Revista Latinoamericana de Metodología de la Investigación Social.
Nº11. Año 6. Abril - Septiembre 2016. Argentina. ISSN 1853-6190. Pp. 21-39.

Precisamos falar sobre métodos quantitativos em Ciência Política*

We need to talk about quantitative methods in Political Science

Dalson Britto Figueiredo Filho, Ranulfo Paranhos, José Alexandre da Silva Júnior y Denisson Silva

Resumo

Esse trabalho apresenta uma introdução à análise de dados. O foco repousa sobre a compreensão intuitiva das técnicas e a interpretação substantiva dos resultados empíricos. Nosso público alvo são estudantes de graduação, pós-graduação e pesquisadores em Ciência Política. Metodologicamente, sintetizamos as principais recomendações da literatura e empregamos simulação básica para ilustrar a utilização das seguintes técnicas: (1) análise de variância (ANOVA) para amostras independentes; (2) análise de componentes principais; e (3) análise de *cluster*.

Palavras-chaves: Métodos Quantitativos; Ciência Política; ANOVA, Análise de Cluster; Análise de Componentes Principais.

Abstract

This paper presents an introduction to data analysis. The focus relies on the intuitive understanding of techniques and the substantive interpretation of the empirical results. Our targeting audiences are undergraduate, graduate students and researchers in Political Science. On methodological grounds, we summarize the main recommendations from the literature and use basic simulation to show the application of the following techniques: (1) analysis of variance (ANOVA) for independent samples; (2) principal component analysis; and (3) cluster analysis.

Keywords: Quantitative Methods; Political Science; ANOVA, Cluster Analysis; Principal Components Analysis.

* Esse trabalho é um produto do projeto Replicabilidade Científica e Metodologia Quantitativa, desenvolvido conjuntamente pela Universidade Federal de Pernambuco (UFPE) e pela Universidade Federal de Alagoas (UFAL). Eventuais imprecisões são monopólio dos autores.

1. Introdução

The social world is exquisitely complex and rich. From the improbable moment of birth, each of our lives is governed by chance and contingency. The statistical models typically used to analyze social data are, by contrast, ludicrously simple. How simple statistical models help us to understand a complex social reality?[1]
John Fox, 2008.

Esse trabalho apresentar uma introdução intuitiva à análise de dados. Trata-se de um trabalho didático, com o foco na compreensão dos principais conceitos, técnicas e interpretação substantiva dos resultados empíricos. Nosso público alvo são estudantes de graduação e pós-graduação em Ciência Política. Metodologicamente, o desenho de pesquisa sintetiza as principais recomendações da literatura e utiliza simulação básica para ilustrar a aplicação de três diferentes técnicas: (1) análise de variância (ANOVA) para amostras independentes, (2) análise de componentes principais e (3) análise de *cluster*. Com esse artigo, esperamos ajudar os interessados no assunto a dar os primeiros passos na utilização dessas ferramentas em suas respectivas áreas de atuação.

O artigo está dividido da seguinte forma. A próxima seção define os conceitos de análise multivariada de dados, nível de mensuração, confiabilidade e validades das medidas, hipótese nula e hipótese alternativa, erros do tipo 1 e tipo 2 e discute a importância das amostras na pesquisa científica. Depois disso, o objetivo é sumarizar as principais características de diferentes técnicas de análise de dados. A última seção sumariza as conclusões.

2. O que é análise multivariada de dados?

Existe uma certa imprecisão sobre a exata definição de análise multivariada, na medida em que diferentes autores utilizam critérios distintos. Para Hair *et al*. (2009: 4) "to be considered truly multivariate, however, all the variable must be random and interrelated in such ways that their different effects cannot meaningfully be interpreted separately"[2]. São exemplos de técnicas multivariadas: análise de cluster (conglomerados), análise de componentes principais, análise de correspondência, análise de variância (ANOVA), análise fatorial, correlação múltipla, análise múltipla de covariância (ANCOVA), análise múltipla de variância (MANOVA), correlação canônica, escalonamento multidimensional, regressão linear múltipla, regressão logística, regressão multinomial, entre outras. Nesse trabalho, definimos análise multivariada como *um conjunto de técnicas estatísticas que permite a análise simultânea de duas ou mais variáveis para uma determinada amostra ou população.* A noção de variável é um conceito fundamental para pesquisa empírica e para a análise multivariada de dados. Nesse artigo definimos variável como um atributo, direta ou indiretamente mensurável, sujeito a variação quantitativa ou qualitativa.

A literatura agrupa as técnicas multivariadas em duas principais categorias: (1) dependência e (2) interdependência. As técnicas de dependência exigem a presença de uma ou mais variáveis dependentes e um conjunto de variáveis independentes. O exemplo típico em Ciência Política é a regressão linear de mínimos quadrados ordinários (MQO). Outro exemplo é a regressão logística que permite modelar a relação entre um conjunto de variáveis independentes e uma variável dependente categórica dicotômica.

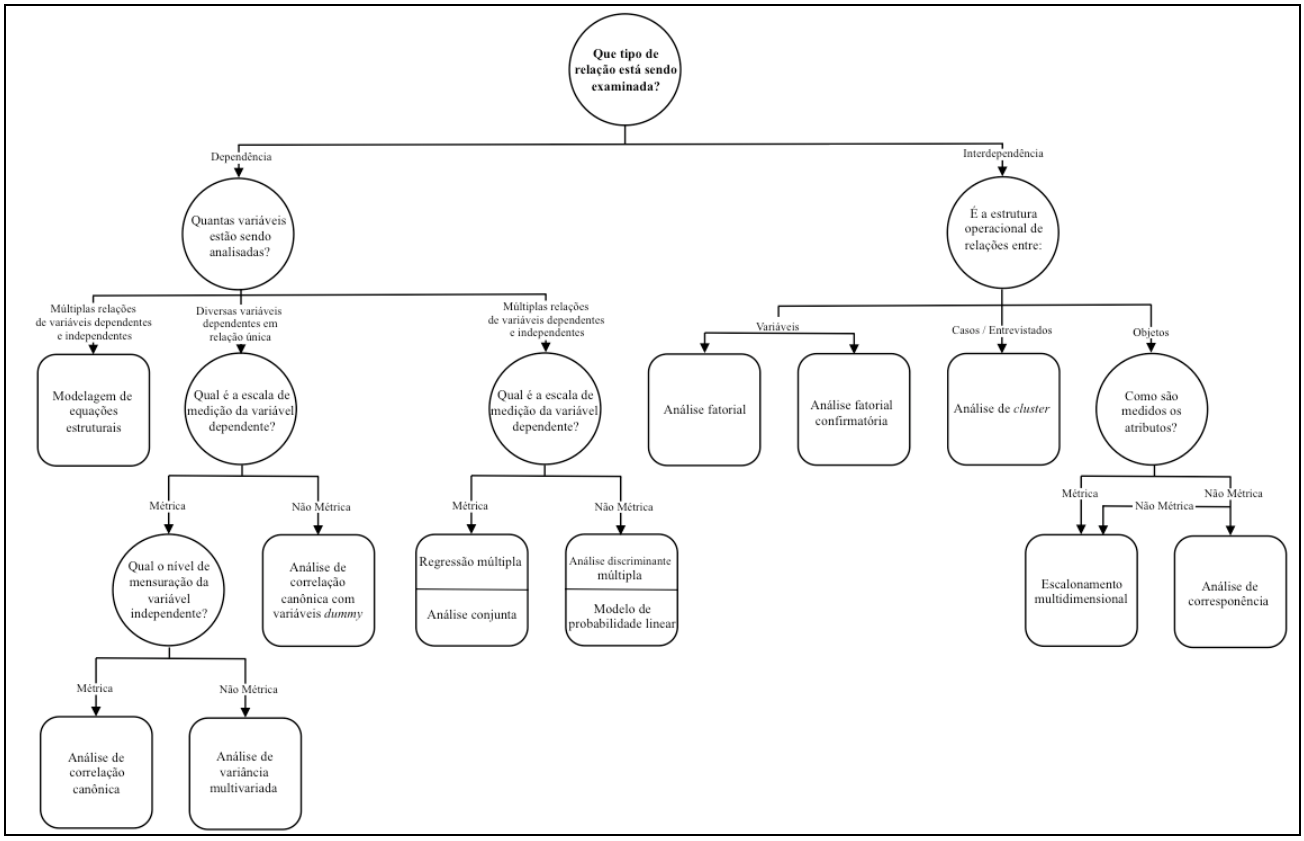
Por sua vez, as técnicas de interdependência não exigem a presença de uma variável dependente. É o caso da análise fatorial em que o pesquisador examina o padrão de correlação

[1] "O mundo social é primorosamente complexo e rico. A partir do momento improvável de nascimento, cada uma de nossas vidas é governado por acaso e contingência. Os modelos estatísticos usados normalmente para analisar os dados sociais são, pelo contrário, ridiculamente simples. Como simples modelos estatísticos nos ajudar a entender a complexa realidade social?" (Tradução própria).

[2] "Ser considerado verdadeiramente multivariada, no entanto, toda a variável deve ser aleatória e inter-relacionados em que as suas diferentes formas, tais efeitos não podem significativamente ser interpretados separadamente" (Tradução própria).

recíproca entre um conjunto de variáveis, procurando por dimensões latentes que possam resumir/explicar a variância compartilhada das variáveis originais. A figura 1 reproduz o modelo proposto por Hair *et al* (2009) para classificar as técnicas multivariadas.

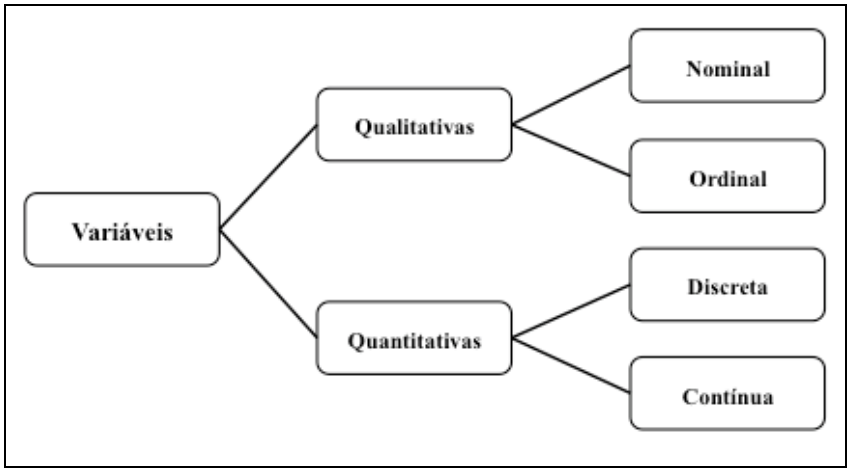
Figura 1 – Seleção da técnica multivariada



Fonte: Elaboração dos autores a partir de Hair et al, 2009.

E como selecionar a técnica multivariada mais adequada? Existem dois principais elementos que devem ser observados: (1) natureza da questão de pesquisa e (2) o nível de mensuração das variáveis. É a partir da questão de pesquisa que deve-se optar pelo tipo de técnica mais adequada aos objetivos do trabalho (dependência *versus* interdependência). O primeiro passo é transformar um conceito em um indicador empiricamente observável (operacionalização da variável). Depois deve-se definir o nível de mensuração dessas variáveis, ou seja, decidir como cada variável vai ser mensurada. A figura 2 ilustra um modelo para entender os diferentes níveis de mensuração.

Figura 2 – Nível de mensuração das variáveis



Fonte: Elaboração dos autores.

As variáveis são classificadas em dois grupos: (1) qualitativas e (2) quantitativas. As variáveis qualitativas ou categóricas não podem assumir valores numéricos e descrevem atributos de interesse em formato de categorias. Se forem independentes entre si, tem-se varáveis nominais. Quando as categorias são hierarquicamente ordenadas, tem-se variáveis ordinais. As quantitativas podem ser contínuas ou discretas. Uma variável é quantitativa contínua quando o seu atributo for um valor numérico, podendo assumir valor zero e passível de fracionamento. Uma variável quantitativa discreta também pode assumir valor zero, mas não faz sentido ser fracionada. O quadro 1 sumariza exemplos de variáveis com diferentes níveis de mensuração.

Quadro 1 – Exemplos de variáveis contínuas, discretas, ordinais e nominais

Contínua	Discreta	Ordinal	Nominal
PIB *per capita*	Número de mortes	*Ranking* de países em uma escala de corrupção	Partido político
Gasto de campanha	Número de votos em uma eleição	Nível de desenvolvimento democrático	Sistema de governo
IDH	Quantidade de cadeiras em um parlamento	Categoria geral em uma escala Likert	Tipo de candidato (*challenger* ou *incumbent*)
Taxa de reeleição	Número de partidos de uma coalizão governativa	Classificação dos países em ordem de desenvolvimento (1º, 2º e 3º mundo)	Sistema Eleitoral

Fonte: Elaboração dos autores.

A utilização correta das técnicas multivariadas depende do nível de mensuração das variáveis. Por exemplo, não é apropriado utilizar um modelo de regressão linear quando a variável dependente é nominal ou ordinal. Similarmente, variáveis quantitativas contínuas não devem ser examinadas em uma tabela cruzada de frequência. Um dos elementos centrais da análise empírica é utilizar corretamente as diferentes ferramentas de acordo com os pressupostos de cada técnica e com os propósitos substantivos da pesquisa.

Depois de definir o que é análise multivariada e discutir a importância do nível de mensuração das variáveis, o próximo passo é compreender os conceitos de confiabilidade e validade. Para Zeller e Carmines (1980), confiabilidade e validade compõem a linguagem básica da mensuração[3]. A definição clássica de Nunnally postula que "reliability concerns the extent to which measurements are repeatable – by the same individual using different measures of the same attribute or by different persons using the same measure of an attribute"[4] (Nunnally, 1967: 172). Uma forma intuitiva de entender o conceito de confiabilidade é imaginar uma balança. Se a cada vez que o mesmo indivíduo subir na balança ela apontar valores diferentes, conclui-se que o instrumento não é confiável. Em síntese, quanto maior a confiabilidade da medida, menor a quantidade de erro aleatório no processo de mensuração, logo, melhor é a qualidade da medida.

No entanto, uma medida altamente confiável não é necessariamente um bom indicador do conceito/fenômeno de interesse. Define-se validade como o grau de correspondência entre o que se mediu e o que se queria medir. Everitt e Skrondal definem validade como "the extent to which a measuring instrument is measuring what was intended"[5] (Everitt e Skrondal, 2010: 365). Nesse artigo entendemos validade como a adequação do instrumento ou a pertinência dos resultados por ele produzidos para medir aquilo que se pretende.

Depois de entender a importância da confiabilidade e da validade no processo de mensuração, o próximo passo é compreender as noções de hipótese nula, hipótese alternativa e erros do tipo 1 e tipo 2. Como regra, a hipótese nula (H_o) postula que não existe relação entre as variáveis ou diferença entre os grupos estudados. Por exemplo, em uma pesquisa sobre a relação entre renda (*x*) e escolaridade (*y*), a hipótese nula assume que não existe relação entre essas variáveis. Contrariamente, a hipótese alternativa postula que existe relação entre renda e escolaridade. O erro do tipo 1 consiste em rejeitar a hipótese nula (H_o) quando ela não deveria ser rejeitada. Ou seja, não existe relação entre *x* e *y* ($r=0$), mas o pesquisador chega à conclusão de que as variáveis estão correlacionadas ($r\neq0$). O erro do tipo 2 consiste em não rejeitar a hipótese nula (H_o) quando ela deveria ser rejeitada. Ou seja, existe relação entre *x* e *y* ($r\neq0$), mas o pesquisador chega à conclusão de que as variáveis são estatisticamente independentes ($r=0$).

Alguns exemplos do cotidiano ajudam a melhor compreender os erros do tipo 1 e tipo 2. Imagine um árbitro auxiliar (bandeirinha) em uma partida de futebol. Ele deve julgar se um determinado jogador está ou não em posição de impedimento. O primeiro passo para identificar os erros do tipo 1 e 2 é escrever as hipóteses nula e alternativa. Nesse exemplo, a hipótese nula postula que *o jogador está em posição legal*, enquanto a hipótese alternativa supõe que *o jogador está em impedimento*. O erro do tipo 1 consiste em rejeitar a hipótese nula quando ela é verdadeira, ou seja, marcar o impedimento quando a posição do jogador era legal. O erro do tipo 2 consiste em não rejeitar a hipótese nula quando ela é falsa, ou seja, não marcar o impedimento quando a posição era ilegal[6].

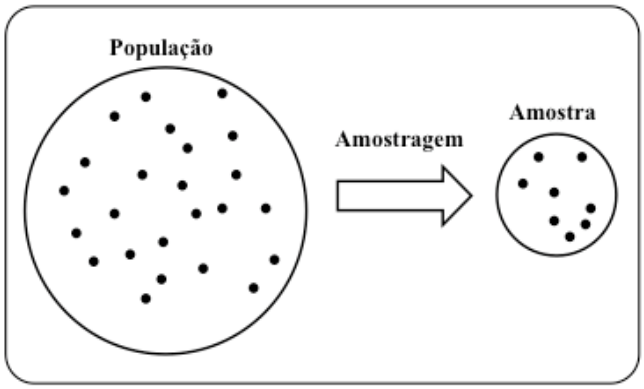
Depois de compreender os conceitos de hipótese nula, alternativa, erros do tipo 1 e 2, o próximo passo é entender a importância das amostras na pesquisa científica. Por que utilizar amostras? Resposta: (1) economia de tempo e (2) economia de recursos. Em geral, catalogar informações sobre todas as observações do universo (censo) pode inviabilizar alguns desenhos de pesquisa. Por exemplo, suponha que uma pesquisa tem como objetivo examinar a intenção de voto em um determinado candidato à Presidência do Brasil. Logisticamente, não faz sentido entrevistar todos os eleitores brasileiros. A pesquisa seria demasiadamente onerosa e demorada. Além disso, o esforço computacional necessário para trabalhar com amostras é menor do que aquele empregado para analisar grandes bases de dados.

Os pesquisadores utilizam amostras para realizar inferências válidas e confiáveis para a população. Lembrando que o conceito de população diz respeito a totalidade de

[3] Para uma introdução à mensuração em Ciências Sociais ver: Zeller e Carmines, 1980.
[4] "Confiabilidade diz respeito ao grau em que as medições são repetíveis - pelo mesmo indivíduo usando diferentes medidas do mesmo atributo ou por pessoas diferentes, utilizando a mesma medida de um atributo". (Tradução própria).
[5] "A medida em que um instrumento de medição é medir o que se pretendia" (Tradução própria).
[6] Existe ainda o erro do tipo 3 que consiste na discrepância entre o foco do trabalho e a questão de pesquisa (Schwartz e Carpenter, 1999).

indivíduos/unidades, enquanto que a amostra refere-se a uma parte da população. A inferência, por sua vez, é o processo pelo qual o pesquisador obtém informações válidas para a população a partir da análise de dados amostrais. A figura 3 ilustra a relação população, amostragem e amostra.

Figura 3 – População, amostragem e amostra



Fonte: Figueiredo Filho *et al,* 2013.

Para que as estimativas amostrais sejam representativas dos parâmetros populacionais, é necessário garantir a aleatoriedade da amostra. Apenas amostras aleatórias garantem que o princípio da equiprobabilidade, ou seja, todos os indivíduos da população tem a mesma chance de participar da amostra. Tecnicamente, a seleção aleatória da amostra tende a garantir a qualidade das estimativas. Qualidade no sentido de assegurar não viesamento e baixa variabilidade. Uma estimativa é não viesada quando ela nem sobreestima nem subestima sistematicamente o valor do parâmetro populacional. A eficiência diz respeito à variabilidade da estimativa: quanto maior a variabilidade, menor a precisão, pior é a estimativa.

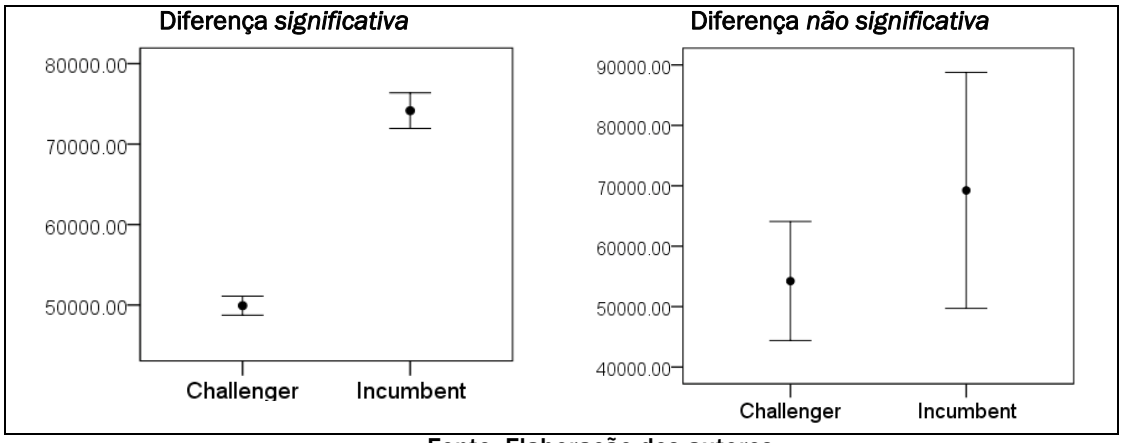
Depois de apresentar a importância das amostras, o próximo passo é descrever a lógica da inferência. Em particular, a inferência estatística é uma ferramenta essencial ao desenvolvimento do conhecimento científico. Independente da área de estudo, é exatamente a inferência estatística que permite utilizar informações limitadas sobre os fatos/fenômenos conhecidos para fazer inferências válidas a respeito de fatos/fenômenos desconhecidos. Na análise multivariada de dados, a utilização da inferência estatística é fundamental para explorar questões desconhecidas, descrever fenômenos e/ou testar hipóteses teoricamente orientadas.

2.1. Análise de variância (ANOVA) para amostras independentes[7]

A ANOVA pode ser melhor compreendida ao se analisar o caso em que a variável independente tem apenas duas categorias ou condições experimentais. Por exemplo, o pesquisador pode estimar a diferença dos gastos de campanha entre *challengers* e *incumbents*. Primeiro, deve-se calcular as médias de cada grupo e depois compará-las. Provavelmente elas serão diferentes, mas essas diferenças podem surgir apenas por variação aleatória. O objetivo principal é saber se essas médias diferem na população que elas foram extraídas. Depois de detectar a diferença entre as médias, o próximo passo é decidir a respeito da sua importância substantiva. Para ilustrar o funcionamento desse raciocínio, simulamos como a variável gastos de campanha varia entre *challengers* e *incumbents*. A figura 4 ilustra essas informações.

Figura 4 – Gastos de campanha (simulação)

[7] Iversen e Norpoth (1987) afirmam que análise de variância é uma denominação inadequada para um conjunto de técnicas estatísticas e modelos para estimar diferenças de médias entre diferentes grupos ou condições experimentais. Nesse trabalho utilizamos comparação de médias, análise de variância e ANOVA como sinônimos. Para um trabalho clássico ver Iversen e Norpoth (1987). Para uma introdução bem humorada ao assunto ver: https://www.youtube.com/watch?v=NKgUPxb9-iw

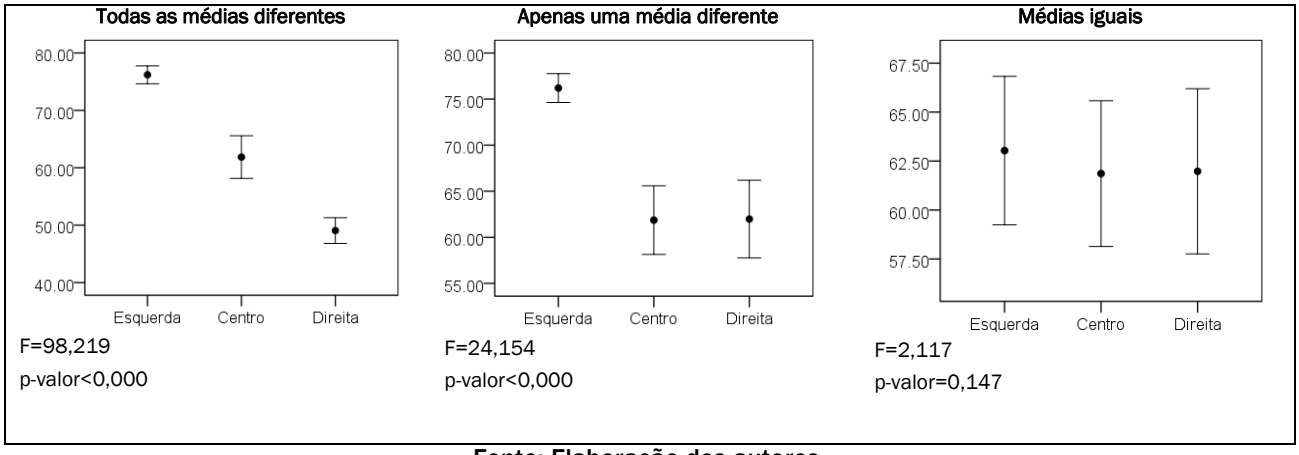


Fonte: Elaboração dos autores.

O gráfico da esquerda ilustra a situação em que a diferença de médias entre *challengers* e *incumbents* é significativa. Em nossa simulação a variável gasto de campanha tem distribuição normal para ambos os grupos. A média de gasto dos *challengers* é de R$ 50.000,00 com desvio padrão de R$ 5.000 e para os *incumbents* a média é de R$ 75.000,00 com desvio padrão de R$ 7.500,00. Considerando a comparação *challengers vs incumbentes* observa-se que média dos *challengers* é menor que a dos *incumbents*. O mais importante: não existe interseção entre os intervalos de confiança de cada grupo. Assim, é possível afirmar que a diferença entre os grupos é estatisticamente significativa (n = 100; dif = 24.229,68; t= -19,41; gl = 74,91; p-valor<0,000). Por sua vez, ao considerar a comparação em que não existe diferença significativa entre os grupos, novamente a média dos *challengers* é menor. No entanto, existe uma interseção entre os intervalos dos grupos, o que pode colocar em xeque a significância estatística da diferença de médias (n = 100; dif = 15.000,51; t= -1,378; gl = 72,39; p-valor = 0,172). Quanto maior a interseção, maior é a probabilidade de não rejeitar a hipótese nula de igualdade entre as médias.

A análise de variância é uma extensão da comparação de médias quando o pesquisador quiser comparar mais de dois grupos/condições experimentais com o objetivo de verificar se existe alguma diferença estatisticamente significativa entre essas médias. Essa técnica requer uma variável dependente quantitativa (discreta ou contínua) e uma variável independente categórica com pelo menos três níveis/categorias. Essas categorias correspondem a diferentes grupos ou condições experimentais. Por exemplo, ao se comparar a disciplina partidária dos deputados (variável dependente), o fator pode ser a ideologia (esquerda, centro e direita). Para facilitar a compreensão da lógica subjacente à ANOVA, a figura 5 ilustra a distribuição de três diferentes amostras.

Figura 5 – Distribuição dos escores de três diferentes amostras



Fonte: Elaboração dos autores.

O gráfico da esquerda ilustra o cenário em que todas as médias são diferentes (μ_e = 75 {10}, μ_c = 62 {23} e μ_d = 48 {15}). A inspeção gráfica revela que não existe interseção entre os intervalos de confiança, logo, deve-se concluir que a diferença entre as médias é estatisticamente significativa (F = 98,219 e p-valor<0,000). O gráfico do centro ilustra o cenário em que apenas uma das médias é estatisticamente diferente das demais. Aqui o pesquisador deve estar atento ao exame gráfico dos resultados já que a tabela informa apenas que um dos grupos é diferente (F = 24,154 e p-valor<0,000). Por fim, tem-se o gráfico em que não existe diferença estatisticamente significativa entre as médias, ou seja, elas devem ser consideradas iguais (F = 2,117 e p-valor = 0,147).

A hipótese nula postula que $\mu_e = \mu_c = \mu_d$, ou seja, a média da disciplina partidária é a mesma para os três grupos de ideologia (*esquerda, centro* e *direita*). Por sua vez, a hipótese alternativa postula que as médias são diferentes ($\mu_e \neq \mu_c \neq \mu_d$). A análise de variância compara dois grupos de variâncias. Uma primeira estimativa é calculada a partir da diferença *entre* os grupos (*between group variance*) e é considerada como reflexo da diferença entre os grupos ou efeito da variável independente. A segunda estimativa é calculada a partir da variância dentro de cada grupo (*within group variance*) e é considerada como reflexo do acaso. A diferença entre essas variâncias é medida por uma razão, tendo o numerador a variância entre os grupos e o denominador a variância dentro dos grupos e segue uma distribuição F. Quanto maior a estatística *F,* maior é a diferença entre a variância entre os grupos e a variância dentro dos grupos, ou seja, maior é o grau de confiança do pesquisador em rejeitar a hipótese nula. Todavia, o teste F informará apenas se o modelo ajustado é significativamente melhor do que o modelo nulo. Em outras palavras, o teste da ANOVA informará apenas se alguma das médias é estatisticamente diferente. Para saber que grupos são diferentes e/ou a magnitude dessas diferenças o pesquisador deve explorar os testes de comparação múltiplas. O quadro 2 apresenta um exemplo de pesquisa utilizando ANOVA para amostras independentes.

Quadro 2 – Exemplo de pesquisa com ANOVA para amostras independentes[8]

O que você precisa	Uma variável independente categórica com pelo menos três categorias/níveis e uma variável dependente quantitativa (contínua ou discreta)
Para que serve	Informa se existem diferenças estatisticamente significativas entre os grupos/níveis da variável independente e/ou condições experimentais
Pressupostos	As variáveis são aproximadamente normais Homocedasticidade (mesma variância) As amostras são aleatórias As observações são independentes

Fonte: Elaboração dos autores.

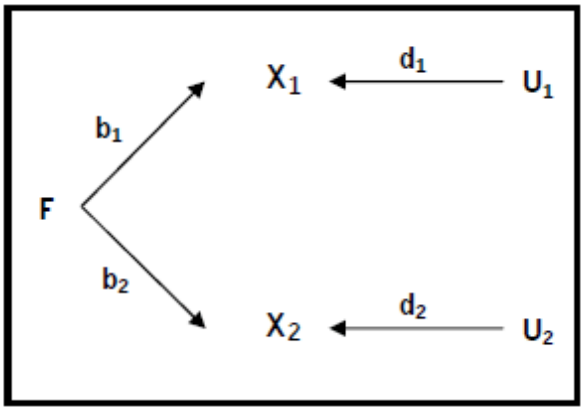
2.2. Análise de componentes principais[9]

A principal função das diferentes técnicas de análise fatorial é reduzir uma grande quantidade de variáveis observadas em um número menor de fatores. Fator é a combinação linear das variáveis (estatísticas) originais. A figura 6 ilustra a relação entre variáveis diretamente observadas e os seus respectivos fatores.

Figura 6 – Modelo das vias para duas variáveis, modelo de um fator

[8] Para solicitar uma ANOVA no SPSS, o leitor deve seguir o seguinte caminho: *Analyze→Compare Means→One Way* ANOVA.No STATA, *Statistics→Linear Model and Related→* ANOVA/MANOVA *→One way* ANOVA.

[9] Essa seção foi elaborada a partir do Figueiredo Filho e Silva Júnior, 2010.



Fonte: Asher, 1983.

X_1 e X_2 são variáveis observadas: X_1 é causado por F e por U_1, já X_2 é causado por F e por U_2. Como F é comum a X_1 e X_2 ele é considerado um fator comum. Contrariamente, tanto U_1 quanto U_2 são considerados fatores únicos já que são restritos a X_1 e X_2, respectivamente (Asher, 1983). Para Kim e Mueller (1978), "a análise fatorial se baseia no pressuposto fundamental de que alguns fatores subjacentes, que são em menor número que as variáveis observadas, são responsáveis pela covariação entre as variáveis" (Kim e Mueller, 1978: 12). Nesse exemplo F, U_1 e U_2 são considerados fatores (não podem ser diretamente observados) enquanto que X_1 e X_2 são as variáveis que o pesquisador observa diretamente.

A literatura identifica duas principais modalidades de análise fatorial: exploratória e confirmatória. A análise fatorial exploratória (AFE) geralmente é utilizada nos estágios embrionários da pesquisa, no sentido de explorar os dados. Nessa fase, procura-se explorar a relação entre um conjunto de variáveis, identificando padrões de correlação. Além disso, a AFE pode ser utilizada para criar variáveis independentes ou dependentes que podem ser empregadas posteriormente em modelos de regressão. Por sua vez, a análise fatorial confirmatória (AFC) é utilizada para testar hipóteses. Nesse caso, o pesquisador guiado por alguma teoria testa em que medida determinadas variáveis são representativas de um conceito/dimensão. O quadro 3 sintetiza o planejamento de uma análise fatorial em três estágios.

Quadro 3 – Planejamento da análise fatorial em três estágios

Procedimento	O que deve ser observado
Verificar a adequabilidade da base de dados	Nível de mensuração das variáveis, tamanho da amostra, razão entre o número de casos e quantidade de variáveis e o padrão de correlação entre as variáveis.
Determinar a técnica de extração e o número de fatores a serem extraídos	O tipo de extração (*principal components, principal factors, image factoring, maximum likelihood factoring, unweighted least squares, generalized least squares*).
Decidir o tipo de rotação dos fatores	Se for ortogonal (*Varimax, Quartimax, Equamax*), se for oblíqua (*Direct Oblimin, Promax*)

Fonte: Elaboração dos autores.

O primeiro estágio é *verificar a adequabilidade da base de dados*. Em relação ao nível de mensuração, a literatura mais conservadora recomenda apenas a utilização de variáveis contínuas ou discretas. Hair *et al* (2005) desaconselham a utilização de variáveis categóricas, mas caso seja necessário, recomendam a inclusão de variáveis *dummies.* Já King (2001) adverte que

determinadas variáveis como sexo e cor nunca devem ser incluídas em um modelo de análise fatorial já que é improvável que algum fator influencie a sua variação. Dessa forma, além dos critérios técnicos é necessário considerar teoricamente como os fatores se relacionam com as variáveis observadas.

Em relação ao número de casos, quanto maior, melhor. Hair *et al* (2005) sugerem que a amostra deve ser superior a 50 observações, sendo aconselhável no mínimo 100 casos para assegurar resultados mais robustos. A razão entre o número de casos e a quantidade de variáveis deve exceder cinco para um ou mais (Hair *et al*, 2005).

Quanto ao padrão de correlação entre as variáveis, a matriz de correlações deve exibir a maior parte dos coeficientes com valor acima de 0,30 (independente do sinal). O teste de Kaiser-Meyer-Olklin (KMO) varia entre 0 e 1. Quanto mais perto de 1, melhor. Pallant (2007) sugere 0,6 como limite razoável. Hair *et al* (2005) sugerem 0,50 como patamar aceitável. Por fim, a estatística *Bartelett Test of Spherecity* (BTS) deve ser estatisticamente significativa ($p<0,05$).

O segundo estágio é determinar a técnica de extração dos fatores (componentes principais, fatores principais, fatoração por imagem; fatoração por verossimilhança; fatoração alfa; mínimos quadrados não ponderados; mínimos quadrados). Aqui vale destacar a diferença entre análise de componentes principais (ACP) e análise fatorial (AF). Ambas as técnicas produzem combinações lineares de variáveis que capturam o máximo possível da variância das variáveis observadas. Na ACP *toda* a variância é utilizada. Na AF apenas a variância *compartilhada*. Na maioria dos casos tanto a ACP, quanto a AF, produzem resultados semelhantes quando o número de variáveis superar 30 e/ou se as comunalidades excederem 0,60 para a maior parte das variáveis.

Apesar de não existir um critério consensual para definir quantos fatores devem ser extraídos, a literatura aponta alguns métodos que auxiliam o pesquisador. Por exemplo, a regra do *eigenvalue* (critério de Kaiser) sugere que devem ser extraídos apenas os fatores com valor acima de um. Isso porque se o fator apresenta baixo autovalor, ele está contribuindo pouco para explicar a variância nas variáveis.

O pesquisador também pode utilizar o critério da variância acumulada para determinar a quantidade de fatores que devem ser extraídos. Hair *et al* (2005) sugerem o patamar de 60% como sendo aceitável. Dessa forma, a extração dos fatores deve continuar até que o referido patamar seja alcançado. Por fim, no caso da análise fatorial confirmatória, além dos critérios estatísticos também é importante apresentar razões teóricas para justificar a extração dos fatores. Nesse sentido, o pesquisador deve justificar em termos conceituais qual é o padrão de relação esperado entre as variáveis observadas e os fatores.

Depois de *verificar a adequabilidade da base de dados* e *determinar a técnica de extração e o número dos fatores,* o pesquisador deve seguir para o terceiro estágio: *decidir o tipo de rotação dos fatores.* O principal objetivo da rotação dos fatores é facilitar a interpretação dos resultados observados. As rotações ortogonais são mais fáceis de reportar e de interpretar. No entanto, o pesquisador deve assumir que os construtos são independentes. Já as rotações oblíquas permitem que os fatores sejam correlacionados. Todavia, são mais difíceis de interpretar. Em geral, as duas formas de rotação produzem resultados bastante semelhantes, principalmente quando o padrão de correlação entre as variáveis é claro. O tipo de rotação ortogonal *Varimax* é o mais utilizado e minimiza o número de variáveis que apresentam altas cargas em cada fator. O quadro 4 sintetiza um exemplo de um desenho de pesquisa que utiliza a referida técnica.

Quadro 4 – Exemplo de um desenho de pesquisa com análise de componentes principais

O que você precisa	Variáveis contínuas e/ou discretas, correlacionadas entre si (r≥0,3)
Para que serve	Reduzir uma grande quantidade de variáveis observadas em um número menor de fatores
Pressupostos	Assume a existência de correlações confiáveis entre a maior parte de variáveis incluídas na análise; A consistência do modelo é afetada por casos ausentes, *outliers* e truncamento de dados; Normalidade univariada e/ou multivariada melhoram a confiabilidade do modelo.

Fonte: Elaboração dos autores.

Para melhor compreender o funcionamento dessa técnica, simulamos cinco variáveis correlacionadas entre si com diferentes níveis de associação. A tabela 1 ilustra essas informações.

Tabela 1 - Matriz de correlação (N=300)

Variáveis observadas	Processo eleitoral e pluralismo	Funcionamento do governo	Participação política	Liberdade civil	Cultura política
Processo eleitoral e pluralism	1,000	0,900 (0,000)	0,800 (0,000)	0,700 (0,000)	0,200 (0,000)
Funcionamento do governo		1,000	0,072 (0,000)	0,630 (0,000)	0,180 (0,002)
Participação política			1,000	0,560 (0,000)	0,160 (0,005)
Liberdade civil				1,000	0,140 (0,015)
Cultura política					1,000

Fonte: Elaboração dos autores. KMO: 0,790

BTS: 1.007,077 (p<0,000)

O primeiro passo é observar o padrão de correlação entre as variáveis. Quanto maior o nível de correlação entre os indicadores observados, mais adequadas são as técnicas de redução de dados. A exceção da *cultura política*, as demais variáveis apresentam r>0,300, o que por sua vez sugere que elas podem ser utilizadas para os nossos propósitos. Os testes de adequação da amostra KMO (0,790) e BTS (p-valor<0,000) também sugerem que a matriz de dados é adequada. A tabela 2 apresenta as comunalidades dos modelo estimado.

Tabela 2 - Comunalidades

Variáveis observadas	Inicial	Extração
Processo eleitoral e pluralismo	1,000	0,920
Funcionamento do governo	1,000	0,842
Participação política	1,000	0,747
Liberdade civil	1,000	0,637
Cultura política	1,000	0,075

Fonte: Elaboração dos autores.

Além do padrão de correlação, deve-se analisar as comunalidades de cada variável. Como pode ser observado, a exceção de *cultura política*, todas as variáveis apresentam comunalidades acima de 0,4. Como a contribuição da cultura política é muito reduzida o pesquisador deve cogitar retirá-la da análise e estimar um novo modelo.

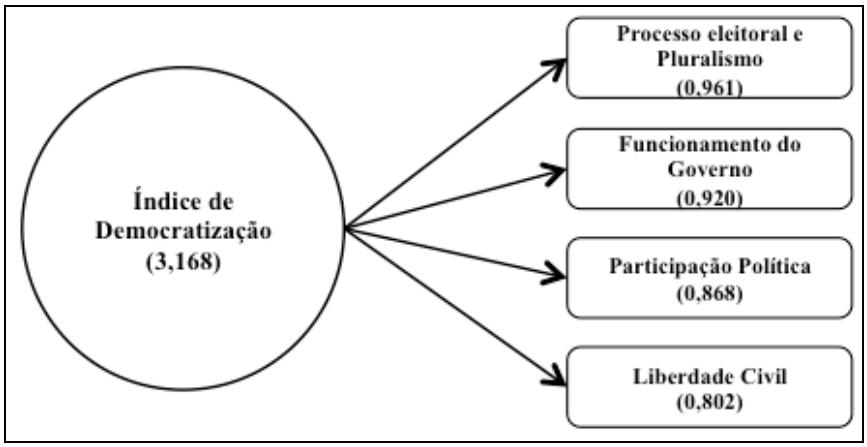
Tabela 3 – Variância total explicada

	Inicial Eigenvalues			*Extração da soma dos quadrados das cargas*		
Componentes	Total	% de variância	% cumulativo	Total	% de variância	% cumulativo
1	3,221	64,410	64,410	3,221	64,410	64,410
2	0,948	18,960	83,370			
3	0,462	9,249	92,619			
4	0,286	5,723	98,342			
5	0,083	1,658	100,000			

Fonte: Elaboração dos autores.

O modelo inicial (com 5 variáveis) carregou 64,410% da variância das variáveis originais. Com a exclusão da variável *cultura política*, esse percentual passou para 79,194%. Dado o ganho informacional, deve-se optar pelo segundo modelo (com 4 variáveis). O novo modelo apresentou KMO de 0,784 com BTS estatisticamente significativo. As comunalidades de cada variável aumentaram (*processo eleitoral e pluralismo* de 0,920 para 0,924; *funcionamento do governo* de 0,842 para 0,847; *participação política* de 0,747 para 0,753 e *liberdade civil* de 0,637 para 0,643). A figura 7 ilustra os valores dos componentes da matriz para cada variável incluída no modelo final.

Figura 7 – Componentes da matriz



Fonte: Elaboração dos autores.

Os valores dos componentes da matriz ilustram a correlação entre cada variável e o componente extraído. O componente tem um autovalor de 3,168 e responde por 79,194% da variância das variáveis originais.

2.3. Análise de cluster (conglomerados)[10]

A análise de *cluster* é uma denominação genérica para um grande grupo de técnicas que podem ser utilizadas para criar uma classificação. Esses procedimentos formam empiricamente *clusters* ou grupos de objetos fortemente similares. Para Hair *et al* (2005), a "análise de conglomerados agrupa indivíduos ou objetos em *clusters* de modo que objetos em um mesmo *cluster* são mais parecidos entre si do que em relação a outros *clusters*" (Hair *et al*, 2005: 555). O principal objetivo da análise de conglomerados é agrupar casos a partir de características que os tornam similares. Para tanto, a análise de conglomerados minimiza a variância dentro do grupo (*within group variance*) e maximiza a variância entre os grupos (*between group variance*).

Uma forma intuitiva de compreender a lógica da análise de conglomerados é imaginar a organização de um supermercado. Em geral, itens semelhantes são agrupados em um mesmo setor: cerveja, vinho e refrigerantes se agrupam no setor de bebidas. Banana, maçã e laranja se agrupam no setor de hortifrutigranjeiro, etc. O principal objetivo da referida técnica é agrupar casos de acordo com o grau de semelhança observado entre eles. Hair *et al* (2005) afirmam que a lógica subjacente à análise de *cluster* é semelhante à lógica da análise fatorial. A diferença básica é que, na análise fatorial, o pesquisador está interessado em representar um conjunto de variáveis observadas a partir de um número menor de fatores. Já na análise de conglomerados o pesquisador procura representar um conjunto de casos a partir de um número menor de grupos (*clusters*). Em uma frase: na análise fatorial, agrupam-se variáveis, na análise de conglomerados, agrupam-se casos. Tecnicamente, o planejamento de uma análise de conglomerados deve seguir cinco estágios: (1) seleção da amostra; (2) escolha das variáveis; (3) definição das medidas de similaridades e métodos de aglomeração; (4) delimitação do número de *clusters*/grupos e (5) validação dos resultados.

O primeiro passo é definir a amostra. Para Hair *et al* (2005), o tamanho da amostra na análise de *cluster* não se relaciona com questões de inferência estatística como em análise de regressão, por exemplo. Ou seja, não se procura estimar em que medida os resultados encontrados na amostra podem ser estendidos à população. Na verdade, o tamanho da amostra deve garantir que os pequenos grupos da população sejam devidamente representados. Além disso, diferente de outras técnicas multivariadas, não existe uma regra geral para especificar o tamanho mínimo da amostra. Nossa recomendação é que ao se elevar a quantidade de variáveis

[10] Essa seção foi elaborada a partir do Figueiredo Filho, Silva Junior e Rocha, 2012.

deve-se aumentar também o número de casos. Outro procedimento importante é a identificação de *outliers*. A presença de casos desviantes pode distorcer a verdadeira estrutura dos dados, produzindo grupos (*clusters*) não representativos[11].

O segundo estágio consiste em decidir que variáveis serão utilizadas para estimar a distância/similaridade entre os casos. Como a análise de *cluster* não diferencia entre variáveis relevantes e irrelevantes, é necessário que essa inclusão seja teoricamente orientada. Hair *et al* (2005) afirmam que devem ser incluídas apenas as variáveis que caracterizem os objetos que serão agrupados e se relacionem especificamente aos objetivos da análise de *cluster*. A inclusão de muitas variáveis dificulta a interpretação substantiva dos resultados. Por esse motivo, o pesquisador deve incluir variáveis que sejam ao mesmo tempo teoricamente relevantes e tenham poder prático de discriminar os grupos de acordo com o fenômeno estudado.

Quanto ao nível de mensuração, Hair *et al* (2005) destacam as medidas correlacionais e as medidas de distância. As correlacionais permitem trabalhar com variáveis categóricas, já as de distância exigem variáveis métricas. Outro ponto importante diz respeito à padronização das variáveis. Alguns especialistas recomendam que variáveis medidas em diferentes escalas devem ser padronizadas (média zero e variância igual a um) para que a comparação entre elas seja inteligível. O problema da ponderação (criar pesos) também divide a opinião dos pesquisadores.

O terceiro estágio consiste em definir a medida de similaridade. Recomendamos que pesquisadores iniciantes utilizem as medidas de similaridade mais convencionais, incorporando diferentes medidas ao longo do seu processo de aprendizado. Uma vez calculada a similaridade, o próximo passo é decidir o método (algoritmo matemático) de aglomeração. Existem três principais abordagens para criar os conglomerados: a) *Hierarchical clustering* (agrupamento hierárquico); b) *K-means clustering* e c) *Two step clustering*.

O agrupamento hierárquico (HCA) é mais apropriado para amostras pequenas (n<250). Na medida em que o tamanho da amostra cresce, a solução do algoritmo tende a ficar mais lenta, podendo, inclusive, travar o computador. Os *clusters* são aninhados, ou seja, não são mutuamente exclusivos. O pesquisador pode escolher a amplitude do número de *clusters* ou a quantidade exata de grupos que devem ser criados. A opção *K-means clustering* é mais indicada para amostras maiores (n>1.000) já que ela não computa a matriz de proximidade de distâncias/similaridade entre todos os casos. Como medida de similaridade, a abordagem *K-means clustering* utiliza a distância Euclidiana e o pesquisador deve especificar antecipadamente o número de grupos (conglomerados) que serão formados (Garson, 2008). A abordagem *Two step clustering* é considerada ideal para grandes bases de dados, já que tanto o agrupamento hierárquico quanto a *K-means clustering* podem apresentar problemas de escalonamento quando a amostra é demasiadamente grande.

O quarto estágio consiste em identificar o número de grupos (K) que serão formados. Deve-se utilizar a teoria para orientar essa escolha. Por exemplo, se trabalhos anteriores sugerem a existência de três grupos, uma possibilidade analítica é replicar o número de grupos com o objetivo de verificar em que medida a solução encontrada é mais ou menos robusta. Na ausência de teoria sobre o assunto, o pesquisador pode adotar uma perspectiva exploratória e repetir a análise variando o número de grupos (K). As diferentes soluções devem ser comparadas à luz da literatura especializada sobre o tema em busca de uma explicação substantiva.

Por fim, o 5º estágio, consiste na validação dos resultados. A validação consiste em garantir que a solução encontrada seja representativa da população, descrevendo um padrão relativamente estável para outras amostras. Um procedimento para executar a validação consiste no particionamento (divisão) da amostra original em outras separadas e comparar as soluções obtidas em ambos os casos, verificando a correspondência dos resultados (Hair *et al,* 2005). Outro caminho é testar a capacidade preditiva da solução gerada a partir da comparação de uma variável aleatória que não tenha sido utilizada na solução inicial de geração dos conglomerados.

[11] Hair et al (2009) sugerem a inspeção do diagrama de perfil para identificar eventuais *outliers*.

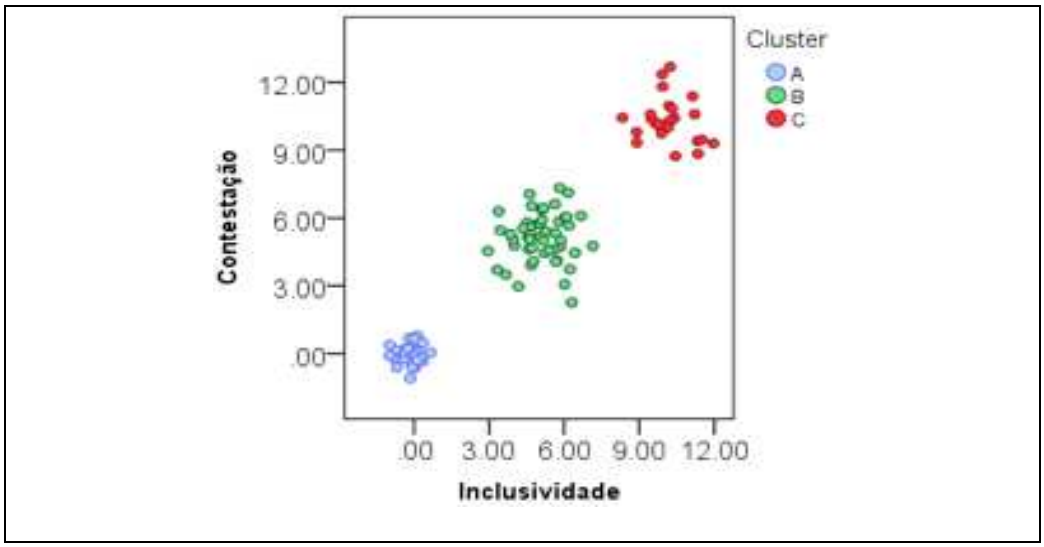
Quadro 5 – Exemplo de um desenho de pesquisa com análise de *cluster*

O que você precisa	Variáveis categóricas ou métricas a depender do método de aglomeração.
Para que serve	Classificar casos em grupos.
Pressupostos	Representatividade da amostra; Ausência de altos níveis de multicolinearidade entre as variáveis

Fonte: Elaboração dos autores.

Para ilustrar a aplicabilidade da análise de cluster optamos por replicar a tipologia proposta por Dahl (1971) para representar as dimensões da Poliarquia: contestação e inclusividade. A figura 8 ilustra a distribuição simulada da *constestação* e *inclusividade* para 100 diferentes países.

Figura 8 – Contestação e inclusividade (simulação)

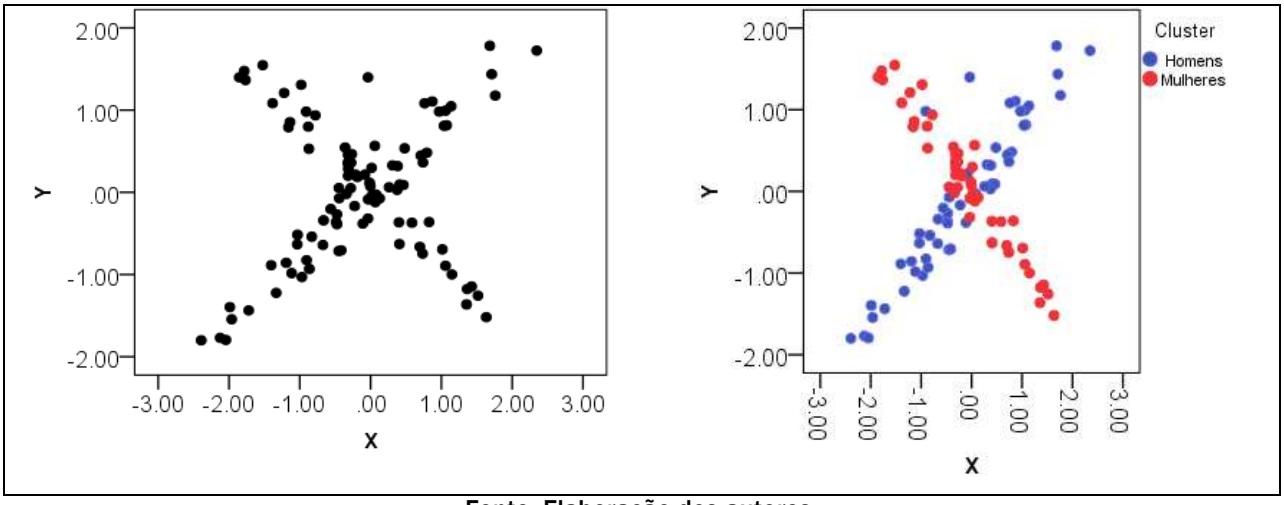


Fonte: Elaboração dos autores.

O desafio da análise de conglomerados é identificar corretamente a estrutura subjacente dos dados e agrupar as observações. Nossos dados simulados podem ser conglomerados em três grupos. O primeiro grupo (azul) agrupa os casos com níveis reduzidos em ambas as dimensões, é o que Dahl (1971) denominou de *hegemonias fechadas*. O segundo grupo (verde) conglomera as observações com nível moderado de *contestação* e *inclusividade*. Por fim, o terceiro *cluster* agrupa os países com altos níveis de *contestação* e *inclusividade*, o que Dahl (1971) chamou de poliarquias.

Outra importante função da análise de conglomerados é identificar grupos de observações que apresentam um padrão diferente de associação em relação a uma variável qualquer. Por exemplo, considere a correlação abaixo:

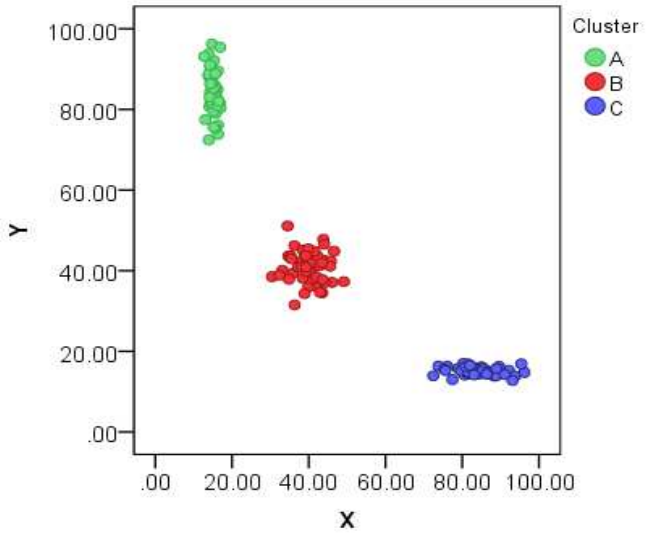
Figura 9 – Correlação entre X e Y (simulação)



Fonte: Elaboração dos autores.

O coeficiente de correlação entre X e Y é de 0,192 (p-valor = 0,055; n= 100). Ou seja, o pesquisador seria levado a concluir que as variáveis são, apenas, fracamente correlacionadas. Todavia, como pode ser observado existem dois diferentes padrões. Para os homens, a correlação é positiva (0,921; p-valor<0,000) enquanto as mulheres apresentam correlação negativa (-0,967; p-valor<0,000). Ao se analisar todos os casos juntos não é possível detectar a verdadeira estrutura subjacente ao fenômeno de interesse. A figura abaixo ilustra outra simulação em que a estrutura dos dados pode levar o pesquisador a realizar inferências equivocadas.

Figura 10 – Correlação entre X e Y (três *clusters*) (simulação)



Fonte: Elaboração dos autores.

A correlação entre X e Y em nosso exemplo simulado é de -0,937 (p-valor<0,000; n = 150). No entanto, a associação apenas ocorre ao se considerar todos os casos simultaneamente. Ao se desagregar a análise por cluster, verifica-se que nenhuma das correlações apresenta significância estatística (cluster A; r = -0,215; p-valor = 0,133; n = 50) (cluster B; r = -0,067; p-valor = 0,645; n = 50) (cluster C; r = -0,215; p-valor = 0,133; n = 50).

3. Considerações Finais

Esse trabalho apresentou uma breve introdução à análise de dados. O foco repousou sobre a compreensão intuitiva de três diferentes técnicas e a interpretação substantiva dos resultados empíricos. Metodologicamente, sintetizamos as principais recomendações da literatura e empregamos simulação básica para ilustrar a utilização das seguintes técnicas: (1) análise de variância (ANOVA) para amostras independentes; (2) análise de componentes principais e (3) análise de *cluster*.

Os recentes avanços computacionais permitem que pesquisadores sem treinamento intensivo em Matemática e/ou Estatística realizem análises sofisticadas. O primeiro passo é compreender os pressupostos que devem ser observados para a correta aplicação de cada técnica. Depois disso, o pesquisador deve se familiarizar com algum pacote estatístico (SPSS, SAS, STATA, STATISTICA, EVIEWS, BIOSTAT, GEODA, ARCVIEW, R, etc.) e aprender as rotinas. O último estágio consiste em interpretar substantivamente os resultados observados, avaliando em que medida os dados corroboram ou refutam sua hipótese de pesquisa.

Com esse artigo, esperamos ajudar os interessados no assunto a dar os primeiros passos na utilização dessas técnicas em suas respectivas áreas de atuação.

4. Referências Bibliográficas

ASHER, H. B. (1983) *Causal Modeling. Beverly Hills*, CA: Sage.

CRAMER, D.& HOWITT, D. L. (2004) *The Sage dictionary of statistics: a practical resource for students in the social sciences*. Sage.

DAHL, R. (1971) *Poliarquia: Participação e Oposição*. São Paulo: Edusp.

EVERITT, B. S.; SKRONDAL, A. (2010) *The Cambridge dictionary of statistics*. Cambridge University Press.

FOX, John (2008) *Applied Regression Analysis and Generalized Linear Models.* Los Angeles, CA: Sage.

FIGUEIREDO FILHO, D. B. ; ROCHA, E. C. de ; SILVA JUNIOR, J. A. ; PARANHOS, R. (2013) "Causalidades e mecanismo em ciência política". *Revista Mediações* (UEL), v. 18, p. 10-27.

FIGUEIREDO FILHO, D. B e SILVA JUNIOR, J. A. (2010) "Visão além do alcance: uma introdução à análise fatorial". *Revista Opinião Pública*, Vol. 16, Nº1, p. 160-185.

FIGUEIREDO FILHO, D. B; SILVA JUNIOR, J. A. e ROCHA, E. C. (2012) "Classificando regimes políticos utilizando análise de conglomerados". *Revista Opinião Pública*, Vol. 18, Nº1, p. 109-128.

GARSON, G. David. (2008) *Statnotes: Topics in Multivariate Analysis*. Disponível em: http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm. Acesso em: 18 fev. 2015.

HAIR, J. F *et al.* (2009) *Multivariate Data Analysis*. 6. ed. Upper Saddle River: Pearson Prentice Hall.

HAIR, J. F. *et al.* (2005) *Análise Multivariada de Dados*. 5. ed. Porto Alegre: Bookman.

IVERSEN, G. R., e NORPOTH, H. (1987). *Analysis of variance* (No. 1). Sage.

KING, G. (2001) How not to lie with statistics [Online]. Disponível em: <http://gking.harvard.edu/files/mist.pdf> Acesso em: 18 fev. 2015.

KIM, J; MUELLER, C. W. (1978) *Factor analysis: Statistical methods and practical issues.* Beverly Hills, CA: Sage.

NUNNALLY, J. (1967) *Psychometric Methods.* New York: MacGraw Hill.

PALLANT, J. (2007) *SPSS Survival Manual*. Open University Press.

SCHWARTZ, S. e CARPENTER, K. M. (1999) The Right Answer for the Wrong Question: Consequences of Type III Error for Public Health Research. *American Journal of Public Health*, 89, 1175-1180.

ZELLER, R. A; CARMINES, E. G. (1980) *Measurement in the social sciences: The link between theory and data.* Cambridge: Cambridge University Press.

Autores.

Dalson Britto Figueiredo Filho

Universidade Federal de Pernambuco (UFPE), Brasil.

Professor do Departamento de Ciência Política da Universidade Federal de Pernambuco (DCP/UFPE), Doutor e Mestre em Ciência Política pelo Departamento de Ciência Política da Universidade Federal de Pernambuco (DCP/UFPE). Brasil

E-mail: dalsonbritto@yahoo.com.br

Ranulfo Paranhos.

Universidade Federal de Alagoas (UFAL) / Universidade Federal de Pernambuco (UFPE), Brasil.

Professor do Instituto de Ciências Sociais de Universidade Federal de Alagoas (ICS/UFAL). Doutorando e Mestre em Ciência Política pelo Departamento de Ciência Política da Universidade Federal de Pernambuco (DCP/UFPE).

E-mail: ranulfoparanhos@me.com

José Alexandre da Silva Junior

Universidade Federal de Goiás (UFG) / Universidade Federal de Pernambuco (UFPE), Brasil.

Professor do Instituto de Ciências Sociais da Universidade Federal de Alagoas (ICS/UFAL). Doutor e Mestre em Ciência Política pelo Departamento de Ciência Política da Universidade Federal de Pernambuco (DCP/UFPE).

E-mail: jasjunior2007@yahoo.com

Denisson Silva.

Universidade Federal de Minas Gerais (DCP/UFMG), Brasil.

Doutorando em Ciência Política pela Universidade Federal de Minas Gerais (DCP/UFMG), Mestre em Sociologia e graduado em Ciências Sociais pela Universidade Federal de Alagoas (ICS/UFAL).

E-mail: denissoncsol@gmail.com

Citado.

FIGUEIREDO FILHO, Dalson Britto; PARANHOS, Ranulfo, SILVA JUNIOR, José Alexandre e SILVA, Denisson (2016). "Precisamos falar sobre Métodos Quantitativos em Ciência Política". *Revista Latinoamericana de Metodología de la Investigación Social - ReLMIS*. Nº11. Año 6. Abril– Septiembre 2016. Argentina. Estudios Sociológicos Editora. ISSN 1853-6190. Pp. 21-39. Disponible en: http://www.relmis.com.ar/ojs/index.php/relmis/article/view/143

Plazos.

Recibido: 13/02/ 2015. Aceptado: 01/11/2015