Unit 6 – Analysis of Variance
Practice Problems (1 of 2)

Solutions

Before you begin. Download from the course website
anova_infants.xlsx

Zelazo et al. (1972) studied the variability in age at first walking in infants. 24 infants were randomly assigned to four groups of equal sample size (6 infants per group), with groups defined by method of reinforcement of walking: (1) active (2) passive (3) no exercise; and (4) 8 week control. The outcome variable measured was age at first walking, in months. The following table lists the study data, by group.

Table – Study Data of Zelazo et al (1972), n=24:

Active Group	Passive Group	No-Exercise Group	8 Week Control
9.00	11.00	11.50	13.25
9.50	10.00	12.00	11.50
9.75	10.00	9.00	12.00
10.00	11.75	11.50	13.50
13.00	10.50	13.25	11.50
9.50	15.00	13.00	12.35

Source: Zelazo et al (1972) "Walking" in the newborn. *Science* 176: 314-315.

Data dictionary/Codebook:

Variable	Label	Type	Coding
group	Group	numeric	1 = active 2 = passive 3 = noex 4 = control
age	Age, months	numeric	Continuous, months
I_active	Indicator group = "active"	numeric	1 if group = 1 ("active") 0 otherwise
I_passive	Indicator group="passive"	numeric	1 if group = 2 ("passive") 0 otherwise
I_noex	Indicator group = "noex"	numeric	1 if group = 3 ("noex") 0 otherwise

#1.

Deviation from means. State the analysis of variance model using deviation from means notation μ and τ_i and σ^2 as appropriate. Define all terms and constraints on the parameters.

Answer:

$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim N(0,\sigma_\varepsilon^2)$ and $\sum_{i=1}^{4} \tau_i = 0$

i = 1, 2, ... K indexes method of reinforcement group;

K = number of groups = 4

j=1, 2, ..., n_i=6 indexes infant within group;

μ = population mean age at first walking, over all groups

μ_i = mean age at first walking for infants in group "i"

$\tau_i = [\mu_i - \mu]$

Y_{ij} = observed age at first walking for the jth infant in group "i"

H_O: $\tau_1=0$, $\tau_2=0$, $\tau_3=0$, and $\tau_4=0$

H_A: At least one $\tau_i \neq 0$

Import excel data

```
library(readxl)
infants <- read_excel("anova_infants.xlsx")
infants <- as.data.frame(infants)
str(infants)
```

```
## 'data.frame':    24 obs. of  7 variables:
##  $ group    : num  1 1 1 1 1 1 2 2 2 2 ...
##  $ age      : num  9 9.5 9.75 10 13 ...
##  $ passive  : num  0 0 0 0 0 0 1 1 1 1 ...
##  $ noex     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ I_active : num  1 1 1 1 1 1 0 0 0 0 ...
##  $ I_passive: num  0 0 0 0 0 0 1 1 1 1 ...
##  $ I_noex   : num  0 0 0 0 0 0 0 0 0 0 ...
```

Create factor variable groupf

```
library(tidyverse)

infants <- infants %>%
          mutate(groupf= recode_factor(group,
                           "1" = "active",
                           "2" = "passive",
                           "3" = "noex",
                           "4" = "control"))

table(infants$group,infants$groupf)                    # check group x groupf crosstab
```

```
##     active passive noex control
##   1      6       0    0       0
##   2      0       6    0       0
##   3      0       0    6       0
##   4      0       0    0       6
```

```
Preliminary: descriptives by group
library(summarytools)                                          # stby() in package {summarytools}

with(infants,                                                  # with(DATAFRAMENAME,
 stby(data = age,                                              # stby(data=OUTCOMEVARIABLE,
 INDICES = groupf,                                             # INDICES=GROUPVARIABLE, must be factor
 FUN = descr,
 stats = c("n.valid", "mean", "sd", "min", "med","max"),       # user chooses statistics to show.
 transpose=TRUE))

## Descriptive Statistics
## age by groupf
## Data Frame: infants
## N: 6
##
##                 N.Valid    Mean   Std.Dev     Min   Median     Max
## ------------- --------- ------- --------- ------- -------- -------
##        active      6.00   10.12      1.45    9.00     9.62   13.00
##       passive      6.00   11.38      1.90   10.00    10.75   15.00
##          noex      6.00   11.71      1.52    9.00    11.75   13.25
##       control      6.00   12.35      0.86   11.50    12.18   13.50
```

#2.

By any means you like, produce a side by side box plot showing the distribution of age at first walking, separately for each of the 4 groups.

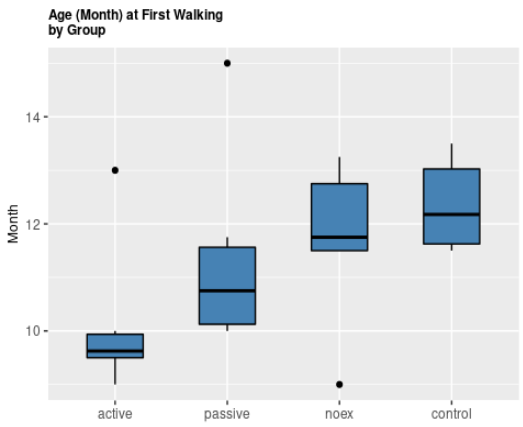
```
Q2: side-by-side plot of outcome (age) by predictor (groupf)
library(ggplot2)

ggplot(data=infants) +                                         # required: dataset
     aes(x=groupf) +                                           # required: x-axis var
     aes(y=age) +                                              # required: y-axis var
     geom_boxplot(color="black",                               # required: plot desired with optional options
                  fill="steelblue",
                  width=0.5) +

    #geom_jitter(shape = 18,                                    # Optional: overlay scatter plot.  Remove # to execute
        #color = "black",
        #position = position_jitter(0.03)) +

    ggtitle("Age (Month) at First Walking\nby Group") +   #  Optional aesthestics:  titles and axis labels
     xlab(" ") +
     ylab("Month") +

     theme(axis.text = element_text(size=9),                   # Optional aesthetics:  font size selections
           axis.title = element_text(size=9),
           plot.title = element_text(size=9, face="bold"))
```



Interpretation

- *In these data, first walking occurs earlier when infants are reinforced*
- *Distributions differ markedly with respect to variability with greatest seen among infants in the passive group and smallest among infants in the control group*

#3.

By any means you like, obtain the entries of the analysis of variance table for this one way analysis of variance. Use your computer output (or excel work or hand calculations or whatever) to complete the following table:

Source	df	Sum of Squares SSQ	Mean Square MSQ	F-Statistic	p-value
Between Groups	3	15.74	5.25	2.40	.10
Within Groups	20	43.69	2.18		
Total, corrected	23	59.43			

```
Q3: One way analysis of variance - table
fit_anova <- aov(age ~ groupf, data=infants)
anova(fit_anova)

## Analysis of Variance Table
##
## Response: age
##           Df Sum Sq Mean Sq F value  Pr(>F)
## groupf     3  15.74  5.2468  2.4018 0.09787 .
## Residuals 20  43.69  2.1845
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#4.

Write a 2-5 sentence report of your description and hypothesis test findings using language as appropriate for a client who is intelligent but is not knowledgeable about statistics. Consider including a figure and/or table that you think is appropriate.

In this sample, the data suggest a trend towards earlier age at first walking with increasing reinforcement and placement. The median age at first walking is greatest among controls (12.35 months) and lowest among infants in the "active" group (10.13 months); see also the box plots. Tests of statistical significance were limited to the overall F test for group differences and this did not achieve statistical significance (p-value = .10), possibly due to the small sample sizes (6 in each group).

Interestingly, examination of the data also suggests that the variability in age at first walking differed, depending on the intervention received. The variability was greater in the three intervention groups ("active", "passive", "no exercise") compared to in the "control" group; this was not statistically significant however (p-value = .45).

Further study, utilizing larger sample sizes and additional hypothesis tests to investigate trend are needed.

#5.

Reference cell coding Repeat your analysis, this time using what you learned in Unit 5 - Normal Theory Regression. Specifically, using appropriately defined indicator variables, perform a multivariable linear regression analysis of these same data! Use your computer output to complete the following table:

Source	df	Sum of Squares SSQ	Mean Square MSQ	F-Statistic	p-value
Due Model	3	15.74	5.25	2.40	.10
Due Error (residual)	20	43.69	2.18		
Total, corrected	23	59.43			

```
Q5: Multiple predictor regression - Modeling categorical using user created design variables
fit_lm1 <- lm(age ~ I_active + I_passive + I_noex, data=infants)      # Categorical levels modeled explicitly using 0/1
anova(fit_lm1)
## Analysis of Variance Table
##
## Response: age
##           Df Sum Sq Mean Sq F value  Pr(>F)
## I_active   1 12.793 12.7934  5.8565 0.02516 *
## I_passive  1  1.712  1.7117  0.7836 0.38656
## I_noex     1  1.235  1.2352  0.5654 0.46083
## Residuals 20 43.690  2.1845
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Q5: Multiple predictor regression - Modeling categorical as factor variable
infants$groupf <- relevel(infants$groupf, ref = "control")          # relevel() with option ref= to set reference

fit_lm2 <- lm(age ~ factor(groupf), data=infants)                   # factor( ) to model categorical levels under the hood
summary(fit_lm2)
##
## Call:
## lm(formula = age ~ factor(groupf), data = infants)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7083 -0.8500 -0.2792  0.5062  3.6250
##
## Coefficients:
##                       Estimate Std. Error t value            Pr(>|t|)
## (Intercept)            12.3500     0.6034  20.468 0.00000000000000694 ***
## factor(groupf)active   -2.2250     0.8533  -2.607              0.0169 *
## factor(groupf)passive  -0.9750     0.8533  -1.143              0.2667
## factor(groupf)noex     -0.6417     0.8533  -0.752              0.4608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.478 on 20 degrees of freedom
## Multiple R-squared:  0.2649, Adjusted R-squared:  0.1546
## F-statistic: 2.402 on 3 and 20 DF,  p-value: 0.09787

anova(fit_lm2)
## Analysis of Variance Table
##
## Response: age
##               Df Sum Sq Mean Sq F value  Pr(>F)
## factor(groupf)  3  15.74  5.2468  2.4018 0.09787 .
## Residuals      20  43.69  2.1845
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#6.

Deviation from means *and* ***reference cell coding*** **are equivalent!** Using your output from your two analyses (1st-analysis of variance, 2nd – regression), obtain the predicted mean of Y =age at first walking twice in two ways.

	Prediction Using One Way Analysis of Variance	Prediction Using Multiple Linear Regression
Active	**10.125**	$\hat{\mu}_1 = (\hat{\beta}_0 + \hat{\beta}_1) = 12.35 - 2.225 = 10.125$
Passive	**11.375**	$\hat{\mu}_2 = (\hat{\beta}_0 + \hat{\beta}_2) = 12.35 - 0.97 = 11.38$
No-Exercise	**11.71**	$\hat{\mu}_3 = (\hat{\beta}_0 + \hat{\beta}_3) = 12.35 - 0.64 = 11.71$
Control (referent)	**12.35**	$\hat{\mu}_4 = \hat{\beta}_0 = 12.35$

```
Q6: Obtain predicted means for deviation from means coding ANOVA
aactive <- predict(fit_anova,data.frame(groupf="active"))
apassive <- predict(fit_anova,data.frame(groupf="passive"))
anoex <- predict(fit_anova,data.frame(groupf="noex"))
acontrol <- predict(fit_anova,data.frame(groupf="control"))

anames <- c("Active", "Passive", "NoEx", "Control")
ahat <- c(aactive,apassive,anoex,acontrol)
means.anova <- data.frame(anames,ahat)
means.anova

##    anames     ahat
## 1  Active 10.12500
## 2 Passive 11.37500
## 3    NoEx 11.70833
## 4 Control 12.35000
```

```
Q6: Obtain predicted means from reference cell coding REGRESSION
active <- predict(fit_lm1,data.frame(I_active=1,I_passive=0,I_noex=0))
passive <- predict(fit_lm1,data.frame(I_active=0,I_passive=1,I_noex=0))
noex <- predict(fit_lm1,data.frame(I_active=0,I_passive=0,I_noex=1))
control <- predict(fit_lm1,data.frame(I_active=0,I_passive=0,I_noex=0))

names <- c("Active", "Passive", "NoEx", "Control")
yhat <- c(active,passive,noex,control)
means.regression <- data.frame(names,yhat)
means.regression

##     names     yhat
## 1  Active 10.12500
## 2 Passive 11.37500
## 3    NoEx 11.70833
## 4 Control 12.35000
```