



UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE ESTATÍSTICA

Curso Livre: Tópicos Avançados em Estatística e Ciência de Dados

Docentes:

- Prof. Dr. Eufrásio de Andrade Lima Neto
- Prof. Dr. Hemílio Fernandes Campos Coêlho
- Prof. Dr. Luiz Medeiros de Araujo Lima Filho
- Prof. Dr. Marcelo Rodrigo Portela Ferreira
- Prof. Dr. Rodrigo Bernardo da Silva
- Prof. Dr. Telmo de Menezes e Silva Filho

Carga horária: 60 horas (4 créditos)

Proposta: O curso será ofertado em formato EAD na plataforma Moodle, com objetivo de apresentar de forma introdutórias temas avançados na área de Estatística e Ciência de Dados. Aspectos computacionais serão considerados bem como a apresentação de seminários (síncronos ou assíncronos) a cerca de tais assuntos.

Requisito Mínimo Desejável: Recomenda-se que os participantes tenham cursado a Disciplina de Regressão I no Bacharelado em Estatística ou possuam conhecimentos de inferência estatística, linguagens de programação R e/ou Python e regressão.

Módulo I - Tópicos Avançados em Estatística:

Semana 1 (6 horas): Tópicos em Séries Temporais: Modelos INGARCH Poisson Mistos

1. Introdução aos modelos de Séries Temporais para dados de contagem;
2. Modelos INGARCH Poisson Mistos;
3. Fórum de dúvidas;
4. Momento síncrono.

Semana 2 (6 horas): Gráficos de controle para proporções contínua

1. Vídeo aula introdutória sobre gráficos de controle;
2. Leitura de artigos científicos relacionados ao tema;
3. Fórum de dúvidas;
4. Momento síncrono.

Semana 3 (6 horas): Tópicos em amostragem: Multiplicidade

1. Definição do problema sob a perspectiva de populações finitas;
2. Multiplicidade: Métodos de estimação; Estimador de Horvitz-Thompson sob multiplicidade; Estimador do tipo regressão aplicado ao conceito de multiplicidade;

3. Fórum de dúvidas;
4. Momento síncrono ou assíncrono (aula ao vivo ou gravada).

Semana 4 (6 horas): Execução de projetos; Fórum de dúvidas;

Semana 5 (6 horas): Entrega do projeto (primeira avaliação).

Módulo II - Tópicos Avançados em Ciência de Dados:

Semana 6 (6 horas): Modelos de Regressão para Dados Tipo-Intervalo

1. Principais modelos de regressão para dados tipo-intervalo;
2. Aspectos computacionais: a biblioteca `iRegression`;
3. Fórum de dúvidas;
4. Momento síncrono ou assíncrono (aula ao vivo ou gravada)

Semana 7 (6 horas): Calibragem de Classificadores

1. Definição de Modelo Calibrado; Visualização de Erros de Calibragem; Calibragem binária e multiclasse; Principais modelos paramétricos e não-paramétricos de
2. Aspectos computacionais: calibrando modelos usando `R` e `Python`;
3. Fórum de dúvidas;
4. Momento síncrono ou assíncrono (aula ao vivo ou gravada)

Semana 8 (6 horas): Covid-19: Tratamento, visualização e predição em `R`

1. Definição do problema; Fontes de dados; Tratamento e Estratégias de Visualização usando `R`;
2. Modelagem S.I.R; Estratégias de predição usando `R`;
3. Fórum de dúvidas;
4. Momento síncrono ou assíncrono (aula ao vivo ou gravada)

Semana 9 (6 horas): Execução de projetos; Fórum de dúvidas;

Semana 10 (6 horas): Entrega do projeto (segunda avaliação).

Metodologia: O curso será ofertado em formato EAD na plataforma Moodle. A cada semana, será abordado um tema associado a Estatística ou Ciência de Dados. Será disponibilizado material para leitura, por meio de artigos científicos ou capítulo de livro, bem como os aspectos computacionais dos temas por meio das linguagens `R` e/ou `Python`. Em cada semana, haverá um fórum de dúvidas, onde o discente poderá esclarecer pontos de interesse com o professor. Ao final de cada semana, haverá um momento síncrono ou assíncrono, onde o professor apresentará um seminário sobre o tema da semana.

Avaliação da aprendizagem: O discente deverá escolher 1 tema de cada Módulo, totalizando 2 temas abordados ao longo do curso. A aprovação está condicionada ao discente obter média aritmética igual ou superior a 5,0. Além disso, também será necessária uma presença maior ou igual a 75% da carga horária do curso.

Referências:

Crawley, Michael J. The R book. John Wiley & Sons, 2012.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, 2009.

- Montgomery, Douglas C. Introduction to Statistical Quality Control. Vol. 6 John Wiley & Sons, 2009.
- Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. Vol. 821. John Wiley & Sons, 2012.
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Lima Neto, E.A. and De Carvalho, F.A.T. Centre and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics & Data Analysis* 52 (3), 1500–1515, 2008.
- Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey methodology*, 33(2), 151-157.
- Singh, A. C., & Mecatti, F. (2011). Generalized multiplicity-adjusted Horvitz-Thompson estimation as a unified approach to multiple frame surveys. *Journal of official statistics*, 27(4), 633.
- Mecatti, F., & Singh, A. C. (2014). Estimation in multiple frame surveys: a simplified and unified review using the multiplicity approach. *Journal de la Société Française de Statistique*, 155(4), 51-69.
- Kull, M., Silva Filho, T. M. and Flach, P. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electron. J. Statist.*, 11(2):5052–5080, 2017.
- Kull, M., Perello-Nieto, M., Kängsepp, M., Silva Filho, T. M., Song, H. and Flach, P. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration, 3 Sep 2019, NeurIPS 2019.