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

Due: Thursday November 7, 2024

Last date to submit late for credit (-20 points): Thursday November 14, 2024

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.

#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.

#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					
Within Groups					
Total, corrected					

#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.

#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					
Due Error (residual)					
Total, corrected					

#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		
Passive		
No-Exercise		
Control		

Supplement - Learn R (nothing to turn in)

Practice with R to perform a one way analysis of variance

Dataset used

`hers_640anova.xlsx`

Packages used:

`ggplot2, summarytools, tidyverse, HH, car, MASS, readxl`

Introduction to The Heart and Estrogen/progestin Replacement Study (HERS)

Source

Hulley et al (1998) Randomized trial of estrogen plus progestin for secondary prevention of heart disease in postmenopausal women. The Heart and Estrogen/progestin Replacement Study. *Journal of the American Medical Association*, **280**(7), 605-613

The Heart and Estrogen/Progestin Replacement Study (HERS) was a randomized clinical trial of hormone therapy (estrogen plus progestin) for the reduction of cardiovascular disease risk in post-menopausal women with established coronary disease. Study participants were n=2,763 women who were: (1) post-menopausal (2) with coronary disease; and (3) with an intact uterus.

This illustration uses a subset of the data with n = 612. Three variables are considered:

Data dictionary/Codebook *(Partial)*

Variable	Label	Type	Codings
sbp	Systolic blood pressure (mm Hg)	numeric	Continuous, range, [45:79]
raceth	Race	numeric	1 = White 2 = African American 3 = Other
physact	Comparative ("compared to other women your age") physical activity	Numeric	1 = much less active 2 = somewhat less active 3 = about as active 4 = somewhat more active 5 = much more active

initialize session

```
setwd("/cloud/project")                                          # Set working directory
getwd()                                                          # Check working directory
options(scipen=999)                                              # Turn off scientific notation
rm(list = ls())                                                  # Clear the Decks
```

import excel source data

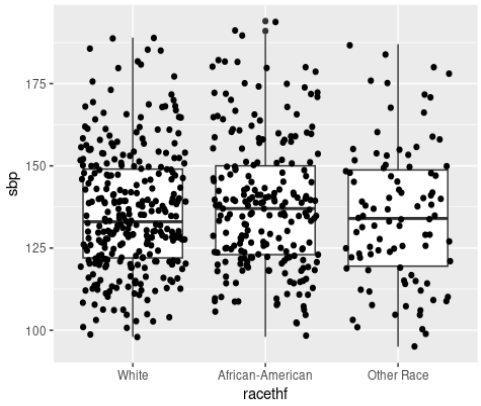
```
library(readxl)
source <- read_excel("hers_640anova.xlsx")
source <- as.data.frame(source)
str(source)

## 'data.frame':    612 obs. of  3 variables:
##  $ raceth : num  3 3 3 3 3 3 3 3 3 3 ...
##  $ physact: num  1 3 2 5 2 1 1 1 1 1 ...
##  $ sbp    : num  132 168 105 159 155 126 107 112 166 150 ...
```

Categorical predictors in R must be type factor. As you like, set the reference level explicitly

```
source$racethf <- factor(source$raceth,                           # Tip: Set explicitly
                         levels=c(1,2,3),
                         labels=c("White", "African-American", "Other Race"))

source$racethf <- relevel(source$racethf, ref="White")            # relevel() with option ref = to set ref
```

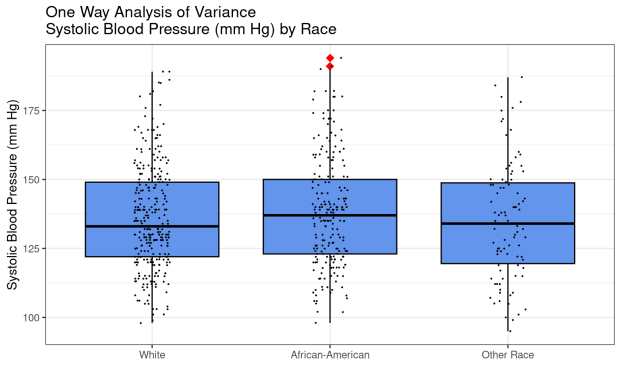
Always look at your data!

```
library(ggplot2)

# Side-by-side box plot w overlay scatter:  basic
ggplot(data=source) +
  aes(x=racethf) +                                                # x = factor predictor
  aes(y=sbp) +                                                    # y = outcome
  geom_boxplot( ) +
  geom_jitter( )
```



Basic plot doesn't look so good. See next page for suggested fixes

```
library(ggplot2)

# Side-by-side box plot w overlay scatter:  with optional aesthetics
ggplot(data=source) +
  aes(x=racethf) +                              # x = factor predictor
  aes(y=sbp) +                                  # y = outcome

  geom_boxplot(color="black",                   # Tip. boxplot first. color=border of box
               fill= "cornflowerblue",          # fill = color of box fill
               outlier.colour="red",           # user sets outlier color
               outlier.shape=18,                # user sets outlier shape (18=diamond)
               outlier.size=3) +                # user sets outlier size

  geom_jitter(color="black",                    # Overlay plot of scatter
               width=.1,                        # width and height to set extent of jitter
               height=.1,
               size=.1) +                       # size to set size of points

  ggtitle("One Way Analysis of Variance\nSystolic Blood Pressure (mm Hg) by Race") +
  xlab("") +
  ylab("Systolic Blood Pressure (mm Hg)") +
  theme_bw()
```



Better! Also, outliers are now shown in red diamonds

Obtain numerical descriptives

```
library(summarytools)

cat("\nDescriptives by group using descr() in {summarytools}\n")
## Descriptives by group using descr() in {summarytools}

with(source,                                                # with(dataframe,
   stby(data = sbp,                                         #      data = yvar
   INDICES =racethf,                                        #      INDICES = factor var
   FUN = descr,
   stats=c("n.valid", "pct.valid", "mean", "sd", "min", "max"),     #      user chooses statistics to show
   #stats=c("common"),                                      #      NOT RUN: another set to show
   transpose=TRUE))                                         #       display descriptives horizontally

## Descriptive Statistics
## sbp by racethf
## Data Frame: source
## N: 300
##
##                          N.Valid   Pct.Valid     Mean   Std.Dev     Min      Max
## ---------------------- --------- ----------- -------- --------- ------- --------
##                  White    300.00      100.00   136.01     18.55   98.00   189.00
##       African-American    218.00      100.00   138.23     19.99   98.00   194.00
##             Other Race     94.00      100.00   135.18     21.26   95.00   187.00

descr() has the advantage of lots of options (choices of statistics, layout, etc)
```

Fit model as an analysis of variance. Show.

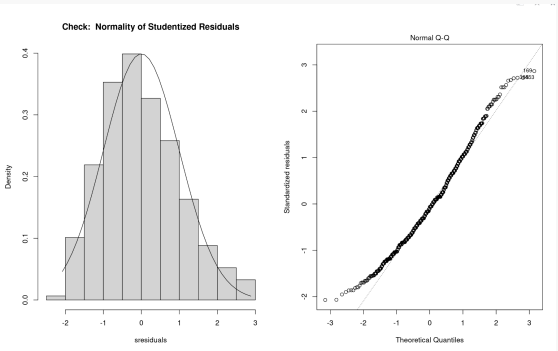
```
m1_anova <- aov(sbp ~ racethf, data=source)                                  # aov(yvar ~ factorvar,data=DATAFRAME)

anova(m1_anova)                                                              # anova(MODELOBJECT) to show results
## Analysis of Variance Table
##
## Response: sbp
##            Df Sum Sq Mean Sq F value Pr(>F)
## racethf     2    871  435.50  1.1448  0.319      Overall F test (Null: equality of means) does NOT reject null (p=.32)
## Residuals 609 231671  380.41

summary(m1_anova)                                                            # summary(MODELFIT) to show results
##              Df Sum Sq Mean Sq F value Pr(>F)
## racethf       2    871   435.5   1.145  0.319
## Residuals   609 231671   380.4

anova( ) provides slightly more information than summary( )
```

Fit model as a regression. Show.

```
m1_regression <- lm(sbp ~ as.factor(raceth), data=source)                    # lm(yvar ~ as.factor(groupvar), data=)

anova(m1_regression)
## Analysis of Variance Table
##
## Response: sbp
##                    Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(raceth)   2    871  435.50  1.1448  0.319
## Residuals         609 231671  380.41

summary(m1_regression)                                                       # summary(MODELFIT) to show results
## Call:
## lm(formula = sbp ~ as.factor(raceth), data = source)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.234 -14.234  -1.624  12.766  55.766
##
## Coefficients:
##                    Estimate Std. Error t value            Pr(>|t|)
## (Intercept)        136.0133     1.1261 120.786 <0.0000000000000002 ***
## as.factor(raceth)2   2.2206     1.7358   1.279               0.201             Not surprising: no effect of racethf = 2
## as.factor(raceth)3  -0.8325     2.3054  -0.361               0.718                  Similarly, no effect of racethf = 3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.5 on 609 degrees of freedom
## Multiple R-squared:  0.003746,   Adjusted R-squared:  0.0004738                Not surprising that R-squared is tiny!
## F-statistic: 1.145 on 2 and 609 DF,  p-value: 0.319
```

Regression diagnostics: normality of residuals - visualization

```
library(MASS)

sresiduals <- studres(m1_anova)                    # Null: studentized residuals are Normal(0,1)
par(mfrow = c(1,2))                                # set graph to be 2 panes (1 row, 2 col)
hist(sresiduals,                                   # histogram of sresiduals (plot density not freq)
     freq=FALSE,
     main="Check:  Normality of Studentized Residuals")
xfit <- seq(min(sresiduals),max(sresiduals),length=40)
yfit <- dnorm(xfit)
lines(xfit,yfit)
plot(m1_anova, which = 2)                              # which=2 qqplot (look for straight line)
```



Not great, but sometimes the cure is worse than the problem. Onward

```
par(mfrow = c(1,1))                                # restore graph to be 1 pane (1 row, 1 col)
```

Regression diagnostics: normality of residuals - hypothesis test of Null: normality

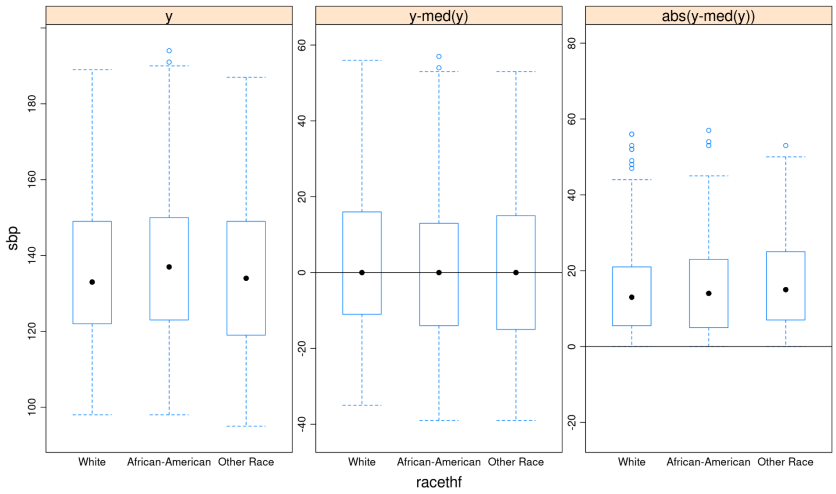
```
shapiro.test(source$fit.resid)
##  Shapiro-Wilk normality test
##
## data:  source$fit.resid
## W = 0.98034, p-value = 0.0000002518          Test of Null: normality of residuals is rejected, possibly due to large n
```

Regression diagnostics: constant variance - hypothesis test of Null: homogeneity

```
library(HH)                                        # hov() for Brown-Forsyth in package {HH}
library(car)                                       # leveneTest() in package {car}

# null:  constant variance, all is well
bartlett.test(sbp ~ racethf, data=source)          # argument must be model formula
##  Bartlett test of homogeneity of variances
##
## data:  sbp by racethf
## Bartlett's K-squared = 3.1766, df = 2, p-value = 0.2043        Test of Null: constant variance is NOT rejected (good!)
```

Regression diagnostics: constant variance - visualization

```
library(HH)                                                    # hovPlot() in package {HH}

hovPlot(sbp ~ racethf, data=source)
```



Plot is consistent with results of tests of common variance (null not rejected)