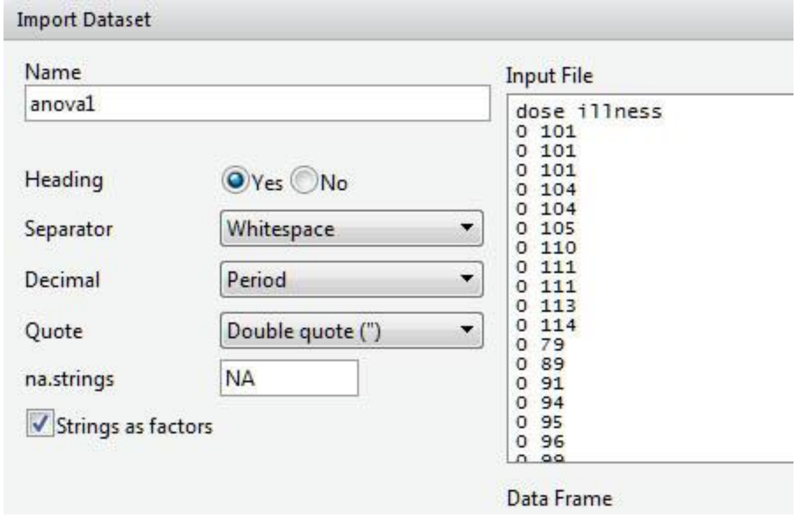
One Way Independent Samples ANOVA and Trend Analysis with R

Download the data file ANOVA1.txt. These are contrived data (I created them with a normal random number generator in the SAS statistical package). We shall imagine that we are evaluating the effectiveness of a new drug (Athenopram HBr) for the treatment of persons with depressive and anxiety disorders. Our independent variable is the daily dose of the drug given to such persons, and our dependent variable is a measure of these persons' psychological illness after two months of pharmacotherapy. We have 20 scores in each of five treatment groups.



Import Dataset

Name: anova1

Heading: ● Yes ○ No

Separator: Whitespace

Decimal: Period

Quote: Double quote (")

na.strings: NA

☑ Strings as factors

Input File

```
dose illness
0 101
0 101
0 101
0 104
0 104
0 105
0 110
0 111
0 111
0 113
0 114
0 79
0 89
0 91
0 94
0 95
0 96
```

Data Frame

Startup RStudio and import the data.

```
> anova1 <- read.table("C:/Users/Vati/Desktop/anova1.txt", header=TRUE, quote="\"")
>   View(anova1)
```

Although the dose variable is numeric, we want R to treat it as a classification variable (aka grouping variable or factor), so we issue this command: anova1$dose <- factor(anova1$dose)

Next, we get a table of the values and sample sizes for dose: table(anova1$dose)

```
0 10 20 30 40
20 20 20 20 20
```

Group means are obtained by: aggregate(anova1$illness, by=list(anova1$dose), FUN=mean)

```
  Group.1      x
1       0 100.80
2      10  85.60
3      20  80.10
4      30  86.55
5      40 100.50
```

And standard deviations by: aggregate(anova1$illness, by=list(anova1$dose), FUN=sd)

```
  Group.1        x
1       0 8.817447
2      10 8.635301
3      20 7.635788
4      30 8.500619
5      40 8.249402
```

The ANOVA by: illness.dose <- aov(illness ~ dose, data=anova1)

summary(illness.dose)

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
dose         4   7073  1768.2   25.19 3.09e-14 ***
Residuals   95   6668    70.2                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

One should report an effect size statistic, and eta-squared is often that reported with an ANOVA. For these data, η^2 = 7073/(7073 + 6668) = .515. Now I want a confidence interval for the η^2. For the confidence interval to be congruent with the test of significant, I should use a confidence coefficient of (1 - 2α). For the usual .05 level for alpha, that will be a confidence coefficient of .90. I install and activate the “MBESS” package and issue this command:

```
ci.pvaf(F.value=25.19, df.1=4, df.2=95, N=100, conf.level=.90)
```

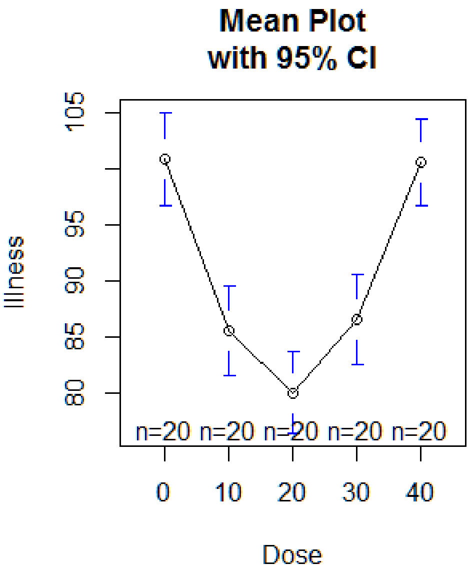
```
[1] "The 0.9 confidence limits (and the actual confidence interval coverage) for the prop
ortion of variance of the dependent variable accounted for by knowing group status are gi
ven as:"
$Lower.Limit.Proportion.of.Variance.Accounted.for
[1] 0.3786666

$Upper.Limit.Proportion.of.Variance.Accounted.for
[1] 0.5868311
```

I want to make a fancy plot, so I Install the package “gplots”

```
package ‘bitops’ successfully unpacked and MD5 sums checked
package ‘gtools’ successfully unpacked and MD5 sums checked
package ‘gdata’ successfully unpacked and MD5 sums checked
package ‘caTools’ successfully unpacked and MD5 sums checked
package ‘gplots’ successfully unpacked and MD5 sums checked
```

```
library(gplots)
plotmeans(anova1$illness ~ anova1$dose, xlab="Dose", ylab="Illness",
main="Mean Plot\nwith 95% CI")
```



Now a Tukey test for pairwise comparisons among the means.

```
TukeyHSD(illness.dose)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = illness ~ dose, data = anova1)

$dose
        diff         lwr        upr     p adj
10-0  -15.20 -22.5672887  -7.832711 0.0000011
20-0  -20.70 -28.0672887 -13.332711 0.0000000
30-0  -14.25 -21.6172887  -6.882711 0.0000053
40-0   -0.30  -7.6672887   7.067289 0.9999622
20-10  -5.50 -12.8672887   1.867289 0.2390435
30-10   0.95  -6.4172887   8.317289 0.9964057
40-10  14.90   7.5327113  22.267289 0.0000018
30-20   6.45  -0.9172887  13.817289 0.1150841
40-20  20.40  13.0327113  27.767289 0.0000000
```

```
40-30  13.95   6.5827113  21.317289 0.0000085
```

For each of the comparisons you are given a 95% confidence interval for the difference in means and an adjusted *p* value. If the confidence interval does not include the value 0, then the adjusted *p* will be less than .05, and the difference in means significant. The best way to display these results is in a table like that below. Start by arranging the groups in order of their means – here I have them listed from highest mean to lowest mean. Then add superscripted letters to the means such that any two means that share a superscripted letter do not differ significant from each other. The pattern here is very simple. Those who received doses of 10, 20, or 30 mg were significantly less ill than those who received doses of 0 or 40 mg.

Table 1

Psychological Illness of Patients

As a Function of Dose of Athenopram

Dose of Drug (mg)	*M*	*SD*	*n*
0	100.89^{A}	8.817	20
40	100.59^{A}	8.249	20
30	86.55^{B}	8.501	20
10	85.69^{B}	8.635	20
20	80.19^{B}	7.636	20

Note. Means sharing a letter in their superscript are not significantly different at the .05 level according to a Tukey HSD test.

I would prefer to use a REGWQ test here, but have been unable to find a package that does that well. It is included in the “toss” package, but one of the dependencies for that package has been removed from the depository, although you can find it elsewhere – but don’t bother. The output is so bad that I can’t figure out how to interpret it.

Trend Analysis

When the ANOVA independent/grouping variable is actually a quantitative variable (measures how much of some attribute was present) rather than a qualitative variable, one might be interested in describing the shape of the relationship between the two variables. The plot makes it pretty clear that the relationship between dose and illness is not linear. A trend analysis, aka polynomial regression, can be helpful. Start out by predicting illness from dose with a bivariate liner regression and then add to the model a second predictor, the square of dose, creating a quadratic model. This will allow one bend in the regression line. Then try adding to that model dose cubed, allowing a second bend and creating a cubic model. Here is how to this in R. First I exited R and then booted it back up and brought the data back in – I did this because I no longer wanted R to treat dose as a factor. Next I created the two new predictor variables, dose-squared and dose-cubed:

```
anova1 <- transform(anova1, d2=dose^2, d3=dose^3)
```

Now I conduct the linear analysis.

```
linear<- lm(illness ~ dose, data=anova1)
```

```
summary(linear)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-26.7100  -8.7188   0.3425   8.3600  23.3600 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 90.64000    2.05091  44.195   <2e-16 ***
dose         0.00350    0.08373   0.042    0.967    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.84 on 98 degrees of freedom
Multiple R-squared:  1.783e-05,       Adjusted R-squared:  -0.01019 
F-statistic: 0.001747 on 1 and 98 DF,  p-value: 0.9667
```

As you can see, the linear relationship is basically nil, the r^2 being 0 to several decimal points. Lets try adding one bend to the regression line”

```
quadratic<- lm(illness ~ dose +d2, data=anova1)
summary(quadratic)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-21.6757  -4.7107   0.3257   5.5754  17.2729 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 100.675714   1.747702  57.605  < 2e-16 ***
dose         -2.003643   0.207029  -9.678 6.65e-16 ***
d2            0.050179   0.004963  10.110  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.305 on 97 degrees of freedom
Multiple R-squared:  0.5131,   Adjusted R-squared:  0.5031 
F-statistic: 51.11 on 2 and 97 DF,  p-value: 6.93e-16
```

Wow, that certainly helped. The R^2 went from about 0 to .51. The p value for d2 shows that this increase in R^2 is significant beyond .001. Well, might a second bend in the curve also help?

```
cubic<- lm(illness ~ dose +d2 +d3, data=anova1)
summary(cubic)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-21.8957  -4.6707   0.3257   5.7954  16.8329 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.009e+02  1.852e+00  54.480  < 2e-16 ***
dose        -2.161e+00  4.711e-01  -4.588 1.35e-05 ***
d2           6.118e-02  2.991e-02   2.045   0.0436 *  
d3          -1.833e-04  4.916e-04  -0.373   0.7100    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.342 on 96 degrees of freedom
Multiple R-squared:  0.5138,   Adjusted R-squared:  0.4986 
F-statistic: 33.82 on 3 and 96 DF,  p-value: 5.282e-15
```

Going to the cubic model produced an R^2 increase of .5138 - .5131 = .0007, a trivial increase which falls way short of significance, p = .71. I’ll stick with the quadratic model.

Presenting the Results

An analysis of variance indicated that dose of Athenopram significantly affected psychological illness of our patients, $F(4, 95) = 25.193$, $MSE = 70.1871$, $p < .001$, $\eta^2 = .515$, 90% CI [.379, .587]. As shown in Table 1, a Tukey HSD test indicated that 10 to 30 mg doses of the drug were associated with significantly better mental health than were doses of 0 or 40 mg. A trend analysis indicated that the data were well fit by a quadratic model, with the quadratic component accounting for a large and significant proportion of the variance in illness ($\eta^2 = .513$, $p < .001$).

Return to Wuensch’s R Lessons