R version 2.12.2 (2011-02-25)

Copyright (C) 2011 The R Foundation for Statistical Computing

ISBN 3-900051-07-0

Platform: i386-pc-mingw32/i386 (32-bit)

….

> # CHAPTER 7, DALGAARD – ANOVA!

#-------------------------------------------------

> ls() #my workspace is empty

character(0)

> attach(red.cell.folate)

Error in attach(red.cell.folate) : object 'red.cell.folate' not found

> search() #ISwR not there!

[1] ".GlobalEnv" "package:stats" "package:graphics"

[4] "package:grDevices" "package:utils" "package:datasets"

[7] "package:methods" "Autoloads" "package:base"

> # Package-> Load package # load ISwR

> attach(red.cell.folate)

> summary(red.cell.folate) # always look at the data first – what do you see?

 folate ventilation

 Min. :206.0 N2O+O2,24h:8

 1st Qu.:249.5 N2O+O2,op :9

 Median :274.0 O2,24h :5

 Mean :283.2

 3rd Qu.:305.5

 Max. :392.0

> # ventilation-a categorical variable!

> # folate - continuous variable

>

> anova(lm(folate~ventilation)) # is there a difference between groups?

Analysis of Variance Table

Response: folate

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

ventilation 2 15516 7757.9 3.7113 0.04359 \*

Residuals 19 39716 2090.3

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> # simple anova is performed in R with the function lm

> # BUT: independent variables must be categorical (factors)

> red.cell.folate #are they categorical?

 folate ventilation

1 243 N2O+O2,24h

2 251 N2O+O2,24h

3 275 N2O+O2,24h

4 291 N2O+O2,24h

5 347 N2O+O2,24h

6 354 N2O+O2,24h

7 380 N2O+O2,24h

8 392 N2O+O2,24h

9 206 N2O+O2,op

10 210 N2O+O2,op

11 226 N2O+O2,op

12 249 N2O+O2,op

13 255 N2O+O2,op

14 273 N2O+O2,op

15 285 N2O+O2,op

16 295 N2O+O2,op

17 309 N2O+O2,op

18 241 O2,24h

19 258 O2,24h

20 270 O2,24h

21 293 O2,24h

22 328 O2,24h

>

> class(red.cell.folate)

[1] "data.frame"

>

>

> anova(lm(folate~ventilation)) #extract the anova table from the model with anova()

Analysis of Variance Table

Response: folate

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

ventilation 2 15516 7757.9 3.7113 0.04359 \*

Residuals 19 39716 2090.3

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>

> #what does this tell us? what is H0?

> #variation between groups is where?

> #variation within groups ?

> # the dependent variable is known as the grouping factor

>

> #what happens if we try anova with independent vars that are not categorical?

>

> attach(juul)

> anova(lm(igf1~tanner)) ## WRONG! describes a linear regression!!!

Analysis of Variance Table

Response: igf1

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

tanner 1 10985605 10985605 686.07 < 2.2e-16 \*\*\*

Residuals 790 12649728 16012

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>

> #so, you need to fix it by categorizing the independent variable...

>

> juul$tanner <- factor(juul$tanner,

+ labels=c("I","II","III","IV","V"))

> detach(juul)

> attach(juul)

> summary(tanner)

 I II III IV V NA's

 515 103 72 81 328 240

> anova(lm(igf1~tanner)) # NOW OK!

Analysis of Variance Table

Response: igf1

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

tanner 4 12696217 3174054 228.35 < 2.2e-16 \*\*\*

Residuals 787 10939116 13900

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> #WHAT IS DIFFERENT?

|  |  |
| --- | --- |
| anova(lm(igf1~tanner)) ## WRONG! Analysis of Variance TableResponse: igf1 Df Sum Sq Mean Sq F value Pr(>F) tanner 1 10985605 10985605 686.07 < 2.2e-16 \*\*\*Residuals 790 12649728 16012  | > summary(tanner) I II III IV V NA's  515 103 72 81 328 240 > anova(lm(igf1~tanner)) # NOW OK!Analysis of Variance TableResponse: igf1 Df Sum Sq Mean Sq F value Pr(>F) tanner 4 12696217 3174054 228.35 < 2.2e-16 \*\*\*Residuals 787 10939116 13900   |

>

> #back to analysis of variance of red.cell.folate data

> # anova showed us that there is a difference between groups,

> # but where is that difference between all groups, or just some?

>

> #use the pairwise t-test

> anova(lm(folate~ventilation))

Analysis of Variance Table

Response: folate

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

ventilation 2 15516 7757.9 3.7113 0.04359 \* # H0: groups have the same mean.

 # H0 can be rejected

Residuals 19 39716 2090.3

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(lm(folate~ventilation))

Call:

lm(formula = folate ~ ventilation)

Residuals:

 Min 1Q Median 3Q Max

-73.625 -35.361 -4.444 35.625 75.375

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 316.63 16.16 19.588 4.65e-14 \*\*\*

ventilationN2O+O2,op -60.18 22.22 -2.709 0.0139 \*

ventilationO2,24h -38.63 26.06 -1.482 0.1548

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45.72 on 19 degrees of freedom

Multiple R-squared: 0.2809, Adjusted R-squared: 0.2052

F-statistic: 3.711 on 2 and 19 DF, p-value: 0.04359

> # intercept is the mean in group I (N2O+O2, 24h)

> # next two lines show difference between that group and group I

> # generally, the first group is treated as a baseline (or control)

>

> # H0: groups have the same mean.

> # H0 can be rejected for Group I and II (N2O+O2,op), p = 0.14

> # H0 not rejected for Group I and III (O2,24h), p = .15

>

> # to compare all groups against each other

|  |  |
| --- | --- |
| > pairwise.t.test(folate, ventilation, + p.adj="bonferroni")Pairwise comparisons using t tests with pooled SD data: folate and ventilation  N2O+O2,24h N2O+O2,opN2O+O2,op 0.042 - O2,24h 0.464 1.000 P value adjustment method: bonferroni  | >pairwise.t.test(folate,ventilation)  #default pooled SD!Pairwise comparisons using t tests with pooled SD data: folate and ventilation  N2O+O2,24h N2O+O2,opN2O+O2,op 0.042 - O2,24h 0.310 0.408 P value adjustment method: holm >  |

> #ch. 7 suggests you use oneway.test(folate~ventilation)

|  |  |
| --- | --- |
| > pairwise.t.test(folate,ventilation,pool.sd=F)Pairwise comparisons using t tests with non-pooled SD data: folate and ventilation  N2O+O2,24h N2O+O2,opN2O+O2,op 0.087 - O2,24h 0.321 0.321 P value adjustment method: holm  | > oneway.test(folate~ventilation)One-way analysis of means (not assuming equal  variances)data: folate and ventilation F = 2.9704, num df = 2.000, denom df = 11.065, p-value = 0.09277 |

> # to graphically represent the data....

> # here you need to understand why SEM (standard error of he means is relevant

>

> xbar <- tapply(folate, ventilation, mean)

> s <- tapply(folate, ventilation, sd) #what does tapply() do?

> s <- tapply(folate, ventilation, sd)

> n <- tapply(folate, ventilation, length)

> s

N2O+O2,24h N2O+O2,op O2,24h

 58.71709 37.12180 33.75648

> n

N2O+O2,24h N2O+O2,op O2,24h

 8 9 5

> sem <- s/sqrt(n)

> sem

N2O+O2,24h N2O+O2,op O2,24h

 20.75963 12.37393 15.09636

> stripchart(folate~ventilation, method="jitter", jitter=0.05, pch=16, vert=T)

> arrows(1:3,xbar+sem,1:3,xbar-sem,angle=90,code=3,length=.1)

> lines(1:3,xbar,pch=4,type="b",cex=2) #see how these three commands build the graph….

>

#how to determine if the distribution of a variable has the same variance in all groups?

#---------------------------------------------------------------------------------------------------------------

> # it is assumed the data are from independent groups

> bartlett.test(folate~ventilation)

 Bartlett test of homogeneity of variances

data: folate by ventilation

Bartlett's K-squared = 2.0951, df = 2, p-value = 0.3508

> # in this case, nothing in the data contradicts assumption of equal variance

>

> #----------------------------------------

> #kruskall wallis is a nonparametric counterpart of one-way anova

> kruskal.test(folate~ventilation)

 Kruskal-Wallis rank sum test

data: folate by ventilation

Kruskal-Wallis chi-squared = 4.1852, df = 2, p-value =

0.1234

> # it uses ranks, not values....

>

> #----------------------------------------------

>

> # two-way anova

>

> #----------------------------------------------

>

> #an important assumption here is that the cross-classified design is **balanced**

> # it is sufficient (though not necessary) that the cell counts be equal

> # there are other balanced designs, see gotelli....

>

> #often, this is where there are multiple measurements on the same experimental unit

>

> # data should be in one vector, with two classifying factors

> attach(heart.rate)

> heart.rate

|  |  |  |  |
| --- | --- | --- | --- |
|  hr subj time1 96 1 02 110 2 03 89 3 04 95 4 05 128 5 06 100 6 07 72 7 08 79 8 09 100 9 0 | 10 92 1 3011 106 2 3012 86 3 3013 78 4 3014 124 5 3015 98 6 3016 68 7 3017 75 8 3018 106 9 30 | 19 86 1 6020 108 2 6021 85 3 6022 78 4 6023 118 5 6024 100 6 6025 67 7 6026 74 8 6027 104 9 60 | 28 92 1 12029 114 2 12030 83 3 12031 83 4 12032 118 5 12033 94 6 12034 71 7 12035 74 8 12036 102 9 120 |

> # so, what do these data represent?

> # how was this defined in R?

> my = data.frame(hr = c(1,2,3,4,5,6), mys=gl(3,1,6), myt=gl(2,3,6,labels=c(0,30)))

 -------------- ------- -------

> my

 hr mys myt

1 1 1 0

2 2 2 0

3 3 3 0

4 4 1 30

5 5 2 30

6 6 3 30

gl(help)

Generate factors by specifying the pattern of their levels.

gl(n, k, length = n\*k, labels = 1:n, ordered = FALSE)

|  |  |
| --- | --- |
| n | an integer giving the number of levels. |
| k | an integer giving the number of replications. |
| length | an integer giving the length of the result. |
| labels | an optional vector of labels for the resulting factor levels. |

> gl(9,1,36)

 [1] 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9 1 2

[30] 3 4 5 6 7 8 9

Levels: 1 2 3 4 5 6 7 8 9

>

> gl(4,9,36,labels=c(0,30,60,120))

 [1] 0 0 0 0 0 0 0 0 0 30 30 30 30 30

[15] 30 30 30 30 60 60 60 60 60 60 60 60 60 120

[29] 120 120 120 120 120 120 120 120

Levels: 0 30 60 120

> length(hr)

[1] 36

> my.heart.rate <- data.frame(my.hr = c(96, 110, 89, 95, 128, 100, 72,

+ 79, 100, 92, 106, 86, 78, 124, 98, 68, 75, 106, 86, 108, 85,

+ 78, 118, 100, 67, 74, 104, 92, 114, 83, 83, 118, 94,

+ 71, 74, 102), my.subj = gl(9,1,36),

+ my.time=gl(4,9,36,labels=c(0,30,60,120)) )

> my.heart.rate

 my.hr my.subj my.time

1 96 1 0

2 110 2 0

3 89 3 0

4 95 4 0

5 128 5 0

6 100 6 0

7 72 7 0

8 79 8 0

9 100 9 0

10 92 1 30

11 106 2 30

12 86 3 30

13 78 4 30

14 124 5 30

15 98 6 30

16 68 7 30

17 75 8 30

18 106 9 30

19 86 1 60

20 108 2 60

21 85 3 60

22 78 4 60

23 118 5 60

24 100 6 60

25 67 7 60

26 74 8 60

27 104 9 60

28 92 1 120

29 114 2 120

30 83 3 120

31 83 4 120

32 118 5 120

33 94 6 120

34 71 7 120

35 74 8 120

36 102 9 120

>

>

>

> anova(lm(hr~subj+time)) #similar to multiple regression -

Analysis of Variance Table

Response: hr

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

subj 8 8966.6 1120.82 90.6391 4.863e-16 \*\*\*

time 3 151.0 50.32 4.0696 0.01802 \*

Residuals 24 296.8 12.37

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>

> interaction.plot(time, subj, hr) #look at the individuals!

> interaction.plot(ordered(time), subj, hr) #look at the individuals!

>

>

> #anova table in regression analysis - what does it tell us?

> #according to dalgaard, anova is but one class of linear model....

> # variation between and within groups for a one-way anova

# generalizes to model variation and residual variation

> # BUT only when the model contains an intercept....

>

>

> attach(thuesen)

> lm.velo <- lm(short.velocity~blood.glucose)

> summary(lm.velo)

Call:

lm(formula = short.velocity ~ blood.glucose)

Residuals:

 Min 1Q Median 3Q Max

-0.40141 -0.14760 -0.02202 0.03001 0.43490

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.09781 0.11748 9.345 6.26e-09 \*\*\*

blood.glucose 0.02196 0.01045 2.101 0.0479 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2167 on 21 degrees of freedom

 (1 observation deleted due to missingness)

Multiple R-squared: 0.1737, Adjusted R-squared: 0.1343

F-statistic: 4.414 on 1 and 21 DF, p-value: 0.0479

> anova(lm.velo)

Analysis of Variance Table

Response: short.velocity

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

blood.glucose 1 0.20727 0.207269 4.414 0.0479 \*

Residuals 21 0.98610 0.046957

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>

> #The F-test gives the same p-value as the t-test for a zero slope.

> # see dalgaard for more details....