

# Tutorial Biologia Evolutiva 2017

*Diogo Melo*

*October 12, 2016*

## Contents

<b>Introdução</b>	<b>1</b>
Genética quantitativa e evolução . . . . .	1
A linguagem R . . . . .	1
Seguindo o tutorial . . . . .	2
Preliminares . . . . .	2
Aos dados! . . . . .	3
<b>Objetivos da Aula 1</b>	<b>3</b>
Médias e variâncias globais . . . . .	3
Médias e variâncias por espécie . . . . .	4
Gráficos . . . . .	4
<b>Leitura da Aula 2</b>	<b>5</b>

## Introdução

Primeiro alguns links:

- Roteiro para elaboração da monografia final
- Apostila da disciplina

## Genética quantitativa e evolução

Neste tutorial vamos passar por todas as etapas de análise de um conjunto de dados em genética quantitativa aplicada à macroevolução. Vamos estudar quatro caracteres quantitativos medidos em cinco espécies de roedores em uma filogenia conhecida. Usando esse conjunto de dados, vamos caracterizar o padrão de modularidade em cada espécie, as diferenças nas morfologias médias, e os processos evolutivos que levaram à diferenciação atual.

O primeiro passo, como sempre, é escolher um ambiente de análise. Neste tutorial vamos usar a linguagem R, mas todos os passos poderiam ser feitos num programa gráfico como excel e Systat, ou em qualquer outra linguagem de programação, como python, perl, java, C ou Haskell. A linguagem R é especialmente conveniente, primeiro por que os monitores sabem usar ela, e segundo por que existe um pacote (escrito pelos monitores) que automatiza várias dos passos da análise.

## A linguagem R

R é um ambiente de programação *open source* criado pensando em análises estatísticas. Você pode ler um pouco mais sobre a linguagem neste site.

## Instalando o R

Se você está utilizando um dos computadores do IB, o R já deve estar instalado. Se não, siga as instruções nesse site para instalar o R, e opcionalmente (mas muito recomendado!) o editor de código RStudio.

## Rodando comandos

Ao longo desse tutorial, você sempre vai ver blocos de código que devem ser executados no ambiente R. Por exemplo, se estivéssemos interessados em somar dois números poderíamos digitar:

```
1 + 1
```

```
## [1] 2
```

E abaixo da operação temos o resultado que o programa retorna.

Você pode também associar valores a variáveis (chamados objetos) e utilizar esses objetos nas suas operações. O operador “=” associa um valor a um objeto:

```
x = 1
```

```
y = 2
```

```
x + y
```

```
## [1] 3
```

## Seguindo o tutorial

Existem dois jeitos de seguir esse tutorial: (1) cegamente copiar e colar os comandos no R ver o resultado; ou (2) entender o que cada parte do comando executado está fazendo e assim já aprender um pouco de R de brinde. Ambas são ok! Nosso objetivo aqui é a biologia evolutiva, o R é só um acessório. Quando alguma operação no R parecer mais esotérica, vamos tentar explicar nos comentários do código ou no texto. Se não ficar claro, consulte os monitores!

## Preliminares

Para começar, precisamos instalar um pacote (chamado *evolqg*) e carregar os dados para a área de trabalho usando a função *data*. Funções no R são como pequenos conjuntos de instruções com um nome. Para executar uma função basta digitar o nome dela, com os argumentos daquela função entre parênteses. Linhas que começam com # são comentários e são ignoradas pelo R. (Para ver mais exemplos iniciais, confira esse tutorial do curso de pós-graduação em R oferecido pela ecologia do IB!)

```
# Instalando o pacote de genética quantitativa evolutiva  
if(!require(evolqg)){install.packages("evolqg")}
```

```
# Carregando o pacote evolqg  
library(evolqg)
```

```
# Carregando o conjunto de dados que vamos utilizar  
data(dentus)
```

```
## Loading required package: plyr
```

## Aos dados!

Com os dados carregados, podemos ver todas as medidas digitando o nome do objeto com os dados.

Aqui vamos ver só as primeiras linhas, mas no seu computador você deve ver o conjunto completo.

```
dentus
```

```
##      humerus      ulna      femur      tibia species
## 1 3.214619  8.036072 13.51779 19.01131      A
## 2 5.222875 10.838375 15.54714 21.01718      A
## 3 5.193021 11.911770 17.24719 22.43190      A
## 4 6.547400 11.293085 14.89040 20.17706      A
## 5 4.724383  9.897135 14.81104 19.59753      A
## 6 4.630147  9.441174 15.24754 19.25937      A
```

Como podemos ver, os dados estão organizados em linhas e colunas. Cada linha representa um indivíduo, e cada coluna representa um caráter. As 4 primeiras colunas são os caracteres quantitativos, e a quinta coluna contém a espécie de cada indivíduo.

Essa mesma planilha está disponível nesse link.

## Objetivos da Aula 1

Em cada aula vamos fazer algumas contas e discutir os resultados. Ao final do curso todos esses resultados devem ser compilado num relatório final.

Nessa aula vamos calcular:

1. Média de cada caráter;
2. Variância de cada caráter;
3. Coeficiente de variação de cada caráter;
4. Gráfico da distribuição de cada caráter.

## Médias e variâncias globais

Vamos começar calculando as médias globais utilizando a função `sapply` e a função `mean`.

Os resultados não aparecem aqui, mas rodando os comandos no R você deve ser capaz de produzi-los!

```
# Como a última coluna contém as espécie,
# vamos utilizar o operador de indexação do R [] para selecionar as colunas de 1 a 4.
# A função sapply passa todas as colunas do primeiro argumento para a função no segundo argumento.
sapply(dentus[,1:4], mean)
```

Podemos utilizar a mesma estratégia para calcular as variâncias e desvios padrões:

```
# Variâncias
sapply(dentus[,1:4], var)

# Desvios padrão
sapply(dentus[,1:4], sd)
```

Para calcular o coeficiente de variação, podemos dividir diretamente o desvio padrão pela média:

```
# Objeto com os desvios padrão
dentus_sd = sapply(dentus[,1:4], sd)
```

```
# Objeto com as médias
dentus_mean = sapply(dentus[,1:4], mean)
```

```
# Cálculo do coeficiente de variação
dentus_sd/dentus_mean
```

Ou podemos criar uma função que calcula o coeficiente de variação diretamente

```
# Função que calcula o coeficiente de variação
cv = function(x) sd(x)/mean(x)
```

```
# Coeficiente de variação
sapply(dentus[,1:4], cv)
```

## Médias e variâncias por espécie

Agora vamos tentar obter esses mesmos valores separando os dados por espécie. A função `aggregate` vai ajudar. Essa função é muito semelhante à `sapply`, mas ela recebe um argumento adicional que permite utilizar uma coluna dos dados para separar os indivíduos em grupos. Por exemplo, para calcular a média por espécie fazemos:

```
aggregate(dentus[,1:4], list(especie = dentus[,5]), mean)
```

##	especie	humerus	ulna	femur	tibia
## 1	A	4.948718	9.949203	15.11119	20.17994
## 2	B	8.449116	17.043157	25.58400	34.03991
## 3	C	10.086848	20.234410	21.54564	28.41128
## 4	D	10.764505	21.917249	27.70168	37.03902
## 5	E	7.396516	14.408397	20.55298	27.52810

Vc consegue seguir essa lógica e calcular as variâncias e coeficientes de variação por espécie?

## Gráficos

É sempre importantíssimo visualizar os dados usando gráficos antes de fazer qualquer análise mais complexa. Vamos explorar várias possibilidades de gráficos que revelam aspectos diferentes da distribuição dos nossos dados.

Para isso vamos carregar alguns pacotes gráficos:

### Carregando pacotes gráficos

```
list_pkgs <- c("ggplot2", "GGally", "reshape2")
new_pkgs <- list_pkgs[!(list_pkgs %in% installed.packages()[,"Package"])]
if(length(new_pkgs) > 0){ install.packages(new_pkgs) }

library(evolqg)
library(ggplot2)
library(GGally)
library(reshape2)
```

## Gráficos univariados

O primeiro e mais simples talvez seja o histograma.

```
ggplot(melt(dentus), aes(value, group = species, fill = species)) +  
  geom_histogram() +  
  facet_wrap(~variable, scale = "free")
```

Podemos também usar boxplots:

```
ggplot(melt(dentus), aes(species, value, fill = species, group = interaction(species, variable))) +  
  geom_boxplot() +  
  facet_wrap(~variable, scale = "free")
```

Ou um chiquérrimo violin plot:

```
ggplot(melt(dentus), aes(species, value, fill = species, group = interaction(species, variable))) +  
  geom_violin() +  
  facet_wrap(~variable, scale = "free")
```

## Gráficos bivariados

Outro aspecto interessante é a relação par a par das variáveis.

Por exemplo, vamos olhar para a distribuição bivariada dos dois primeiros caracteres, incluindo também uma reta de regressão por espécie:

```
ggplot(dentus, aes(humerus, ulna, group = species, color = species)) +  
  geom_point() +  
  geom_smooth(method = "lm", aes(group = species))
```

Vc poderia fazer esse gráfico manualmente para todos os pares de caracteres, mas a função `ggpairs` resume tudo isso de uma vez:

```
ggpairs(dentus, mapping = aes(color = species),  
        columns = c("humerus", "ulna", "femur", "tibia"))
```

## Leitura da Aula 2

- The Ecological Significance of Correlation Pleiades (pdf)

**Abstract:** The phenomenon of correlation pleiades, i.e., the presence of correlations between some quantitative characteristics of certain parts of an organism and at the same time by the absence of correlations between these and the other parts demonstrates the independence of certain developmental processes with respect to other processes within the organism. The study of correlation sets makes it possible to establish the degree of such independence. The adoption of a quantitative criterion provides an adequate basis for the comparative method in the study of stabilizing selection. Nineteen species of herbaceous plants have been studied in order to determine the variability in the degree of dependence between the dimensions of different parts of an organism. It has been established that twelve species have correlation pleiades, while seven species have no correlation pleiades with respect to the dimensions of analogous parts. In all the instances of existence of pleiades it was the flower that had become free of the correlation interdependences within the plant organism. Existence of correlation pleiades is associated with the mode of pollen transfer, involving the presence of specific insect pollinators, with adaptations to localize the pollen deposit on some definite part of the body of the pollinating insect, and with at least one of the structural features providing for such localization (zygomorphism,

presence of tubular parts, lateral position of flowers, oligomerization and a rigid fixation of the number of homologous parts of a flower). Insects were the selective factor stabilizing the dimensions of those parts of flowers which participate in the process of the precise localization of pollen on the insect body. Stabilizing selection of flower dimensions created correlation pleiades for reproductive organs, as well as for vegetative parts of the same organism. The general principle underlying the origin of developmental homeostasis in the process of evolution is the discrepancy between the agencies participating in the formation of the character and the selective forces determining its function. The same principle is responsible for the origin of correlation pleiades.

$$Cor(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y} = \frac{1}{N} \sum_{i=1}^N \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$