# Prática em R - Aula 07

Análise exploratória e visualização de dados

Prof. Fabio Cop (*fcferreira@unifesp.br*)

Instituto do Mar - Unifesp

2026-03-29

## Conteúdo

**Curso:** Bacharelado Interdisciplinar em Ciências do Mar
**Unidade Curricular (UC):** Probabilidade e Estatística
**Atividade:** Prática no Laboratório de Informática

# 1 Morfometria de passeriformes — distribuição e comparações entre grupos

Os dados de morfometria de passeriformes de Zuur et al. (2009), arquivo `Sparrows.txt`, registram medidas de comprimento de asa (`wingcrd`), comprimento do tarso (`tarsus`), comprimento do crânio (`head`) e peso corporal (`wt`), além de variáveis categóricas de sexo (`Sex`) e classe etária (`Age`). O arquivo usa separador de tabulação e exige o argumento `sep = "\t"` na leitura.

## 1.1 Carregamento e inspeção inicial

***Orientação***

Carregue o arquivo `Sparrows.txt` com a função `read.table()`, especificando `header = TRUE` para que a primeira linha seja interpretada como nomes de colunas e `sep = "\t"` para o separador de tabulação.

Após o carregamento, use as seguintes funções para inspecionar o arquivo:

- `str()` — exibe o tipo de cada coluna e os primeiros valores.
- `summary()` aplicado às colunas `wingcrd`, `tarsus`, `head` e `wt` — produz o resumo numérico das variáveis morfométricas.
- `table()` aplicado às colunas `Sex` e `Age` separadamente — conta o número de observações em cada categoria.

**O que observar**

- No resultado de `str()`, verifique quais variáveis são numéricas e quais são representadas como inteiros codificando categorias. A distinção importa para o mapeamento de cores em `ggplot2`.
- No `summary`, compare o intervalo de cada variável morfométrica. Qual variável apresenta maior amplitude relativa em relação ao seu valor médio?
- A distribuição de machos e fêmeas no dataset é equilibrada ou um dos sexos concentra a maioria das observações?

## 1.2 Histograma e escolha de intervalos

O histograma divide o intervalo de uma variável contínua em classes e conta as observações em cada classe. O parâmetro `bins` controla o número de intervalos e influencia a percepção da forma da distribuição: intervalos demais criam irregularidades que dificultam identificar o padrão geral, e intervalos de menos suavizam a distribuição a ponto de ocultar detalhes relevantes.

***Orientação***

Construa três histogramas do peso corporal (`wt`) usando `ggplot()` com `geom_histogram()`, variando o argumento `bins` entre os valores 10, 30 e 60. Em cada gráfico:

- Mapeie `wt` ao eixo $x$ dentro de `aes()`.
- Use os argumentos `fill` e `color` de `geom_histogram()` para definir a cor de preenchimento das barras e a cor da borda.
- Adicione rótulos aos eixos com `labs()` e aplique `theme_bw()`.

Para exibir os três gráficos lado a lado, carregue o pacote `patchwork` e combine os objetos gerados com o operador `+` ou `/`.

**O que explorar**

- Compare os três histogramas. Em qual versão a forma geral da distribuição fica mais clara? Em qual surgem barras isoladas que parecem ruído em vez de padrão real?

- A distribuição do peso corporal parece simétrica em torno de um valor central ou apresenta assimetria? A presença de valores atípicos (barras isoladas nas extremidades) é visível em alguma das três versões?

### 1.3 Diagrama de caixa com pontos individuais sobrepostos

O diagrama de caixa resume a distribuição de uma variável contínua por meio de cinco descritores: o primeiro quartil ($Q_1$), a mediana, o terceiro quartil ($Q_3$), as hastes (que se estendem até 1,5 vezes o intervalo interquartílico além das bordas da caixa) e os pontos além das hastes, que são potenciais valores atípicos. A sobreposição dos pontos individuais com `geom_jitter()` torna visível a distribuição real dentro de cada grupo.

***Orientação***

Construa um diagrama de caixa do peso corporal por sexo usando `ggplot()` com `geom_boxplot()`. Mapeie `factor(Sex)` ao eixo $x$ e ao argumento `fill` dentro de `aes()`. Em `geom_boxplot()`, defina `outlier.shape = NA` para suprimir a marcação automática de valores atípicos (os pontos individuais serão exibidos pela camada seguinte) e use `alpha` para tornar a caixa semitransparente.

Adicione uma camada `geom_jitter()` sobre o diagrama de caixa. Ajuste `width` para controlar a dispersão horizontal dos pontos e `alpha` para controlar a transparência. O argumento `size` reduz o tamanho dos pontos quando o número de observações é grande.

Finalize com `labs()` para nomear os eixos e a legenda, `theme_bw()` para o tema e `theme(legend.position = "none")` para remover a legenda redundante.

**O que observar**

- A mediana de machos e fêmeas está na mesma altura ou uma é visivelmente maior do que a outra?
- As caixas dos dois grupos se sobrepõem ou estão separadas? A sobreposição indica que a distinção entre os sexos com base no peso corporal é parcial.
- Os pontos individuais revelam alguma concentração ou padrão dentro de cada grupo que a caixa não mostra?

## 2 Morfometria de passeriformes — relações bivariadas e multivariadas

Com a distribuição de cada variável e as comparações entre grupos examinadas, o passo seguinte é explorar as relações entre pares de variáveis morfométricas. O diagrama de dispersão mostra a direção, a força e a forma de cada relação individualmente. O gráfico de pares organiza todas as combinações possíveis em uma grade.

### 2.1 Diagrama de dispersão com variável de grupo

O mapeamento de uma variável categórica para a cor dos pontos acrescenta uma terceira dimensão ao diagrama de dispersão, permitindo verificar se a relação entre as variáveis contínuas varia entre os grupos.

***Orientação***

Construa um diagrama de dispersão usando `ggplot()` com `geom_point()`. Mapeie `tarsus` ao eixo $x$, `wingcrd` ao eixo $y$ e `factor(Sex)` ao argumento `color`, todos dentro de `aes()`. Use `alpha` em `geom_point()` para reduzir a sobreposição visual dos pontos. Nomeie os eixos e a legenda com `labs()` e aplique `theme_bw()`.

**O que observar**

- A relação entre comprimento do tarso e comprimento de asa é aproximadamente linear ou apresenta curvatura visível?

- Machos e fêmeas ocupam regiões distintas do espaço tarso-asa ou as nuvens de pontos se sobrepõem amplamente?
- A dispersão em torno da relação geral é uniforme ao longo dos valores de `tarsus` ou aumenta com o tamanho? Uma dispersão crescente com o preditor é chamada de heterocedasticidade.

## 2.2 Gráfico de pares das variáveis morfométricas

O gráfico de pares exibe, em uma grade, todos os diagramas de dispersão possíveis entre $p$ variáveis, além da distribuição de cada variável na diagonal. Com quatro variáveis morfométricas, a grade contém $\frac{4\times3}{2} = 6$ pares distintos.

***Orientação***

Carregue o pacote `GGally`. Caso ele não esteja instalado no ambiente do projeto, execute `renv::install("GGally")` uma única vez antes de prosseguir.

Use a função `select()` do pacote `dplyr` para extrair as colunas `wingcrd`, `tarsus`, `head` e `wt` do dataset. Passe o resultado para `ggpairs()` com um `aes()` contendo `alpha = 0.3`. O operador pipe nativo do R (`|>`) facilita o encadeamento das etapas.

**O que observar**

- Identifique o par de variáveis com o maior coeficiente de correlação exibido na parte superior da grade. O diagrama de dispersão correspondente mostra pontos próximos de uma linha reta?
- Alguma variável apresenta distribuição claramente assimétrica na diagonal da grade? Uma assimetria marcante pode indicar a necessidade de transformação antes da modelagem.
- Há pares de variáveis com coeficiente de correlação próximo de zero? O que isso implica sobre a informação que cada uma acrescenta a um modelo de predição?

## 2.3 Gráfico de pares separado por classe etária

O argumento `color` no `aes()` do `ggpairs()` separa as observações por grupo em todos os painéis da grade ao mesmo tempo.

***Orientação***

Repita o procedimento da seção anterior, agora incluindo a coluna `Age` no `select()`. Mapeie `factor(Age)` ao argumento `color` dentro do `aes()` de `ggpairs()`, mantendo `alpha = 0.3`.

Para comparar padrões, repita a análise substituindo `Age` por `Sex` no mapeamento de cor.

**O que explorar**

- As correlações entre variáveis morfométricas variam entre classes etárias? Compare os coeficientes de correlação exibidos na parte superior da grade para cada grupo.
- As distribuições na diagonal diferem em localização ou dispersão entre as classes etárias? Uma separação clara na localização indica que o tamanho corporal varia com a idade.
- Substitua `Age` por `Sex` para comparar machos e fêmeas. Em qual agrupamento as diferenças entre grupos são mais visíveis?

# 3 Alometria de mariscos — transformação logarítmica

O dataset de Zuur et al. (2009), arquivo `Clams.txt`, contém comprimento corporal (`LENGTH`) e biomassa medida como massa seca livre de cinzas (`AFD`) de mariscos, além de versões já transformadas (`LNLENGTH` e `LNAFD`) e o mês de coleta (`MONTH`). A relação comprimento-biomassa segue uma lei de potência: na escala original, ela aparece como curva com crescimento acelerado. Na escala logarítmica, converte-se em uma relação linear, com o coeficiente alométrico $\beta_1$ como inclinação da reta ajustada.

## 3.1 Relação na escala original

***Orientação***

Carregue o arquivo `Clams.txt` com `read.table()`, usando `header = TRUE` e `sep = "\t"`. Inspecione as variáveis disponíveis com `str()`.

Construa um diagrama de dispersão usando `ggplot()` com `geom_point()`, mapeando `LENGTH` ao eixo $x$ e `AFD` ao eixo $y$. Use `alpha` para transparência e defina uma cor com o argumento `color` em `geom_point()`. Nomeie os eixos com `labs()` e aplique `theme_bw()`.

**O que observar**

- A relação entre comprimento e biomassa é aproximadamente linear ou segue uma curva com crescimento acelerado?
- A dispersão dos pontos em torno da tendência central é uniforme ao longo do eixo `LENGTH` ou aumenta com o comprimento? Esse padrão de dispersão crescente é característico de relações de potência na escala original.

## 3.2 Relação na escala logarítmica

O dataset `Clams.txt` já contém as colunas `LNLENGTH` e `LNAFD`, com os logaritmos naturais do comprimento e da biomassa. A adição de uma reta de regressão com `geom_smooth()` permite avaliar visualmente se a relação log-log é bem descrita por uma linha reta.

***Orientação***

Construa um novo diagrama de dispersão usando as colunas `LNLENGTH` no eixo $x$ e `LNAFD` no eixo $y$. Adicione uma camada `geom_smooth()` com os argumentos `method = "lm"` e `se = FALSE` para sobrepor uma reta de regressão linear sem banda de confiança. Use o argumento `color` em `geom_smooth()` para diferenciar a cor da reta em relação aos pontos.

**O que explorar**

- Compare este gráfico com o da escala original. A dispersão em torno da relação parece mais uniforme ao longo do eixo $x$ na escala logarítmica?
- A reta ajustada por `geom_smooth(method = "lm")` descreve bem a nuvem de pontos nessa escala? O que isso indica sobre a adequação da lei de potência para descrever a relação comprimento-biomassa dos mariscos?

## 3.3 Variação sazonal com facetas

A variável `MONTH` indica o mês de coleta de cada marisco. `facet_wrap()` cria um painel separado para cada mês, com as mesmas escalas em todos os painéis, permitindo comparar o padrão alométrico ao longo do ano.

***Orientação***

Parta do gráfico da escala logarítmica construído na seção anterior (com `geom_point()` e `geom_smooth(method = "lm", se = FALSE)`). Acrescente `facet_wrap(~ MONTH)` como uma camada adicional para dividir o gráfico em painéis por mês de coleta. Mantenha os mesmos rótulos de eixo com `labs()` e `theme_bw()`.

**O que observar**

- A inclinação da reta ajustada é semelhante em todos os meses ou varia de forma visível ao longo do ano? Uma variação na inclinação indicaria que o padrão alométrico muda sazonalmente.
- Todos os meses têm o mesmo número de observações? Painéis com poucos pontos produzem retas ajustadas com maior incerteza.

# 4 Bioluminescência oceânica — variação entre estações

O dataset de Zuur et al. (2009), arquivo `ISIT.txt`, registra contagens de fontes bioluminescentes (`Sources`) em diferentes profundidades de amostragem (`SampleDepth`) em estações do Atlântico Norte. Cada estação (`Station`) corresponde a um local geográfico com múltiplas amostras coletadas em diferentes profundidades. A relação profundidade-bioluminescência pode variar entre estações em função de diferenças oceanográficas locais. Examinar se esse padrão é consistente entre locais motiva abordagens de modelagem que estimam a variação entre grupos de forma explícita.

## 4.1 Padrão global profundidade-bioluminescência

***Orientação***

Carregue o arquivo `ISIT.txt` com `read.table()`, usando `header = TRUE` e `sep = "\t"`. Inspecione as variáveis disponíveis com `str()`.

Construa um diagrama de dispersão usando `ggplot()` com `geom_point()`, mapeando `SampleDepth` ao eixo $x$ e `Sources` ao eixo $y. Usealphaesizeemgeom_point()para lidar com a sobreposição de pontos. Nomeie os eixos comlabs()e apliquetheme_bw()‘.

**O que observar**

- Existe uma tendência visível entre profundidade e número de fontes bioluminescentes no padrão global?
- A dispersão dos pontos em torno da tendência geral é pequena ou grande? Alta dispersão pode indicar que as estações diferem entre si na relação profundidade-bioluminescência.

## 4.2 Variação entre estações com facetas

***Orientação***

Parta do gráfico construído na seção anterior. Acrescente `facet_wrap(~ Station)` para criar um painel separado por estação amostral. Mantenha os mesmos mapeamentos estéticos, rótulos de eixo e tema.

**O que observar**

- A relação profundidade-bioluminescência é consistente entre as estações ou algumas apresentam padrões claramente distintos das demais?
- A profundidade máxima amostrada varia entre estações, o que resulta em painéis com escalas de $x$ aparentemente diferentes. Em quais estações a relação é mais pronunciada?
- Compare a variação dentro de cada painel (variação interna à estação) com a variação entre os painéis (variação entre estações). Em qual escala a variabilidade é maior?

Zuur, Alain F., Elena N. Ieno, Neil J. Walker, Anatoly A. Saveliev, e Graham M. Smith. 2009. *Mixed Effects Models and Extensions in Ecology with R*. Springer.