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 20(12), 2211–2220.
- Reilly, S. (2017). The Need to Help Journalists with Data and Information Visualization. *IEEE Computer Graphics and Applications*, 37(2), 8–10.
- Rogers, P. J. (2007). Theory-Based Evaluation: Reflections Ten Years On. 114, 63–81.
- Rosenthal, R., Rosnow, R. L. (1989). Statistical Procedures and the Justification of Knowledge in Psychological Science. *Journal of Experimental Psychology: General*, Vol 44(10), pp. 1276-1284. American Psychological Association: Washington-DC.
- Simonsohn, U., Simmons, J. P., Nelson, L. D. (2014). P-Curve: A Key to the File-Drawer. *Journal of Experimental Psychology: General*, Vol 143(2), pp. 534-547. American Psychological Association: Washington-DC.
- Tomita, K. (2015). Principles and Elements of Visual Design: A Review of the Literature on Visual Design of Instructional Materials. *Educational Studies*, Nº 57, pp. 167-174. International Christian University: Tóquio.
- Tufte, E. (2013). *Beautiful Evidence*. Graphics Press. Cheshire-CN. 4ª impressão.
- Tufte, E. (2016). *Visual Explanations: Images and quantities, evidence and narrative*. Graphics Press. Cheshire-CN. 11ª impressão.
- Tufte, E. (2017). *Envisioning Information: Images and quantities, evidence and narrative*. Graphics Press. Cheshire-CN. 15ª reimpressão.
- Vivacqua, A. S., Dias, A. F. da S., & França, J. B. dos S. (2019). Promoting data literacy in brazilian schools. *Proceedings of the IX Latin American Conference on Human Computer Interaction*, 1–4.

Sobre os autores

Jaqueline Vasconcelos Braga

<jaqvasconcelos@gmail.com>

Mestranda, UnB, Brasil

Tiago Barros Pontes e Silva

<tiagobarros@unb.br>

Doutor, UnB, Brasil

Virgínia Tiradentes Souto

<v.tiradentes@gmail.com>

Doutora, UnB, Brasil

Editora Associada responsável/Assigned Associate Editor

Helena Barbosa

Artigo enviado em/Submission date 16/09/2020

Artigo aprovado em/Approval date 23/11/2020