Statistično načrtovanje poskusov
Analize primerov z R

A. Blejec

January 11, 2012

Contents

1 Enostavno slučajnostno vzorčenje

Pripravimo podatke za primer iz SNP (Blejec, 1971), stran 73.

```
> y <- c(8, 7, 6, 5, 7, 8, 9, 7, 9, 12, 10, 13, 18, 16,
+     12, 15, 13, 14, 17, 13, 7, 6, 5, 4, 3, 12, 5, 7,
+     6, 9, 8, 7, 9, 8, 10)
> n <- 5
> A <- factor(rep(paste("A", 1:7, sep = ""), each = n))
> data <- data.frame(A, id = rep(1:n, 7), y)
> reshape(data, idvar = "A", timevar = "id", direction = "wide")
    A y.1 y.2 y.3 y.4 y.5
1  A1   8   7   6   5   7
6  A2   8   9   7   9  12
11 A3  10  13  18  16  12
16 A4  15  13  14  17  13
21 A5   7   6   5   4   3
26 A6  12   5   7   6   9
31 A7   8   7   9   8  10
```

Preverimo povzetke

```
> data.frame(group = aggregate(A, by = list(A), function(x) as.character(x[1]),
+     2], n = aggregate(y, by = list(A), length)[, 2],
+     sum = aggregate(y, by = list(A), sum)[, 2], mean = aggregate(y,
+         by = list(A), mean)[, 2])
  group n sum mean
1    A1 5  33  6.6
2    A2 5  45  9.0
3    A3 5  69 13.8
4    A4 5  72 14.4
5    A5 5  25  5.0
6    A6 5  39  7.8
7    A7 5  42  8.4
```

1.1 Analiza variance

```
> aovFit <- aov(y ~ A, data = data)
> aovFit
Call:
   aov(formula = y ~ A, data = data)

Terms:
                        A Residuals
Sum of Squares  375.9429  117.2000
Deg. of Freedom        6        28

Residual standard error: 2.045902
Estimated effects may be unbalanced
> aovTable <- summary(aovFit)
> aovTable
            Df Sum Sq Mean Sq F value    Pr(>F)
A            6 375.94  62.657  14.969 1.333e-07 ***
Residuals   28 117.20   4.186
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diagnostične slike so prikazane na sliki 5
Iz tabele za analizo variance lahko izluščimo posamezne dele tabele:

```
> str(aovTable[[1]])
Classes 'anova' and 'data.frame':       2 obs. of  5 variables:
 $ Df     : num  6 28
 $ Sum Sq : num  376 117
 $ Mean Sq: num  62.66 4.19
 $ F value: num  15 NA
 $ Pr(>F) : num  1.33e-07 NA
```

Tako lahko izluščimo varianco ostanka in število ponovitev

```
> (se2 <- aovTable[[1]]$"Mean Sq"[2])
[1] 4.185714
> df <- (aovTable[[1]])$Df
> (n <- df[2]/(df[1] + 1) + 1)
```

```
> par(mfrow = c(2, 2))
> plot(aovFit)
```

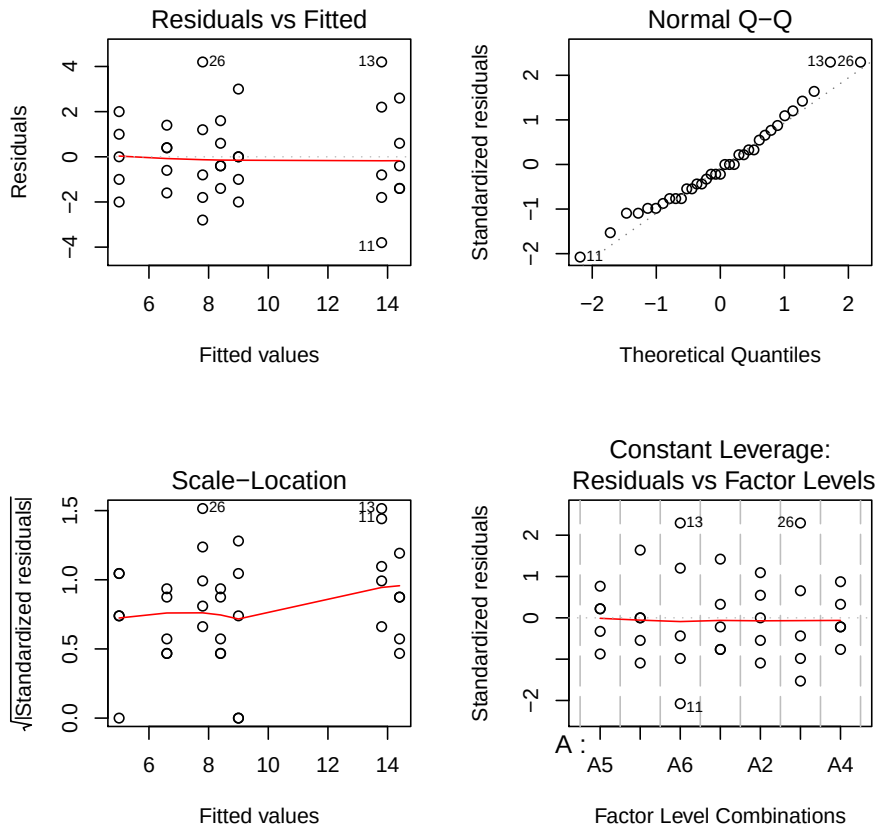

Figure 1: DIagnostične slike za analizo variance

```
[1] 5
```

$$s_e^2 = 4.186$$

Pripravimo lahko tudi uporabne povzetke povprečij in učinkov.

```
> means <- model.tables(aovFit, "means", se = TRUE)
> means
Tables of means
Grand mean

9.285714

 A
A
  A1   A2   A3   A4   A5   A6   A7
 6.6  9.0 13.8 14.4  5.0  7.8  8.4

Standard errors for differences of means
           A
       1.294
replic.    5
```

Združena standardna napaka za primerjavo razlik povprečij je v posebnem primeru enakega števila ponovitev poskusa

```
> sqrt(2 * se2/n)
[1] 1.293942
```

Iz tabele povprečij lahko izračunamo učinke, to je odstopanja povprečij skupin od skupnega povprečja

```
> ucinek <- means$tables$A - means$tables$"Grand mean"
> ucinek
A
       A1        A2        A3        A4        A5        A6        A7
-2.685714 -0.285714  4.514286  5.114286 -4.285714 -1.485714 -0.885714
```

Standardna napaka učinkov

```
> sqrt(se2/n)
[1] 0.9149551
```

Lahko pa tudi takole:

```
> effects <- model.tables(aovFit, "effects", se = TRUE)
> effects
Tables of effects

 A
A
    A1     A2     A3     A4     A5     A6     A7
-2.686 -0.286  4.514  5.114 -4.286 -1.486 -0.886

Standard errors of effects
            A
        0.915
replic.     5
```

1.1.1 Tukey HSD

```
> hsd <- TukeyHSD(aovFit, ordered = TRUE)
> hsd$A <- hsd$A[order(hsd$A[, "diff"], decreasing = TRUE),
+     ]
> hsd
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = y ~ A, data = data)

$A
      diff        lwr       upr     p adj
A4-A5  9.4  5.2954479 13.504552 0.0000013
A3-A5  8.8  4.6954479 12.904552 0.0000042
A4-A1  7.8  3.6954479 11.904552 0.0000321
A3-A1  7.2  3.0954479 11.304552 0.0001103
```

```
A4-A6  6.6  2.4954479 10.704552 0.0003805
A3-A6  6.0  1.8954479 10.104552 0.0013038
A4-A7  6.0  1.8954479 10.104552 0.0013038
A3-A7  5.4  1.2954479  9.504552 0.0043724
A4-A2  5.4  1.2954479  9.504552 0.0043724
A3-A2  4.8  0.6954479  8.904552 0.0140689
A2-A5  4.0 -0.1045521  8.104552 0.0598394
A7-A5  3.4 -0.7045521  7.504552 0.1558159
A6-A5  2.8 -1.3045521  6.904552 0.3456781
A2-A1  2.4 -1.7045521  6.504552 0.5246841
A7-A1  1.8 -2.3045521  5.904552 0.8018267
A1-A5  1.6 -2.5045521  5.704552 0.8737534
A2-A6  1.2 -2.9045521  5.304552 0.9647873
A6-A1  1.2 -2.9045521  5.304552 0.9647873
A7-A6  0.6 -3.5045521  4.704552 0.9991198
A2-A7  0.6 -3.5045521  4.704552 0.9991198
A4-A3  0.6 -3.5045521  4.704552 0.9991198
```

Rezultat bolje prikaže slika 2

```
> plot(hsd, las = 1)
```

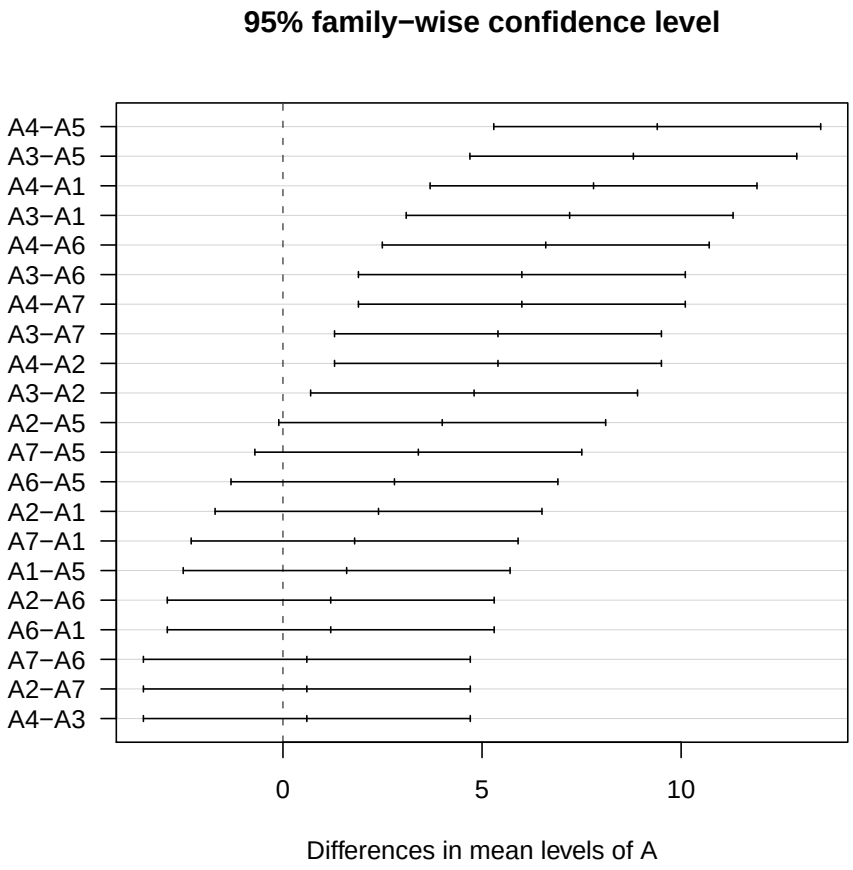

Figure 2: Intervali zaupanja za Tukey HSD

1.2 Linearni model

Ker nimamo nobenega kontrolnega nivoja, želimo analizo brez določitve začetne vrednosti.

```
> lmFit <- lm(y ~ 0 + A, data = data)
> lmFit
Call:
lm(formula = y ~ 0 + A, data = data)

Coefficients:
 AA1   AA2   AA3   AA4   AA5   AA6   AA7
 6.6   9.0  13.8  14.4   5.0   7.8   8.4
```

Analiza koeficientov ni posebno smiselna, saj nas zanima primerjava na skupno povprečje, ali pa razlike med stanji.

Analiza s korekcijo na skupno povprečje:

```
> lmFit <- lm(y - mean(y) ~ 0 + A, data = data)
> lmFit
Call:
lm(formula = y - mean(y) ~ 0 + A, data = data)

Coefficients:
    AA1      AA2      AA3      AA4      AA5      AA6      AA7
-2.6857  -0.2857   4.5143   5.1143  -4.2857  -1.4857  -0.8857


> summary(lmFit)
Call:
lm(formula = y - mean(y) ~ 0 + A, data = data)

Residuals:
   Min     1Q Median     3Q    Max
  -3.8   -1.4   -0.4    1.1    4.2

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
AA1  -2.6857     0.9150  -2.935  0.00659 **
AA2  -0.2857     0.9150  -0.312  0.75715
AA3   4.5143     0.9150   4.934 3.32e-05 ***
AA4   5.1143     0.9150   5.590 5.55e-06 ***
AA5  -4.2857     0.9150  -4.684 6.58e-05 ***
AA6  -1.4857     0.9150  -1.624  0.11562
AA7  -0.8857     0.9150  -0.968  0.34131
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.046 on 28 degrees of freedom
Multiple R-squared: 0.7623,        Adjusted R-squared: 0.7029
F-statistic: 12.83 on 7 and 28 DF,  p-value: 2.898e-07
```

Primerjava razlik med skupinami

```
> lmFit2 <- lm(y ~ 0 + A, data = data)
> lmFit2
Call:
lm(formula = y ~ 0 + A, data = data)

Coefficients:
 AA1   AA2   AA3   AA4   AA5   AA6   AA7
 6.6   9.0  13.8  14.4   5.0   7.8   8.4

> summary(lmFit2)
Call:
lm(formula = y ~ 0 + A, data = data)

Residuals:
   Min     1Q Median     3Q    Max
  -3.8   -1.4   -0.4    1.1    4.2

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
AA1    6.600      0.915   7.213 7.50e-08 ***
AA2    9.000      0.915   9.837 1.38e-10 ***
AA3   13.800      0.915  15.083 5.69e-15 ***
AA4   14.400      0.915  15.738 1.95e-15 ***
AA5    5.000      0.915   5.465 7.80e-06 ***
AA6    7.800      0.915   8.525 2.88e-09 ***
AA7    8.400      0.915   9.181 6.13e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.046 on 28 degrees of freedom
Multiple R-squared: 0.9666,        Adjusted R-squared: 0.9583
F-statistic: 115.8 on 7 and 28 DF,  p-value: < 2.2e-16
```

Matrika primerjav vseh skupin med seboj

```
> m <- nrow(data)/n
> mm <- diag(m)
> contr <- matrix(unlist(sapply(1:(m - 1), function(j) sapply(j:(m -
+     1), function(i) mm[, i + 1] - mm[, j]))), nrow = m)
> contr
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
[1,]   -1   -1   -1   -1   -1   -1    0    0    0     0     0     0
[2,]    1    0    0    0    0    0   -1   -1   -1    -1    -1     0
[3,]    0    1    0    0    0    0    1    0    0     0     0    -1
[4,]    0    0    1    0    0    0    0    1    0     0     0     1
[5,]    0    0    0    1    0    0    0    0    1     0     0     0
[6,]    0    0    0    0    1    0    0    0    0     1     0     0
[7,]    0    0    0    0    0    1    0    0    0     0     1     0
     [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21]
[1,]     0     0     0     0     0     0     0     0     0
[2,]     0     0     0     0     0     0     0     0     0
[3,]    -1    -1    -1     0     0     0     0     0     0
[4,]     0     0     0    -1    -1    -1     0     0     0
[5,]     1     0     0     1     0     0    -1    -1     0
[6,]     0     1     0     0     1     0     1     0    -1
[7,]     0     0     1     0     0     1     0     1     1
```

Primerjave med skupinami kažejo razlike med povprečji posameznih skupin. Skupine so preurejene glede na naraščajoče vrednosti povprečij. Grafični prikaz je na Sliki 3

```
> d <- matrix(NA, m, m)
> ordr <- order(lmFit2$coefficients)
> d[lower.tri(d)] <- lmFit2$coefficients[ordr] %*% contr
> dimnames(d) <- list(levels(A)[ordr], levels(A)[ordr])
> d <- t(d)[, m:1]
> print(d, na.print = "")
```

```
    A4  A3  A2  A7  A6  A1 A5
A5 9.4 8.8 4.0 3.4 2.8 1.6
A1 7.8 7.2 2.4 1.8 1.2
A6 6.6 6.0 1.2 0.6
A7 6.0 5.4 0.6
A2 5.4 4.8
A3 0.6
A4
```

1.2.1 Tukey HSD

Tukey HSD postavi enotno mejo za najmanjšo značilno razliko med povprečji skupin.

```
> alpha <- 0.05
> df.residual <- aov(lmFit2)$df.residual
> q <- qtukey(1 - alpha, m, df = df.residual)
> q
```

```
[1] 4.486069
```

```
> se.ybar <- sqrt(summary(aov(lmFit2))[[1]]$"Mean Sq"[2]/n)
> W <- q * se.ybar
> W
```

```
[1] 4.104552
```

S tveganjem $\alpha = 0.05$ je resnično značilna razlika $W = 4.1$.
Poglejmo, katere skupine so značilno različne:

```
> sigDif <- d > W
> print((sigDif + 0), na.print = "")
```

```
   A4 A3 A2 A7 A6 A1 A5
A5  1  1  0  0  0  0
A1  1  1  0  0  0
A6  1  1  0  0
A7  1  1  0
A2  1  1
A3  0
A4
```

```
> dx <- 10
> gm <- 1:m
> gm[ordr] <- 1:m
> xGroup <- rep(gm * dx, each = n) - n/2
> at <- rep((1:n), m) + xGroup
> plot(at, y, axes = FALSE, xlab = "Group")
> points(at, aov(lmFit2)$fitted.values, col = "red", pch = "_",
+     cex = 1.2)
> axis(2)
> segments(at, aov(lmFit2)$fitted.values, at, y, col = "blue")
> mtext(levels(A), side = 1, at = unique(xGroup) + n/2)
```

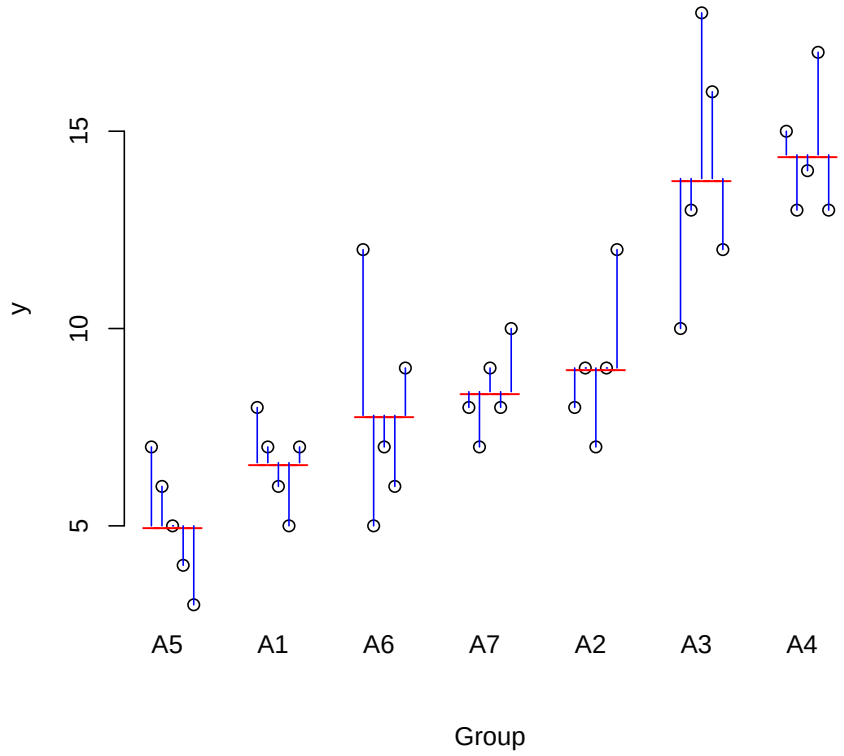

Figure 3: Podatki (○), ocenjene vrednosti v skupinah (vodoravne črte) in odkloni (navpične črte)

2 Čisto slučajnostni poskus z neenakim številom ponovitev

2.1 Problem in podatki

Problem je opisan v SNP (Blejec, 1971), stran 91.

Vpliv pakiranja na prodajo krompirja. Trije postopki so:

- A1: predpakirano v prozornih plastičnih vrečkah
- A2: pakirano v močnih papirnatih vrečah
- A3: raztresen v razstavnih košarah

Kot poskusno gradivo smo izbrali 20 trgovin istega tipa in na slučajnosten način izbrali izmed njih
$n_1 = 5$ trgovin, ki naj bi prodajale po postopku A1
$n_1 = 10$ trgovin, ki naj bi prodajale po postopku A2
$n_1 = 5$ trgovin, ki naj bi prodajale po postopku A3

Kot kriterialni znak je vzeta prodana količina na 10.000 din skupnega prometa. S to opredelitvijo smo iz kriterielnega znaka izločilivelikost trgovine oziroma zaledja, ki bistveno vpliva na višino prodaje.

Rezultati poskusa so:

```
> y <- c(49, 36, 47, 23, 40, 29, 26, 30, 39, 45, 13, 32,
+     18, 38, 40, 12, 16, 23, 28, 16)
> ns <- c(5, 10, 5)
> nGroups <- length(ns)
> group <- rep(1:length(ns), ns)
> A <- factor(paste("A", group, sep = ""))
> id <- unlist(sapply(ns, function(x) 1:x))
> data <- data.frame(A, group, id, y)
> data
```

```
    A group id  y
1  A1     1  1 49
2  A1     1  2 36
3  A1     1  3 47
4  A1     1  4 23
5  A1     1  5 40
6  A2     2  1 29
7  A2     2  2 26
8  A2     2  3 30
9  A2     2  4 39
10 A2     2  5 45
11 A2     2  6 13
12 A2     2  7 32
13 A2     2  8 18
14 A2     2  9 38
15 A2     2 10 40
16 A3     3  1 12
17 A3     3  2 16
18 A3     3  3 23
19 A3     3  4 28
20 A3     3  5 16
```

```
> print(reshape(data[, -2], idvar = "A", timevar = "id",
+     direction = "wide"), na.print = "")
```

```
    A y.1 y.2 y.3 y.4 y.5 y.6 y.7 y.8 y.9 y.10
1  A1  49  36  47  23  40  NA  NA  NA  NA   NA
6  A2  29  26  30  39  45  13  32  18  38   40
16 A3  12  16  23  28  16  NA  NA  NA  NA   NA
```

Preverimo povzetke

```
> data.frame(group = aggregate(A, by = list(A), function(x) as.character(x[1])
+     2], n = aggregate(y, by = list(A), length)[, 2],
+     sum = aggregate(y, by = list(A), sum)[, 2], mean = aggregate(y,
+         by = list(A), mean)[, 2])
```

```
  group  n sum mean
1    A1  5 195   39
2    A2 10 310   31
3    A3  5  95   19
```

2.2 Analiza variance

```
> aovFit <- aov(y ~ A, data = data)
> aovFit
```

```
Call:
   aov(formula = y ~ A, data = data)

Terms:
                    A Residuals
Sum of Squares  1020      1508
Deg. of Freedom    2        17

Residual standard error: 9.41838
Estimated effects may be unbalanced
```

```
> aovTable <- summary(aovFit)
> aovTable
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
A            2   1020  510.00  5.7493 0.01238 *
Residuals   17   1508   88.71
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diagnostične slike so prikazane na sliki 5
Iz tabele za analizo variance lahko izluščimo posamezne dele tabele:

```
> str(aovTable[[1]])
```

```
Classes 'anova' and 'data.frame':       2 obs. of  5 variables:
 $ Df     : num  2 17
 $ Sum Sq : num  1020 1508
 $ Mean Sq: num  510 88.7
 $ F value: num  5.75 NA
 $ Pr(>F) : num  0.0124 NA
```

```
> dx <- 10
> gm <- 1:nGroups
> ordr <- 1:nGroups
> gm[ordr] <- 1:nGroups
> xGroup <- (group - 1) * dx
> at <- id + xGroup
> plot(at, y, axes = FALSE, xlab = "Group")
> points(at, aov(aovFit)$fitted.values, col = "red", pch = "_",
+     cex = 1.2)
> axis(2)
> segments(at, aov(aovFit)$fitted.values, at, y, col = "blue")
> mtext(levels(A), side = 1, at = unique(xGroup) + ns/2)
```

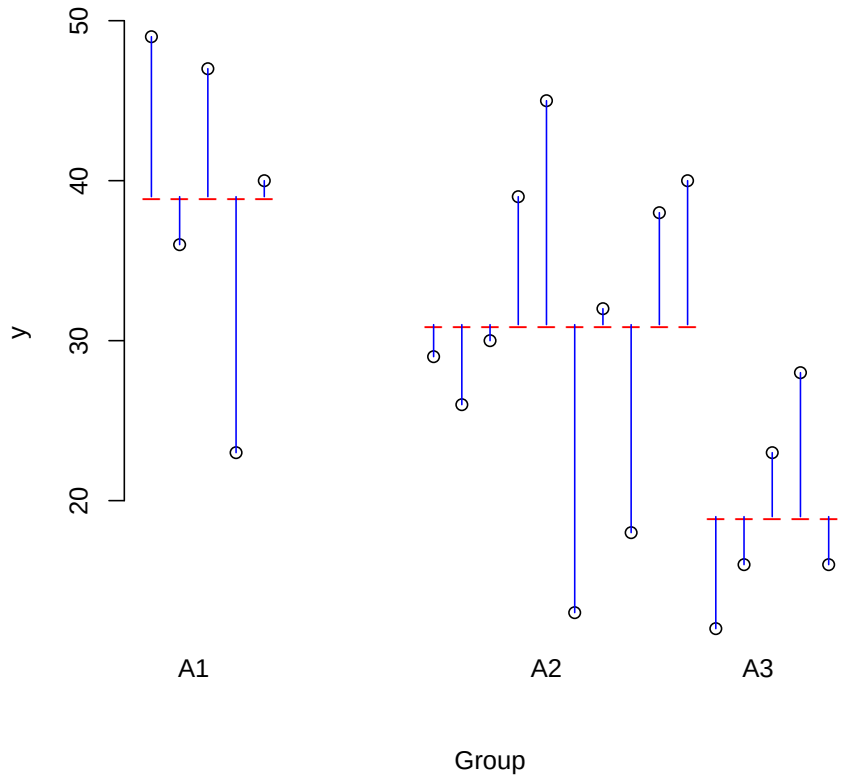

Figure 4: Podatki (○), ocenjene vrednosti v skupinah (vodoravne črte) in odkloni (navpične črte)

Tako lahko izluščimo varianco ostanka.

```
> (se2 <- aovTable[[1]]$"Mean Sq"[2])
[1] 88.70588
```

Število ponovitev v skupinah ni enako, zato ga iz stopinj prostosti ne moremo izluščiti.

$$s_e^2 = 88.706$$

Pripravimo lahko tudi uporabne povzetke povprečij in učinkov.

```
> par(mfrow = c(2, 2))
> plot(aovFit)
```

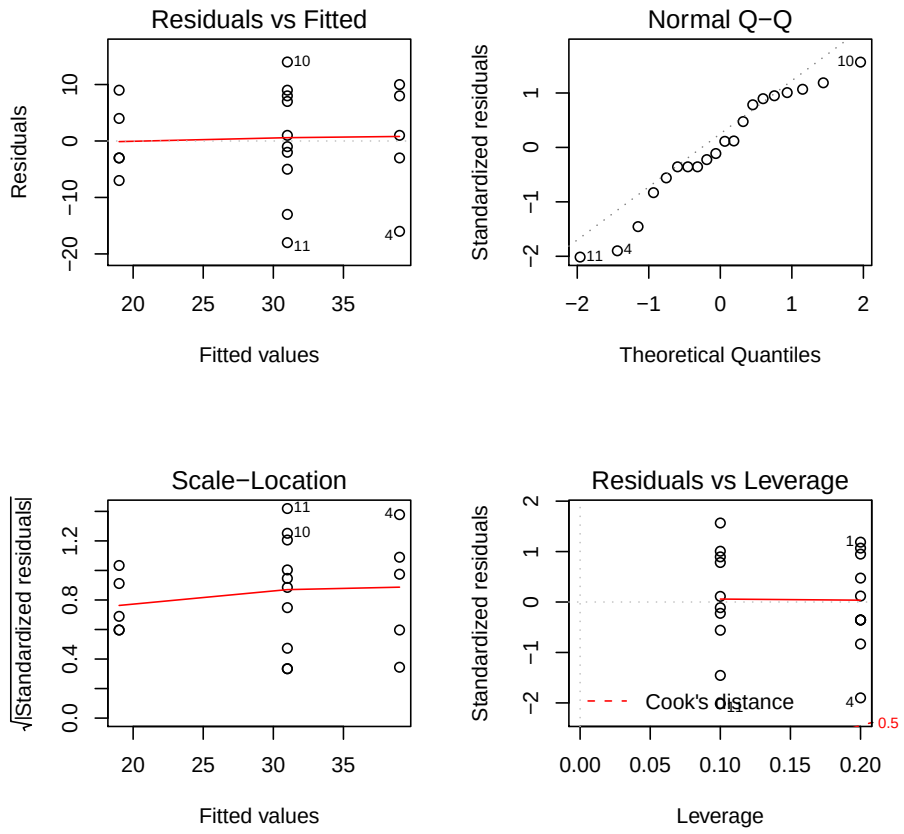

Figure 5: Diagnostične slike za analizo variance

```
> means <- model.tables(aovFit, "means", se = TRUE)
> means
Tables of means
Grand mean

30

 A
   A1 A2 A3
   39 31 19
rep  5 10  5
```

Združena standardna napaka za primerjavo razlik povprečij je v posebnem primeru enakega števila ponovitev poskusa

```
> sqrt(2 * se2/ns)
[1] 5.956707 4.212028 5.956707
```

Iz tabele povprečij lahko izračunamo učinke, to je odstopanja povprečij skupin od skupnega povprečja

```
> ucinek <- means$tables$A - means$tables$"Grand mean"
> ucinek
A
 A1  A2  A3
  9   1 -11
```

Standardna napaka učinkov

```
> sqrt(se2/n)
[1] 4.212028
```

Lahko pa tudi takole:

```
> effects <- model.tables(aovFit, "effects", se = TRUE)
> effects
Tables of effects

 A
    A1 A2  A3
     9  1 -11
rep  5 10   5
```

2.2.1 Tukey HSD

```
> hsd <- TukeyHSD(aovFit, ordered = TRUE)
> hsd$A <- hsd$A[order(hsd$A[, "diff"], decreasing = TRUE),
+     ]
> hsd
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = y ~ A, data = data)

$A
      diff       lwr      upr     p adj
A1-A3   20  4.718921 35.28108 0.0099030
A2-A3   12 -1.233803 25.23380 0.0790724
A1-A2    8 -5.233803 21.23380 0.2932649
```

Rezultat bolje prikaže slika 6

```
> plot(hsd, las = 1)
```

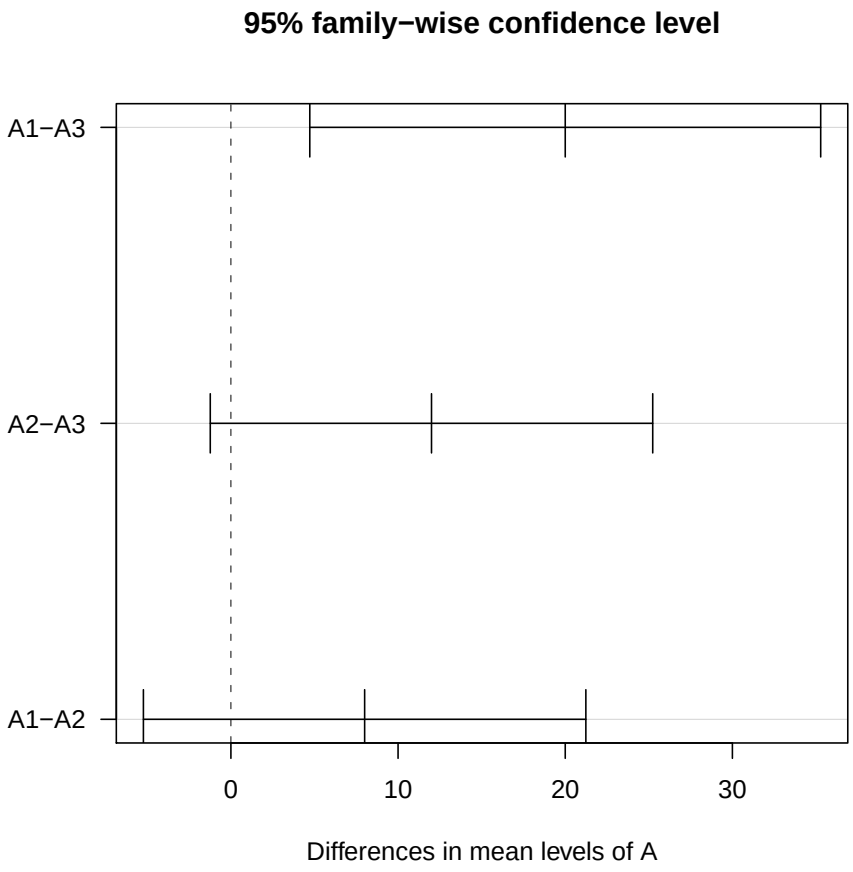

Figure 6: Intervali zaupanja za Tukey HSD

References

Blejec, M. (1971). *Statistično načrtovanje poskusov*, Volume 53/71. Inštitut za ekonomiko in organizacijo podjetja RCEF, Univerza v Ljubljani. 1, 10

SessionInfo

Windows XP (build 2600) Service Pack 3

- R version 2.10.0 (2009-10-26), `i386-pc-mingw32`
- Locale: `LC_COLLATE=Slovenian_Slovenia.1250`, `LC_CTYPE=Slovenian_Slovenia.1250`, `LC_MONETARY=Slovenian_Slovenia.1250`, `LC_NUMERIC=C`, `LC_TIME=Slovenian_Slovenia.1250`
- Base packages: base, datasets, graphics, grDevices, methods, splines, stats, utils
- Other packages: Hmisc 3.7-0, patchDVI 1.4.1545, survival 2.35-8
- Loaded via a namespace (and not attached): cluster 1.12.1, grid 2.10.0, lattice 0.18-3, tools 2.10.0

Project path: `C:/_Y/R/StatPred`

View as vignette

Project files can be viewed by pasting this code to R console:

```
> projectName <-"StatPred";  mainFile <-"Vzorcenje"

> commandArgs()
> library(tkWidgets)
> openPDF(file.path(dirname(getwd()), "doc", paste(mainFile,
+     "PDF", sep = ".")))
> viewVignette("viewVignette", projectName, file.path("../doc",
+     paste(mainFile, "RNW", sep = ".")))
```