Instituto de Estudos Sociais e Políticos
Universidade do Estado do Rio de Janeiro

Introdução à Análise de Dados

Lego I
1o Semestre de 2018

Prof. Dr. Ricardo Ceneviva
ceneviva@iesp.uerj.br

Dr. Thiago Moreira da Silva
thiagomoreira@iesp.uerj.br

Objetivos do Curso

Este é um curso de introdução à análise de dados para alunos de pós-graduação em Ciência Política e Sociologia. O objetivo principal do curso é fornecer uma introdução básica de métodos estatísticos para a análise de dados sociais e políticos. A análise de dados, e a estatística que a fundamenta, podem ser difinidas suscintamente como a arte de fazer conjecturas numéricas sobre questões intrigantes (Freedman et al, 2009). Questões que há muito despertam o interesse das Ciências Sociais, tais como: existe alguma relação entre o tipo de regime político e o nível de desenvolvimento econômico dos países? Candidatos que concorrem à reeleição na função do cargo político tem mais chances de sucesso eleitoral? Mulheres têm menores rendimentos no mercado e trabalho se comparadas a homens de mesma escolaridade e que ocupam cargos semelhantes? Estas são perguntas que podem ser respondidas com o o auxílio de métodos e técnicas estatísticas de pesquisa. Neste curso serão apresentados os princípios básicos de pesquisa quantitativa em ciência política e sociologia. Como definir e mensurar conceitos, tipos de dados e variáveis, construir e interpretar gráficos e tabelas, além de noções usualmente utilizadas nas Ciências Sociais como médias, variação e correlação serão discutidos com cuidado ao longo desse curso.

Ao final do semestre espera-se que os alunos sejam capazes de produzir e analisar seus dados, além de ler e criticar textos que apliquem as técnicas discutidas nas aulas. Para tanto, o curso combina aulas expositivas e sessões práticas no laboratório, onde os alunos serão familiarizados com o manuseio de *softwares* de análise quantitativa de dados. Os comandos desenvolvidos estarão disponíveis no seguinte endereço de *github*: `https://github.com/thiago-ms-cp/`. O conteúdo do curso não é exaustivo, contudo, pretende-se encorajar os alunos a continuar estudando métodos mais sofisticados para a análise de dados.

A quem esse curso se destina?

O foco do curso são os alunos de pós-graduação de Ciência Política e Sociologia e os exemplos e casos analisados nas sessões teóricas e práticas serão retirados, principalmente, de artigos e trabalhos dessas duas disciplinas. Embora o curso tenha um conteúdo bastante técnico, sua abordagem será conceitual e aplicada, em vez de técnica e teórica. Isto é, as aulas expositivas e sessões de laboratório se concentrarão na intuição básica por trás de cada um dos modelos estudados e na sua aplicação prática nas ciências sociais. O curso não tem pré-requisitos. Porém, assume-se que os alunos estejam familiarizados com os conteúdos abordados no curso de nivelamento oferecido na semana que precede o início das aulas; ou seja, conhecimentos básicos de matemática.

Avaliação

Como se trata de um curso aplicado, a participação em sala de aula terá grande importância. Além das atividades em laboratório e das listas de exercícios, os alunos serão encorajados a coletar, analisar e apresentar seus próprios dados ou a replicar algum trabalho ou artigo já publicado. Para tanto, será proposto um projeto final que consistirá em uma apresentação de uma análise de algum banco de dados escolhido pelo aluno, ou no desenvolvimento de um projeto de pesquisa original que envolva análise de dados.

1. Os alunos deverão entregar 5 (cinco) listas de exercícios ao longo do curso. Cada lista de exercício vale 10% da nota final do aluno. Não serão aceitas listas de exercícios entregues fora do prazo.
2. Projeto final valendo 50% da nota final do aluno.

Software para Análise de Dados

Neste curso, tanto nas sessões de laboratório como para a realização das listas de exercícios, serão usados *softwares* para a análise de dados. Hoje, há vários programas computacionais concebidos especialmente para a análise estatística de dados sociais e políticos, além de outros de uso mais geral. Parte importante desse curso destina-se à introdução ao uso de *softwares* para a análise estatística de dados. Espera-se que ao final do semestre o aluno seja capaz de realizar operações básicas para a implementação, análise, disgnóstico e apresentação gráfica dos modelos estatísticos estudados ao longo do curso. Dentre as várias opções de *softwares* para a análise de dados, as sessões práticas em laboratório fornecerão treinamento básico para o manuseio de dois programas computacionais: R e Stata. Caberá ao aluno escolher a opção que mais lhe convém.

Cabe ressaltar que *softwares* e aplicativos para análises estatísticas estão sempre sendo modificados e aprimorados, de modo que qualquer coisa que se diga hoje a respeito deles pode estar incorreta quando uma próxima versão do aplicativo for disponibilizada aos usuários.

R

O R é uma linguagem para computação estatística e gráficos. Uma das grandes vantagens do R é que se trata de um *Software* Livre, que pode ser baixado gratuitamente na página oficial do Projeto R na internet: http://cran.r-project.org/. Recentemente um artigo do *New York Times* (Data Analysts Captivated by Rs Power, 6 de janeiro de 2009) caracterizou o R como:

> *a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca [...] whether being used to set ad prices, find new drugs more quickly or fine-tune financial models. Companies as diverse as Google, Pfizer, Merck, Bank of America, the InterContinental Hotels Group and Shell use it. [...] The great beauty of R is that you can modify it to do all sorts of things, said Hal Varian, chief economist at Google. And you have a lot of prepackaged stuff thats already available, so youre standing on the shoulders of giants.*

Como mencionado, o R não é apenas um pacote estatístico; trata-se de uma linguagem para computação estatística e gráfica. Como tal, é muito mais flexível, rápido e poderoso do que os demais *softwares* de análise de dados comerciais disponíveis no mercado, como o SPSS, Stata ou SAS, além de ser gratuito. O R é também mais difícil de ser usado, já que exige alguma noção de programação. Há uma imensa variedade de materiais de apoio, apostilas, e recursos (em português e inglês) destinados à aprendizagem do R disponíveis gratuitamente na Internet. Alguns desses recursos serão comentados e usados nas sessões de laboratório.

Stata

O Stata é um software comercial muito popular para análise estatística de dados. Este *software* foi criado em 1985 pela StataCorp. O Stata é largamente empregado, tanto por instituições do setor privado como na academia, particularmente, por pesquisadores que lidam com gerenciamento e manipulação de bancos de dados, o que engloba áreas tão diversas como: biologia, epidemiologia , economia, sociologia, demografia e ciência política, entre outras.

Os recursos do Stata incluem não apenas o gerenciamento e manipulação de bancos de dados, mas também, análises estatísticas, gráficos, simulações, além de oferecer ferramentas de programação personalizada para execução de tarefas específicas. O Stata oferece uma interface gráfica o que permite ao usuário a análise via programação ou utilizando os menus (*point-and-click*). O aprendizado no Stata é relativamente fácil, sendo superado apenas pelo SPSS, que não será usado no curso devido a seu elevado custo comercial.

Ms Excel

O Excel, programa de planilhas eletrônicas da Microsoft, embora não seja um *software* concebido especificamente para análise de dados, permite a realização de análises estatísticas diversas. O Excel oferece a opção "Análise de Dados no menu "Ferramentas que inclui, entre outras possibilidades, o cálculo de estatísticas descritivas, a construção de histogramas e outros gráficos, a realização de testes de comparações de médias, ANOVA e regressão linear. Usualmente, essa opção não se encontra ativa no programa e, portanto, não aparece no menu de opções. Para ativá-la, é preciso clicar em Ferramentas / Suplementos / Ferramentas de análise.

Uma das principais vantagens do Excel, com relação ao R ou ao Stata, é que se trata de um dos programas de computador mais populares hoje no mercado. Ele é o programa de planilhas eletrônicas dominante para as plataformas Windows e Mac e já vem incluído como parte do Microsoft Office. Assim sendo, ele está instalado na maioria dos computadores pessoais e é, sem dúvida o programa mais popular entre esses três que serão utilizados no curso. Entretanto, o Excel não tem muita flexibilidade e a execução dos comandos não é simples, seja usando as ferramentas nele inclusas ou os programas que a ele podem ser integrados.

Programa

***Aula 1**: "Introdução: Teoria e Dados"*

Apresentação do curso: o modelo científico nas ciências sociais, estudos observacionais e estudos experimentais.

Leitura obrigatória: não há leitura obrigatória para essa aula

Leitura complementar:

Kellstedt & Whitten (2009) capítulo 1: *"The Scientific Study of Politics"* e capítulo 2: *"The Art of Theory Building"*

Box-Steffensmeier, Janet M., Henry E. Brady, and David Collier. "Political science methodology." The Oxford handbook of political methodology. 2008.

***Aula 2**: "Conceitos, Medidas e Hipóteses"*

Medidas: Os conceitos das ciências sociais podem ser medidos? Quais as principais dificuldades da mensuração? Mensuração, tipos de dados e variáveis.

Leitura obrigatória:

Goertz, Gary. "Concepts, Theories, and Numbers: A Checklist for Constructing, Evaluating, and Using Concepts or Quantitative Measures." The Oxford Handbook of Political Methodology.

Pedhazur & Schmelkin (1991). capítulo 1: *"Measurement and Scientific Inquiry"*

Leitura complementar:

Pereira (2004) capítulo 1: "O Dado Qualitativo" (pp.29-42) e capítulo 2: "Definição de Medidas"

Agresti, Franklin & Klingenberg (2017) capítulo 4: *"Gathering Data"*

Pollock (2012) capítulo 1: "The Definition and Measurement of Concepts" (pp.6-27).

***Aula 3**: "Definindo e Medindo suas Variáveis"*

Validade e confiabilidade na operacionalização e mensuração de conceitos nas ciências sociais e seus problemas.

Leitura obrigatória:

Kellstedt & Whitten (2009) capítulo 5: *"Getting to Know your Data: Evaluating Measurement and Variations"*

Leitura complementar:

Pedhazur & Schmelkin (1991). capítulos 2 - 5 : *"Criterion-Related Validation", "Construct Validation", e " Reliability"*

Freedman, Pisani & Purves (2007) capítulo 6: *"Measurement Error"*

***Aula 4**: "Estatísticas Descritivas: Medidas de Tendência Central e Medidas de Dispersão"*

Medidas de tendência central: média, moda e mediana. Medidas de dispersão: variância, desvio padrão e outras medidas de variabilidade dos dados em torno da média. Como usar gráficos para examinar e entender seus dados.

Leitura obrigatória:

Agresti, Franklin & Klingenberg (2017) capítulo 2: *"Exploring Data with Graphs and Numerical Summaries"*

Leitura complementar:

Kellstedt & Whitten (2009) capítulo 6: *"Descriptive Statistics and Graphs"*

Monogan, James (2015), capítulo 4.1. Political Analysis using R. Springer

***Aula 5**: "Gráficos: analisando os dados de maneira descritiva"*

Como usar gráficos para examinar e entender os dados. Nesse tópico cobriremos os principais quadros de representação usados nas ciências sociais: histogramas, *scatter plots*, gráficos de barra, gráficos de linhas, *boxplots* e mapas.

Leitura obrigatória:

Agresti, Franklin & Klingenberg (2017) capítulo 2: *"Exploring Data with Graphs and Numerical Summaries"*

Leitura complementar:

Kellstedt & Whitten (2009) capítulo 6: *"Descriptive Statistics and Graphs"*

Freedman, Pisani & Purves (2007) capítulos 3 e 4: *"The Histogram" e "The Average and the Standard Deviation"*

Wickham, Hadley (2016). ggplot2: Elegant Graphics for Data Analysis. Springer.

***Aula 6**: "Associação: Contingência, Correlação e Regressão."*

Leitura obrigatória:

Agresti, Franklin & Klingenberg (2017) capítulo 3: *"Association: Contingency, Correlation, and Regression"*

Leitura complementar:

Freedman, Pisani & Purves (2007) capítulo 8 e 9: *"Correlation" e "More about Correlation"*

***Aula 7**: "Distribuições de Probabilidade"*

Como a probabilidade quantifica a aleatoriedade? As principais distribuições de probabiblidade.

Leitura obrigatória:

Agresti, Franklin & Klingenberg (2017) capítulo 5: *"Probability in Our Daily Lives"* e capítulo 6 *"Probability Distributions"*

Leitura complementar:

Kaplan (2011) capítulo 11 *"Modeling Randomness"*

Freedman, Pisani & Purves (2007) capítulo 13: *"What are the Chances?"*

***Aula 8**: "Os Fundamentos da Inferência Estatística"*

Inferência e Causalidade nas Ciências Sociais. O que é inferência estatística? O teste de significância, amostras e populações.

Leitura obrigatória:

Agresti, Franklin & Klingenberg (2017) capítulo 8: *"Statistical Inference: Confidence Intervals"*

Leitura complementar:

Kellstedt & Whitten (2009) capítulo 6: *"Probability and Statistical Inference"* (120-132)

Pollock (2012) capítulo 5: "Foundations of Statistical Inference" (pp.122-152).

***Aula 9**: "Teste de Hipóteses"*

Leitura obrigatória:

Agresti, Franklin & Klingenberg (2017) capítulo 9: *"Statistical Inference: Significance Tests About Hypotheses"*.

Leitura complementar:

Kaplan (2011) capítulo 13 *"The Logic of Hypothesis Testing"*.

Kellstedt & Whitten (2009) capítulo 7: *"Bivariate Hipotesis Testing"* (120-132)

***Aula 10**: "Comparações Envolvendo Médias e Proporções"*

Como comparar dois grupos? O teste t para duas amostras. Testando hipóteses. O teste de diferenças entre médias e o teste de diferenças entre proporções.

Leitura obrigatória:

Agresti, Franklin & Klingenberg (2017) capítulo 10: *"Comparing Two Grups"*.

Leitura complementar:

Kaplan (2011) capítulo 4: *"Group-wise Models"*.

***Aula 11**: "Analisando a Associação entre Dados Categóricos"*

Associação e independência. Frequência observada e frequência esperada. Correção de Yates. Teste exato de Fisher.

Leitura obrigatória:

Agresti, Franklin & Klingenberg (2017) capítulo 11: *"Analyzing the Association Between Categorical Variables"*.

Leitura complementar:

Sirkin (2006) capítulo 12: "The Qui-Square Test" (pp.397-437).

***Aula 12**: "Analisando a Associação entre Dados Quantitativos: Análise de Regressão"*

O modelo de regressão linear simples. Estimação via MQO. Interpretação da reta de regressão. Correlação de Pearson.

Leitura obrigatória:

Kennedy (2009) capítulo 3: "O Modelo Clássico de Regressão Linear" (pp.38-49).

Leitura complementar:

Agresti, Franklin & Klingenberg (2017) capítulo 12: *"Analyzing the Association Between Quantitative Variables: Regression Analysis"*.

***Aula 13**: "Entendendo e Interpretando o Modelo de Regressão"*

O modelo de regressão linear multipla. Interpretação dos coeficientes da Regressão. Resíduos. ANOVA.

Leitura obrigatória:

Kennedy (2009) capítulo 4: "Estimação de Intervalo e Teste de Hipótese" (pp.50-65) e capítulo 5: "Especificação' (pp.70-88) .

Leitura complementar:

Agresti, Franklin & Klingenberg (2017) capítulo 13: *"Multiple Regression"*.

Kellstedt & Whitten (2009) capítulo 10: *"Multiple Regression Models I: The Basics"* (183-200).

Bibliografia

Os bancos de dados e os *scripts* dos programas computacionais de análise estatística utilizados nas sessões de laboratório e nas listas de exercícios serão distribuídos por e-mail e, posteriormente, postados na pasta compartilhada da disciplina no site: www.dropbox.com

Dropbox.com. As leituras exigidas para as aulas dividem-se em obrigatórias e complementares. Espera-se que os alunos já tenham feito a leitura dos textos indicados como obrigatórios antes das respectivas aulas. Cabe observar que o conteúdo dessa disciplina tem uma caráter fortemente cumulativo. Ou seja, a compreensão dos tópicos abordados numa determinada aula dependem do domínio dos conteúdos abordados nas aulas passadas. Uma bibliografia complementar será apresentada ao final de cada aula como sugestão de leitura para aqueles que desejem se aprofundar no tema.

Livros textos

1. Agresti, Alan, Christine Franklin and Bernhard Klingenberg. *The art and science of learning from data.* Upper Saddle River, NJ: Prentice Hall, 2007.
2. Bussab, Wilton e Morettin, Pedro. *A Estatística Básica.* 6ł edição. São Paulo: Saraiva, (2009).
3. Freedman, David, Robert Pisani, and Roger Purves. *"Statistics (International Student Edition)."* Pisani, R. Purves.4th edition.NY, USA: WW Norton & Company 720 (2007).
4. Gelman, Andrew, and Jennifer Hill. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.(2007).
5. Kennedy, Peter. *Manual de Econometria.* Rio de Janeiro: Elsevier, 2010.
6. Levin, Jack e Fox, James A. *Estatística para Ciências Humanas.* 9ł edição. São Paulo: Prentice Hall, 2004.
7. Monogan, James III. 2015. *Political Analysis Using R.* Springer.
8. Pereira, Julio Ignacio Piovani. *Análise de Dados Qualitativos: Estratégias Metodológicas para as Ciências de Saúde, Humanas e Sociais.* 3ł edição. São Paulo: EDUSP, 2004.
9. Pedhazur, Elazar J.;Schmelkin , Liora Pedhazur.*Measurement, design, and analysis: An integrated approach.* Psychology Press, New York, 1991.
10. Wooldridge, Jeffrey M. 2010. *Introdução à Econometria: Uma Abordagem Moderna.* Tradução da 4 ed. norte-americana. São Paulo: Cengage Learning.

Leitura Complementar

1. Kellstedt, Paul M. & Whitten, Guy D. *The Fundamentals of Political Research.* Cambridge: Cambridge University Press, 2009.
2. Pollock, Philip H. *The Essentials of Political Analysis.* 4th ed. Washington: CQ Press, 2012.

3. Sirkin, R. Mark. *Statistics for the Social Sciences*. 3rd ed. Thousand Oaks: Sage, 2006.