

# Tópicos em Zoologia 1

## 2024/02

### Módulo 01

# Introdução ao Curso

## Lógica de Análise de Dados

Depto de Zoologia  
15 de outubro de 2024

# Usos da Estatística em Biologia

- Visualização e Descrição dos Dados
- Transformar idéias em hipóteses testáveis
- Testes de Hipóteses e comparação de amostras
- Testes de Hipóteses e ajuste de modelos
- Modelagem e construção de cenários no tempo e no espaço além do universo amostral disponível

# Expectativas do Semestre

- Conhecimento básico dos métodos descritivos, tipos de variáveis, R, QGIS
- Conhecimento de testes de hipóteses por comparações de amostras, capacidade de compreender abordagens nos artigos científicos, o que são distribuições
- Ajustes por mínimos quadrados e modelos lineares e linearizados. Regressão e Análise de Variância.
- Análise de dados espaciais, interpretação de imagens de satélites
- Métodos avançados para modelar a natureza: redes neurais, algoritmos genéticos, autômatos celulares

# Abordagem do Curso

- Foco na discussão de dados biológicos e estatística: idéias, coleta de dados, testes de hipóteses, métodos de análise
- Quatro seções conceituais:
- Seção 1: coleta, visualização e descrição de dados por métodos gráficos e quantitativos (estatística descritiva)
- Seção 2: comparações de amostras e populações, testes de hipóteses, estatística frequencista
- Seção 3: ajustes de modelos por métodos estatísticos - Análise de variância, regressão, mínimos quadrados, máxima verossimilhança, random forest, splines
- Seção 4: análises multivariadas para redução de variáveis, ordenação, classificação geração de cenários, modelagem por inteligência artificial

# Lógica de Análise de Dados

- Em nossas aulas discutiremos o uso de análises estatísticas, incluindo a caracterização de distribuições e testes de hipóteses sobre tais características
- Primeiro devemos selecionar os dados que nos interessam como indicadores de parâmetros biológicos.
- Em seguida usamos a estatística para descrever as distribuições de dados e testar hipóteses a respeito das distribuições.
- Finalmente usamos os resultados dos testes estatísticos para fazer inferências sobre os processos biológicos dos quais os dados originais foram gerados
- Ref para hoje: Andrade e Ogari cap. 2

# Tipos de dados para análises

- **Nominal, Categóricos** - valores não ordenados (classes)
- **Ordinal** - valores ordenados sem quantificar as diferenças permite teste de ranking, scores
- **Intervalo** - variáveis contínuas, permite analisar variâncias, erro
- **Razão** - valor zero é referência absoluta não arbitrária
- **Frequências** - contagens (mesmos métodos dos categóricos)

# Exemplos de dados em biologia

- Nominal - atributo ex nome de espécie
- Ordinal - reflete ordenação, ranking. Exemplo: ordem de preferência das presas por um predador
- Intervalo - dados numéricos com intervalos contínuos e equivalentes. Exemplo: temperatura fahrenheit, celsius. Zero é um valor convencional, não é absoluto.
- Razão - dados numéricos em que a origem (zero) tem valor absoluto: exemplo temperatura kelvin (zero grau), tempo
- Quantitativas discretas (números inteiros)

# Sistemas Complexos

(Claudia Pahl-Wostl 1995)

- Extrapolações lineares não são factíveis
- Prever os limites e transições é extremamente difícil
- Fatores relevantes são difíceis de reconhecer devido à sua pequena importância em situações estáveis.
- Relações causa-efeito quase inexistentes. Efeitos dependem do estado atual e do contexto.
- Exemplos: dinâmica de ecossistemas, queimadas, espécies invasoras, sociedades humanas
- Métodos: universos digitais no computador. A vida biológica é digital (4 bases)



# Sistemas Complexos

- Apresentam características de auto-organização
- Existem no limite do caos
- Não são previsíveis individualmente, mas seu comportamento segue regras gerais.
- Como estudar? por métodos estatísticos e por modelagem. Embora a trajetória individual do sistema não seja previsível, o conjunto de trajetórias tem limites (atratores)
- Nas sociedades humanas, os sistemas complexos são estudados por meio de análise histórica. É possível demonstrar o encadeamento de variáveis que produziu o resultado observado, a posteriori.

# Sistemas Complexos

- Há estabilidade em sistemas complexos?
- Como diferem de comportamento caótico?
- Estudos de simulação mostram resultados interessantes. Exemplo: redes de interação em ecossistemas.
- Se há interações fortes entre todos os componentes do ecossistema, e feedbacks positivos, o sistema é instável
- Em sistemas que tendem à estabilidade, ao longo do tempo, as interações fortes se restringem a um número pequeno de componentes (espécies dominantes) e a mediação se dá principalmente por retroalimentação negativa (feedback negativo).

# Sistemas Complexos

- A escolha das medidas é fundamental para entender os processos que ocorrem.
- Por exemplo, uma medida ordinal dá uma idéia da ordem de importância, mas não explicita as magnitudes das diferenças entre os componentes do sistema.
- Já as medidas de intervalo ou razão permitem posicionar melhor a situação de cada componente
- Uma escola de samba é um exemplo ótimo de um sistema complexo (veja as definições anteriores)
- Vamos ver como os resultados variam de ano para ano. E como o uso de medidas de intervalo permite compreender o fenômeno muito melhor do que usando medidas ordinais.

# Exemplos Dados Ordinal Intervalo

## Classif Escolas de Samba Rio 2017

- 1 Portela
- 2 Mocidade
- 3 Salgueiro
- 4 Mangueira
- 5 Grande Rio
- 6 Beija Flor
- 7 Imperatriz
- 8 União da Ilha
- 9 São Clemente
- 10 Vila Isabel



Creative Commons / Ben Tavener

# Exemplos Dados Ordinal Intervalo

## Escolas de Samba Rio 2017 2018

● 1 Portela	269,9	269,4	4
● 2 Mocidade	269,8	269,3	6
● 3 Salgueiro	269,7	269,5	3
● 4 Mangueira	269,6	269,3	5
● 5 Grande Rio	269,4	266,8	rebaix
● 6 Beija Flor	269,2	269,6	1
● 7 Imperatriz	268,5	268,8	8
● 8 União da Ilha	267,8	267,3	10
● 9 São Clemente	267,4	266,9	11
● 10 Vila Isabel	267,4	268,1	9

# Fenômeno de Regressão à Média

- Resultados extremos não são replicáveis consistentemente em grupos coesos se forem resultados de alta complexidade e muitas variáveis
- Times esportivos
- Grupos artísticos
- Grupos políticos
- Indivíduos nestes sistemas
- A tendência é a volta à média depois de resultados extremos (positivos ou negativos)
- Determinados tipos de variáveis como ordenação ou percentagens não informam os dados

# Análise exploratória de dados

- Faremos exercício prático no R
- Elementos essenciais: gráficos de distribuição, dispersão, classificação, tendência central, estimativa e erro de estimativa, variância
- Métodos geralmente adotados:
- Gráficos contínuos ou por classes
- Tabelas de Frequências
- Medidas de tendência central: média, moda, mediana
- Variância, desvio padrão da amostra, coeficiente de variação
- Índices de associação (correlação de pearson), regressão



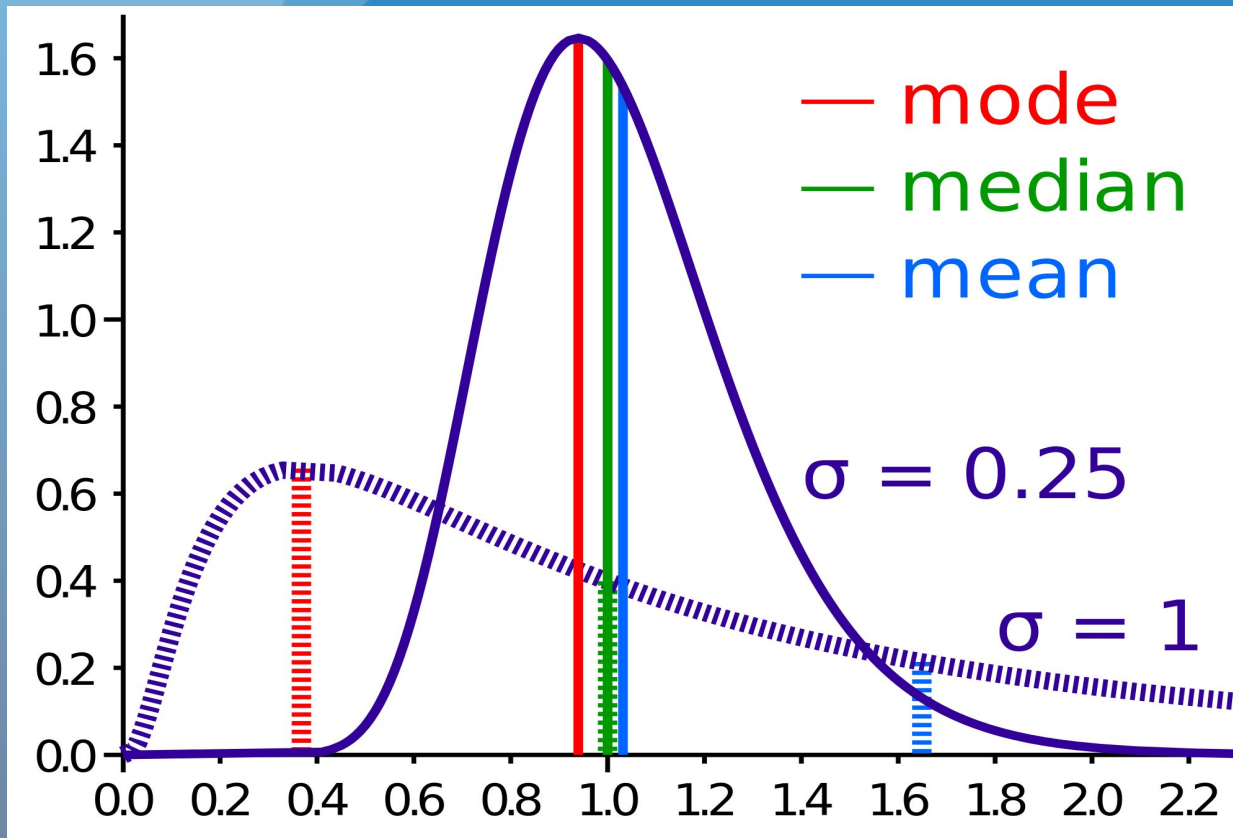
# Estatísticas descritivas úteis

- Matriz de dados
- Mediana (divide distribuição em duas)
- Quartis - dividem a distribuição em quartos
- Moda (frequência ou valor mais abundante)
- Variância e desvio padrão (medida de dispersão da pop)
- Erro padrão da média (medida de dispersão na estimativas da média)
- Intervalo de confiança - intervalo que cobre x% da distribuição



# Estatísticas descritivas úteis

- Média, moda, mediana



# Exemplo de Pesquisa Moderna

- Trabalho na Serra da Mesa, Goiás. Laboratório Prof Reuber, UnB/EFL.



## Lizards on newly created islands independently and rapidly adapt in morphology and diet

Mariana Eloy de Amorim<sup>a,b,1</sup>, Thomas W. Schoener<sup>b,1</sup>, Guilherme Ramalho Chagas Cataldi Santoro<sup>c</sup>, Anna Carolina Ramalho Lins<sup>a</sup>, Jonah Piovia-Scott<sup>d</sup>, and Reuber Albuquerque Brandão<sup>a</sup>

<sup>a</sup>Laboratório de Fauna e Unidades de Conservação, Departamento de Engenharia Florestal, Universidade de Brasília, Brasília DF, Brazil CEP 70910-900;

<sup>b</sup>Evolution and Ecology Department, University of California, Davis, CA 95616; <sup>c</sup>Departamento de Pós-Graduação em Zoologia, Instituto de Biologia, Universidade de Brasília, Brasília DF, Brazil CEP 70910-900; and <sup>d</sup>School of Biological Sciences, Washington State University, Vancouver, WA 98686-9600

Contributed by Thomas W. Schoener, June 21, 2017 (sent for review December 31, 2016; reviewed by Raymond B. Huey and Dolph Schluter)

**Rapid adaptive changes can result from the drastic alterations humans impose on ecosystems. For example, flooding large areas for hydroelectric dams converts mountaintops into islands and leaves surviving populations in a new environment. We report differences in morphology and diet of the termite-eating gecko *Gymnodactylus amarali* between five such newly created islands**

**study, because it was the most common lizard species in the area at the time of the field study.**

We evaluated the effects of isolation (actually, insularization) on diet and morphology of *G. amarali* populations on islands formed by the Serra da Mesa reservoir. We collected data on lizard diet and morphology on five islands, as well as five nearby



## Data from: Lizards on newly created islands independently and rapidly adapt in morphology and diet

Eloy de Amorim M, Schoener TW, Santoro GRCC, Lins ACR, Piovio-Scott J, Brandão RA

Date Published: August 10, 2017

DOI: <https://doi.org/10.5061/dryad.3nk78>

[Submit data now](#)

[How and why?](#)

## Search for data

Enter keyword, DOI, etc.

[Go](#)

[Advanced search](#)

## Files in this package

Content in the Dryad Digital Repository is offered "as is." By downloading files, you agree to the [Dryad Terms of Service](#). To the extent possible under law, the authors have waived all copyright and related or neighboring rights to this data.  

Title	<b>Data used for analysis of niche breadth</b>
Downloaded	21 times
Description	Data are from <i>Gymnodactylus amarali</i> individuals collected from field sites near the Serra da Mesa Reservoir in central Brazil. The following column headings are used. area: Each field site was either an island created by the filling of the reservoir or part of the adjacent mainland site: a unique identifier for each field site lizard: a unique identifier for each lizard captured as part of the study termite.length_mm: the length of individual termites (in millimeters) found in the stomach of lizards used in the study
Download	<a href="#">NicheBreadth.csv (15.28 Kb)</a>
Details	<a href="#">View File Details</a>

## Be part of Dryad

We encourage organizations to:

[Become a member](#)

[Sponsor data publishing fees](#)

[Integrate your journal\(s\)](#), or

All of the above

# O que é o R

- Linguagem computacional de alto nível voltada para manipulação e análise de dados
- Versão de código livre e aberto da linguagem S
- Desenvolvida por consórcio global de pessoas e organizações (R Project)
- A linguagem base é suplementada e estendida por “pacotes” com rotinas, funções e dados voltados para disciplinas e aplicações específicas.
- Linguagem interpretada, não compilada, portanto tem restrições de tamanho de conjunto de dados e velocidade de computação.

# As 4 abordagens para usar R

- Importar e organizar dados e objetos
- Funções e operações
- Visualização e descrição de dados
- Ajuste de modelos e análises estatísticas

# Colocando dados no R

- Manual usando comandos R e arquivos de texto ou clipboard
- Usando menu do Rstudio
- Carregando pacotes contendo conjuntos de dados
- Executando scripts do R para local ou internet

# Pacotes no R

- Os pacotes em R são elementos de programação executáveis que contêm rotinas pré-escritas, permitindo:
- Utilizar funções desenvolvidas para aplicações específicas
- Executar análises, plotar gráficos, etc, em formatos e para necessidades personalizadas
- Integrar vários produtos da linguagem R em um arquivo único: dados, funções, variáveis, rotinas
- Minimizar o trabalho de executar trabalhos repetidos
- Distribuir métodos analíticos de forma confiável e replicável



# O que é o Rstudio

- Interface Gráfica para a linguagem R
- Cada janela permite um tipo de acesso à linguagem
- 4 janelas básicas:
  - Script ou markdown ou notebook
  - Comando
  - Saída/ajuda
  - Variáveis de estado
- Versões windows, mac, linux, servidor, cloud
- Software gratuito para uso individual, pago na versão empresarial
- Modelo comercial / apoio comunitário



# Como funciona o Rstudio

- Oferece janelas para visualizar ao mesmo tempo diversas interfaces do R
- Sem as janelas o usuário teria só uma forma de visualização: a linha de comando
- Uma das janelas é a linha de comando
- Uma janela permite editar e executar os scripts, markdown ou notebooks (o R é uma linguagem interpretada, opera linha por linha, o script é só uma sequência de comandos)
- Uma janela permite administrar pacotes, acessar o help, visualizar saídas gráficas, e outros
- Uma janela mostra as variáveis em uso
- Menu do Rstudio permite executar alguns comandos sem ter de digitar por extenso

# Softwares Estatísticos Conhecidos

- R
- Systat, SPSS, SAS, MVSP
- Bioestat - Ayres (gratuito, distribuído pela Soc Civ Mamirauá)
- Vários sites de análise online
- Cuidado: cada software usa um algoritmo próprio para implementar análises, podendo estar sujeitos a erros de aproximação ou “bugs”. Verificar a documentação e notícias na internet.
- É importante padronizar a análise para permitir replicação. Procure publicar seus dados originais junto com os artigos, e use programas de amplo uso quando possível