

XVIII ESCOLA DE MODELOS DE REGRESSÃO
21 A 24/11/2023, Belém, PA

PROPOSTA DE MINICURSO

Título: Análise Estatística de Dados Genômicos

Instrutores: Andressa Cerqueira, UFSCar, acerqueira@ufscar.br

Daiane Zuanetti, UFSCar, dzuanetti@ufscar.br

Júlia Pavan Soler, USP, pavan@ime.usp.br

Colaborador: Jaqueline L. Pereira, FSP-USP, jaquelinelopes@usp.br

Resumo: Há longa data a Estatística e a Genética têm compartilhado conhecimento e ferramentas, as quais têm promovido avanços em ambas as áreas. A Genômica, por exemplo, tem permitido observar certos tipos de dados moleculares pela primeira vez e inaugura a estrutura de big-data nos dados biológicos, derrubando fronteiras entre a Estatística, Biologia e Informática. Em um sentido amplo, as novas tecnologias estão revolucionando a coleta, armazenamento, análise e interpretação de dados. Entretanto, o entusiasmo com essa nova perspectiva de pesquisa tem resultado em situações desconfortáveis nas quais os pesquisadores estão “se afogando em dados, mas ainda sedentos de domínio sobre os mesmos”. Neste Minicurso, cobriremos um racional de análise de dados genômicos que consideramos necessários para ter sucesso nessa área. Partiremos de conceitos básicos envolvendo p-valores e tamanhos de efeitos genéticos dispostos em gráficos Manhattan e vulcão, respectivamente, e avançaremos a tópicos de colapsagem de variáveis de efeitos fixos e aleatórios, bem como enriquecimento e integração de dados, relacionados à análise de grandes bancos de dados genômicos oriundos tanto de estudos com indivíduos independentes como com grau de parentesco. Recursos principalmente do R serão fornecidos para execução das análises propostas e um subconjunto de dados reais extraídos do projeto ISA-Nutrição 2015 (apoiado pela FAPESP #2017/05125-7) será usado para motivação e ilustração.

Formato: Presencial. O conteúdo será apresentado e conduzido por meio de um tutorial usando recursos do R *markdown* que incluirá códigos do R para execução das análises propostas. Subconjuntos de dados reais serão disponibilizados para que o participante possa reproduzir e acompanhar o racional das análises. O Minicurso seguirá a dinâmica de um *Lab de Estatística*, em que o problema factual e soluções metodológicas afins serão brevemente introduzidas, e o participante, por rodar os códigos e ter acesso a dados, terá uma melhor intuição e vivência dos conceitos envolvidos.

Conteúdo: Este Minicurso cobrirá os seguintes tópicos:

1. Introdução, padrões de influência genética em doenças (variantes comuns e raras / doenças comuns e raras), tipos de estudos (observações independentes e estruturadas em famílias).
2. Processamento de dados de SNP-array: MAF, Equilíbrio de Hardy-Weinberg, LD (desequilíbrio de ligação); Matriz GRM.

3. Ancestralidade (global): análises de componentes principais enriquecidas com dados externos (1000 Genomas).
4. Análises de Associação:
 - Unilocos: GWAS com efeitos genéticos fixos e aleatórios.
 - Multilocos: haplótipos, blocos de SNP (campos Markovianos), colapsagem (em efeitos fixos ou aleatórios), escore de risco poligênico.
5. Integração de dados ômicos: Análises duais (em espaços $n \times p$, $p \times p$ e $n \times n$).

Breve CV:

Andressa Cerqueira é professora adjunta do Departamento de Estatística da UFSCar, São Carlos, SP, desde 2020 e é membro do Young Researchers Committee da Bernoulli Society desde 2022. Obteve o título de Doutora em Estatística pela Universidade de São Paulo em 2018 e realizou doutorado sanduíche na Université Paul Sabatier em Toulouse, França, (2015-2016) com bolsa da FAPESP. Foi pós-doutoranda no departamento de Estatística da University of Michigan, USA (2018-2019). Atua na área de análise estatística de redes complexas com enfoque em redes biológicas. Seus principais interesses de pesquisa são formulação de modelos estatísticos para redes complexas, métodos de agrupamento em redes e inferência paramétrica.

Daiane Zuanetti é professora adjunta do Departamento de Estatística da UFSCar, São Carlos, SP, Brasil. Possui graduação em Estatística pela UFSCar (2003), mestrado em Estatística pela UFSCar (2006) e doutorado em Estatística pelo PIPGEs (2016), com período sanduíche na Universidade do Texas em Austin. Antes do doutorado, trabalhou por sete anos no mercado financeiro no banco Itaú Unibanco. Tem pesquisado na área de Probabilidade e Estatística, com ênfase em seleção e estimação de modelos de mistura independentes ou dependentes, mapeamento QTL, inferência Bayesiana paramétrica e não paramétrica e métodos computacionais MCMC com aplicação, principalmente, em Genômica e Biologia Molecular.

Júlia P Soler é Professora do Departamento de Estatística do IME-USP desde 1995, integralmente dedicada ao ensino e pesquisa em Estatística. Tem projetos colaborativos com o Incor-USP e Faculdade de Saúde Pública-USP, Mayo Clinic (Rochester, USA) e Pharmacy Institute (University of Groningen, Netherlands), principalmente direcionados para o desenvolvimento e aplicação de métodos estatísticos no planejamento, processamento e análise de dados multiômicos. É orientadora na pós-graduação em Estatística e em Bioinformática do IME/USP. Desafios de sua pesquisa envolvem principalmente a formulação de métodos multivariados de redução de dimensionalidade e de integração de bancos de dados, bem como de aprendizado de estruturas causais entre variáveis, em espaços duais, com dependência familiar entre observações e de alta dimensionalidade.

Jaqueline L. Pereira é nutricionista (2011), mestre (2014) e doutora (2018) em Nutrição em Saúde Pública pela Faculdade de Saúde Pública da USP, com estágio em pesquisa na Universidade de Harvard (2017). Tem experiência na área de epidemiologia nutricional, avaliação do consumo alimentar, inquéritos de saúde, análise de banco de dados e variabilidade da dieta. Atualmente, é pós-doutoranda no Departamento de Nutrição da FSP/USP e pesquisadora do Grupo de Avaliação do Consumo Alimentar da USP (GAC).

Seus principais interesses de pesquisa são qualidade da dieta e nutrigenética na obesidade e outros fatores de risco cardiometabólico.