stcp-karadimitriouANOVAR

The following resources are associated: Excel dataset ’Diet.csv’, ANOVA in R script file, Summarising continuous variables in R, Statistical Hypothesis testing, Checking normality in R

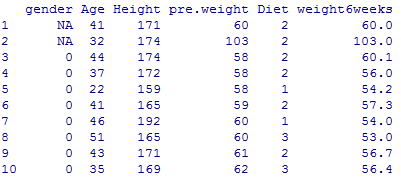
One-way (between-groups) ANOVA in R

Dependent variable: Continuous (scale),

Independent variable: Categorical (at least 3 unrelated/ independent groups)

Common Applications: Used to detect a difference in means of 3 or more independent groups.  It can be thought of as an extension of the independent t-test for and can be referred to as ‘between-subjects’ ANOVA.

Data: The data set Diet.csv contains information on 78 people who undertook one of three diets. There is background information such as age, gender (Female=0, Male=1) and height.



Diet 1, 2 or 3

Female = 0

Download the diet data set and save it to your computer. To open the file use the read.csv() command. You will need to change the command depending on where you have saved the file.dietR<-read.csv("D:\\diet.csv",header=T)

Tell R to use the diet dataset until further notice using attach(dataset) so 'Height' can be used instead of dietR$Height. Tell R that 'Diet' is a factor using as.factor(variable).

attach(dietR)

Diet<-as.factor(Diet)

Research question: Which of three diets was best for losing weight?

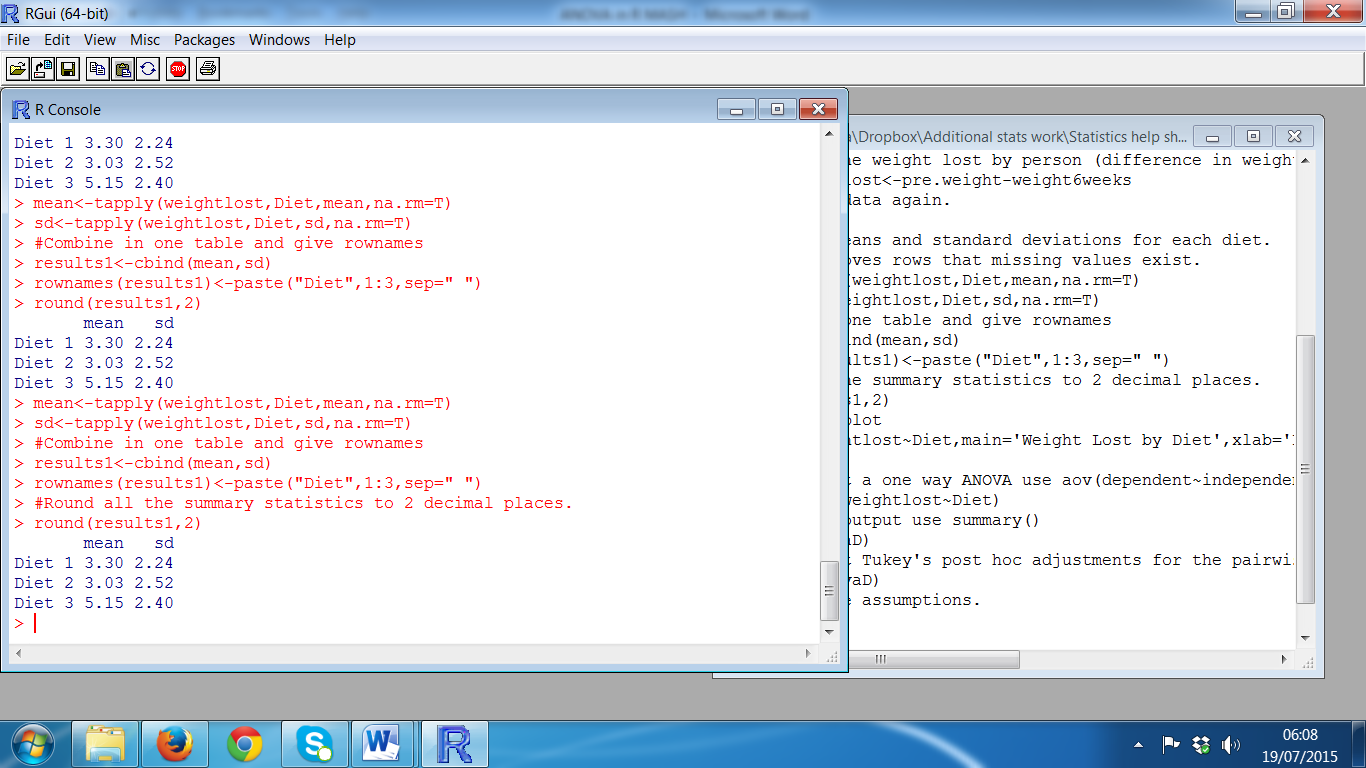
The dependent variable is weight lost (scale) and the independent variable (group) is diet. Calculate the weight lost by person (difference in weight before and after the diet) and add the variable to the dataset. Then attach the data again.

dietR$weightlost<-pre.weight-weight6weeks

attach(dietR)

Summary Statistics

Before carrying any analysis, summarise weight lost by diet using a box-plot or interval plot and some summary statistics. Do the group means and standard deviations look similar or very different? Calculate means and standard deviations for weight lost by diet using tapply(dependent, independent, summary statistic required, na.rm=T) na.rm=T removes rows where missing values exist. See the Summarising Continuous variables resource for more details.

mean<-tapply(weightlost,Diet,mean,na.rm=T)

sd<-tapply(weightlost,Diet,sd,na.rm=T)

Combine in one table and give the rows the names of the diets

results1<-cbind(mean,sd)

rownames(results1)<-paste("Diet",1:3,sep=" ")

Round all the summary statistics to 2 decimal places.

round(results1,2)

To produce a boxplot of weight lost by diet:

boxplot(weightlost~Diet,main='Weight Lost by Diet',xlab='Diet', ylab='Weight Lost')

Diet 3 seems better than the other diets as the mean weight lost is greater. The standard deviations are similar so weight lost within each group is equally spread out. One of the assumptions for ANOVA is that the variances (SD2) must be similar. If the largest is more than twice the smallest, the assumption has been violated.

ANOVA stands for ‘Analysis of variance’ as it uses the ratio of between group variation to within group variation, when deciding if there is a statistically significant difference between the groups. *Within group variation* measures how much the individuals vary from their group mean. Each difference between an individual and their group mean is called a residual. These residuals are squared and added together to give the sum of the squared residuals or the within group sum of squares (SSwithin). *Between group variation* measures how much the group means vary from the overall mean (SSbetween).

# Steps in R and output

# To carry out a one way ANOVA use aov(dependent~independent, give the ANOVA model a name e.g. anovaD and use summary() to see the output.

# anovaD<-aov(weightlost~Diet)

# summary(anovaD)

F = Test statistic

MSDiet = 35.55 =6.197

MSerror 5.74

# 

P = p-value = sig

= P(F > 6.197)

**p = 0.00323**

# 

# 

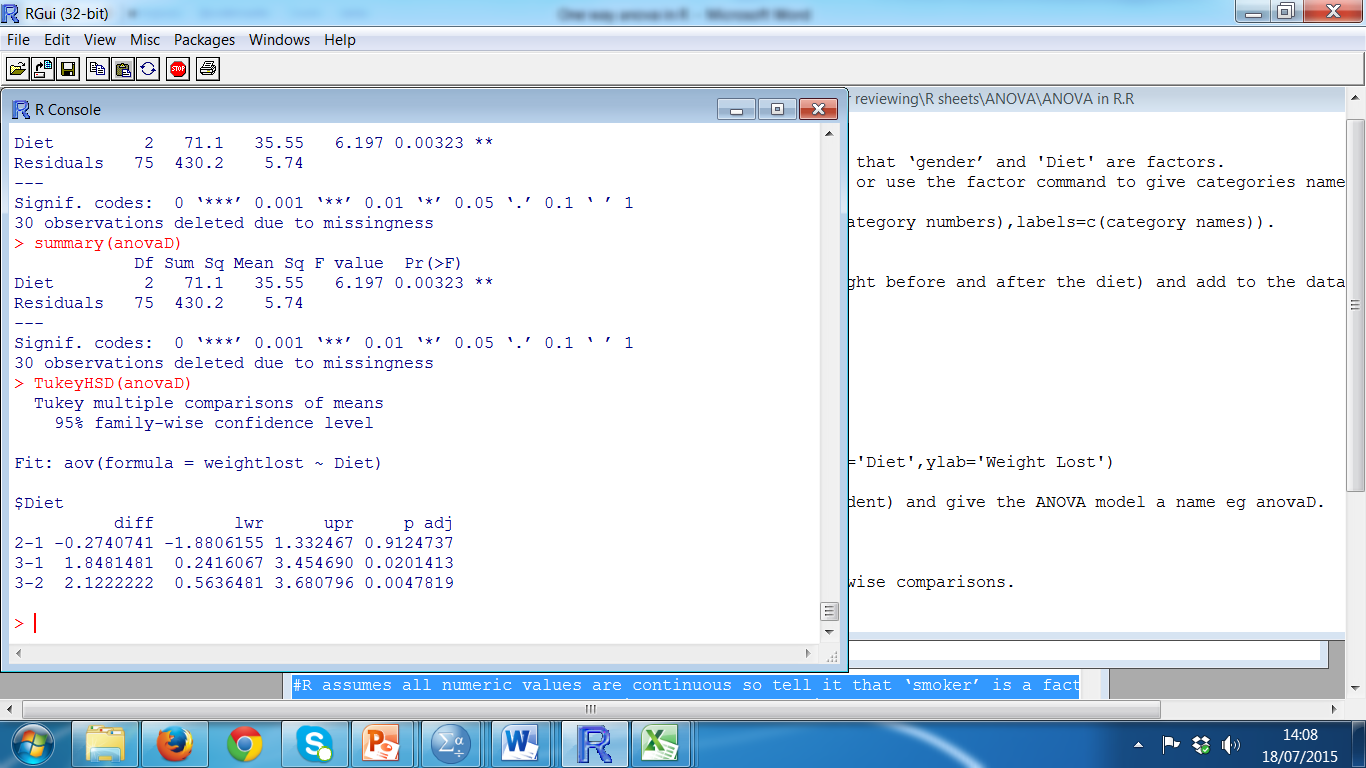
# When writing up the results, it is common to report certain figures from the ANOVA table.

**F(dfbetween, dfwithin)= Test Statistic, p =** 🡺 **F(2, 75)= 6.197, p =0.003**

There was a significant difference in mean weight lost [F(2,75)=6.197, p = 0.003] between the diets.

**Post Hoc Tests**

# ANOVA tests the null hypothesis ‘all group means are the same’ so the resulting p-value only concludes whether or not there is a difference between one or more pairs of groups. If the ANOVA is significant, further ‘post hoc’ tests have to be carried out to confirm where those differences are. The post hoc tests are often t-tests with an adjustment to account for the multiple testing. *Tukey’s* is the most commonly used post hoc test but check if your discipline uses something else. Use the command TukeyHSD(anovaD).



The p-values are in the p adj column, the mean differences in diff and the confidence interval for the difference in lwr and upr. Report each of the three pairwise comparisons e.g. there was a significant difference between diet 3 and diet 1 (p = 0.02). Use the mean difference between each pair e.g. people on diet 3 lost on average 1.85 kg more than those on diet 1 or use individual group means to conclude which diet is best.

Checking the assumptions for one-way ANOVA

|  |  |  |
| --- | --- | --- |
| **Assumptions** | **How to check** | **What to do if the assumptions is not met** |
| Residuals should be normally distributed | Use histogram, QQ plots and normality tests as diagnostic tools (see the Checking **normality in R resource** for more details) | If the residuals are very skewed, the results of the ANOVA are less reliable so the Kruskall- Wallis test should be used instead (see the **Kruskall-Wallis in R** resource) |
| Homogeneity (equality) of variance: The variances (SD squared) should be similar for all the groups | If largest SD is more than twice the smallest, assumption is not met. The Levene's test of equality of variances can also be used. If p - value > 0.05, equal variances can be assumed and the ANOVA results are valid | If p - value < 0.05, the results of the  ANOVA are less reliable. The Welch test is more appropriate and can be accessed via library(car)  oneway.test(weightlost~Diet)  The Games Howell post hoc test should be used instead of Tukeys but doesn’t exist in R |

Checking the assumptions for this data

Ask for the standardised residuals (difference between each individual and their group mean) and give them a name (res).

res<-anovaD$residuals

Produce a histogram of the residuals.

hist(res, main="Histogram of standardised residuals",xlab="Standardised residuals")

The standard deviations for the groups are similar, so the assumption of equal variances has been met. If you wish to carry out a Levene’s test as well you can do this through the additional packages car or lawstat which need to be loaded using library(car)or library(lawstat)

If this command does not work, you will need to go to the Packages --> Install package(s) and select the UK (London)CRAN mirror. Then look for the package 'car' or ‘lawstat’ and click. A lot of extra menus will download as well. Then try library(car)again*.*

Note: There are some issues with Rstudio not having all the aspects of the car package so even though the package will download, you may not be able to use some commands.

Once car is loaded, carry out Levene's test using leveneTest(weightlost~Diet).

For lawstat use levene.test(weightlost,Diet)

|  |  |
| --- | --- |
| **Homogeneity Assumption** | **Normality Assumption** |
| As p - value (0.5377) > 0.05, equal variances can be assumed | The residuals are normally distributed |

Reporting ANOVA

A one-way ANOVA was conducted to compare the effectiveness of three diets. Normality checks and Levene’s test were carried out and the assumptions met.

There was a significant difference in mean weight lost [F(2,75)=6.197, p = 0.003] between the diets. Post hoc comparisons using the Tukey test were carried out. There was a significant difference between diets 1 and 3 (p = 0.02) with people on diet 3 lost on average 1.85 kg more than those on diet 3. There was also a significant difference between diets 2 and 3 difference (p = 0.005) with people on diet 3 lost on average 2.12 kg more than those on diet 2.