Boxplots

Alan T. Arnholt
Department of Mathematical Sciences
Appalachian State University
arnholt@math.appstate.edu

Spring 2006 R Notes

*Copyright © 2006 Alan T. Arnholt*
Boxplots

Overview of Boxplots

Problem

Application

The R Script
A popular method of representing the information in the five-number summary is the **boxplot**. To show spread, a box is drawn from the lower hinge ($H_L$) to the upper hinge ($H_U$) with a vertical line drawn through the box to indicate the median or second quartile ($Q_2$).
Whiskers, Fences, and Adjacent Values

- A “whisker” is drawn from $H_U$ to the largest data value that does not exceed the upper fence. This value is called the adjacent value.
Whiskers, Fences, and Adjacent Values

- A “whisker” is drawn from $H_U$ to the largest data value that does not exceed the upper fence. This value is called the adjacent value.
- The upper fence is defined as $Fence_U = H_U + 1.5 \times H_{spread}$ where $H_{spread} = H_U - H_L$. 
Whiskers, Fences, and Adjacent Values

- A “whisker” is drawn from $H_U$ to the largest data value that does not exceed the upper fence. This value is called the adjacent value.
- The upper fence is defined as $Fence_U = H_U + 1.5 \times H_{spread}$ where $H_{spread} = H_U - H_L$.
- A whisker is also drawn from $H_L$ to the smallest value that is larger than the lower fence where the lower fence is defined as $Fence_L = H_L - 1.5 \times H_{spread}$.
Whiskers, Fences, and Adjacent Values

- A “whisker” is drawn from $H_U$ to the largest data value that does not exceed the upper fence. This value is called the adjacent value.
- The upper fence is defined as $Fence_U = H_U + 1.5 \times H_{spread}$ where $H_{spread} = H_U - H_L$.
- A whisker is also drawn from $H_L$ to the smallest value that is larger than the lower fence where the lower fence is defined as $Fence_L = H_L - 1.5 \times H_{spread}$.
- Any value smaller than the lower fence or larger than the upper fence is considered an outlier and is generally depicted with a hollow circle.
Figure 1 on page 14 illustrates a boxplot for the variable `fat` from the data frame `Bodyfat`. Figure 2 on page 16 shows progressively more complicated boxplots using the `boxplot()`.

- To create a boxplot with R, use the command `boxplot(x)`. 

Figure 1 on page 14 illustrates a boxplot for the variable fat from the data frame *Bodyfat*. Figure 2 on page 16 shows progressively more complicated boxplots using the `boxplot()`.

- To create a boxplot with R, use the command `boxplot(x)`.  
- x is either a numeric vector, or a single list containing vectors.
Figure 1 on page 14 illustrates a boxplot for the variable `fat` from the data frame `Bodyfat`. Figure 2 on page 16 shows progressively more complicated boxplots using the `boxplot()`.

- To create a boxplot with R, use the command `boxplot(x)`.
- `x` is either a numeric vector, or a single list containing vectors.
- It is also possible to pass a formula to `boxplot()` of the type `y ~ grp`, where `y` is a numeric vector of data values to be split into groups according to the grouping variable `grp` (usually a factor).
Figure 1 on page 14 illustrates a boxplot for the variable fat from the data frame Bodyfat. Figure 2 on page 16 shows progressively more complicated boxplots using the `boxplot()` function.

- To create a boxplot with R, use the command `boxplot(x)`.
- `x` is either a numeric vector, or a single list containing vectors.
- It is also possible to pass a formula to `boxplot()` of the type `y ~ grp`, where `y` is a numeric vector of data values to be split into groups according to the grouping variable `grp` (usually a factor).
- By default, boxplots in R have a vertical orientation.
Figure 1 on page 14 illustrates a boxplot for the variable `fat` from the data frame `Bodyfat`. Figure 2 on page 16 shows progressively more complicated boxplots using the `boxplot()` function.

- To create a boxplot with R, use the command `boxplot(x)`.
- `x` is either a numeric vector, or a single list containing vectors.
- It is also possible to pass a formula to `boxplot()` of the type `y ~ grp`, where `y` is a numeric vector of data values to be split into groups according to the grouping variable `grp` (usually a factor).
- By default, boxplots in R have a vertical orientation.
- To create a horizontal boxplot with R, use the optional argument `horizontal=TRUE`.
Figure 1 on page 14 illustrates a boxplot for the variable *fat* from the data frame **Bodyfat**. Figure 2 on page 16 shows progressively more complicated boxplots using the `boxplot()`.

- To create a boxplot with R, use the command `boxplot(x)`.
- `x` is either a numeric vector, or a single list containing vectors.
- It is also possible to pass a formula to `boxplot()` of the type `y ~ grp`, where `y` is a numeric vector of data values to be split into groups according to the grouping variable `grp` (usually a factor).
- By default, boxplots in R have a vertical orientation.
- To create a horizontal boxplot with R, use the optional argument `horizontal=TRUE`.
- Common arguments for `boxplot()` include `col=` to set the box color and `notch=TRUE` to add a notch to the box to highlight the median.
Boxplot Illustrated

![Boxplot Illustration](image)

Figure: Graph depicting the five-number summary in relationship to original data and the boxplot
Code for Boxplots

```r
site <- "http://www1.appstate.edu/~arnholta/PASWS/DATA/Bodyfat"
Bodyfat <- read.table(file=url(site),header=T)
attach(Bodyfat)
Bodyfat[1:5,]
> par(mfrow=c(2,2))
> boxplot(fat)
> boxplot(fat~sex,horizontal=TRUE,)
> boxplot(fat~sex,horizontal=TRUE,col=c("pink","blue"),
  + varwidth=TRUE)
> boxplot(fat~sex,horizontal=FALSE,col=c("pink","blue"),
  + varwidth=TRUE, notch=TRUE,main="Boxplot of Fat by Gender")
> legend(x="bottomleft", legend=c("Females", "Males"),
  + fill=c("pink", "blue"))
> par(mfrow=c(1,1))
```
The Boxplots

Figure: Vertical and horizontal boxplots with and without color
Simpson’s Paradox

The boxplots in Figure 3 on the following page and Figure 4 on page 20 are similar to those found on page 57 of BSDA.

```r
> library(BSDA)
> attach(Simpson)
> boxplot(gpa~gender,names=c("Males","Females"),
+ col=c("blue","pink"), ylab="Grade Point Average",
+ main="Side-by-Side Boxplots of GPA by Gender",
+ notch=TRUE)
```
**Side-by-Side Boxplots of GPA by Gender**

![Side-by-Side Boxplots of GPA by Gender](image)

**Figure:** Side-by-side boxplots of GPA by Gender
Code for Figure 4 on the following page

```R
> boxplot(gradept~gender2, col=rep(c("blue","pink"),3),
+ names=c("MBBA","FBBA","MSOC","FSOC","MTRA","FTRA"),
+ notch=TRUE,main="",ylab="Grade Point Average",
+ varwidth=TRUE)
> axis(side=3, at=c(1.5,3.5,5.5),
+ labels=c("basketball","soccer","track"),col.axis="blue",
+ font=2)
> mtext("Figure 1.32 from BSDA Improved",side=3,line=2.5,
+ cex=1.25, col="blue")
```
Duplication of Figure 1.32 from BSDA

Figure 1.32 from BSDA Improved

Figure: Graphical illustration of Simpson’s paradox
• Go to my web page **Script for Boxplots**
• Homework: problems 1.81 - 1.92
• See me if you need help!