STAT 8010 R Session 4

Whitney Huang

6/9/2023

Contents

Session Objectives

- To gain experience with R, a programming language and free software environment for statistical computing and graphics.
- To perform two sample t-test and paired t-test using `R`
- To conduct *ANOVA* and multiple comparisons using `R`

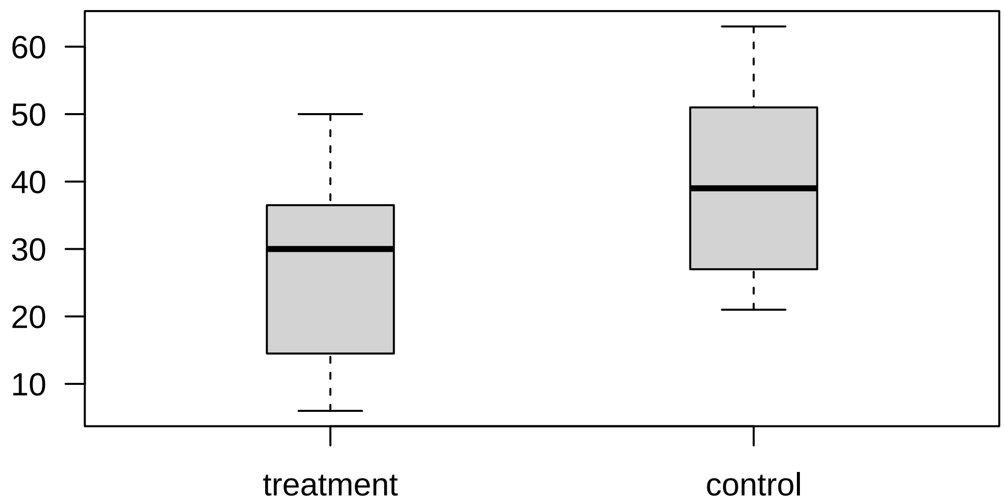
Two sample t-test

Tapeworm example

```
treatment <- c(18, 43, 28, 50, 16, 32, 13, 35, 38, 33, 6, 7)
control <- c(40, 54, 26, 63, 21, 37, 39, 23, 48, 58, 28, 39)
dat <- data.frame(cbind(treatment, control))
summary(dat)
```

```
##    treatment        control
##  Min.   : 6.00   Min.   :21.00
##  1st Qu.:15.25   1st Qu.:27.50
##  Median :30.00   Median :39.00
##  Mean   :26.58   Mean   :39.67
##  3rd Qu.:35.75   3rd Qu.:49.50
##  Max.   :50.00   Max.   :63.00
```

```
boxplot(dat, boxwex = 0.3, las = 1)
```




```
apply(dat, 2, mean)
```

```
## treatment   control
##  26.58333  39.66667
```

```
apply(dat, 2, sd)
```

```
## treatment   control
##  14.36193  13.85859
```

```
var.test(treatment, control)
```

```
##
##  F test to compare two variances
```

```
##
## data:  treatment and control
## F = 1.074, num df = 11, denom df = 11, p-value = 0.9079
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3091686 3.7306092
## sample estimates:
## ratio of variances
##           1.073959
```

```
# Assuming $\sigma_{1}=\sigma_{2}$
t.test(treatment, control, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  treatment and control
## t = -2.2709, df = 22, p-value = 0.03329
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -25.031761  -1.134906
## sample estimates:
## mean of x mean of y
##  26.58333  39.66667
```

```
# Assuming $\sigma_{1}\neq\sigma_{2}$
t.test(treatment, control, var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  treatment and control
## t = -2.2709, df = 21.972, p-value = 0.03331
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -25.032642  -1.134025
## sample estimates:
## mean of x mean of y
##  26.58333  39.66667
```

```
# Left-tailed test
t.test(treatment, control, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  treatment and control
## t = -2.2709, df = 21.972, p-value = 0.01665
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -3.189613
## sample estimates:
## mean of x mean of y
##  26.58333  39.66667
```

Two sample t test with only sample statistics

```r
t.test.from.summary.data <- function(mean1, sd1, n1, mean2, sd2, n2, ...) {
    data1 <- scale(1:n1)*sd1 + mean1
    data2 <- scale(1:n2)*sd2 + mean2
    t.test(data1, data2, ...)
}

t.test.from.summary.data(19.45, 4.3, 37, 18.2, 2.2, 31)
```

```
##
##  Welch Two Sample t-test
##
## data:  data1 and data2
## t = 1.5435, df = 55.507, p-value = 0.1284
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3726447  2.8726447
## sample estimates:
## mean of x mean of y
##     19.45     18.20
```

```r
## Check
(df = ((4.3^2)/37 + (2.2^2)/31)^2 / (((4.3^2)/37)^2 / 36 +  ((2.2^2)/31)^2 / 30))
```

```
## [1] 55.50703
```

```r
(se <- sqrt(4.3^2 / 37 + 2.2^2 / 31))
```

```
## [1] 0.8098511
```

```r
(tstat <- (19.45 - 18.2) / se)
```

```
## [1] 1.543494
```

```r
(Pvalue <- 2 * (1 - pt(1.5435, df)))
```

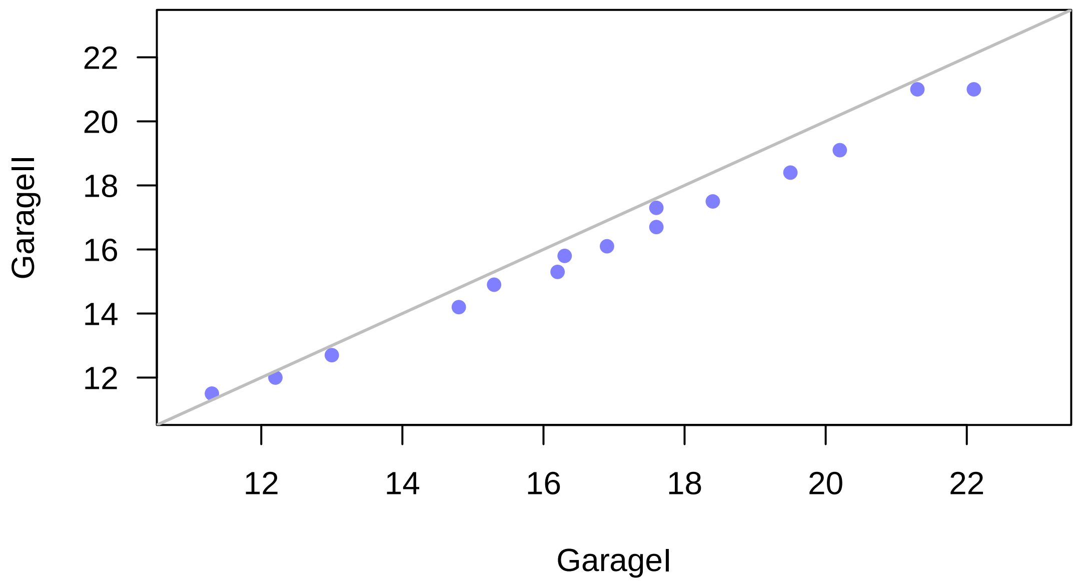
```
## [1] 0.128392
```

Paired t-Test

```r
repair <- c(17.6, 17.3, 20.2, 19.1, 19.5, 18.4, 11.3, 11.5,
            13.0, 12.7, 16.3, 15.8, 15.3, 14.9, 16.2, 15.3,
            12.2, 12.0, 14.8, 14.2, 21.3, 21.0, 22.1, 21.0,
            16.9, 16.1, 17.6, 16.7, 18.4, 17.5)
GarageI <- repair[seq(1, 29, 2)]
GarageII <- repair[seq(2, 30, 2)]
dat <- cbind(GarageI, GarageII)
apply(dat, 2, mean)
```

```
##  GarageI GarageII
## 16.84667 16.23333
```

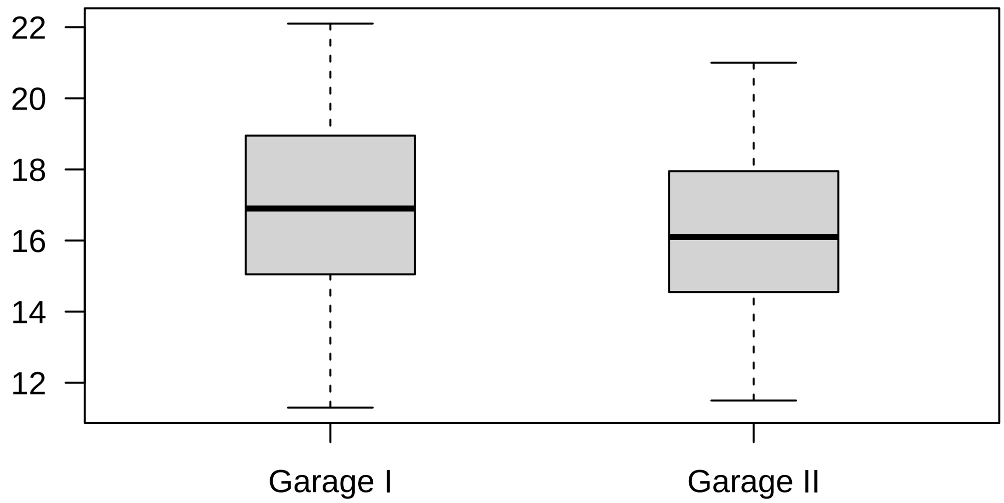
```
library(scales)
plot(GarageI, GarageII,
     pch = 16, col = alpha("blue", 0.5), las = 1, xlim = c(11, 23),
     ylim = c(11, 23))
abline(0, 1, col = "gray", lwd = 1.5)
```



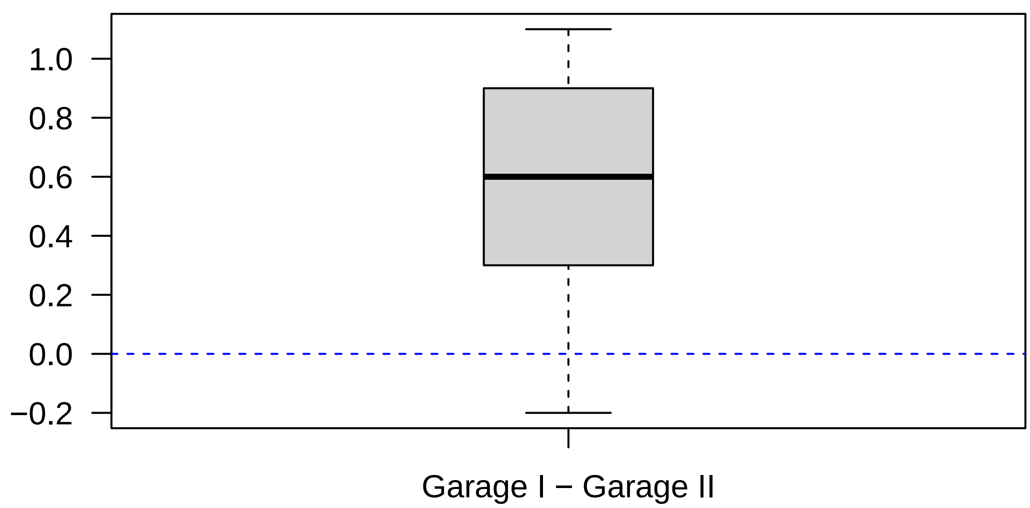

```
t.test(GarageI, GarageII, alternative = c("greater"), var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  GarageI and GarageII
## t = 0.54616, df = 27.797, p-value = 0.2947
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -1.29749      Inf
## sample estimates:
## mean of x mean of y
##  16.84667  16.23333
```

```
boxplot(GarageI, GarageII, boxwex = 0.4, xaxt = "n", las = 1)
axis(1, at = 1:2, labels = c("Garage I", "Garage II"))
```




```r
boxplot(GarageI - GarageII, boxwex = 0.4, xaxt = "n", las = 1)
axis(1, at = 1, labels = "Garage I - Garage II")
abline(h = 0, col = "blue", lty = 2)
```



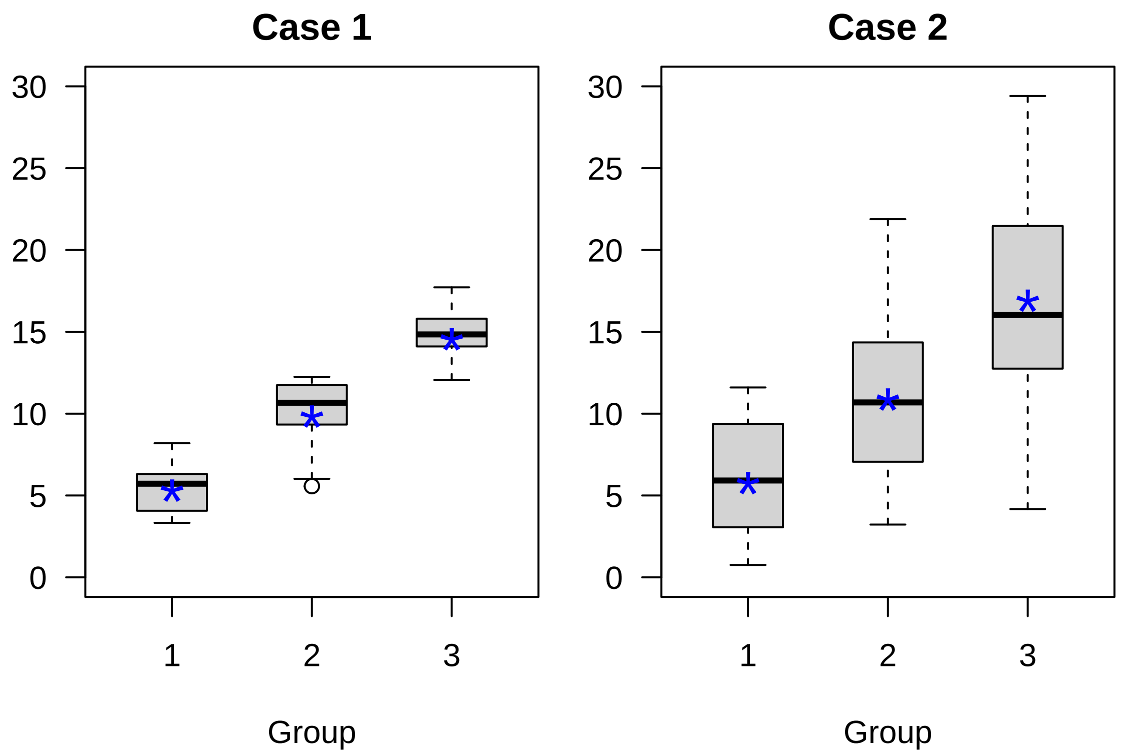

```r
t.test(GarageI, GarageII, alternative = c("greater"), paired = T)
```

```
##
##  Paired t-test
##
## data:  GarageI and GarageII
## t = 6.0234, df = 14, p-value = 1.563e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4339886       Inf
## sample estimates:
## mean of the differences
##               0.6133333
```

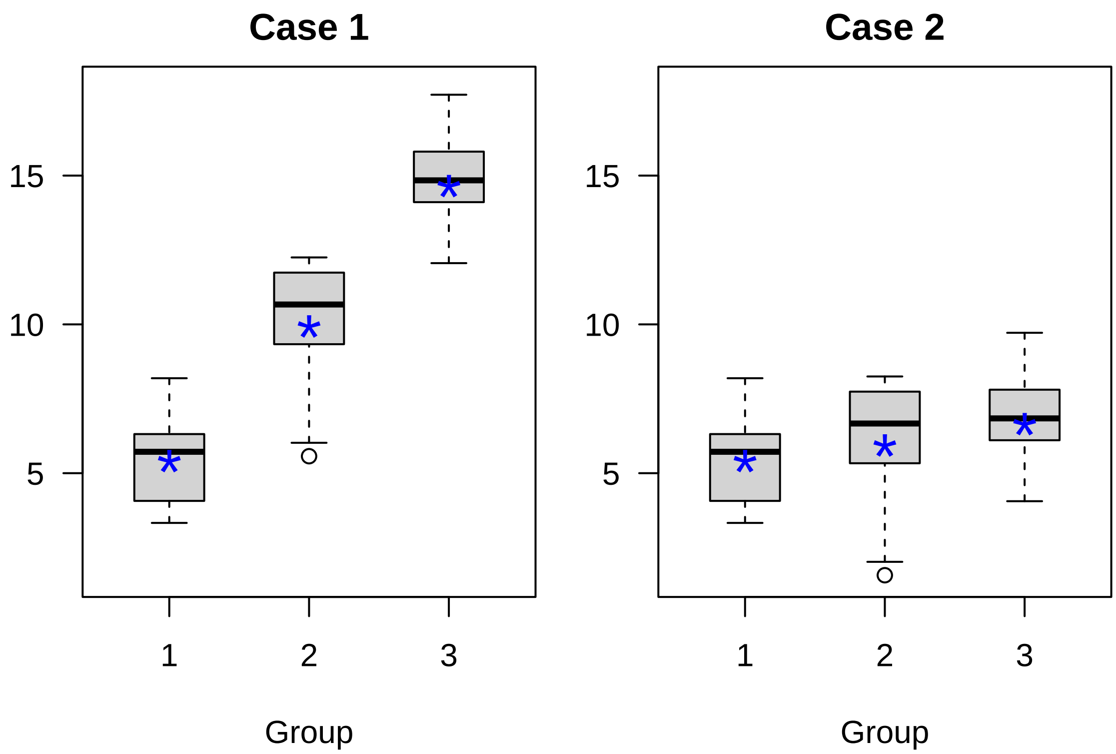
AONVA

Toy Examples

```r
set.seed(1)
base1 <- rnorm(n = 36, sd = 2)
base2 <- rnorm(n = 36, sd = 6)
dat1 <- base1 + c(rep(5, 12), rep(10, 12), rep(15, 12))
dat2 <- base2 + c(rep(5, 12), rep(10, 12), rep(15, 12))
dat3 <- base1 + rep(5:7, each = 12)
level <- as.factor(rep(1:3, each = 12))
dat1 <- data.frame(x = dat1, Group = level)
dat2 <- data.frame(x = dat2, Group = level)
dat3 <- data.frame(x = dat3, Group = level)
library(dplyr)
g1summary <- dat1 %>%
select(x, Group) %>%
group_by(Group) %>%
summarise(mean = mean(x), sd1 = sd(x))

g2summary <- dat2 %>%
select(x, Group) %>%
group_by(Group) %>%
summarise(mean = mean(x), sd1 = sd(x))

g3summary <- dat3 %>%
select(x, Group) %>%
group_by(Group) %>%
summarise(mean = mean(x), sd1 = sd(x))


par(mfrow = c(1, 2), mar = c(4.1, 2.1, 2.1, 1.1))
boxplot(x ~ Group, data = dat1, las = 1, boxwex = 0.5,
        ylab = "", ylim = c(0, 30), main = "Case 1")
for (i in 1:3) points(i, g1summary$mean[i], pch = "*",
                      col = "blue", cex = 2)
boxplot(x ~ Group, data = dat2, las = 1, boxwex = 0.5,
        ylab = "", ylim = c(0, 30), main = "Case 2")
for (i in 1:3) points(i, g2summary$mean[i], pch = "*",
                      col = "blue", cex = 2)
```




```
par(mfrow = c(1, 2), mar = c(4.1, 2.1, 2.1, 1.1))
boxplot(x ~ Group, data = dat1, las = 1, boxwex = 0.5,
        ylab = "", ylim = c(1.5, 18), main = "Case 1")
for (i in 1:3) points(i, g1summary$mean[i], pch = "*",
                      col = "blue", cex = 2)
boxplot(x ~ Group, data = dat3, las = 1, boxwex = 0.5,
        ylab = "", ylim = c(1.5, 18), main = "Case 2")
for (i in 1:3) points(i, g3summary$mean[i], pch = "*",
                      col = "blue", cex = 2)
```




```
model1 <- lm(x ~ Group, data = dat1)
model2 <- lm(x ~ Group, data = dat2)
anova(model1)
```

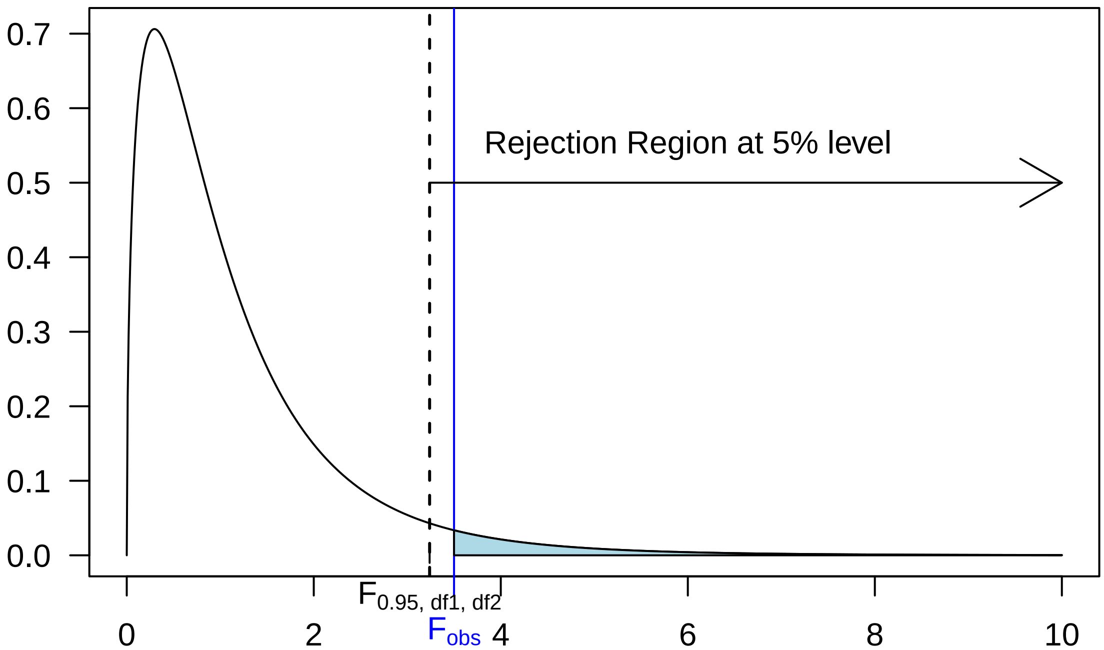
```
## Analysis of Variance Table
##
## Response: x
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Group      2 512.54 256.271  75.443 4.902e-13 ***
## Residuals 33 112.10   3.397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: x
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Group      2  747.35  373.67  11.323 0.0001802 ***
## Residuals 33 1089.03   33.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

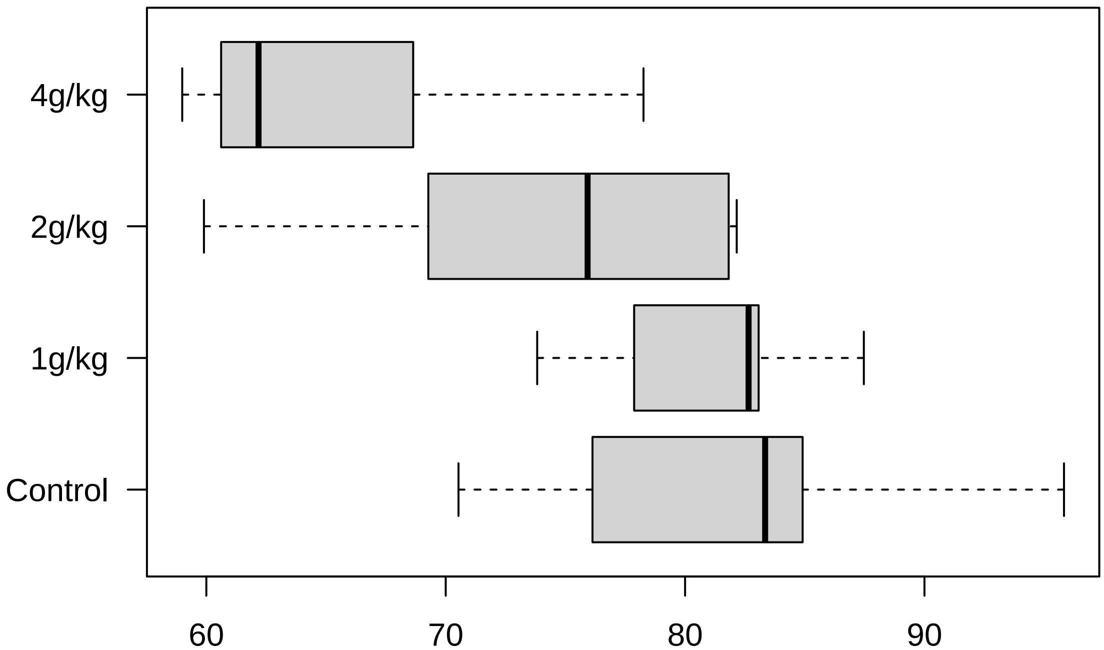
F Distribution

```r
par(mar = c(4.1, 2.6, 1.1, 1.1))
curve(df(x, 3, 16), from = 0, to = 10, n = 1001, las = 1,
      xlab = "", ylab = "")
abline(v = 3.5, col = "blue")
abline(v = qf(0.95, 3, 16), lty = 2, lwd = 1.5)
xg <- seq(3.5, 10, 0.01)
yg <- df(xg, 3, 16)
polygon(c(xg[xg >= 3.5], rev(c(xg[xg >= 3.5]))), c(yg[xg >= 3.5], rep(0, length(yg[xg >= 3.5]))),
        col = "lightblue")
axis(1, at = 3.5, labels = expression(F["obs"]), col = "blue", col.axis = "blue")
axis(1, at = qf(0.95, 3, 16), line = -0.85, labels = expression(F[paste(0.95, ", df1", ", df2")]))
arrows(qf(0.95, 3, 16), 0.5, 10)
text(6, 0.55, "Rejection Region at 5% level")
```



Effects of Ethanol on Sleep Time Example

```r
### Data setup
set.seed(124)
g1 <- rnorm(5, 83, 9); g2 <- rnorm(5, 76, 9.5); g3 <- rnorm(5, 73, 9.2); g4 <- rnorm(5, 70, 9)
dat <- cbind(Response = c(g1, g2, g3, g4), Treatment = as.factor(rep(1:4, each = 5)))
dat <- data.frame(dat)
dat$Treatment <- as.factor(dat$Treatment)
par(mar = c(4.1, 4.1, 1.1, 1.1))
boxplot(Response ~ Treatment, data = dat, horizontal = T, yaxt = "n", ylab = "", xlab = "")
axis(2, at = 1:4, labels = c("Control", "1g/kg", "2g/kg", "4g/kg"), las = 1)
```




```
### Data Summary
summary <- dat %>%
select(Response, Treatment) %>%
group_by(Treatment) %>%
summarise(mean = mean(Response),
          sd1 = sd(Response))
lm <- lm(Response ~ Treatment, dat)
anova(lm)
```

```
## Analysis of Variance Table
##
## Response: Response
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## Treatment  3  861.13 287.044  4.2542 0.02173 *
## Residuals 16 1079.56  67.472
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
### Pairwise t-test
t.test(dat$Response[1:5], dat$Response[6:10], var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  dat$Response[1:5] and dat$Response[6:10]
## t = 0.24012, df = 8, p-value = 0.8163
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.09426  12.44081
## sample estimates:
## mean of x mean of y
##  82.15052  80.97724
```

```
t.test(dat$Response[1:5], dat$Response[6:10], var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  dat$Response[1:5] and dat$Response[6:10]
## t = 0.24012, df = 6.2015, p-value = 0.818
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.68922  13.03577
## sample estimates:
## mean of x mean of y
##  82.15052  80.97724
```

Facebook Example

```
dat <- read.csv("FacebookFriends.csv")
head(dat); str(dat)
```

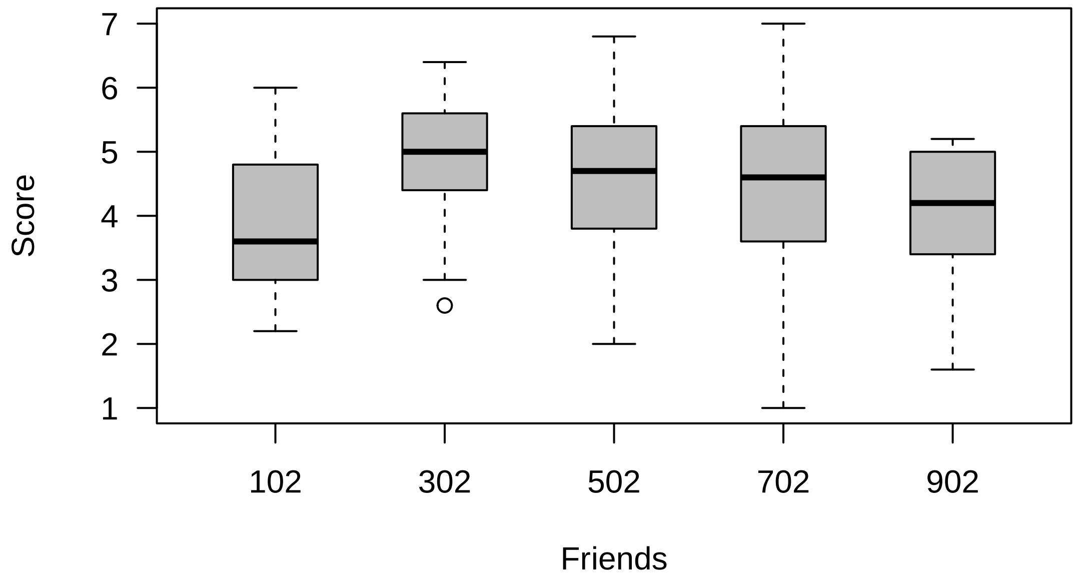
```
##   Friends Participant Score
## 1     102           1   3.8
## 2     102           2   3.6
## 3     102           3   3.2
## 4     102           4   2.4
## 5     102           5   4.8
## 6     102           6   3.0
```

```
## 'data.frame':    134 obs. of  3 variables:
##  $ Friends    : int  102 102 102 102 102 102 102 102 102 102 ...
##  $ Participant: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Score      : num  3.8 3.6 3.2 2.4 4.8 3 4.2 3.6 3.2 3 ...
```

```
dat$Friends <- as.factor(dat$Friends)
str(dat)
```

```
## 'data.frame':    134 obs. of  3 variables:
##  $ Friends    : Factor w/ 5 levels "102","302","502",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Participant: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Score      : num  3.8 3.6 3.2 2.4 4.8 3 4.2 3.6 3.2 3 ...
```

```
boxplot(Score ~ Friends, data = dat, las = 1, col = "gray", boxwex = 0.5)
```



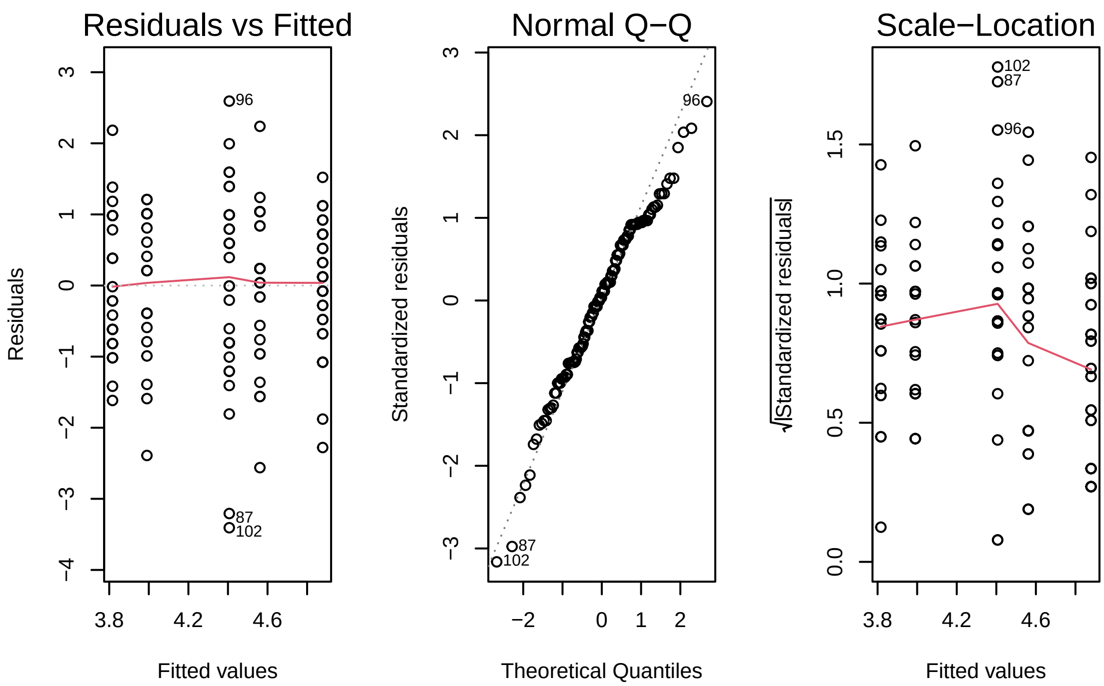

```
library(dplyr)
summary <- dat %>%
select(Score, Friends) %>%
group_by(Friends) %>%
summarise(mean = mean(Score),
          sd1 = sd(Score))
summary
```

```
## # A tibble: 5 x 3
##   Friends  mean   sd1
##   <fct>   <dbl> <dbl>
## 1 102      3.82 0.999
## 2 302      4.88 0.851
## 3 502      4.56 1.07
## 4 702      4.41 1.43
## 5 902      3.99 1.02
```

```
lm <- lm(Score ~ Friends, dat)
anova(lm)
```

```
## Analysis of Variance Table
##
## Response: Score
##            Df Sum Sq Mean Sq F value  Pr(>F)
## Friends     4  19.89  4.9726   4.142 0.00344 **
## Residuals 129 154.87  1.2005
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow = c(1, 3))
plot(lm, which = 1:3)
```




```r
aov <- aov(Score ~ Friends, dat)
aov
```

```
## Call:
##    aov(formula = Score ~ Friends, data = dat)
##
## Terms:
##                  Friends Residuals
## Sum of Squares  19.89023 154.86679
## Deg. of Freedom        4       129
##
## Residual standard error: 1.095681
## Estimated effects may be unbalanced
```

```r
anova(lm(Score ~ Friends, dat))
```

```
## Analysis of Variance Table
##
## Response: Score
##            Df Sum Sq Mean Sq F value  Pr(>F)
## Friends     4  19.89  4.9726   4.142 0.00344 **
## Residuals 129 154.87  1.2005
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple Comparisons

Fisher's LSD

```r
library(agricolae)
LSD_none <- LSD.test(aov ,"Friends", p.adj = "none")
LSD_none$groups
```

```
##        Score groups
## 302 4.878788      a
## 502 4.561538     ab
## 702 4.406667    abc
## 902 3.990476     bc
## 102 3.816667      c
```

```r
LSD_bon <- LSD.test(aov ,"Friends", p.adj = "bonferroni")
LSD_bon$groups
```

```
##        Score groups
## 302 4.878788      a
## 502 4.561538     ab
## 702 4.406667     ab
## 902 3.990476      b
## 102 3.816667      b
```

Tukey's HSD

```r
HSD <- TukeyHSD(aov, conf.level = 0.95)
HSD$Friends
```

```
##               diff        lwr         upr       p adj
## 302-102  1.0621212  0.2488644  1.87537798 0.003889635
## 502-102  0.7448718 -0.1132433  1.60298691 0.121456224
## 702-102  0.5900000 -0.2402014  1.42020143 0.288431585
## 902-102  0.1738095 -0.7320145  1.07963355 0.984016816
## 502-302 -0.3172494 -1.1121910  0.47769215 0.804080046
## 702-302 -0.4721212 -1.2368466  0.29260420 0.432633745
## 902-302 -0.8883117 -1.7345313 -0.04209203 0.034535577
## 702-502 -0.1548718 -0.9671402  0.65739661 0.984391504
## 902-502 -0.5710623 -1.4604793  0.31835479 0.391768065
## 902-702 -0.4161905 -1.2787075  0.44632652 0.669927748
```

CRD and RCBD

Create the data set

```r
x <- c(52, 47, 44, 51, 42, 60, 55, 49, 52, 43, 56, 48, 45, 44, 38)
trt <- rep(c("A", "B", "C"), each = 5)
blk <- rep(1:5, 3)
dat <- data.frame(x = x, trt = trt, blk = as.factor(blk))
```

Two-way ANOVA

```r
aov <- aov(x ~ trt + blk, data = dat)
lm <- lm(x ~ trt + blk, data = dat)
anova(lm)
```

```
## Analysis of Variance Table
##
## Response: x
##           Df Sum Sq Mean Sq F value    Pr(>F)
## trt        2   89.2   44.60  7.6239 0.0140226 *
## blk        4  363.6   90.90 15.5385 0.0007684 ***
## Residuals  8   46.8    5.85
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One-way ANOVA

```r
lm2 <- lm(x ~ trt, data = dat)
anova(lm2)
```

```
## Analysis of Variance Table
##
## Response: x
##           Df Sum Sq Mean Sq F value Pr(>F)
## trt        2   89.2    44.6  1.3041 0.3073
## Residuals 12  410.4    34.2
```

Interaction plot: assessing the additivity assumption

```r
interaction.plot(dat$trt, dat$blk, x, las = 1, col = 1:5)
```