Statistics with R – Advanced Level

Section 1

Mean Difference

Lesson 1 – ANCOVA

```
vit = read.csv("vitamin-a.csv")

View(vit)

#########
### how to perform the one-way analysis of covariance
(ANCOVA)
#########

#########
### Basic assumptions:

# the response variable do not present outliers
# the relationship between the dependent variable and the
covariate is linear
# there is no relationship between the covariate and the
factor*
# the residuals of the response variable are normally
distributed*
# there is homogeneity of variances*
# there is homoskedasticity*

### we will check only the assumptions marked with an
asterisk (*)
```

```
##########

### dependent variable: effort resistance
### factor: dose of vitamin
### covariate: employees' age

####################

### how to run the ANCOVA, get the adjusted means
### and do multiple comparisons of the adjusted means

### load the car package
### so we can compute the type III sum of squares

require(car)

### run the ANCOVA

model <- aov(effort~dose+age, data=vit)
ancova <- Anova(model, type="III")
print(ancova)

### compute the adjusted means

require(effects)

effect("dose", model)

#### perform multiple comparisons between adjusted means

require(multcomp)

mcomp <- glht(model, linfct=mcp(dose="Tukey"))

### linfct (linear function) specifies the hypotheses to be
tested
### here we use the mcp (multiple comparisons) specifying
the option Tukey

### get the differences and their statistical significance

summary(mcomp)

### get the confidence interval for the differences
```

```
confint(mcomp)
```

Lesson 2 - ANCOVA - checking assumptions

```
vit = read.csv("vitamin-a.csv")

View(vit)

#########
### the analysis of covariance - checking the assumptions
#########

#########
### Basic assumptions:

# the response variable do not present outliers
# the relationship between the dependent variable and the
covariate is linear
# there is no relationship between the covariate and the
factor*
# the residuals of the response variable are normally
distributed*
# there is homogeneity of variances*
# there is homoskedasticity*

### we will check only the assumptions marked with an
asterisk (*)
##########

##########  check for the homogeneity of regression slopes
########## i.e. independence between factor and covariate

### we compute the interaction between age and dose of
vitamin

model <- aov(effort~age*dose, data=vit)
av <- Anova(model, type="III")
print(av)

############ check for the normality of residuals

### get the residuals and standardize them
```

```
res <- residuals(model)

zres <- scale(res)

shapiro.test(zres)

########## check for the homogeneity of variances

require(car)

leveneTest(vit$effort, vit$dose)

########## check for homoskedasticity

### get the predicted values of the dependent variable

pred <- predict(model)

### build the scatterplot (predicted vs. residuals)

require(ggplot2)

ggplot()+geom_point(aes(x=pred, y=zres))
```

Lesson 3 - Within-subjects ANOVA

```
diet <- read.csv("diet1.csv")

View(diet)

#########
### the within-subjects (repeated measures) analysis of
variance
#########

#########
### Basic assumptions:

# the variables are approximately normally distributed
# the variables do not present significant outliers
# there is sphericity*
```

```
### we will check only the assumptions marked with an
asterisk (*)
##########

### we will determine whether there is a significant
difference
### between the average subjects' weights at the three diet
moments:
### beginning, middle, end

### dependent variable: weight (measured three times)
### factor: time

###### running the within-subjects ANOVA supposes several
steps

### build a matrix and a dataframe with the factor levels

moments_mat <- c("beginning", "middle", "end")

print(moments_mat)

moments_frm <- data.frame(moments_mat)

View(moments_frm)

### build a matrix with the values of the measure (weight)

weight_mat <- cbind(diet$weight_beg, diet$weight_mid,
diet$weight_end)

print(weight_mat)

#### get the means of the groups (these will be compared
through the ANOVA)

model <- lm(weight_mat~1)

summary(model)

### now do the within-subjects analysis

require(car)
```

```
model2 <- Anova(model, idata = moments_frm, idesign =
~moments_mat, type="III")

### the options idata and idesign are used to define the
factor levels
### in the repeated-measure analyses

summary(model2, multivariate=F)

### the option multivariate = F prevents the display of the
MANOVA results
```

Lesson 4 - Within-subjects ANOVA - paired comparisons

```
diet <- read.csv("diet1.csv")

View(diet)

#########
### the within-subjects analysis of variance - multiple
comparisons
#########

### to perform the multiple (paired) comparisons
### we must reshape the data frame first
### (put it in the "long data" format)

require(reshape2)

dietm <- melt(diet)

View(dietm)

### give the columns some suggestive names

colnames(dietm) <- c("group", "weight")

### build an ANOVA model

model <- aov(weight~group, data=dietm)

### perform the Tukey test
```

```
TukeyHSD(model)

### perform the Bonferroni paired comparisons

pairwise.t.test(dietm$weight, dietm$group, p.adjust.method
= "bonferroni")
```

Lesson 5 - Within-within subjects ANOVA

```
diet <- read.csv("diet2.csv")

View(diet)

#########
### the within-within-subjects analysis of variance
#########

#########
### Basic assumptions:

# the variables are approximately normally distributed
# the variables do not present significant outliers
# there is sphericity*

### we will check only the assumptions marked with an
asterisk (*)
##########

### dependent variable: weight
### factors: time (beginning, middle, end) and physical
exercises (with and without exercises)

### first you must prepare a data frame with the combined
factor levels
### like this one:

fact <- read.csv("factors-within-within.csv")

View(fact)

### create a matrix with all the dependent variables
```

```
weight <- cbind(diet$weight_beg, diet$weight_mid, 
diet$weight_end, 
             diet$weight_beg_ex, diet$weight_mid_ex, 
diet$weight_end_ex)

### get the means of the dependent variables

model <- lm(weight~1)
summary(model)

### run the ANOVA

require(car)

model2 <- Anova(model, idata=fact, idesign=~Exercise*Time, 
type="III")

summary(model2, multivariate=F)   ## we do not need the 
MANOVA results

### Exercise and Time are the variable names in the fact 
data frame

### since the interaction effect is significant, we must 
compute
### the simple main effects of the factors time and 
exercise
```

Lesson 6 - Within-within subjects ANOVA - main effects (1)

```
diet = read.csv("diet2.csv")

View(diet)

#########
### the simple main effects of the factor exercise
#########

### the simple main effects of the factor exercise 
represent the effects
### of this factor at every level of the factor time, i.e
### at every moment of the diet: beginning, middle, end
```

```
### concretely, they consist of three differences

### weight with physical exercises - weight without physical exercises, at the beginning of the diet
### weight with physical exercises - weight without physical exercises, in the middle of the diet
### weight with physical exercises - weight without physical exercises, at the end of the diet

### we will evaluate the first difference only (at the beginning of the diet)

### from the diet data frame, we extract the columns we need

diet2 <- diet[,c("weight_beg", "weight_beg_ex")]

View(diet2)

### create the dataframe with the levels of the factor exercise

xr <- c("no", "yes")

xr_frm <- data.frame(xr)

View(xr_frm)

### create a matrix with the columns of the data frame diet2

xr_mat <- cbind(diet2$weight_beg, diet2$weight_beg_ex)

### create the linear model to get the means of the dependent variables

model <- lm(xr_mat~1)

#### create the within-subjects model

require(car)

model2 <- Anova(model, idata=xr_frm, idesign=~xr, type="III")
```

```
summary(model2, multivariate=F)


######## get the simple Tukey comparisons
######## to find out how big the difference is

### reshape the initial data set

require(reshape2)

dietm <- melt(diet2)

View(dietm)

### give the columns some suggestive names

colnames(dietm) <- c("group", "weight")

### build an ANOVA model

model3 <- aov(weight~group, data=dietm)

### compute the paired comparison tests

TukeyHSD(model3)

### the same procedure is to be applied for the other two
differences
```

Lesson 7 - Within-within subjects ANOVA - main effects (2)

```
diet = read.csv("diet2.csv")

View(diet)

#########
### the simple main effects of the factor time
#########

### the simple main effects of the factor time represent
the effects
```

```
### of this factor at every level of the factor exercise, i.e
### with and without physical exercises

### concretely, they consist of two differences

### the difference between the weight at the beginning, in the middle and at the end of the diet, WITHOUT exercises
### the difference between the weight at the beginning, in the middle and at the end of the diet, WITH exercises

### we already evaluated the first set of differences, in the lecture about within-subjects ANOVA

### now we will evaluate the second set

### build a dataframe with the levels of the factor time

moments <- c("beginning", "middle", "end")

moments_frm <- data.frame(moments)

View(moments_frm)

### build a matrix with the values of the measure (weight)

moments_mat <- cbind(diet$weight_beg_ex, diet$weight_mid_ex, diet$weight_end_ex)

#### get the means of the dependent variables

model <- lm(moments_mat~1)

### now do the within-subjects analysis

require(car)

model2 <- Anova(model, idata = moments_frm, idesign = ~moments, type="III")

summary(model2, multivariate=F)


############# get the Tukey pairwise comparisons
```

```
############# to see how big the differences are

### from the data frame diet, extract the columns we need

diet2 <- diet[,c("weight_beg_ex", "weight_mid_ex",
"weight_end_ex")]

### reshape the new data frame

dietm <- melt(diet2)

View(dietm)

### give the columns some suggestive names

colnames(dietm) <- c("group", "weight")

### build an ANOVA model

model3 <- aov(weight~group, data=dietm)

### compute the paired comparison tests

TukeyHSD(model3)
```

Lesson 8 - Mixed ANOVA

```
diet <- read.csv("diet3.csv")

View(diet)

#########
### the mixed analysis of variance
#########

#########
### Basic assumptions:

# the dependent variables are normally distributed
# the dependent variables do not present outliers
# there is homogeneity of variances (for the between-
subjects factor)*
```

```
# there is homogeneity of covariances (for the between-
subjects factor)*
# there is sphericity (for the within-subjects factor)*

### we will check only the assumptions marked with an
asterisk (*)
##########

### we will determine whether there is a significant
difference in average weight
### between the three moments of the diet, for both male
and female subjects

### within-subjects factor: time (beginning, middle, end)
### between-subjects factor: gender (male, female)

###############

###### check the assumption of equal variances (for each
dependent variable)

require(car)

leveneTest(diet$weight_beg, diet$gender)

leveneTest(diet$weight_mid, diet$gender)

leveneTest(diet$weight_end, diet$gender)

####### check the assumption of equal covariances (Box's M
test)

require(biotools)

### from the diet data frame, extract the dependent
variables

diet2 <- diet[c(2,3,4)]

View(diet2)

boxM(diet2, diet$gender)

################# get to the ANOVA
```

```
### prepare and load a new data frame, with all the
combinations of the factors

fact <- read.csv("factors-mixed.csv")

View(fact)

### create a new data frame with the male subjects only

dietm <- diet[diet$gender=="male",]

View(dietm)

### rename the columns conveniently

colnames(dietm) <- c("gender", "weight_beg_male",
"weight_mid_male", "weight_end_male")

### create a new data frame with the female subjects only

dietf <- diet[diet$gender=="female",]

View(dietf)

### rename the columns

colnames(dietf) <- c("gender", "weight_beg_female",
"weight_mid_female", "weight_end_female")

### create a matrix with all the dependent variables

weight <- cbind(dietm$weight_beg_male,
dietm$weight_mid_male,
              dietm$weight_end_male,
dietf$weight_beg_female, dietf
              $weight_mid_female, dietf$weight_end_female)

View(weight)

### get the means of all the dependent variables

model <- lm(weight~1)
```

```
summary(model)

### finally, create the ANOVA model

model2 <- Anova(model, idata=fact,
idesign=~time+gender*time, type="III")

summary(model2, multivariate=F)    ## we don't want the
MANOVA results

### gender and time are the variables of the data frame
fact

### since the interaction effect is statistically
significant,
### we are going to compute the simple main effects of the
factors
```

Lesson 9 - Mixed ANOVA - main effects

```
diet = read.csv("diet3.csv")

View(diet)

#########
### the mixed analysis of variance - simple main effects
#########

### the simple main effects of the variable time represent
### the mean differences of weight between the three
moments of the diet
### for each gender separately

### to compute them, we will run a within-subjects ANOVA
for each gender category

######## for the male subjects

### create a new data frame with the male subjects only

dietm <- diet[diet$gender=="male",]

### build a dataframe with the levels of the factor time
```

```
moments <- c("beginning", "middle", "end")

moments_frm <- data.frame(moments)

View(moments_frm)

### build a matrix with the values of the measure (weight)

weight_male <- cbind(dietm$weight_beg, dietm$weight_mid,
dietm$weight_end)

#### get the means of the dependent variables

model <- lm(weight_male~1)

### now do the within-subjects analysis

model2 <- Anova(model, idata = moments_frm, idesign =
~moments, type="III")

summary(model2, multivariate=F)

### the same procedure will be used for the female subjects

#######################

### the simple main effects of the variable gender
represent
### the mean differences of weight between the male and
female subjects
### for each moment of the diet: beginning, middle, end

### they consist of three pairs of differences

### average male weight - average female weight, at the
beginning of the diet
### average male weight - average female weight, in the
middle of the diet
### average male weight - average female weight, at the end
of the diet

### we will evaluate these differences using the
independent sample t test
```

```
### the first difference (beginning of the diet)

t.test(diet$weight_beg~diet$gender, var.equal=T)

### the second difference (middle of the diet)

t.test(diet$weight_mid~diet$gender, var.equal=T)

### the third difference (end of the diet)

t.test(diet$weight_end~diet$gender, var.equal=T)
```

Lesson 10 - Friedman test

```
diet = read.csv("diet1.csv")

View(diet)

#########
### the Friedman test
#########

### we will compare the median weights at the three moments
of the diet
### using the Friedman test

### create a matrix from the dataframe

weight <- cbind(diet$weight_beg, diet$weight_mid,
diet$weight_end)

#### apply the friedman.test function to the matrix

friedman.test(weight)
```