Vilniaus Universitetas

Regresinė analizė

Laboratorinis darbas

Darbą atliko:

Vainius Gataveckas, Matas Gaulia, Dovydas Martinkus

Duomenų Mokslas

3 kursas 2 gr.

Vilnius, 2021

# Naudoti metodai

Darbas atliktas naudojant R, SAS ir Python.

Naudoti R paketai:

*tidyverse* – duomenų nuskaitymas, sutvarkymas, *ggplot2* paketas grafikams.

*janitor*

*car*

*lmtest*

*RcmdrMisc*

*lm.beta*

*psych*

*ppcor*

Naudoti Python paketai:

# Duomenys ir jų šaltiniai

Šalių gyventojų vidutinė gyvenimo trukmė pagal sveikatos rodiklius.

Duomenų šaltinis - Kaggle. Prieiga per internetą: <https://www.kaggle.com/kumarajarshi/life-expectancy-who>

Originalus šaltinis - WHO

# Atliktos analizės aprašymas

**1. Naudojant R**

**library**(tidyverse)

**library**(faux)

**library**(readr)

**library**(car)

**library**(agricolae)

*# Duomenų simuliavimas*

*# Naudojamos straipsnyje aprašytos charakteristikos*

mu\_l <- c(13.82, 13.59, 10.55, 25.64)

sigma\_l <- c(2.15, 1.68, 1.57, 5.18)

mu\_r <- c(14.09, 13.09, 10.41, 22.68)

sigma\_r <- c(1.72, 1.87, 2.17, 4.60)

construct\_df <- **function**(hand, mean, sd) {

pmap(list(rnorm\_multi(22, 4, mean, sd, r = 0.5), sd, mean), wanted\_mean\_sd) %>%

set\_names("t1", "t2", "t3", "t4") %>%

as\_tibble() %>%

mutate(handedness = hand)

}

wanted\_mean\_sd <- **function**(x, sd, mean) {

(x - mean(x)) / sd(x) \* sd + mean

}

df\_right <- construct\_df("right", mu\_r, sigma\_r)

biggest <- sort(df\_right$t4, decreasing = TRUE)[1:4]

**for** (i **in** biggest) {

df\_right$t4[which(df\_right$t4 == i)] <- df\_right$t4[which(df\_right$t4 == i)] + abs(rnorm(1, 6))

}

t4 <- df\_right$t4

df\_right$t4 <- (t4 - mean(t4)) / sd(t4) \* 4.60 + 22.68

df <- rbind(construct\_df("left", mu\_l, sigma\_l), df\_right)

df <- df %>%

mutate(

age = rnorm(44, 17.32, 1.07),

sex = sample(c(rep("male", 20), rep("female", 24)), 44)

)

Simuliuoti duomenys išsaugoti ir pateikti data.csv faile.

df <- read\_csv("data.csv")

options(contrasts = c("contr.sum", "contr.poly"))

variance\_check <- **function**(x) {

eval(substitute(leveneTest(x ~ handedness \* sex, data = df)))

}

anova\_model <- **function**(x) {

eval(substitute(aov(x ~ handedness \* sex, data = df)))

}

*# Tiriamieji grafikai*

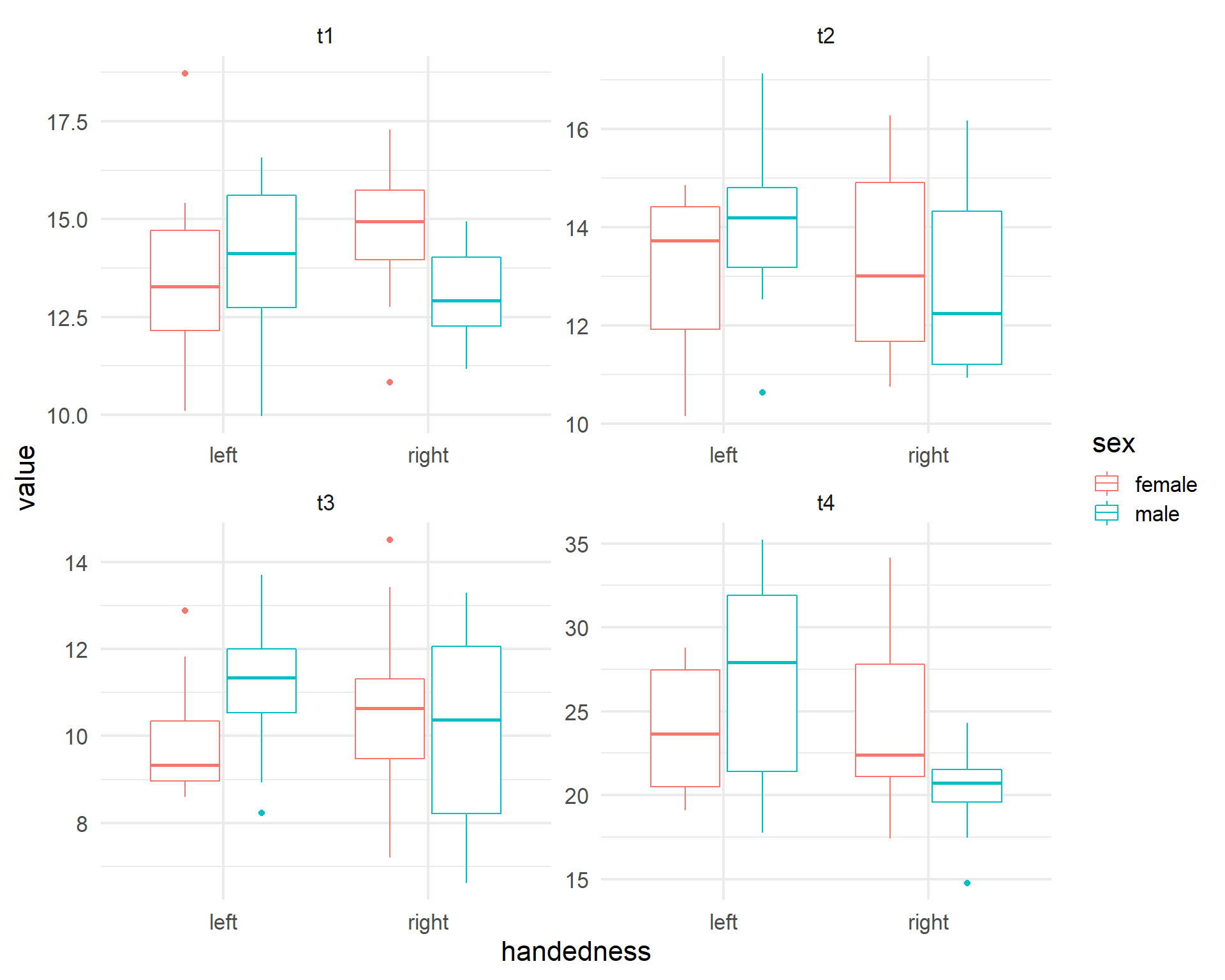
df\_pivoted <- df %>% pivot\_longer(1:4)

ggplot(df\_pivoted, aes(handedness, value, color = sex)) +

geom\_boxplot() +

theme\_minimal(base\_size = 16) +

facet\_wrap(vars(name), scales = "free")



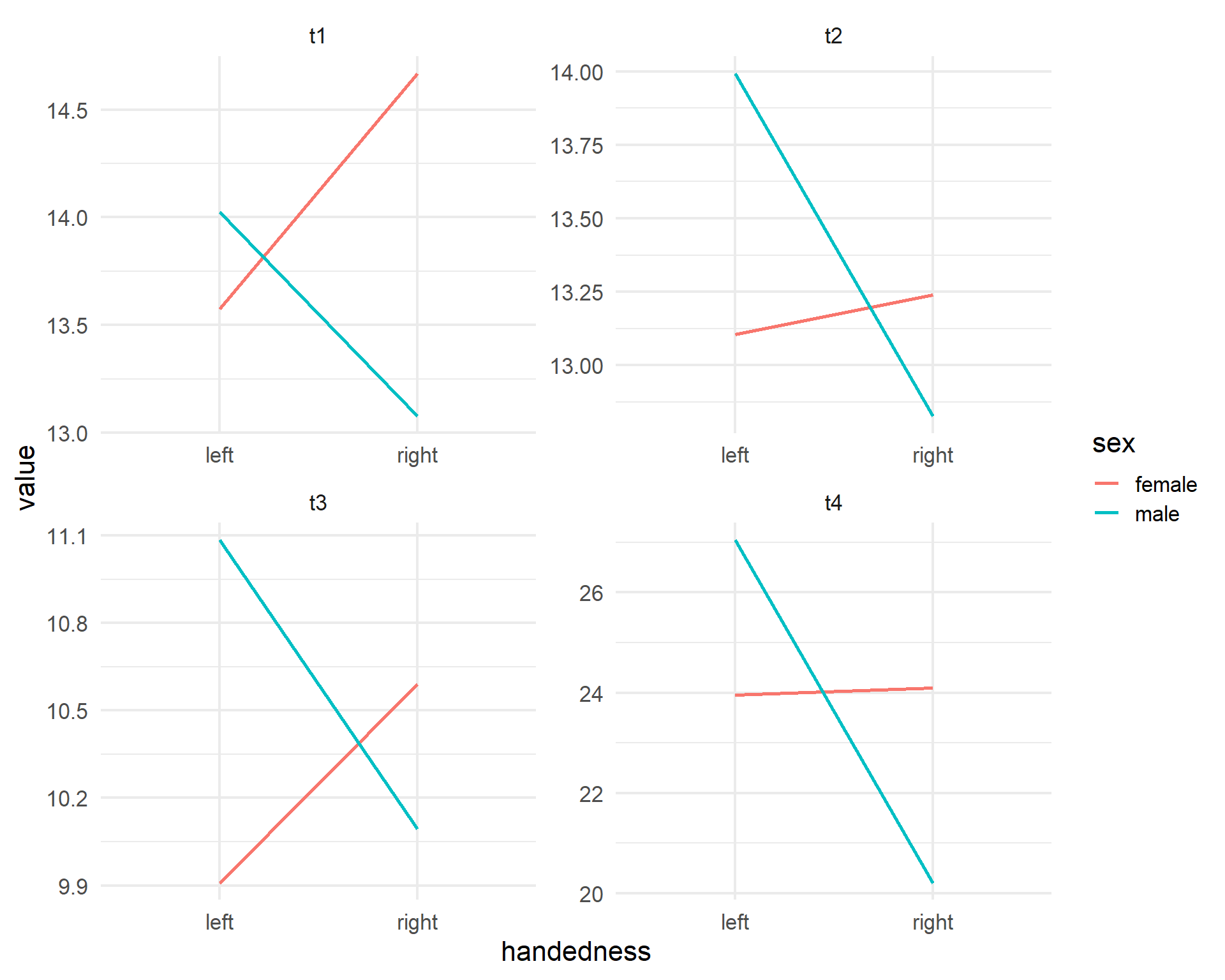
*# Vidurkių grafikas*

ggplot(df\_pivoted, aes(handedness, value, color = sex, group = sex)) +

stat\_summary(fun = "mean", geom = "line", size = 1) +

theme\_minimal(base\_size = 16) +

facet\_wrap(vars(name), scales = "free")



*# Dispersijų lygybės testas*

variance\_checks <- list(variance\_check(t1), variance\_check(t2), variance\_check(t3), variance\_check(t4))

variance\_checks

## [[1]]

## Levene's Test for Homogeneity of Variance (center = median)

## Df F value Pr(>F)

## group 3 0.8848 0.4572

## 40

##

## [[2]]

## Levene's Test for Homogeneity of Variance (center = median)

## Df F value Pr(>F)

## group 3 0.4793 0.6985

## 40

##

## [[3]]

## Levene's Test for Homogeneity of Variance (center = median)

## Df F value Pr(>F)

## group 3 1.7389 0.1745

## 40

##

## [[4]]

## Levene's Test for Homogeneity of Variance (center = median)

## Df F value Pr(>F)

## group 3 1.8124 0.1604

## 40

Hipotezė apie dispersijų lygybę neatmetama ne vienam testui.

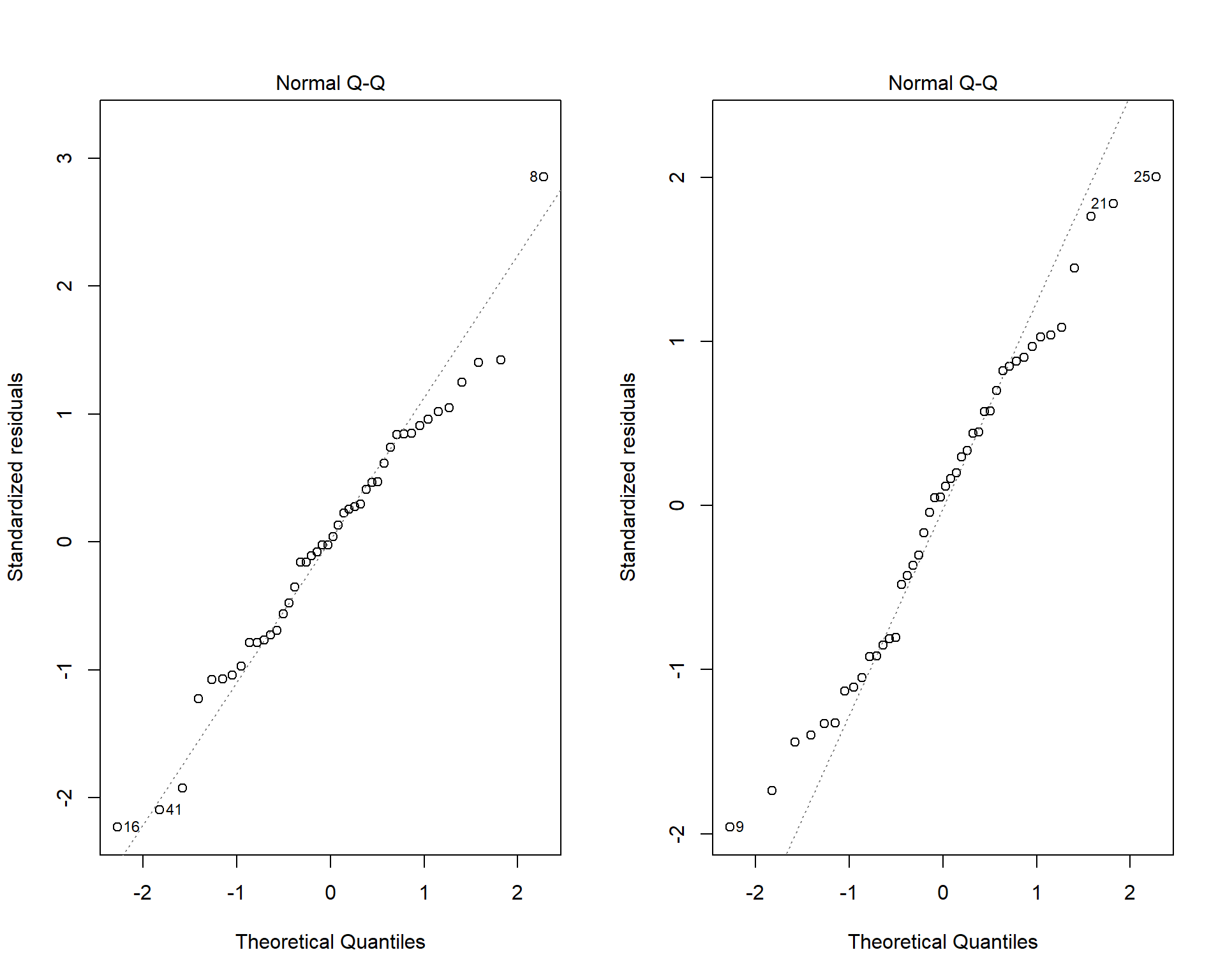
*# Sukuriami dispersinės analizės modeliai kiekvienam testui*

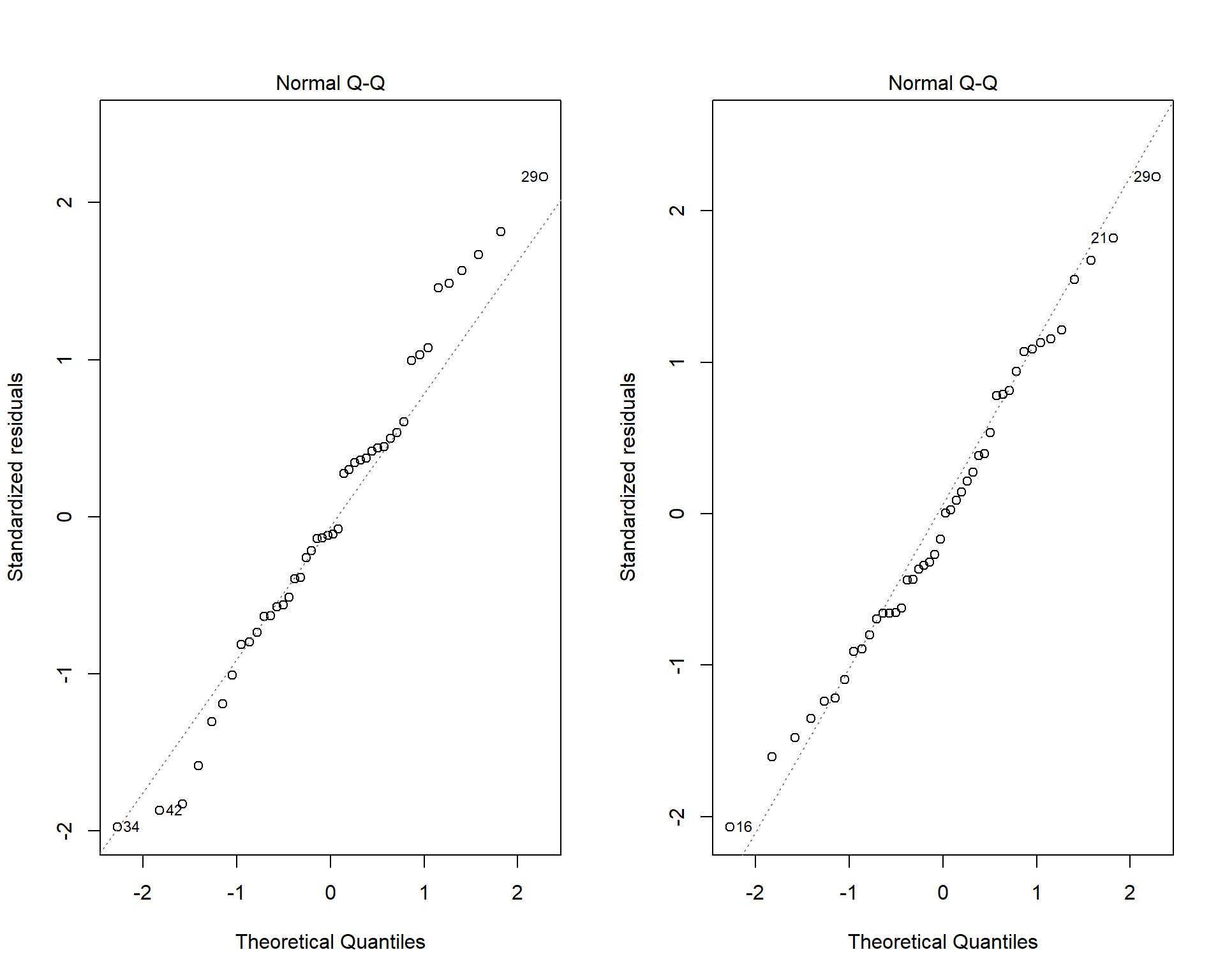
models <- list(anova\_model(t1), anova\_model(t2), anova\_model(t3), anova\_model(t4))

*# Tikrinamas liekanų normalumas (dispersinės analizės prielaida)*

op <- par(mfrow = c(1, 2))

map(models, ~ plot(.x, which = 2))





Liekanos visiems testas stipriai nesiskiria nuo normalumo.

*# Nesubalansuotas eksperimento planas -> naudojamos Type III kv. sumos*

map(models, ~ Anova(.x, type = "III"))

## [[1]]

## Anova Table (Type III tests)

##

## Response: t1

## Sum Sq Df F value Pr(>F)

## (Intercept) 8065.8 1 2221.9073 < 2e-16 \*\*\*

## handedness 0.1 1 0.0156 0.90108

## sex 3.4 1 0.9457 0.33665

## handedness:sex 11.0 1 3.0212 0.08988 .

## Residuals 145.2 40

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## [[2]]

## Anova Table (Type III tests)

##

## Response: t2

## Sum Sq Df F value Pr(>F)

## (Intercept) 7443.3 1 2334.5324 <2e-16 \*\*\*

## handedness 2.8 1 0.8795 0.3540

## sex 0.6 1 0.1862 0.6684

## handedness:sex 4.5 1 1.3986 0.2439

## Residuals 127.5 40

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## [[3]]

## Anova Table (Type III tests)

##

## Response: t3

## Sum Sq Df F value Pr(>F)

## (Intercept) 4574.0 1 1289.7625 <2e-16 \*\*\*

## handedness 0.3 1 0.0706 0.7918

## sex 1.2 1 0.3433 0.5612

## handedness:sex 7.4 1 2.0750 0.1575

## Residuals 141.9 40

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## [[4]]

## Anova Table (Type III tests)

##

## Response: t4

## Sum Sq Df F value Pr(>F)

## (Intercept) 23915.5 1 1088.1923 < 2e-16 \*\*\*

## handedness 118.1 1 5.3749 0.02562 \*

## sex 1.7 1 0.0752 0.78531

## handedness:sex 127.9 1 5.8217 0.02050 \*

## Residuals 879.1 40

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Kiekvienam iš 4 testų atskirai atlikta dvifaktorinė dispersinė analizė su fiksuotais faktoriais pasirenkant ranką ir lytį kaip nepriklausomus kintamuosius. Ketvirtame teste rastos statistiškai reikšmingos rankos F(1,40) = 5.37 , p = 0.025 ir rankos/lyties sąveikos F(1,40) = 5.82, p = 0.020 įtakos. Kituose testuose statistiškai reikšmingos faktorių įtakos nerasta.

**library**(agricolae)

HSD.test(models[[4]], trt = c("handedness", "sex"), console = TRUE, unbalanced = TRUE)

##

## Study: models[[4]] ~ c("handedness", "sex")

##

## HSD Test for t4

##

## Mean Square Error: 21.97729

##

## handedness:sex, means

##

## t4 std r Min Max

## left:female 23.95501 3.705637 10 19.07796 28.78542

## left:male 27.04416 5.937883 12 17.76119 35.20957

## right:female 24.09143 4.873529 14 17.40837 34.13473

## right:male 20.21000 2.900647 8 14.77681 24.32336

##

## Alpha: 0.05 ; DF Error: 40

## Critical Value of Studentized Range: 3.790685

##

## Groups according to probability of means differences and alpha level( 0.05 )

##

## Treatments with the same letter are not significantly different.

##

## t4 groups

## left:male 27.04416 a

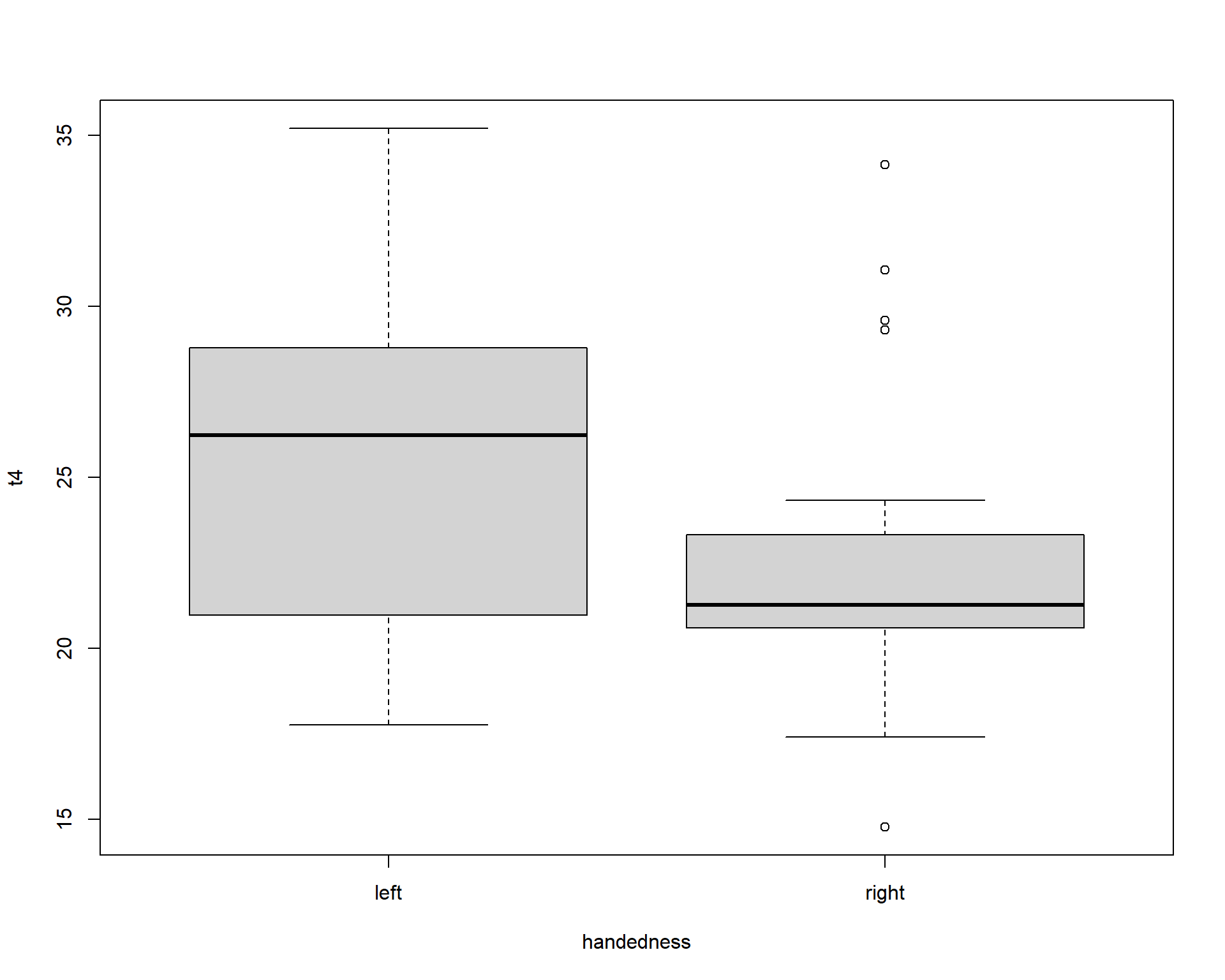
## right:female 24.09143 ab

## left:female 23.95501 ab

## right:male 20.21000 b

Naudojantis porinių kontrastų analize ketvirtam testui rastas statistiškai reikšmingas skirtumas tarp kairiarankių ir dešiniarankių vyrų.

boxplot(t4 ~ handedness, data = df)



Ketvirto testo duomenyse tarp dešiniarankių rastos keturios išskirtys.

Šio testo duomenys transformuoti taip, kaip tai atlikta straipsnyje. Dispersinė analizė atliekama pakartotinai.

# Duomenų transformacija

df2 <- df

t4 <- df$t4

limit <- mean(t4[df$handedness=="right"])+ 1.5\*sd(t4[df$handedness=="right"])

df2$t4 <- sqrt(ifelse(t4>limit,limit,t4))

model\_trans <- aov(t4 ~ handedness \* sex, df2)

Anova(model\_trans, type = "III")

## [[1]]

## Anova Table (Type III tests)

##

## Response: t4

## Sum Sq Df F value Pr(>F)

## (Intercept) 977.83 1 5923.2061 < 2e-16 \*\*\*

## handedness 0.96 1 5.8408 0.02031 \*

## sex 0.09 1 0.5601 0.45858

## handedness:sex 0.76 1 4.6264 0.03758 \*

## Residuals 6.60 40

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Dispersine transformuotų duomenų analize gauti tie patys statistiškai reikšmingi rezultatai kaip prieš transformaciją: statistiškai reikšmingos rankos F(1,40) = 5.84, p = 0.020 ir rankos/lyties sąveikos F(1,40) = 4.62, p= 0.037 įtakos. Kairiarankių vyrų rezultatai statistiškai reikšmingai geresni už dešiniarankių vyrų (Tjukio metodu α=0.05 ).

pairwise\_test <- HSD.test(model\_trans, trt = c("handedness", "sex"), console = TRUE, unbalanced = TRUE)

pairwise\_test

##

## Study: model\_trans ~ c("handedness", "sex")

##

## HSD Test for t4

##

## Mean Square Error: 0.165085

##

## handedness:sex, means

##

## t4 std r Min Max

## left:female 4.881060 0.3804443 10 4.367833 5.365205

## left:male 5.056641 0.4577357 12 4.214402 5.438750

## right:female 4.847775 0.4135532 14 4.172334 5.438750

## right:male 4.484798 0.3322381 8 3.844062 4.931872

##

## Alpha: 0.05 ; DF Error: 40

## Critical Value of Studentized Range: 3.790685

##

## Groups according to probability of means differences and alpha level( 0.05 )

##

## Treatments with the same letter are not significantly different.

##

## t4 groups

## left:male 5.056641 a

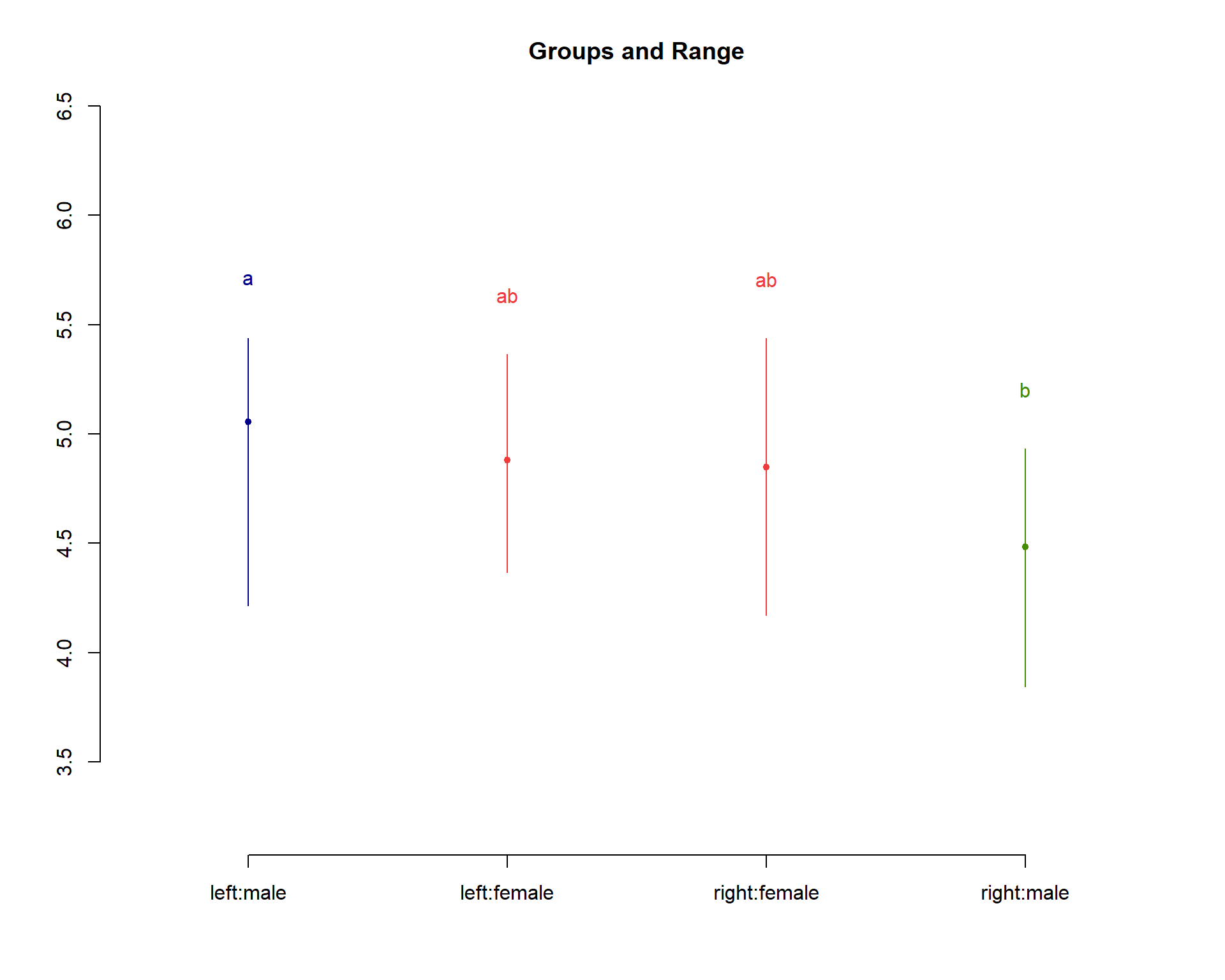
## left:female 4.881060 ab

## right:female 4.847775 ab

## right:male 4.484798 b

Plot(pairwise\_test)

Nubraižomas 95% pasikliovimo grafikas rankos/lyties sąveikos skirtingų lygmenų įtakai:



**2. Naudojant SAS**

PROC IMPORT DATAFILE='/home/u45871880/data.csv'

DBMS=CSV

OUT=data;

GETNAMES=YES;

RUN;

PROC SORT data=data;

BY sex;

RUN;

data colours;

length value FillColor LineColor $30;

Id='X'; Value="male"; FillColor='#799fcb'; LineColor='#799fcb'; output;

Id='X'; Value="female"; FillColor='#f9665e'; LineColor='#f9665e'; output;

run;

/\* Tiriamieji grafikai \*/

%macro box;

%do m=1 %to 4;

proc sgplot data=data dattrmap=colours;

vbox t&m/group=sex category=handedness attrid=X;

%end;

%mend;

%box;

RUN;

/\* Vidurkių grafikai \*/

%macro box;

%do m=1 %to 4;

proc sgplot data=data dattrmap=colours;

vline handedness / response=t&m group=sex stat=mean attrid=X;

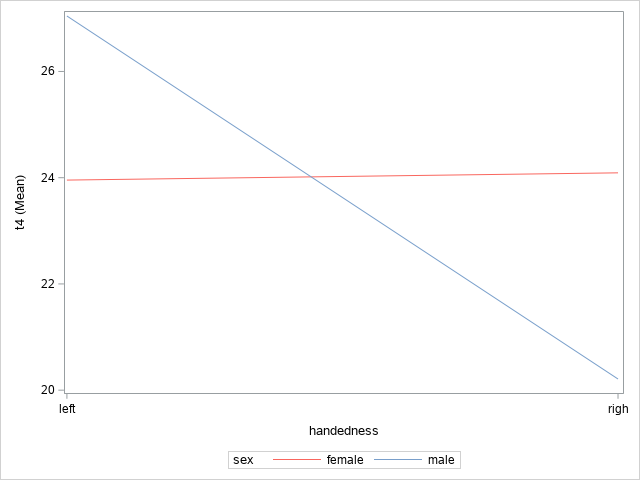
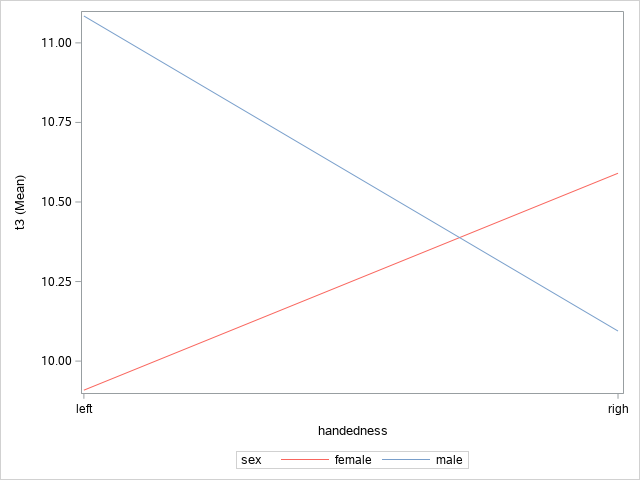
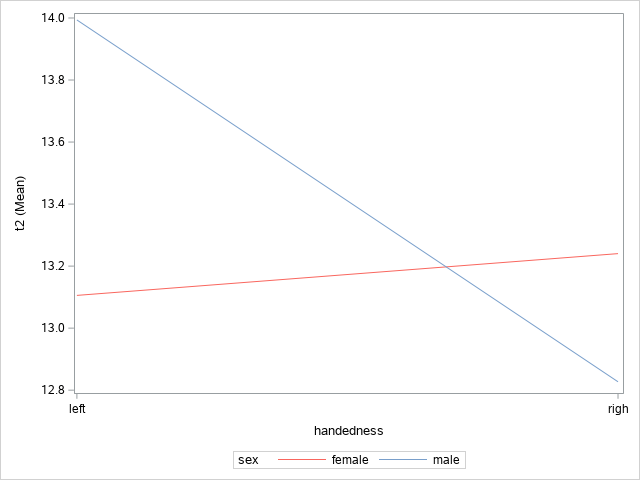
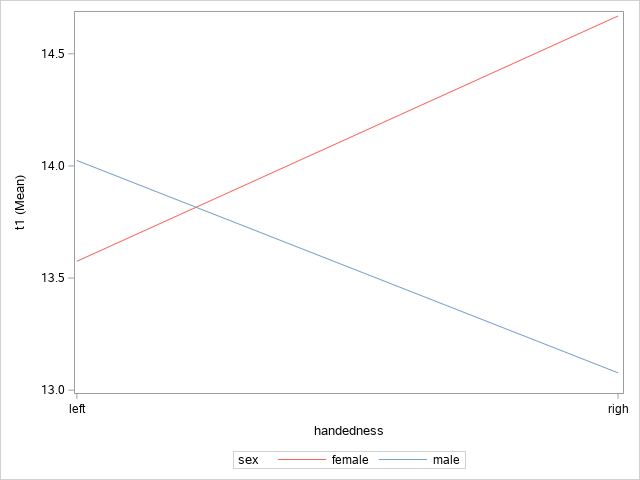
run;

%end;

%mend;

%box;

RUN;



/\* Dispersinė analizė \*/

%macro box;

%do m=1 %to 4;

proc glm data = data plots=diagnostics;

class handedness sex;

model t&m = handedness sex handedness\*sex;

lsmeans handedness\*sex /adjust=TUKEY linestable plots=None;

%end;

%mend;

%box;

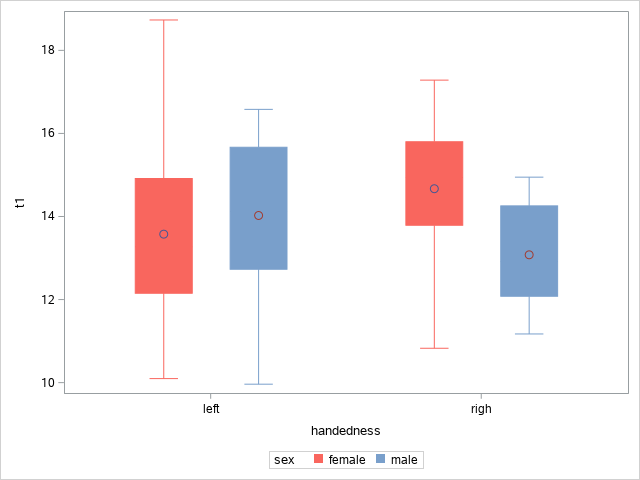
RUN;

**The GLM Procedure**

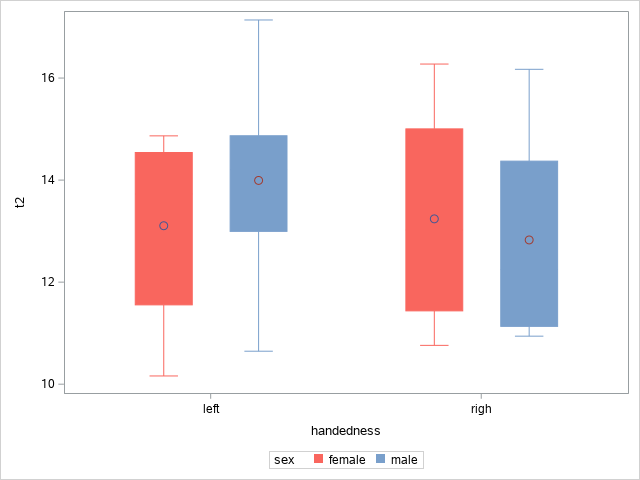
| **Class Level Information** | | |
| --- | --- | --- |
| **Class** | **Levels** | **Values** |
| **handedness** | 2 | left righ |
| **sex** | 2 | female male |

|  |  |
| --- | --- |
| **Number of Observations Read** | 44 |
| **Number of Observations Used** | 44 |

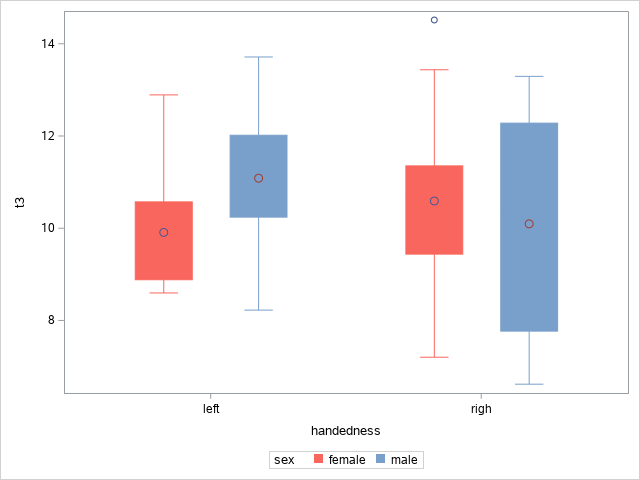
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **handedness** | 1 | 0.05679745 | 0.05679745 | 0.02 | 0.9011 |
| **sex** | 1 | 3.43312163 | 3.43312163 | 0.95 | 0.3367 |
| **handedness\*sex** | 1 | 10.96730369 | 10.96730369 | 3.02 | 0.0899 |



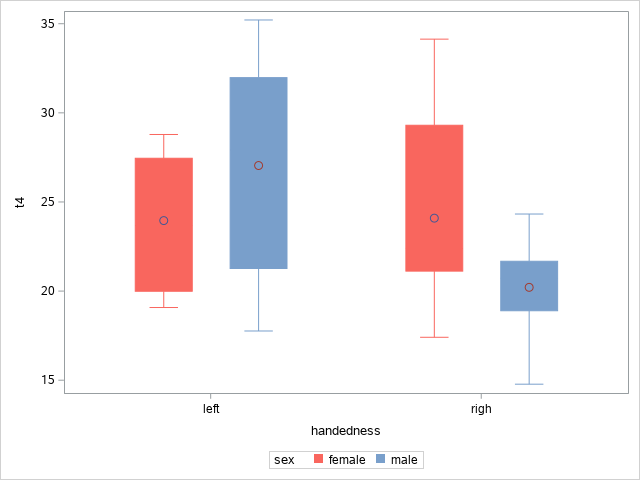
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **handedness** | 1 | 2.80425930 | 2.80425930 | 0.88 | 0.3540 |
| **sex** | 1 | 0.59375922 | 0.59375922 | 0.19 | 0.6684 |
| **handedness\*sex** | 1 | 4.45919765 | 4.45919765 | 1.40 | 0.2439 |



| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **handedness** | 1 | 0.25036993 | 0.25036993 | 0.07 | 0.7918 |
| **sex** | 1 | 1.21733123 | 1.21733123 | 0.34 | 0.5612 |
| **handedness\*sex** | 1 | 7.35893993 | 7.35893993 | 2.08 | 0.1575 |



| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **handedness** | 1 | 118.1258936 | 118.1258936 | 5.37 | 0.0256 |
| **sex** | 1 | 1.6529039 | 1.6529039 | 0.08 | 0.7853 |
| **handedness\*sex** | 1 | 127.9456061 | 127.9456061 | 5.82 | 0.0205 |



| **Tukey-Kramer Grouping for LS-Means of handedness\*sex** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **LS-means with the same letter are not significantly different.** | | | | | |
|  | | **t4 LSMEAN** | **handedness** | **sex** | **LSMEAN Number** |
|  | A | 27.04416 | **left** | **male** | 2 |
|  | A |  |  |  |  |
| B | A | 24.09143 | **righ** | **female** | 3 |
| B | A |  |  |  |  |
| B | A | 23.95501 | **left** | **female** | 1 |
| B |  |  |  |  |  |
| B |  | 20.21000 | **righ** | **male** | 4 |

Kaip ir naudojant R kiekvienam iš 4 testų atskirai atlikta dvifaktorinė dispersinė analizė su fiksuotais faktoriais pasirenkant ranką ir lytį kaip nepriklausomus kintamuosius. Ketvirtame teste rastos statistiškai reikšmingos rankos F(1,40) = 5.37 , p = 0.025 ir rankos/lyties sąveikos F(1,40) = 5.82, p = 0.020 įtakos. Kituose testuose statistiškai reikšmingos faktorių įtakos nerasta.

Naudojantis porinių kontrastų analize ketvirtam testui rastas statistiškai reikšmingas skirtumas tarp kairiarankių ir dešiniarankių vyrų.

/\* Transformuoti duomenys \*/

PROC IMPORT DATAFILE='/home/u45871880/data2.csv'

DBMS=CSV

OUT=data2;

GETNAMES=YES;

RUN;

proc glm data = data2 plots=diagnostics;

class handedness sex;

model t4 = handedness sex handedness\*sex;

lsmeans handedness\*sex /adjust=TUKEY linestable plots=None;

run;

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **handedness** | 1 | 0.96423252 | 0.96423252 | 5.84 | 0.0203 |
| **sex** | 1 | 0.09247206 | 0.09247206 | 0.56 | 0.4586 |
| **handedness\*sex** | 1 | 0.76375261 | 0.76375261 | 4.63 | 0.0376 |

| **Tukey-Kramer Grouping for LS-Means of handedness\*sex** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **LS-means with the same letter are not significantly different.** | | | | | |
|  | | **t4 LSMEAN** | **handedness** | **sex** | **LSMEAN Number** |
|  | A | 27.04416 | **left** | **male** | 2 |
|  | A |  |  |  |  |
| B | A | 24.09143 | **righ** | **female** | 3 |
| B | A |  |  |  |  |
| B | A | 23.95501 | **left** | **female** | 1 |
| B |  |  |  |  |  |
| B |  | 20.21000 | **righ** | **male** | 4 |

Gauti rezultatai sutampa su rezultatais gautais naudojant R: Dispersine transformuotų duomenų analize gauti tie patys statistiškai reikšmingi rezultatai kaip prieš transformaciją: statistiškai reikšmingos rankos F(1,40) = 5.84, p = 0.020 ir rankos/lyties sąveikos F(1,40) = 4.62, p= 0.037 įtakos. Kairiarankių vyrų rezultatai statistiškai reikšmingai geresni už dešiniarankių vyrų (Tjukio metodu α=0.05 ).

3**. Naudojant Python**

**import** warnings

warnings**.**filterwarnings("ignore")

**import** pandas **as** pd

**import** seaborn **as** sns

**import** numpy **as** np

**import** matplotlib.pyplot **as** plt

**from** scipy.stats **import** levene

**import** statsmodels.api **as** sm

**from** statsmodels.formula.api **import** ols

**import** pylab

**import** scipy.stats **as** stats

**from** bioinfokit.analys **import** stat

**def** split(df, col):

**return** [

df[df["handedness"] **==** "left"][df["sex"] **==** "female"][col],

df[df["handedness"] **==** "left"][df["sex"] **==** "male"][col],

df[df["handedness"] **==** "right"][df["sex"] **==** "female"][col],

df[df["handedness"] **==** "right"][df["sex"] **==** "male"][col]

]

**def** vartest(df,col):

s **=** split(df,col)

stat, p **=** levene(s[0],s[1],s[2],s[3])

print("F value:", round(stat,4), "Pr(>F)", round(p,4))

**def** anova(df, col):

stats**.**probplot(df[col], dist**=**"norm", plot**=**pylab)

pylab**.**show()

model **=** ols(col **+** ' ~ sex \* handedness', data**=**df)**.**fit()

anova\_table **=** sm**.**stats**.**anova\_lm(model, typ**=**3)

**return** anova\_table

data **=** pd**.**read\_csv("data.csv")

data **=** data**.**sort\_values(["sex", "handedness"])

mypal **=** {sex: '#f9665e' **if** sex **==** "female" **else** '#799fcb' **for** sex **in** data["sex"]**.**unique()}

ft **=** data

ft["group"] **=** ft["handedness"] **+** ft["sex"]

fig, axes **=** plt**.**subplots(2, 2,figsize**=**(15,12))

fig**.**suptitle("Tiriamieji grafikai")

sns**.**boxplot(ax **=** axes[0,0],x**=**"handedness", y**=**"t1", hue**=**"sex", data**=**data, palette**=**mypal)

sns**.**boxplot(ax **=** axes[0,1],x**=**"handedness", y**=**"t2", hue**=**"sex", data**=**data, palette**=**mypal)

sns**.**boxplot(ax **=** axes[1,0],x**=**"handedness", y**=**"t3", hue**=**"sex", data**=**data, palette**=**mypal)

sns**.**boxplot(ax **=** axes[1,1],x**=**"handedness", y**=**"t4", hue**=**"sex", data**=**data, palette**=**mypal)

Paveikslėlis, kuriame yra žinutė, kryžiažodis, pirmosios pagalbos rinkinys, rezultatų lentelė

Automatiškai sugeneruotas aprašymas

means **=** data**.**groupby(['sex','handedness'])**.**mean()

fig, axes **=** plt**.**subplots(2, 2,figsize**=**(15,12))

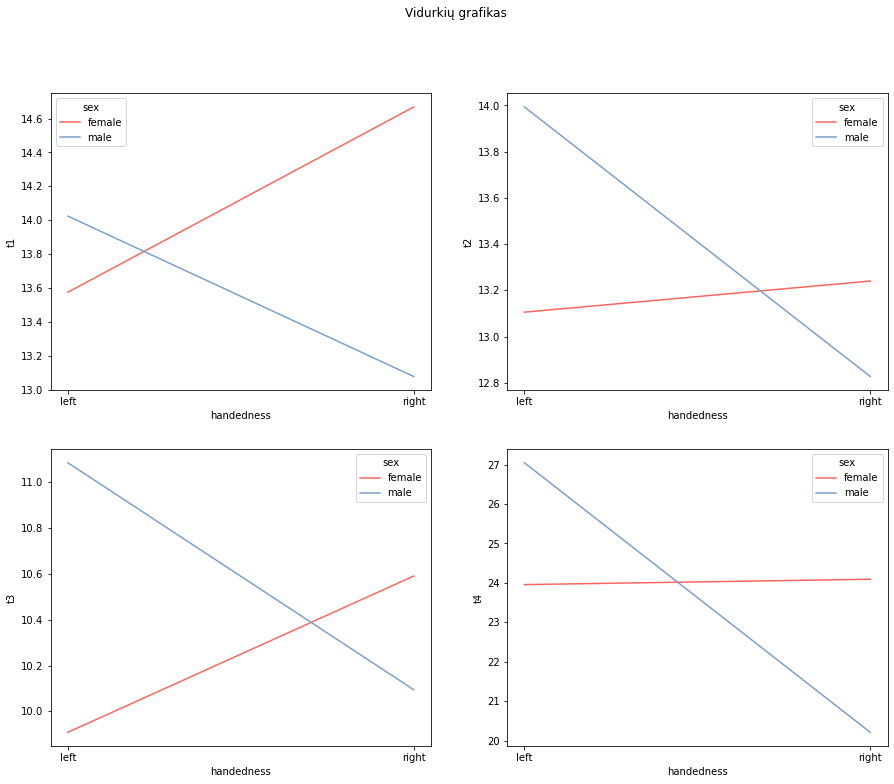
fig**.**suptitle("Vidurkių grafikas")

sns**.**lineplot(ax **=** axes[0,0],x**=**"handedness", y**=**"t1", hue**=**"sex", data**=**means, palette**=**mypal)

sns**.**lineplot(ax **=** axes[0,1],x**=**"handedness", y**=**"t2", hue**=**"sex", data**=**means, palette**=**mypal)

sns**.**lineplot(ax **=** axes[1,0],x**=**"handedness", y**=**"t3", hue**=**"sex", data**=**means, palette**=**mypal)

sns**.**lineplot(ax **=** axes[1,1],x**=**"handedness", y**=**"t4", hue**=**"sex", data**=**means, palette**=**mypal)



vartest(data,"t1")

F value: 0.8848 Pr(>F) 0.4572

vartest(data,"t2")

F value: 0.4793 Pr(>F) 0.6985

vartest(data,"t3")

F value: 1.7389 Pr(>F) 0.1745

vartest(data,"t4")

F value: 1.8124 Pr(>F) 0.1604

sum\_sq df F PR(>F)

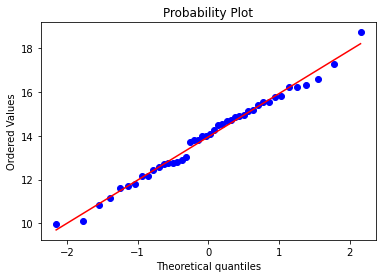
Intercept 1842.757078 1.0 507.629040 2.428428e-24

sex 1.102086 1.0 0.303594 5.847024e-01

handedness 6.979565 1.0 1.922679 1.732417e-01

sex:handedness 10.967304 1.0 3.021191 8.987678e-02

Residual 145.205016 40.0 NaN NaN



sum\_sq df F PR(>F)

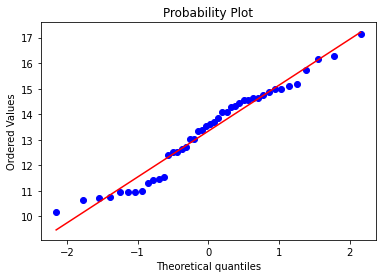
Intercept 1717.564555 1.0 538.700233 8.037321e-25

sex 4.301995 1.0 1.349286 2.522877e-01

handedness 0.105805 1.0 0.033185 8.563713e-01

sex:handedness 4.459198 1.0 1.398591 2.439417e-01

Residual 127.533975 40.0 NaN NaN



sum\_sq df F PR(>F)

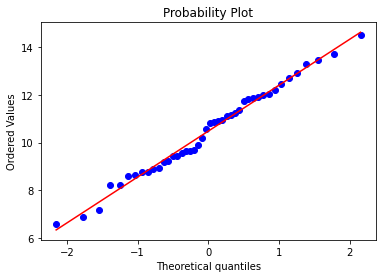
Intercept 981.812039 1.0 276.846236 1.413343e-19

sex 7.541214 1.0 2.126432 1.525851e-01

handedness 2.710706 1.0 0.764351 3.871895e-01

sex:handedness 7.358940 1.0 2.075035 1.575086e-01

Residual 141.856657 40.0 NaN NaN



sum\_sq df F PR(>F)

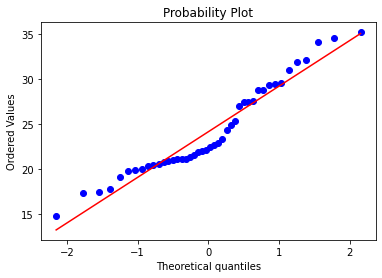
Intercept 5738.425909 1.0 261.107038 3.929689e-19

sex 52.051729 1.0 2.368432 1.316855e-01

handedness 0.108554 1.0 0.004939 9.443204e-01

sex:handedness 127.945606 1.0 5.821718 2.050468e-02

Residual 879.091725 40.0 NaN NaN



res = stat()

res.tukey\_hsd(df=ft, res\_var='t4', xfac\_var='group', anova\_model='t4 ~ group')

res.tukey\_summary

group1 group2 Diff Lower Upper q-value p-value

0 leftfemale rightfemale 0.136416 -5.066654 5.339486 0.099392 0.900000

1 leftfemale leftmale 3.089145 -2.291556 8.469846 2.176434 0.425942

2 leftfemale rightmale 3.745010 -2.215856 9.705877 2.381714 0.345613

3 rightfemale leftmale 2.952729 -1.990949 7.896407 2.264224 0.390736

4 rightfemale rightmale 3.881426 -1.688128 9.450981 2.641903 0.257923

5 leftmale rightmale 6.834156 1.098309 12.570002 4.516826 0.013984