

Using for statistical analyses

Robert Bauer

Warnemünde, 05/16/2012



Day 6 - Agenda:

- ▶ non-parametric alternatives to t-test and ANOVA (incl. post hoc tests)
 - ▶ Wilcoxon Rank Sum/Mann-Whitney U-Test
 - ▶ Kruskal-Wallis Rank Sum Test
- ▶ repeated measures ANOVA

Wilcoxon Rank Sum Test

- ▶ alternative to t-test (for non-normal distributed data)
 - ▶ for one or two sample analyses
 - ▶ for two samples also known as Mann-Whitney U-Test
- ▶ Assumptions
 - ▶ variance homogeneity
 - ▶ independent data
 - ▶ at least ordinal scaled data
- ▶ functions for both tests:
`wilcox.test()`
`wilcox.exact()` # from `exactRankTests`-package;
necessary if data contains ties or more than 50
distinct values

One-Sample Tests

```
# One sample tests
setwd('~/.Dropbox/R_course/day5') # set working directory

# load dataframe
file <- "Clone1.csv"
datasheet <- read.table(file, header=T, sep=',', dec=".")
head(datasheet)

#### H0: mean == 0
# a) Student's t-test
t.test(datasheet$growth.rate)

# b) Wilcoxon Rank Sum Test
wilcox.test(datasheet$growth.rate)
# if p-value < 0.05: reject H0, else: keep H0
```

```
Wilcoxon signed rank test
data:  datasheet$growth.rate
V = 300, p-value = 1.192e-07
alternative hypothesis: true location is not equal to 0
```

One-Sample Tests

```
#### H0: mean == 2
# a) Student's t-test
t.test(datasheet$growth.rate, mu=2)
t.test(datasheet$growth.rate, mu=2, alternative="two.sided")
t.test(datasheet$growth.rate, mu=2, alternative="t")

# b) Wilcoxon Rank Sum Test
wilcox.test(datasheet$growth.rate, mu=2)
wilcox.test(datasheet$growth.rate, mu=2, alternative="t")

#### H0: mean <= 2
# a) Student's t-test
t.test(datasheet$growth.rate, mu=2, alternative="g")

# b) Wilcoxon Rank Sum Test
wilcox.test(datasheet$growth.rate, mu=2, alternative="g")
```

One-Sample Tests

```
#### H0: mean >= 2
# a) Student's t-test
t.test(datasheet$growth.rate, mu=2, alternative="l")

# b) Wilcoxon Rank Sum Test
wilcox.test(datasheet$growth.rate, mu=2, alternative="l")

#### H0: mean == 3
# a) Student's t-test
t.test(datasheet$growth.rate, mu=3)

# b) Wilcoxon Rank Sum Test
wilcox.test(datasheet$growth.rate, mu=3)

#### calculate confidence intervals and pseudo-median
wilcox.test(datasheet$growth.rate, mu=3, conf.int=TRUE)
```

Two-Sample comparisons

```
# compare means between sexes
Data = datasheet
Data$sex <- factor(Data$sex)
formula <- growth.rate~sex
ylabel <- "growth rate"
xlabel <- "sex"

boxplot(formula, data=Data, ylab=ylabel, xlab=xlabel)

# Shapiro-Wilk test for Normal Distribution
# better for small sampling sizes (<50)
# H0: normal distributed data
shapiro.test(Data$growth.rate[Data$sex == "f"])
shapiro.test(Data$growth.rate[Data$sex == "m"])
# if p-value < 0.05: reject H0, else: keep H0

# Levene's test for homogeneity of variance across groups
# H0: variances are equal (homogeneity of variance)
# H1: variances differ
install.packages('car') # install required package
library(car)            # load package
leveneTest(formula, data=Data)
# if Pr(>F) < 0.05: reject H0, else: keep H0
```

Two-Sample comparisons - parametric tests

```
# a) Two-Sample t-test
# H0: average growth rates are equal between sexes
# H1: average growth rates differ between sexes
t.test(formula, alternative="two.sided",
        paired=F, var.equal=T, data=Data)
# if p-value < 0.05: reject H0, else: keep H0

# b) One-Way ANOVA
# H0: average growth rates are equal between sexes
anova(lm(formula, data=Data))      # version 1
summary(aov(formula, data=Data))   # version 2
# if Pr(>F) < 0.05: reject H0, else: keep H0
```


Two-Sample comparisons - non-parametric tests

```
# a) Two sample Wilcoxon Rank Sum Test (= Mann-Whitney U-Test)
# H0: average growth rates are equal between sexes
# H1: average growth rates differ between sexes
wilcox.test(formula, alternative="two.sided",
             paired=F, var.equal=T, data=Data, conf.int=T)
# if p-value < 0.05: reject H0, else: keep H0
```

Wilcoxon rank sum test with continuity correction

data: growth.rate by sex

W = 77, p-value = 0.7987

alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:

-0.4693695 0.6574221

sample estimates:

difference in location

0.1012352

Kruskal-Wallis Rank Sum Test

- ▶ alternative to ANOVA (for non-normal distributed data)
- ▶ Assumptions
 - ▶ variance homogeneity
 - ▶ independent data
 - ▶ at least ordinal scaled data
- ▶ function:
`kruskal.test()`

Two-Sample comparisons - non-parametric tests

```
# a) Two sample Wilcoxon Rank Sum Test (= Mann-Whitney U-
    Test)
# H0: average growth rates are equal between sexes
# H1: average growth rates differ between sexes
wilcox.test(formula, alternative="two.sided",
             paired=F, var.equal=T, data=Data, conf.int=T)
# if p-value < 0.05: reject H0, else: keep H0

# b) Kruskal-Wallis Rank Sum Test (One-Way non-parametric
    ANOVA)
# H0: average growth rates are equal between sexes
kruskal.test(formula, data=Data)
# if Pr(>F) < 0.05: reject H0, else: keep H0
```

Kruskal-Wallis rank sum test

data: growth.rate by sex

Kruskal-Wallis chi-squared = 0.0833, df = 1, p-value = 0.7728

Exercise

A botanist tried to analyze the effects of different light regimes (“Red”, “Green”) on the growth rates of soybeans.

- ▶ Import the data from the soybean.xls file (datasheet: soybean1)
- ▶ Define a hypothesis
- ▶ Choose an appropriate test (one vs. two sided; parametric vs. non-parametric)

Two-Sample comparisons

```
# load dataframe
file <- "soybean1.csv"
datasheet <- read.table(file, header=T, sep=',', dec=".")
head(datasheet)
attach(datasheet)
formula = Height~Light

boxplot(formula, ylab="Height", xlab="Light")

# Shapiro-Wilk test for Normal Distribution
# better for small sampling sizes (<50)
# H0: normal distributed data
shapiro.test(subset(Height, Light == "Green"))
shapiro.test(subset(Height, Light == "Red"))
# if p-value < 0.05: reject H0, else: keep H0

# Levene's test for homogeneity of variance across groups
# H0: variances are equal (homogeneity of variance)
# H1: variances differ
install.packages('car') # install required package
library(car)           # load package
leveneTest(formula, data=datasheet)
# if Pr(>F) < 0.05: reject H0, else: keep H0
```

Two-Sample comparisons

```
wilcox.test(formula, alternative="two.sided",  
            paired=F, var.equal=T, data=Data, conf.int=T)
```

Wilcoxon rank sum test with continuity correction

data: Height by Light

W = 272, p-value = 0.1302

alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:

-0.1999954 1.5000204

sample estimates:

difference in location

0.7000338

Warning messages: 1: In wilcox.test.default(x = c(8.6, 5.9, 4.6,
9.1, 9.8, 10.1, 6, 10.4, :

cannot compute exact p-value with ties

2: In wilcox.test.default(x = c(8.6, 5.9, 4.6,
9.1, 9.8, 10.1, 6, 10.4, :

cannot compute exact confidence intervals with ties

Two-Sample comparisons

```
install.packages('exactRankTests')  
library(exactRankTests)  
  
wilcox.exact(formula, alternative="t",  
              paired=F, var.equal=T, data=Data, conf.int=T)
```

```
Exact Wilcoxon rank sum test  
data: Height by Light W = 272, p-value = 0.1296  
alternative hypothesis: true mu is not equal to 0  
95 percent confidence interval:  
 -0.2  1.6  
sample estimates:  
difference in location  
          0.75
```

Three-Sample comparisons

```
# load dataframe
file <- "soybean2.csv"
datasheet <- read.table(file, header=T, sep=',', dec=".")
head(datasheet)
attach(datasheet)
formula = Height~Light

boxplot(formula, ylab="Height", xlab="Light")

# Shapiro-Wilk test for Normal Distribution
# better for small sampling sizes (<50)
# H0: normal distributed data
shapiro.test(subset(Height, Light == "Green"))
shapiro.test(subset(Height, Light == "Red"))
# if p-value < 0.05: reject H0, else: keep H0

# Levene's test for homogeneity of variance across groups
# H0: variances are equal (homogeneity of variance)
# H1: variances differ
install.packages('car') # install required package
library(car)           # load package
leveneTest(formula, data=datasheet)
# if Pr(>F) < 0.05: reject H0, else: keep H0
```


Three-Sample comparisons

```
# Kruskal-Wallis Rank Sum Test (non-parametric ANOVA)
# H0: average growth rates are equal between sexes
# option 1:
kruskal.test(formula, data=datasheet)

# option 2:
Red <- subset(Height, Light == "Red")
Green <- subset(Height, Light == "Green")
Yellow <- subset(Height, Light == "Yellow")

kruskal.test(list(Red, Green, Yellow))
# if  $\Pr(>F) < 0.05$ : reject H0, else: keep H0
```

Three-Sample comparisons

```
# post hoc tests
# non parametric tests

# method a)
wilcox.test(Red, Green)
wilcox.test(Red, Yellow)
wilcox.test(Yellow, Green)

pvalues <- rep(NA, 3)
pvalues[1] <- wilcox.test(Red, Green)$p.value
pvalues[2] <- wilcox.test(Yellow, Green)$p.value
pvalues[3] <- wilcox.test(Yellow, Red)$p.value

p.adjust(pvalues, method="bonf")

# method b)
pairwise.wilcox.test(Height, Light, p.adj = "bonf")

# ties: One or more equal values or sets of equal values in
       the data set.
```

Three-Sample comparisons

```
# method c)
pvalues <- rep(NA, 3)
pvalues[1] <- wilcox.exact(Red, Green, paired=F)$p.value
pvalues[2] <- wilcox.exact(Yellow, Green)$p.value
pvalues[3] <- wilcox.exact(Yellow, Red)$p.value

p.adjust(pvalues, method="bonf")

# further options:
# wilcox_test from the "coin"-package
```

Three-Sample comparisons

```
# recall options for parametric post hoc tests
ANOVA <- aov(formula, data=datasheet)
summary(ANOVA)

# option 1: Tukey comparisons
# method a)
library(multcomp)
TUKEY <- glht(ANOVA, linfct=mcp(Light="Tukey"),
              interaction_average=TRUE)
summary(TUKEY)

# method b)
TukeyHSD(ANOVA)

# option 2: pairwise t-tests
pairwise.t.test(Height, Light, p.adj = "bonf")
```

One-Way repeated measures ANOVA (rmANOVA)

parametric:

```
summary(aov(values~factor + Error(subject/factor))
```

non-parametric:

```
friedman.test(values~factor | subject)
```

Advantage:

Unlike the common, randomized ANOVA, the rmANOVA accounts for variability due to subjects (blocks)!

Two-Sample comparisons

```
# repeated measures analyses
# Example 1: Sea urchins

setwd('~/.Dropbox/R_course/day6') # set working directory
file <- "Seeigel_abhaengigeStichproben.csv"
datasheet <- read.csv(file, header=T, dec=".", sep=',')
head(datasheet)

Data <- data.frame(datasheet$grass_t0, datasheet$grass_1
  month)
Data <- stack(Data) # rearrange dataframe

# comparing two repeated samplings: paired t-test
formula = values~ind
t.test(formula, alternative="two.sided",
  paired=T, var.equal=T, data=Data)

wilcox.exact(formula, alternative="two.sided",
  paired=T, var.equal=T, data=Data, conf.int=T)
```

Three-Sample comparisons

```
# comparing three samples: repeated measures ANOVA
Data2 <- data.frame(datasheet$grass_t0, datasheet$grass_1
  month, datasheet$grass_1year)
Data2 <- stack(Data2) # rearrange dataframe

attach(Data2)
Gebiet <- rep(1:length(datasheet$Gebiet),3)

# a) parametric repeated measures ANOVA
ANOVA.rep <- aov(values~ind + Error(Gebiet/ind))
summary(ANOVA.rep)

pairwise.t.test(values, ind, p.adj = "bonf", paired=T)

# b) non parametric repeated measures ANOVA
friedman.test(values ~ ind | Gebiet)

pairwise.wilcox.test(values, ind, p.adj = "bonf", paired=T)
```

Exercise

The coverage of corals of a lagoon was analyzed along 50 transects during 9 subsequent years. Did the coverage change over time?

- ▶ Import the data from the Coral_covearge.xls file
- ▶ Define a hypothesis
- ▶ Choose an appropriate test (one vs. two sided; parametric vs. non-parametric)

Multiple-Sample comparisons

```
# repeated measures analyses
# Exercise: Coral Coverage

setwd('~/.Dropbox/R_course/day6') # set working directory
file <- "Coral_Coverage.csv"
datasheet <- read.table(file, header=T, sep=',', dec=".") #
  load dataframe

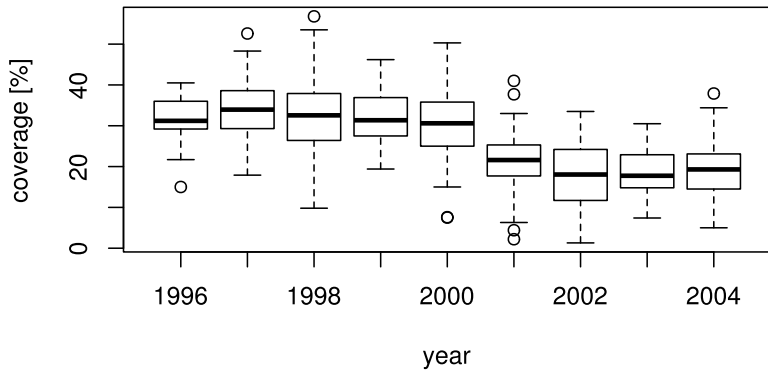
head(datasheet)
datasheet <- stack(datasheet)
head(datasheet)

transect <- rep(1:50,9)

attach(datasheet)

boxplot(values~ind, data=datasheet, names=1996:2004,
  ylab="coverage [%]", xlab="year")
```

Multiple-Sample comparisons



Multiple-Sample comparisons

```
# a) parametric repeated measures ANOVA
ANOVA.rep <- aov(values ~ ind + Error(transect/ind), data=
  datasheet)
summary(ANOVA.rep)

pairwise.t.test(values, ind, p.adj = "bonf", paired=T, data=
  datasheet)

# b) non parametric repeated measures ANOVA
friedman.test(values ~ ind | transect, data=datasheet)

pairwise.wilcox.test(values, ind, p.adj = "bonf", paired=T,
  data=datasheet)
```

Multiple-Sample comparisons

```
library(multcomp)
values <- datasheet$values
groups <- factor(datasheet$ind)
group.levels <- levels(groups)

n <- 1:length(group.levels)
names(n) <- group.levels

contrasts <- contrMat(n, type = "Sequen") # Sequen contrasts
pvalues <- rep(NA, length(contrasts[,1])) # creates vector
      for pvalue assignment

for (i in 1:length(pvalues))
{
  x <- subset(values, groups == group.levels[which(contrasts
    [i,] == 1)])
  y <- subset(values, groups == group.levels[which(contrasts
    [i,] == -1)])
  post.hoc.test.results <- t.test(y,x,alternative="two.sided",
    paired=T)
  pvalues[i] <- as.numeric(post.hoc.test.results$p.value)
}

pvalues <- p.adjust(pvalues, method="bonf")
cbind(contrasts, pvalues)
```