data_analysis_1.R

khagen

Sat Feb 22 11:29:49 2020

```r
# $$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
#                                    Data Analysis with R
#                                     Khagendra Adhikari
#                                          2016
# $$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$

# Use packages
library(DAAG)
```

```
## Loading required package: lattice
```

```r
library(lattice)
library(ggplot2)
library(reshape2)
library(latticeExtra)
```

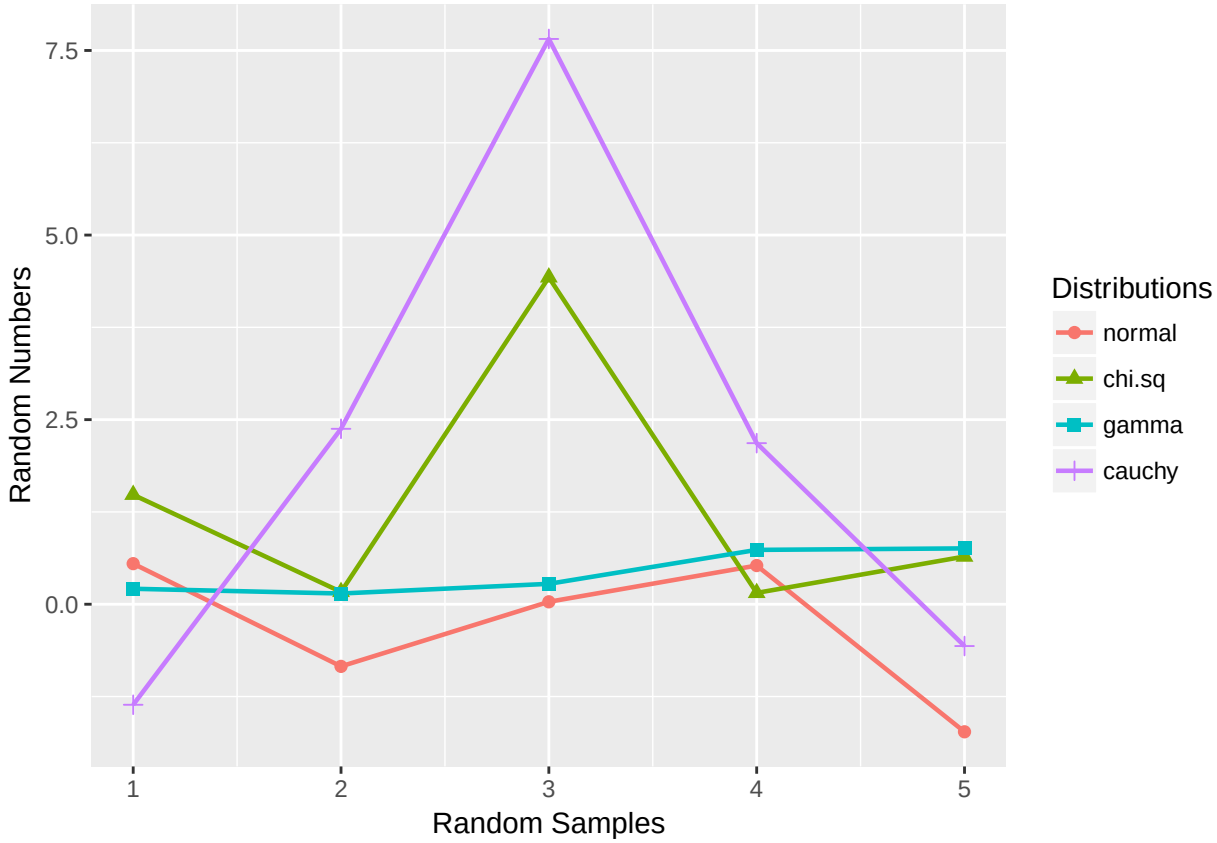
```
## Loading required package: RColorBrewer

##
## Attaching package: 'latticeExtra'

## The following object is masked from 'package:ggplot2':
##
##     layer
```

```r
library(boot)
```

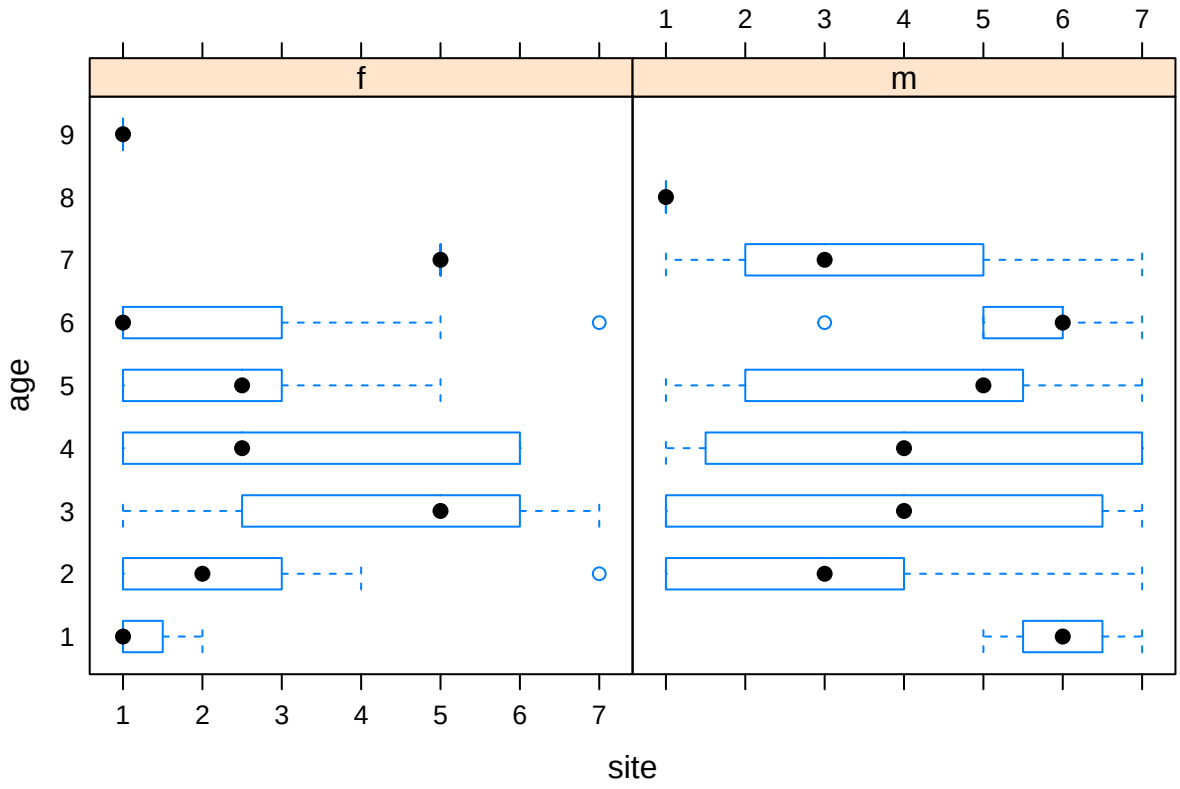
```
##
## Attaching package: 'boot'

## The following object is masked from 'package:lattice':
##
##     melanoma
```

```r
# ***************************************************************************************
# Q.1.
# Draw four random samples, each of size 5. One from Normal distribution, one from Chi-
# square distribution, one from Gamma-distribution and one from Cauchy-distribution.
# Plot the samples so that all of them are in one graph sheet. Use different point
# characters (symbols) and colors in each of the plot.
# (You can use your own selection of distributional parameters when drawing the samples)
# ****************************************************************************************

set.seed(50)
df <- data.frame(x.value = 1:5, normal = rnorm(5), chi.sq = rchisq(5,3),
                  gamma = rgamma(5,1), cauchy = rcauchy(5))

require(reshape2)
melt.df <- melt(df,  id.vars = 'x.value',
                variable.name = 'Distributions', value.name = 'Random.Numbers')
```

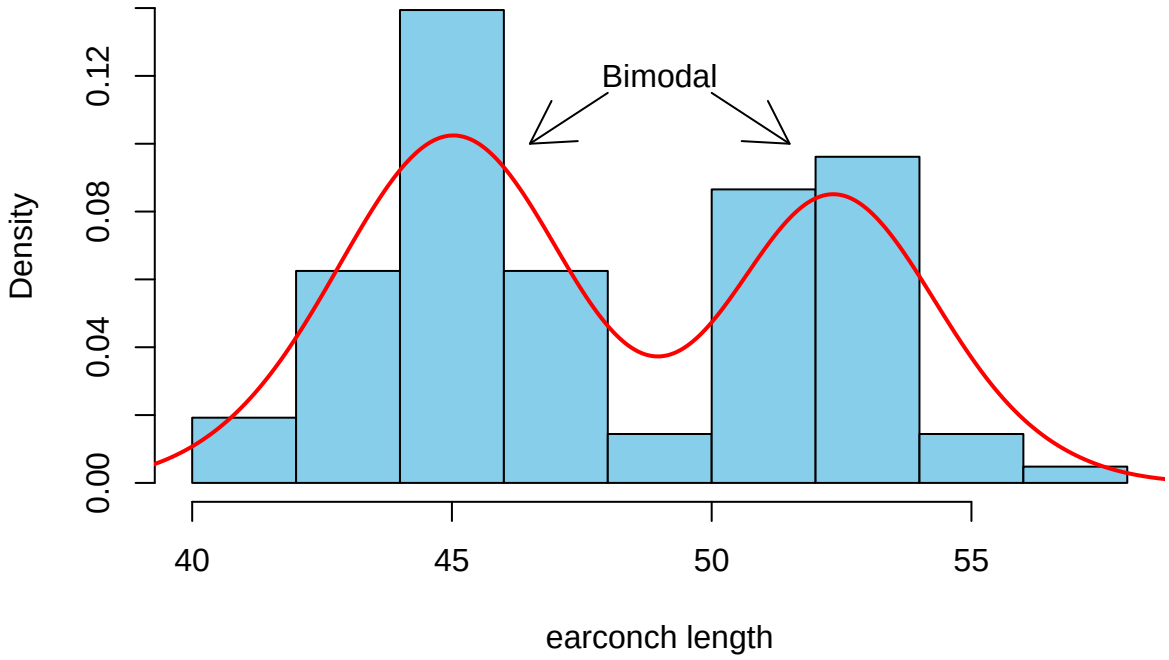
```
require(ggplot2)
ggplot(melt.df, aes(x.value, Random.Numbers )) +
       geom_line(aes(colour = Distributions), lwd = 0.8) +
       geom_point(aes(shape = Distributions, colour = Distributions), size = 2)+
       labs(x= 'Random Samples', y = 'Random Numbers')
```



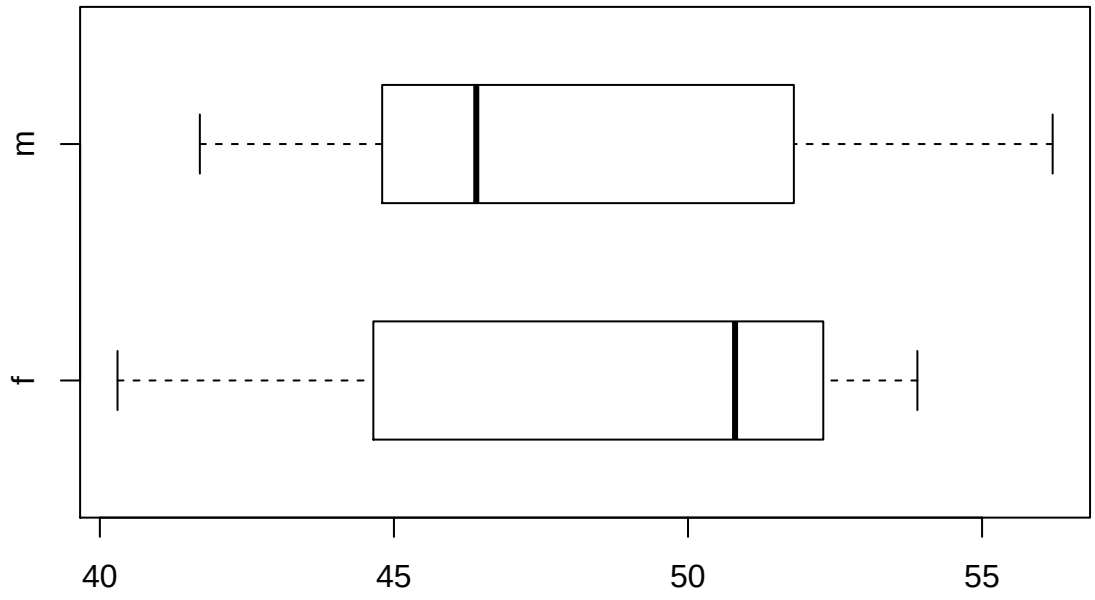

```
# Q.2.
# Data frame possum (DAAG package)
# The possum data frame consists of nine morphometric measurements on each
# of 104 mountain brushtail possums, trapped at seven sites from Southern
# Victoria to central Queensland.

# ********************************************************************************
# Q.2.a
# Use the lattice function bwplot () to display the distribution of ages for each
# combination of site and sex. Show the different sites on the same panel, with
# different panels for different sexes.
# ********************************************************************************

bwplot(age ~ site | sex, data = possum)
```




```
# ********************************************************************************
# Q.2.b
# plot a histogram and density estimation curve on the earconch measurement.
# The distribution should appear bimodal (two peaks). This is a simple indication
# of clustering, possibly due to sex differences.
# ********************************************************************************

hist(possum$earconch, main = "", xlab = "earconch length", col ='skyblue', prob = TRUE)
lines(density(possum$earconch), col = 'red', lwd = 2)
text(49, 0.12, "Bimodal")
arrows(48, 0.115, x1 = 46.5, y1 = 0.10)
arrows(50, 0.115, x1 = 51.5, y1 = 0.10)
```




```
# ***************************************************************************************
# Q.2.c
# Obtain side-by-side boxplots of the male and female earconch measurements. How
# do these measurement distributions differ? Can you predict what the corresponding
# histogram would like? Plot them to check your answer.
# ***************************************************************************************

boxplot(earconch ~ sex, data = possum, boxwex = 0.5, horizontal = TRUE)
```

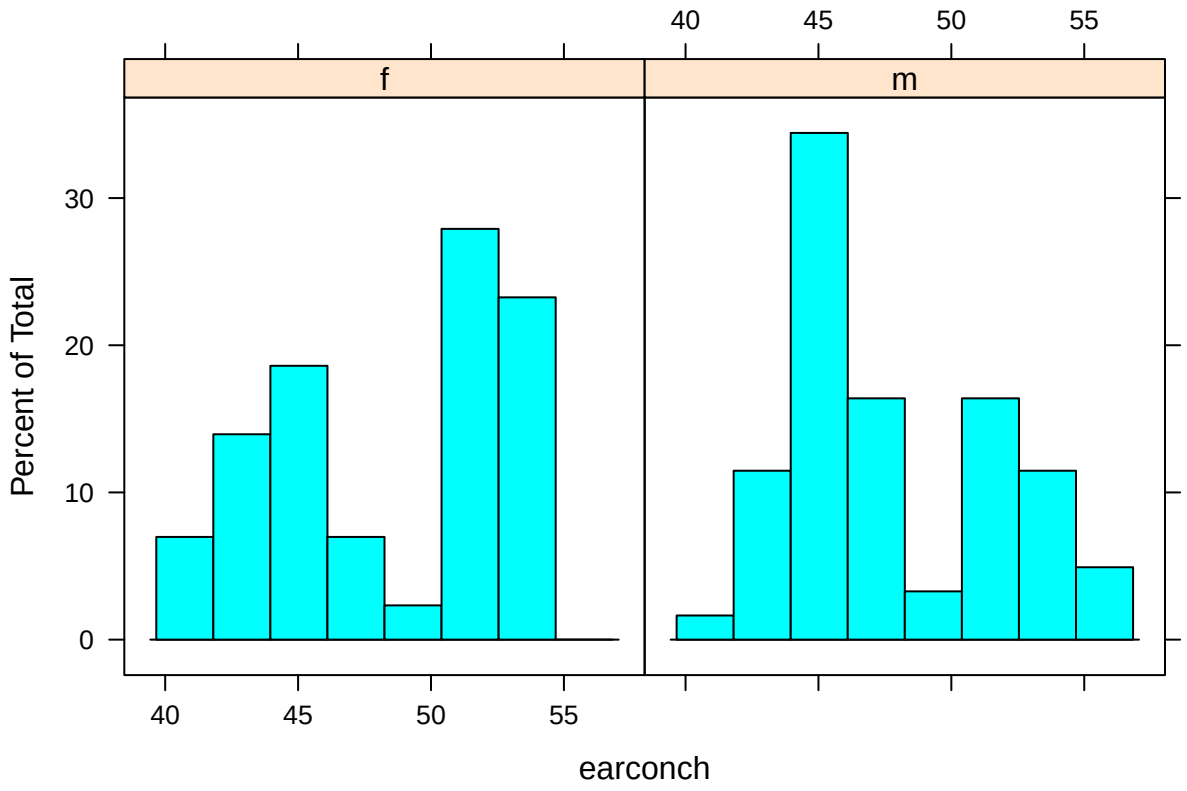



```
# A.2.c
"
  The side-by-side boxplots of the male and female enconch measurement shows that
  the median earconch length of female is greater than that of male. In contrast,
  female has minimum earconch length and male has maximum earconch length.
```

```
  The corresponding histogram should have more frequencies in the region of
  long earconch length for the female and vice versa for the male.
"
```

```
## [1] "\n  The side-by-side boxplots of the male and female enconch measurement shows that\n  the media
```

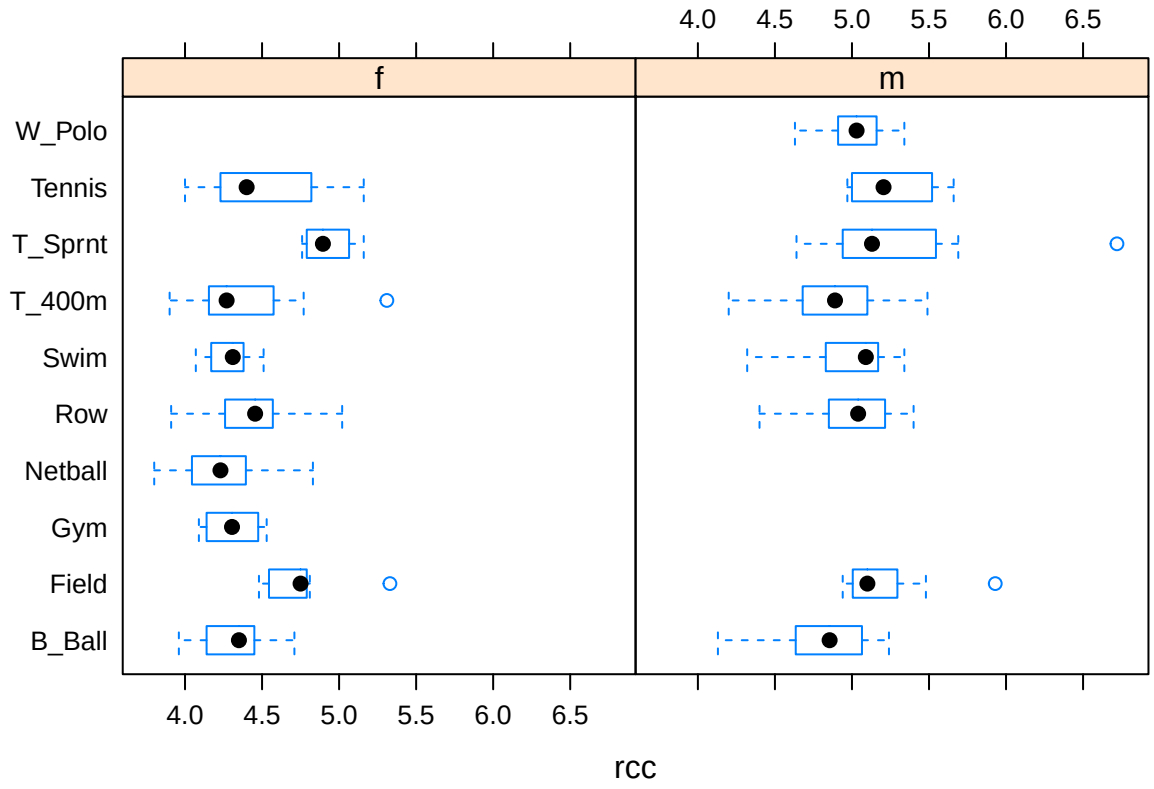
```
histogram(~ earconch | sex, data = possum)
```



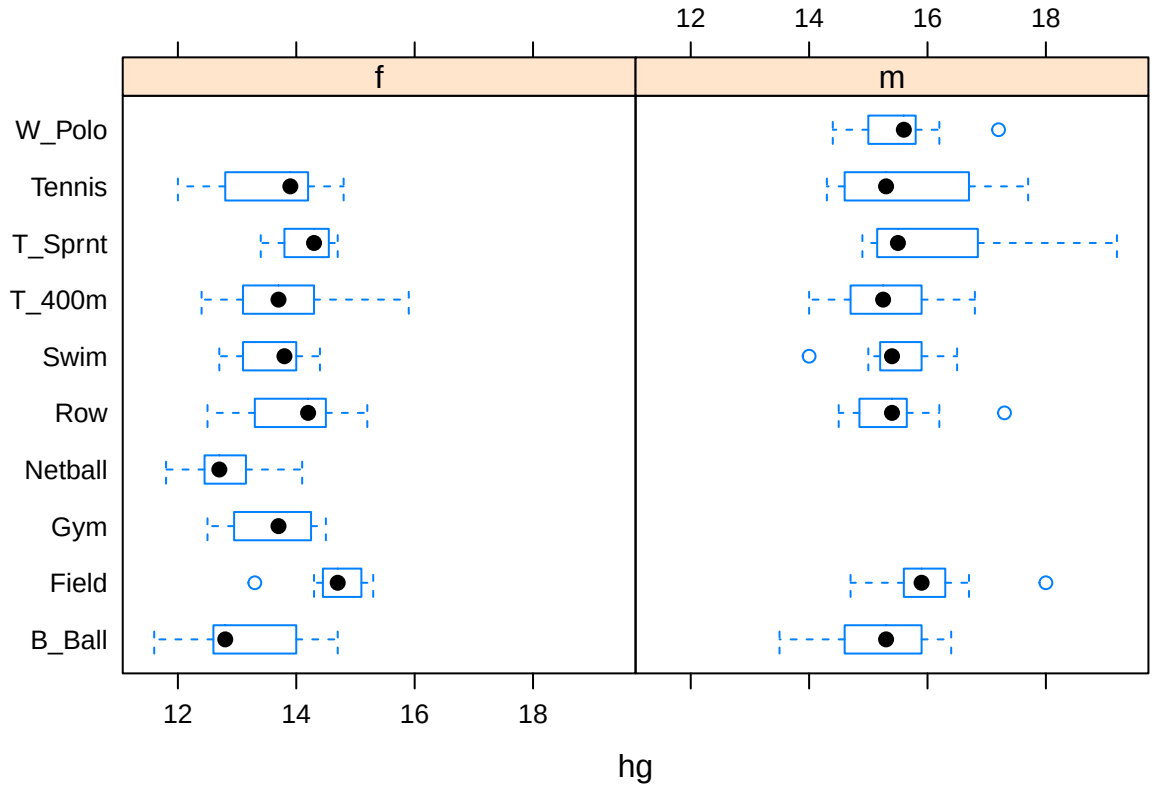

```
"
  This histogram shows that the bimodality in the distribution is not becouse of sex.
"
```

```
## [1] "\n  This histogram shows that the bimodality in the distribution is not becouse of sex.\n"
```

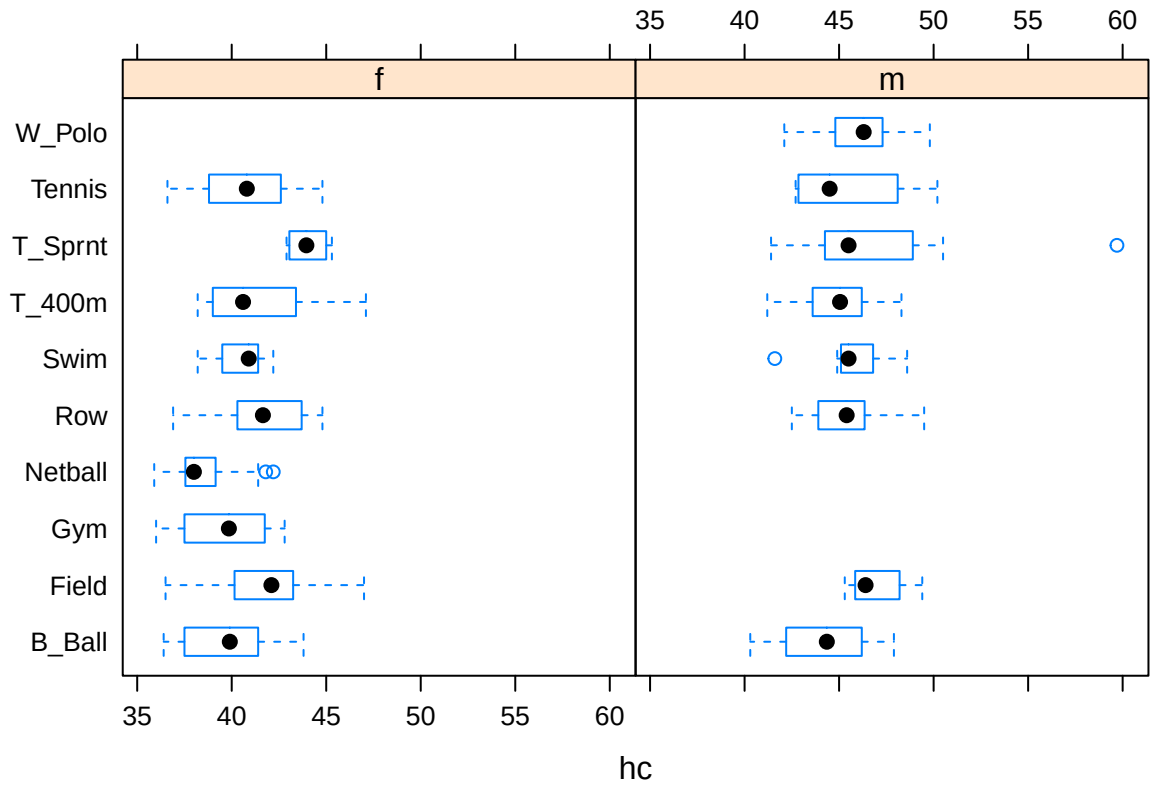
```
# ********************************************************************************
# Q.3
# For the data frame ais (DAAG package), draw graphs that show how the values of the
# hematological measures (red cell count, hemoglobin concentration, hematocrit, white
# cell count and plasma ferritin concentration) vary with the sport and sex of the athlete.
# ********************************************************************************

bwplot(sport ~ rcc | sex, data = ais)
```



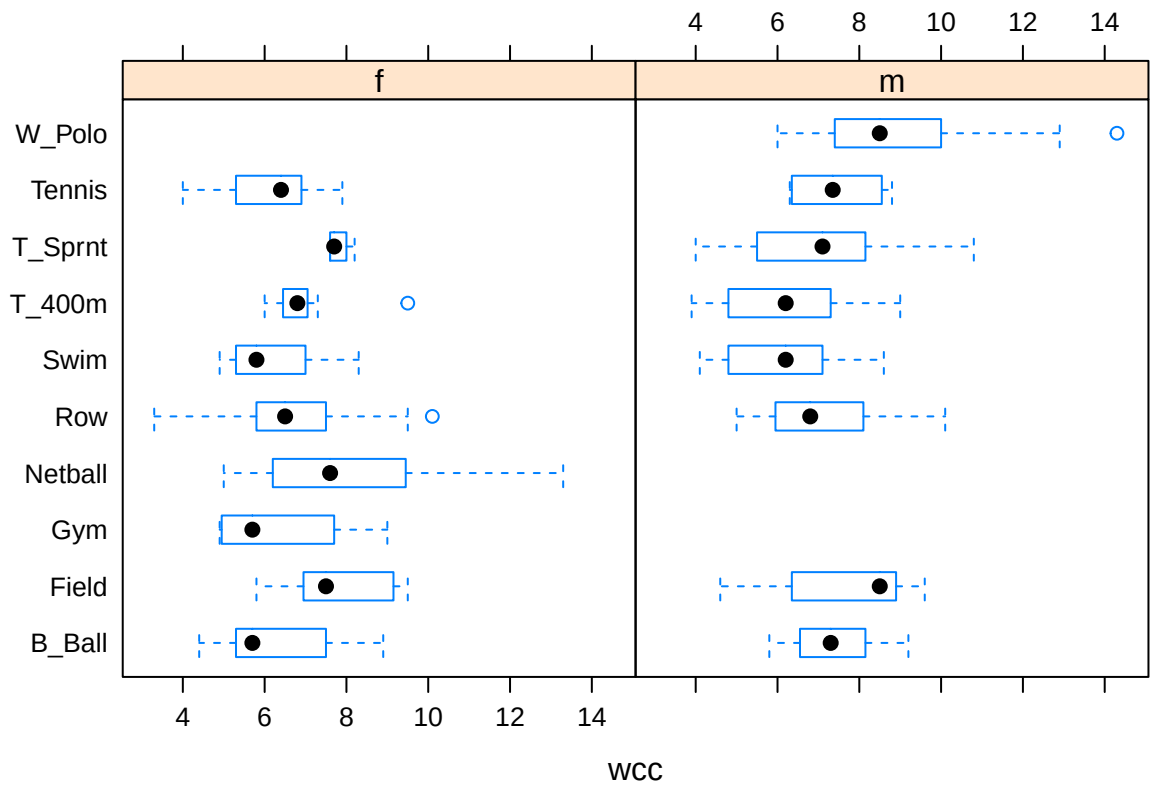

```
bwplot(sport ~ hg | sex, data = ais)
```



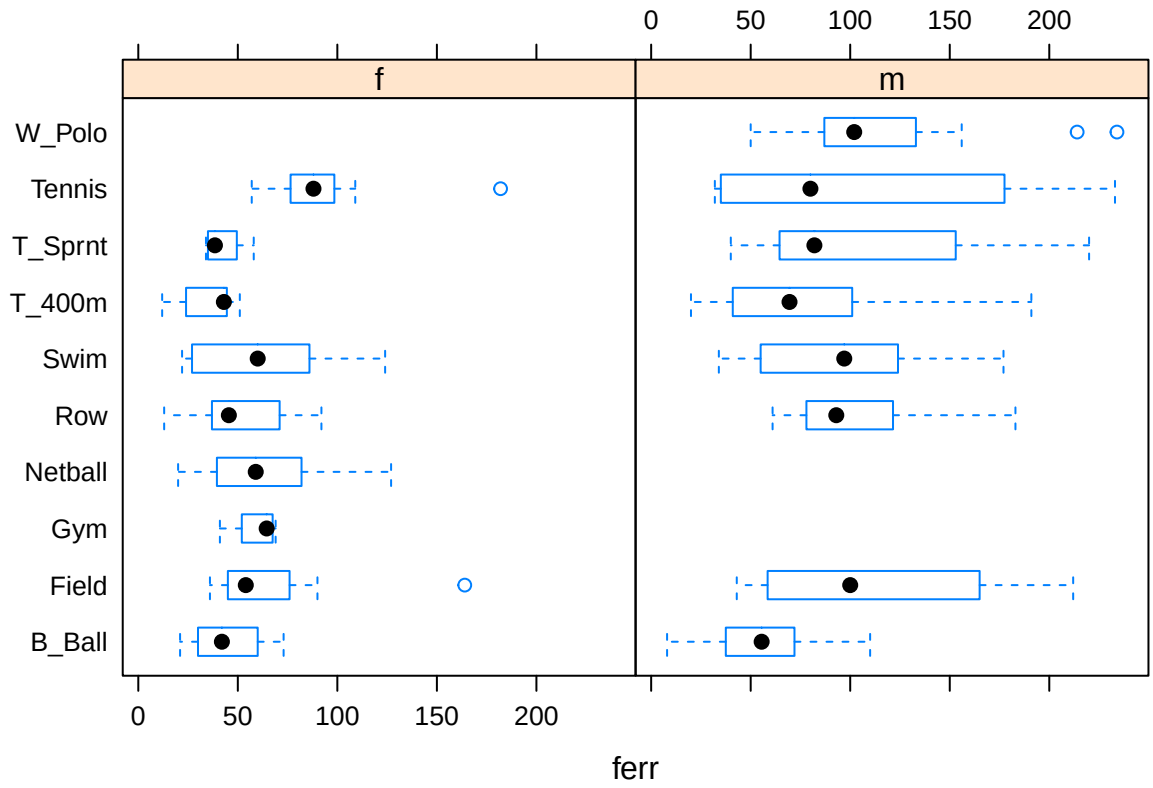
```
bwplot(sport ~ hc | sex, data = ais)
```



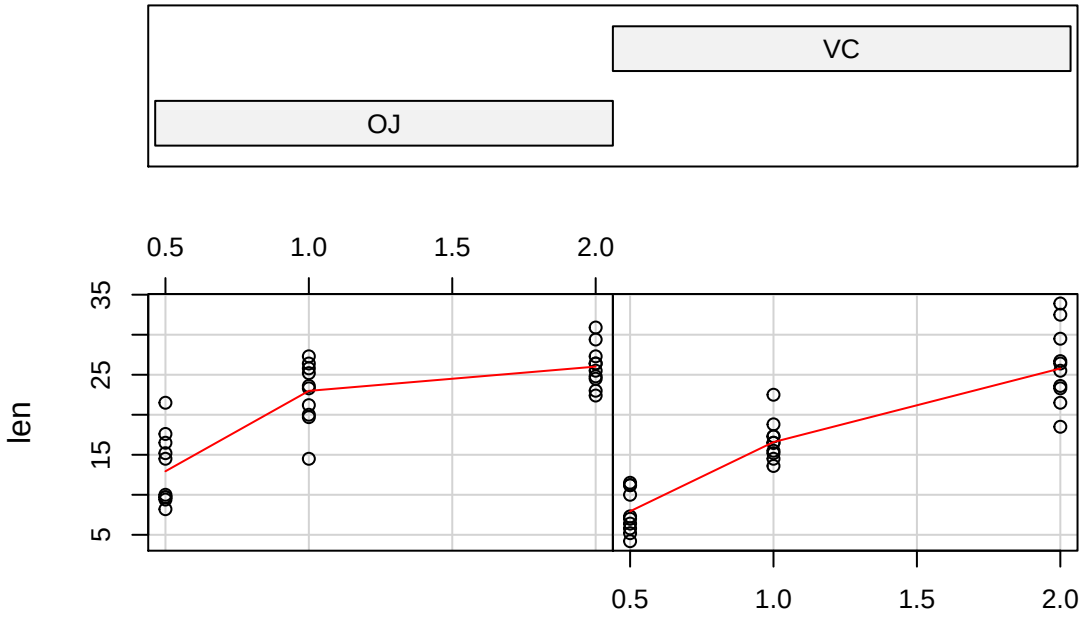

```
bwplot(sport ~ wcc | sex, data = ais)
```



```
bwplot(sport ~ ferr | sex, data = ais)
```




```
# Q.4.
# DATA
# The Effect of Vitamin C on Tooth Growth in Guinea Pigs.
# The response is the length of odontoblasts (cells responsible for tooth growth)
# in 60 guinea pigs. Each animal received one of three dose levels of vitamin
# C (0.5, 1, and 2 mg/day) by one of two delivery methods,
# (orange juice or ascorbic acid (a form of vitamin C and coded as VC).

# For the built-in data ToothGrowth, the plot from the following code indicates some
# tooth length differences between 'VC' and 'OJ' groups and each dose level.

require(graphics)
coplot(len ~ dose | supp, data = ToothGrowth, panel = panel.smooth,
       xlab = "ToothGrowth data: length vs dose, given type of supplement")
```



ToothGrowth data: length vs dose, given type of supplement

```
# ********************************************************************************************
# Q.4.a
# Compare the difference between 'VC' and 'OJ' groups by t-test. First you have
# to test can the variances of the groups be assumed equal, then based on the result
# you can do t-test correctly.
# ********************************************************************************************
#  Use F test two compare to variances.

var.test(ToothGrowth$len[ToothGrowth$supp  == "VC"],
         ToothGrowth$len[ToothGrowth$supp == "OJ"])
```

```
##
##  F test to compare two variances
##
## data:  ToothGrowth$len[ToothGrowth$supp == "VC"] and ToothGrowth$len[ToothGrowth$supp == "OJ"]
## F = 1.5659, num df = 29, denom df = 29, p-value = 0.2331
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.745331 3.290028
## sample estimates:
## ratio of variances
##           1.565937
```

```
"
  Here, variances are equal based on p-vlaue and 95 percent confidence interval.
  So, we do two samples t-test with equal variance.
"
```

```
## [1] "\n  Here, variances are equal based on p-vlaue and 95 percent confidence interval.\n  So, we do
```

```
t.test(ToothGrowth$len[ToothGrowth$supp == "VC"],
       ToothGrowth$len[ToothGrowth$supp == "OJ"],
       variance.equal = TRUE)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "VC"] and ToothGrowth$len[ToothGrowth$supp == "OJ"]
## t = -1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.5710156  0.1710156
## sample estimates:
## mean of x mean of y
##  16.96333  20.66333
```

```
"
  We can not reject the null hypothesis. It shows that orange juice (OJ) and
  ascorbic acid (VC) have not different impact on the tooth growth of Guinea Pigs.
"
```

```
## [1] "\n  We can not reject the null hypothesis. It shows that orange juice (OJ) and\n  ascorbic acid
```

```
#*******************************************************************************
# Q.4.b
# Construct the 95% confidence interval for the mean of the tooth length of guinea
# pigs under each dose level.
# *******************************************************************************

# Here, we do all the t-test assuming unequal variance (default) for each dose level.

dose.level <- unique(ToothGrowth$dose)
for (dl in dose.level){
  print(sprintf ("For the given dose level dl = %0.1f",dl))

   print(t.test(ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose == dl],
                ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose == dl]))
}
```

```
## [1] "For the given dose level dl = 0.5"
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose ==  and ToothGrowth$len[ToothGrow
## t = -3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.780943 -1.719057
## sample estimates:
## mean of x mean of y
##      7.98     13.23
##
## [1] "For the given dose level dl = 1.0"
##
##  Welch Two Sample t-test
##
```

```
## data:  ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose ==  and ToothGrowth$len[ToothGrow
## t = -4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.057852 -2.802148
## sample estimates:
## mean of x mean of y
##     16.77     22.70
##
## [1] "For the given dose level dl = 2.0"
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose ==  and ToothGrowth$len[ToothGrow
## t = 0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.63807  3.79807
## sample estimates:
## mean of x mean of y
##     26.14     26.06
```

```
# ********************************************************************************
# Q.4.c
# Interpret your results.
# ********************************************************************************
# A.4.c
"
  For lowest dose level 0.5 and intermediate dose level 1.0, we can reject the
  null hypothesis. So, one suppliment is more effective than the other.

  For highet dose leve 2.0, we can not reject the null hypothesis and it shows that
  the effect of both supplyments are similar.
"
```

```
## [1] "\n  For lowest dose level 0.5 and intermediate dose level 1.0, we can reject the\n  null hypoth
```

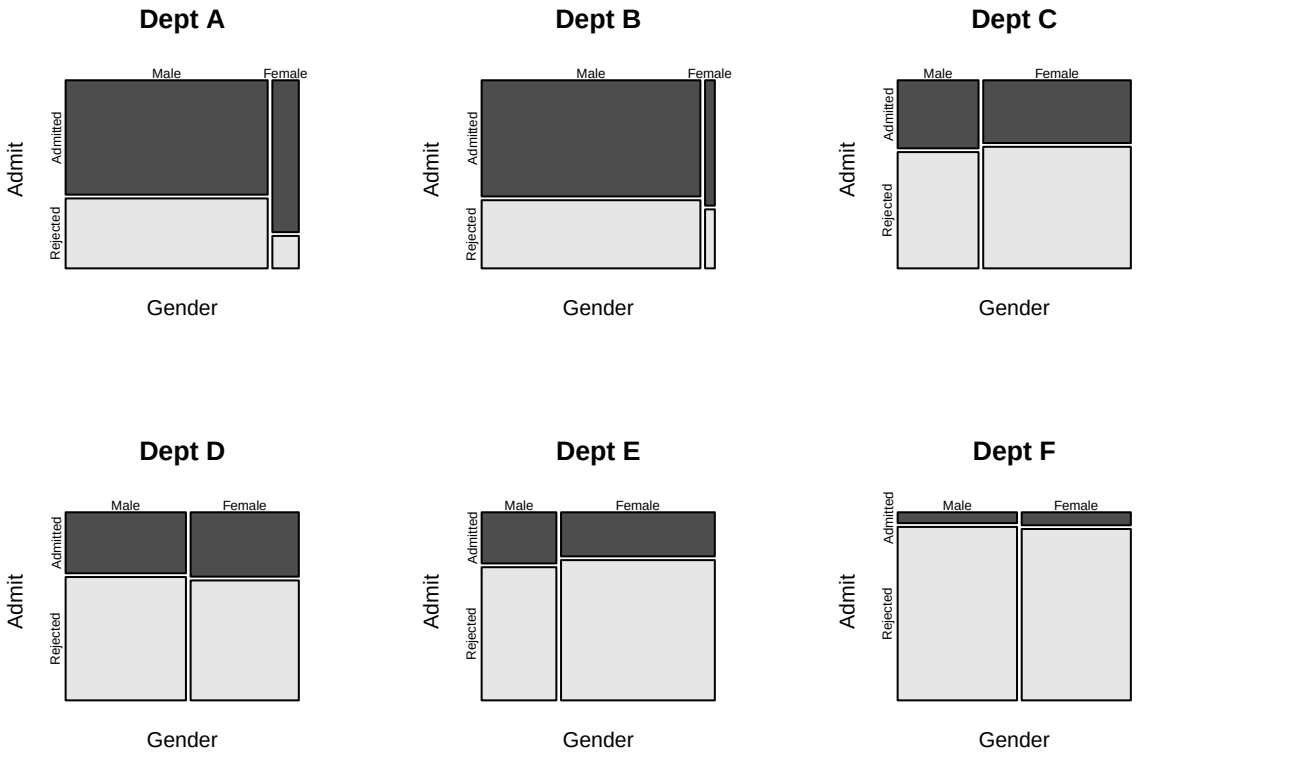
```
# Q.5
# The multi-way table UCBAdmissions has admission frequencies, by sex, for the six
# largest departments at the University of California at Berkeley in 1973. The following
# gives a table that adds the 2 × 2 tables of admission data overall all departments:
## For each combination of margins 1 and 2, calculate the sum
UCBtotal <- apply(UCBAdmissions, c(1,2), sum)
# ********************************************************************************
# Q.5.a
# What are the names of the two dimensions of this table?
# ********************************************************************************

dimnames(UCBtotal)
```

```
## $Admit
## [1] "Admitted" "Rejected"
##
## $Gender
## [1] "Male"   "Female"
```

```
# A.5.a
"
  The names of the two dimensions are Admit and Gender.
"
```

```
## [1] "\n  The names of the two dimensions are Admit and Gender.\n"
```

```
# ************************************************************************************************
# 5.b
# From the table UCBAdmissions, create mosaic plot for each faculty separately.
# (If necessary refer to the code given in the help page for USBAdmissions.)
# ************************************************************************************************

par(mfrow=c(2,3))
for(i in 1:6) mosaicplot(t(UCBAdmissions[,,i]),main = paste("Dept", LETTERS[i]), color = TRUE)
```




```
par(mfrow=c(1,1))

# ************************************************************************************************
# 5.c
# Compare the information in the table UCBtotal with the result from applying the
# function mantelhaen.test( ) to the table UCBAdmissions. Compare the two sets
# of results, and comments on difference.
# ************************************************************************************************
UCBtotal
```

```
##          Gender
## Admit      Male Female
##   Admitted 1198    557
##   Rejected 1493   1278
```

```r
mantelhaen.test(UCBAdmissions)
```

```
##
##  Mantel-Haenszel chi-squared test with continuity correction
##
## data:  UCBAdmissions
## Mantel-Haenszel X-squared = 1.4269, df = 1, p-value = 0.2323
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.7719074 1.0603298
## sample estimates:
## common odds ratio
##         0.9046968
```

```r
# A.5.c
"
  The information in the UCBtotal table shows that the male admitted rate is higher
  than that of female admitted rate. But, mantelhaen.test( ) to the table UCBAdmissions
  shows no evidence for association between gender and admission.
"
```

```
## [1] "\n  The information in the UCBtotal table shows that the male admitted rate is higher\n  than th
```

```r
# ********************************************************************************************
# 5.d
# The Mantel-Haenzel test is valid only if the male to female odds ratio for admission
# is similar across departments. The following code calculates the relevant odds ratios:
  apply(UCBAdmissions, 3, function(x) (x[1,1]*x[2,2])/(x[1,2]*x[2,1]))
```

```
##         A         B         C         D         E         F
## 0.3492120 0.8025007 1.1330596 0.9212838 1.2216312 0.8278727
```

```r
# Is the odds ratio consistent across department? Which department(s) stand(s)
# out as different? What is the nature of the difference?
# ********************************************************************************************

# A.5.d
  "
    The odds ratio is not consistent across the department, lowest for department A
    and highest for department E.
  "
```

```
## [1] "\n    The odds ratio is not consistent across the department, lowest for department A\n    and h
```

```r
# ********************************************************************************************
# Q.6.
  # P142 15 For constructing bootstrap confidence intervals for the correlation coefficient,
  # it is advisable to work with the Fisher z-transformation of the correlation coefficient.
  # The following lines of R code show how to obtain a bootstrap confidence interval
  # for the z-transformed correlation between chest and belly in the possum data frame.
  # The last step of the procedure is to apply the inverse of the z-transformation to the
  # confidence interval to return it to the original scale. Run the following code and
  # compare the resulting interval with the one computed without transformation. Is the
  # z-transformation necessary here?

  z.transform <- function(r) .5 * log((1 + r)/(1 - r))
  z.inverse <- function(z) (exp(2 * z) -1)/(exp(2 * z) + 1)
```

```
  possum.fun <- function(data, indices) {
    chest <- data$chest[indices]
    belly <- data$belly[indices]
    z.transform(cor(belly, chest))}
  possum.boot <- boot(possum, possum.fun, R = 1000)
  z.inverse(boot.ci(possum.boot, type = "perc")$percent[4:5])
```

```
## [1] 0.4641521 0.7078934
```

```
  # See help(bootci.object)

# ***************************************************************************************************
# Without Fisher z-transformation, the bootstrap confidence interval is
  possum.fun0 <- function(data, indices) {
    chest <- data$chest[indices]
    belly <- data$belly[indices]
    cor(belly, chest)}
  possum.boot0 <- boot(possum, possum.fun0, R = 1000)
  boot.ci(possum.boot0, type = "perc")$percent[4:5]
```

```
## [1] 0.4695294 0.7101498
```

```
"
  There is no significant difference in the bootsrap confidence interval for the
  correlation coefficient between with transformation and without transformation.
"
```

```
## [1] "\n  There is no significant difference in the bootsrap confidence interval for the\n  correlati
```

```
# $$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
```