CAPÍTULO 2

Análisis descriptivo y pruebas de supuestos para análisis de varianza

Tatiane Caroline Grella, José Bruno Malaquias, Jéssica Karina da Silva Pachú, Milton Fernando Cabezas Guerrero

Resumen

En este capítulo, describimos de una manera simple y objetiva elementos esenciales para el análisis descriptivo en R usando boxplot y también exploramos algunas líneas de comando que son esenciales para las pruebas de normalidad y homogeneidad de varianzas, para un posterior análisis de varianza y la prueba de comparación de dos medias, en este caso prueba t, y comparación múltiple de medias - prueba de Tukey. Todas las líneas de comando se presentan paso a paso de forma comentada y detallada. Usaremos los siguientes paquetes: **readxl** y **ExpDes.pt**. Los comandos a ejecutar en R se muestran en color azul.

Palabras claves: boxplot; normalidad; homocedasticidad; variables continuas; pruebas de comparação; ANOVA.

1. Lectura de archivos en R

Para acceder a la base de datos y el script usados como ejemplo en este capítulo, haga **clic aquí**.

Para iniciar el análisis, es necesario descargar la base de datos y el script (disponible arriba) y luego verificar en qué parte de la computadora (directorio) se encuentra el archivo a analizar.

Para eso, usemos el comando **getwd**, que verificará en qué directorio estás trabajando: **getwd().**

Si necesita cambiar el directorio, simplemente siga los pasos:

Session -> Set working Directory -> Choose Directory

Luego de elegir un directorio es posible ver qué archivos existen en él, para eso usamos la función: **list.files()** la cual mostrará la lista de archivos dentro de su directorio.

Después de elegir el directorio, leemos el archivo que se analizará.

Para leer el archivo de Excel, necesitaremos el paquete: **readxl** [1].

Una forma elegante de cargar el paquete es usar la siguiente línea de comando (esta línea también funciona si el paquete no está instalado, ya que la instalación se llevará a cabo automáticamente):

if(!require("readxl")) install.packages("readxl"); require(readxl)

El signo de exclamación indica denegación, por lo que la línea de comando anterior se traduce como: *"si el paquete requerido (readxl) no está instalado, instale el paquete, luego cárguelo"*.

2. Análisis descriptivo con bloxplot

Luego de cargar el paquete, necesitamos leer el archivo, para eso usaremos la línea de comando:

df<-read_excel("BD1.xls ", sheet = 1)

- df es el nombre que se le da al dataframe (puede poner el nombre que desee);
- read_excel es el comando para leer el archivo de Excel;
- **BD1** es el nombre de su archivo dentro de la carpeta seleccionada;
- xls es la extensión del archivo con el que está trabajando;
- sheet = 1 se refiere a qué hoja de su archivo de Excel desea analizar.

Leamos el Archivo: **BD1.xls**.

En este ejemplo didáctico se analizó el efecto de dos tratamientos sobre el peso de los insectos (Tabla 1). El diseño experimental fue completamente al azar y con solo 3 repeticiones.

Tabla 1. Peso (g) de una especie de insecto hipotética sometida a dos tratamientos.

TRATAMIENTO	REPETICIÓN	PESO
A	1	0,0750
A	2	0,0810
A	3	0,0820
B	1	0,0780
B	2	0,0790
B	3	0,0800

Para ver el encabezado usamos la función: **head(df)**

Para ver la base de datos usamos la función: **View(df)**

Para expresar los datos en un boxplot y construirlo, usamos la función:

```
boxplot(PESO~TRATAMIENTO, data=df)
```

Después de la construcción, es posible cambiar varios elementos, como:

- nombre en los ejes, usando el comando:

```
boxplot(PESO~TRATAMIENTO, data=df,
        xlab = "Tratamientos",
        ylab = "Matéria seca (g)")
```

En este ejemplo, los nombres de los ejes son "Tratamientos" "Materia seca (g)"

- nombre en los ejes y con límite inferior y superior.

```
boxplot(PESO~TRATAMIENTO, data=df,
        xlab = "TratamIentos",
        ylab = "Matéria seca (g)"
        , ylim = c(0.07,0.085))
```

- nombre en los ejes y con límite superior e inferior y sumando la media dentro de boxplot.

```
boxplot(PESO~TRATAMIENTO, data=df, col= "white",
      xlab= "Tratamientos", ylab= "Matéria Seca (g)", ylim = c(0.07,0.085))
points(1:nlevels(TRATAMIENTO), tapply(PESO, TRATAMIENTO, mean),
data=df)
abline(mean(TRATAMIENTO), data=df)
```

Para exportar la imagen en el formato deseado una sugerencia es el formato tiff con 300 dpi, que suele ser el formato solicitado por las publicaciones científicas.

Para cambiar los tamaños, use: **cex.main** (título), **cex.lab** (etiquetas) y **cex.axis** (tamaño de los ejes).

```
tiff("Figura_BoxPlot_.tiff", width=12, height=8, units="in", res=300)
boxplot(PESO~TRATAMIENTO, data=df, col= "white",
      xlab= "Tratamiento", ylab= "Matéria Seca (g)",
      ylim = c(0.07,0.085),
      cex.main=1.5, cex.lab=1.5, cex.axis=1.5)
        dev.off()
```

NOTA: tenga en cuenta su directorio, ya que la imagen se exportará directamente a esta carpeta.
Para elegir el directorio use **Control + Shift + H** o simplemente use **getwd()**

La Figura 1 es el boxplot exportado. Cada flecha expresa un parámetro del análisis descriptivo, es decir, los valores mínimo, máximo y cuartil (1º, 2º y 3º).

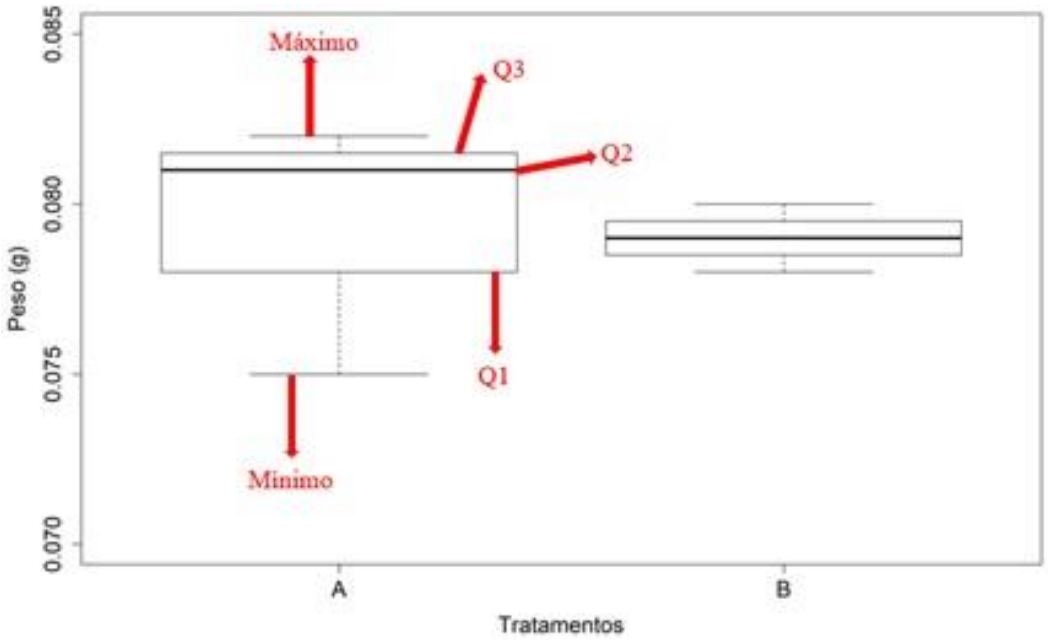

Figura 1. Diagrama de caja que representa el efecto de dos tratamientos sobre el peso (g) de una especie de insecto hipotética. **Q1**: primer cuartil. **Q2**: segundo cuartil o mediana. **Q3**: tercer cuartil.

Usando la línea de comando a continuación, será posible ver los mismos valores expresados en el diagrama de caja, es decir, valor mínimo (Min), primer cuartil (1st Qu.), Mediana (Mediana), media (Media), tercer cuartil (3rd Qu.) y máximo (Max), por lo que se mostrará un resumen del análisis descriptivo del peso (g) de una hipotética especie de insecto sometida a cada uno de los dos tratamientos.

tapply(df$PESO, df$TRATAMIENTO, summary)

```
$A
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.07500 0.07800 0.08100 0.07933 0.08150 0.08200

$B
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0780  0.0785  0.0790  0.0790  0.0795  0.0800
```

3. Prueba de los supuestos del modelo ANOVA

Usaremos la misma base de datos usada previamente, así que lea el archivo: BD1.xls (Tabla 1) usando el script que está disponible haciendo clic aquí. Como se muestra en la Tabla 1 y se comentó anteriormente, en este ejemplo didáctico se analizó el efecto de dos tratamientos sobre el peso de los insectos. El diseño experimental fue completamente al azar y con solo 3 repeticiones.

Cargue el paquete **readxl** para leer el archivo de Excel:

require(readxl) si tiene preguntas sobre la carga e instalación de paquetes, consulte el tema "Lectura de archivos en R".

Leer la base de datos: **df<- read_excel("BD1.xlsx", sheet = 1)**

Ver encabezado: **head(df)**

Ver la base de datos: **View(df)**

Antes de realizar el análisis **ANOVA**, es necesario probar la normalidad y homogeneidad de las varianzas.

Para probar la normalidad, usamos la prueba de Shapiro Wilk, usando:

shapiro.test(df$PESO)

Prueba de Shapiro: Si el valor *p* es mayor que 0.05 (5%), hay normalidad de los datos.

Para probar la homogeneidad, en un estudio realizado en DCA, con un solo factor, utilizamos la prueba de Bartlett, utilizando:

bartlett.test(df$PESO, df$TRATAMIENTO)

Prueba de Bartlett: si el valor *p* es mayor que 0.05, las varianzas son homogéneas.

NOTA: si sus datos son variables continuas y no cumplen con la normalidad y/u homogeneidad, serán necesarias transformaciones.

4. Prueba *t*

Continuaremos usando la misma base de datos que se utilizó anteriormente, pero con la modificación del script al que se puede acceder haciendo clic aquí, luego leer el archivo: **BD1.xls** (Tabla 1). Como solo hay dos tratamientos, aplicaremos una prueba *t*. Pero antes de eso, necesitaremos probar los supuestos de normalidad y homogeneidad de las varianzas. Para obtener más detalles, consulte el tema: “Prueba de los supuestos del modelo Anova”.

Cargue el paquete **readxl** para leer el archivo de Excel:
require(readxl) si tiene preguntas sobre la carga e instalación de paquetes, consulte el tema "Lectura de archivos en R".

Leer la base de datos: **df<- read_excel("BD1.xlsx", sheet = 1)**

Ver encabezado: **head(df)**

Ver la base de datos: **View(df)**

Prueba de Shapiro Wilk: **shapiro.test(df$PESO)**

Prueba de Bartlett para un estudio realizado en DCA - - con un solo factor:

bartlett.test(df$PESO, df$TRATAMIENTO)

Para aplicar la prueba *t*, use la función **t.test.**

- El peso es la variable de respuesta;
- El tratamiento es la variable independiente.

Puedes usar las opciones: "**two.sided**", "**less**" ou "**greater**".

t.test(PESO ~ TRATAMIENTO, data = df, alternative = "two.sided")

5. Prueba de Tukey

Usemos la base de datos presente en el Archivo: **Anova1.xlsx** (Tabla 2). El diseño experimental fue completamente al azar con 4 repeticiones. Como hay más de dos tratamientos, aplicaremos una prueba de comparaciones múltiples, en este caso la prueba de Tukey. Recuerde probar los supuestos de normalidad y homogeneidad de las varianzas, para obtener más detalles, consulte el tema: "Prueba de los supuestos del modelo ANOVA".

Tabla 2. Diámetro (mm) del *pronotum* de una hipotética especie de insecto sometida a tres tratamientos

TRATAMIENTO	REPETICION	DIAMETRO
X	1	30
X	2	40
X	3	20
X	4	67
Y	1	20
Y	2	20
Y	3	35
Y	4	45
Z	1	60
Z	2	40
Z	3	50
Z	4	30

Cargue el paquete **readxl** para leer el archivo de Excel:

require(readxl) si tiene preguntas sobre la carga e instalación de paquetes, consulte el tema "Lectura de archivos en R".

Leer la base de datos: **df<- read_excel("BD1.xlsx", sheet = 1)**

Ver encabezado: **head(df)**

Ver la base de datos: **View(df)**

Prueba de Shapiro Wilk: **shapiro.test(df$DIAMETRO)**

Prueba de Bartlett para un estudio realizado en DCA, con un solo factor:

bartlett.test(df$DIAMETRO, df$TRATAMIENTO)

Para este análisis usaremos el paquete ExpDes [2]: **require(ExpDes)**

Para enviar los comandos escritos (data frame) a la memoria: **attach(df)**

La prueba se realizó en un DCA, por lo que usaremos la función **crd**

#en **mcomp**, escoja el método: "**tukey**"

crd(TRATAMENTO, DIAMETRO, quali = TRUE, mcomp = "tukey", nl=FALSE, sigT = 0.05, sigF = 0.05)

Este es el resultado del análisis:

```
De acordo com o teste F, as medias nao podem ser consideradas diferentes.
------------------------------------------------------------------------
  Niveis Medias
1      x  39.25
2      y  30.00
3      z  45.00
------------------------------------------------------------------------
```

Del análisis, queda claro que no hay evidencia de diferencias entre tratamientos. Si desea presentar los resultados usando letras, simplemente asigne la misma letra a las 3 medias, aunque esto sea considerado redundante.

6. Referencias de los paquetes utilizados

[1] WICKHAM H., BRYAN J. readxl: Read Excel Files. R package version 1.3.1. 2019. https://CRAN.R-project.org/package=readxl

[2] FERREIRA E.B., CAVALCANTI P.P., NOGUEIRA D.A. ExpDes.pt: Pacote Experimental Designs (Portuguese). R package version 1.2.0. 2018. https://CRAN.R-project.org/package=ExpDes.pt

7. Referencias recomendadas

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

MATLOFF, Norman. The art of R programming: A tour of statistical software design. No Starch Press, 2011.

PETERNELLI, Luiz Alexandre; MELLO, MP de. Conhecendo o R: uma visão estatística. Viçosa: UFV, v. 1, 2011.

VENABLES W.N., RIPLEY B. D. Modern Applied Statistics with S. Fourth Edition. Springer, New York. 2002.