Worksheet 8 ANOVA

Let’s perform an ANOVA on the soil data we looked at in class, first in Excel and then in R.

Recall, an agricultural researcher wishes to investigate the effect of soil type (sand, clay or loam) on crop yield. We have an experiment in which crop yields per unit area were measured from 10 randomly selected fields on each of the soil types. All fields were sown with the same variety of seed and provided with the same fertilizer and pest control inputs. The question is whether soil type significantly affects crop yield, and if so, to what extent (taken from Crawley p.503).

|  |  |  |
| --- | --- | --- |
| **Sand** | **Clay** | **Loam** |
| 6 | 17 | 13 |
| 10 | 15 | 16 |
| 8 | 3 | 9 |
| 6 | 11 | 12 |
| 14 | 14 | 15 |
| 17 | 12 | 16 |
| 9 | 12 | 17 |
| 11 | 8 | 13 |
| 7 | 10 | 18 |
| 11 | 13 | 14 |

Hypothesis to be tested:

H₀ : the mean crop yields on each soil type are equal i.e. μ₁=μ₂=μ₃

HA : at least one of the mean crop yields on each soil type are not equal

We will test this assumption at the 5% significance level so α = 0.05

Excel

1. In Excel open a blank workbook and arrange the data in columns with the headings ‘Sand’, ‘Clay’ and ‘Loam’.
2. In the bottom row calculate the mean for each soil type ( ) and the overall mean ()
3. Calculate SST
4. Calculate SSF
5. Calculate SSE (remember SSE = SST – SSF)
6. Create ANOVA table
7. Test the significance of calculated F-ratio by comparing it to the critical F-value or by calculating the p-value.

To obtain the critical value using R we can the quantile function qf of the F distribution for 5% significance we use the quantile 0.95.

qf(.95, df1 = 2, df2 = 27)

To obtain the p- value associated with the calculated F-ratio of 4.24 we can use:

pf(4.24, df1 = 2, df2 = 27, lower.tail = FALSE)

R

First import the soildata.txt file to R.

results<-read\_tsv((here(soil\_data.txt))

Note that the data is not in the correct format, each level of the soil factor (Sand, Clay and Loam) has its own column.

Sand Clay Loam

6 17 13

10 15 16

8 3 9

6 11 12

Can you use the gather() function to get the data in the correct format shown below?

Yield Soil

6 Sand

10 Sand

8 Sand

6 Sand

First let’s calculate the mean yields and the standard deviation for each of the soil types. We can use the group\_by function followed by the summarise() function.

results %>% group\_by(Soil) %>%

summarise(mean\_yield = mean(Yield), sd\_yield = sd(Yield))

We obtain the following output:

# A tibble: 3 x 3

Soil mean\_yield sd\_yield

<chr> <dbl> <dbl>

1 Clay 11.5 3.92

2 Loam 14.3 2.67

3 Sand 9.90 3.54

To carry out a one way ANOVA, we need to ensure that the Soil variable is defined to be a factor, we can check:

is.factor(results$Soil)

[1] FALSE

To discover the *names* of the factor levels, we use the levels function:

levels(results$Soil)

[1] “Sand”, “Clay”, “Loam”

To discover the number of levels of a factor, we use the nlevels function:

nlevels(results$Soil)

[1] 3

To visualise the effect of soil type on yield we can use a box and whisker plot of Yield vs. Soil

boxplot(results$Yield~results$Soil)

What does the plot tell us about the data?

* Median yield is lowest on sand and highest on loam.
* There is considerable variation within each soil type.
* There is a potential outlier on clay.
* It looks as though yield on loam is significantly greater than yield on sand.

We can test the potential outlier using grubbs.test()

grubbs.test(results$Yield[results$Soil == "Clay"])

The point is not significant at the 5% significance level, however we will examine its influence with regard to the model at the diagnostics stage.

Assumptions of ANOVA

When using ANOVA on a data set we should check that the following assumptions hold:

1. random sampling
2. variances of different treatments are equal
3. the error terms are independent from observation to observation
4. the error terms are normally distributed with zero mean and the same variance

The assumption that random sampling has been carried out must be satisfied at the experimental/data collection stage. Before proceeding with the ANOVA we will check that homogeneity of variance is satisfied. This can be done using Fisher’s F Test (to compare two variances) [var.test], Bartlett's Test [Bartlett.test], Levene’s Test [leveneTest], or the Fligner-Killeen test of homogeneity of variances.

[Aside: for more detail on the comparison of variances Crawley p. 354]

Let’s calculate the variance for the different soil types:

results %>% group\_by(Soil) %>%

summarise( var\_yield = var(Yield))

To test whether the variances are the same for the different treatments we test the null hypothesis:

H0 = the variances of the different soil types are the same.

HA = the variances of the different soil types are not the same.

fligner.test(results$Yield~results$Soil)

Fligner-Killeen test of homogeneity of variances

data: results$y by results$Soil

Fligner-Killeen:med chi-squared = 0.3651, df = 2, p-value =0.8332

Here the p-value is 0.8332 so we fail to reject the null hypothesis and conclude that the variances are not significantly different from one another and we may apply the ANOVA. We can check the rest of the assumptions after we have carried out the ANOVA.

The ANOVA model

In R the one way ANOVA model is specified by:

aov(y~x)

where y is the response variable and x is the explanatory variable.

It is good practice to assign the model a name so that it can be called on again with ease e.g.

model\_treat<- aov(results$Yield~results$Soil)

summary(model\_treat)

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

results$Soil 2 99.2 49.60 4.245 0.025 \*

Residuals 27 315.5 11.69

---

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

We see the values that we calculated in Excel. The errors are referred to as Residuals. In the last column, instead of a critical F Value we have a **p-value** which tells us that the probability of obtaining the F value of 4.245 or more if the null hypothesis is true (i.e. the probability of making a type I error) is 0.025 or a 1 in 40 chance. The asterisk indicates that the difference between the soil means is significant at the 5% level but not at the 1% level. Note that R does not print the SST or the associated degrees of freedom.

We can obtain estimates for the ANOVA model using the summary.lm() command. The default command will output *treatment* *contrasts* that compare each treatment to a baseline:

Coefficients:

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

(Intercept) 11.500 1.081 10.638 3.7e-11 \*\*\*

results$SoilLoam 2.800 1.529 1.832 0.0781 .

results$SoilSand -1.600 1.529 -1.047 0.3046

---

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

Residual standard error: 3.418 on 27 degrees of freedom

Multiple R-squared: 0.2392, Adjusted R-squared: 0.1829

F-statistic: 4.245 on 2 and 27 DF, p-value: 0.02495

The estimate for the intercept is the average yield for the Clay treatment.

The estimate for Loam gives the difference in mean yield between the Clay treatment and the Loam treatment.

The estimate for Sand gives he difference in mean yield between the Clay treatment and the Loam treatment.

We can write down the treatment effects model as:

Where for the Clay treatment and .

For the Loam treatment and and .

For the Sand treatment and .

Note that the p-values reported show that mean yield for Loam and Sand was not significantly different to the mean yield for Clay at the 5% significance level.

We can view the treatment contrasts in matrix form:

contrasts(results$Soil)

Loam Sand

Clay 0 0

Loam 1 0

Sand 0 1

We can see that the Loam column represents the dummy variable in the model above, and the Sand column represents the dummy variable.

To obtain the effects model where each treatment is compared to the overall mean, we can specify *sum* *contrasts*:

contrasts(results$Soil) <- contr.sum

Now we can rerun the ANOVA model again:

model\_sum<- aov(results$Yield~results$Soil)

summary(model\_sum)

Note that there is no change to the ANOVA table but the treatment estimates are different:

summary.lm(model\_sum)

Coefficients:

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

(Intercept) 11.9000 0.6241 19.067 <2e-16 \*\*\*

results$Soil1 -0.4000 0.8826 -0.453 0.6540

results$Soil2 2.4000 0.8826 2.719 0.0113 \*

---

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

Residual standard error: 3.418 on 27 degrees of freedom

Multiple R-squared: 0.2392, Adjusted R-squared: 0.1829

F-statistic: 4.245 on 2 and 27 DF, p-value: 0.02495

Notice that for this output, the intercept coefficient corresponds to the overall mean.

The Soil1 coefficient corresponds to the difference between the overall mean and the mean of the Clay treatment.

The Soil2 coefficient corresponds to the difference between the overall mean and the mean of the Loam treatment.

What about the Sand treatment? To figure out how to calculate the mean of the Sand treatment from the output, we need to look at the contrasts:

contrasts(results$Soil)

[,1] [,2]

Clay 1 0

Loam 0 1

Sand -1 -1

We can write down the model as:

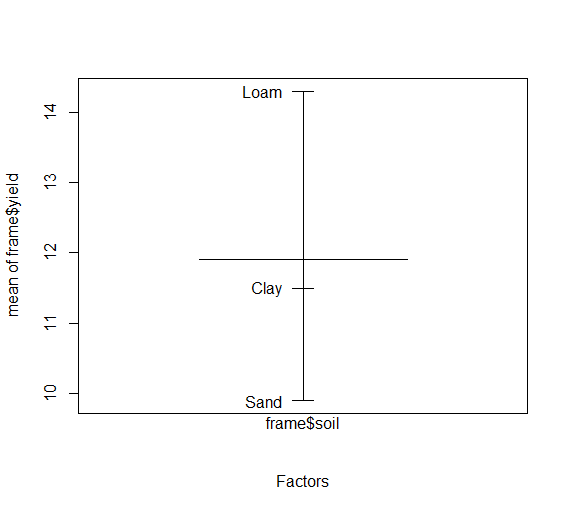
Where for the Clay treatment and

For the Loam treatment and

For the Sand treatment and

We can view the treatment effects graphically using plot.design (which takes a formula rather than a model object):

plot.design(results$Yield~results$Soil)



To see the treatment effects in tabular form we can use the command model.tables:

model.tables(model\_treat)

Tables of effects

results$Soil

results$Soil

Clay Loam Sand

-0.4 2.4 -2.0

To return the contrasts to the default setting:

contrasts(results$Soil) <- NULL

Effect Size

To measure the effect size associated with a one way ANOVA, we use (Eta squared).

For the yield data set or 23%. We can say that 23.9% of the variation in yield was caused by soil type.

**Pairwise Comparisons**

For the yield example, the ANOVA found that there was a significant difference between the three soil types but did not tell us which types of soil had significantly different yields. To compare each pair of soil types we can use Tukey's honest significant difference (HSD).

TukeyHSD(model\_treat)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = results$Yield ~ results$Soil)

$`results$Soil`

diff lwr upr p adj

Loam-Clay 2.8 -0.9903777 6.5903777 0.1785489

Sand-Clay -1.6 -5.3903777 2.1903777 0.5546301

Sand-Loam -4.4 -8.1903777 -0.6096223 0.0204414

The diff coefficients report the difference between the mean yields for each of the pairwise comparisons, the lwr and upr coefficients report the 95% confidence interval around the difference and the p-value tells us whether this difference is statistically different to zero. Here we see that the mean yield of the Sand and Loam treatments were significantly different. To see this information graphically we can use:



We see that only the confidence interval for the difference in means between Sand and Loam soil types does not contain 0. This tells us that the difference between treatment means for Sand and Loam is significantly different to 0, i.e. the mean yield for Sand is significantly different to the mean yield for Loam.

Model Validation

Next we check the assumptions of the aov model (this is referred to as model validation).

par(mfrow = c(2,2))

plot(model\_treat)



The plot(model\_treat) command produces a series of commands, spread over four pages (here compressed to a single page using the par(mfrow = c(2,2)) command).

The first graph (top left) shows a plot of the **residuals against the fitted values**.

The residuals should be random with a mean of 0 and constant variation.

We see three sets of residuals plotted at each of the treatment means (the fitted values) and there appears to be no pattern in the variance of the residuals (i.e. the variance is not increasing or decreasing with the mean).

The second graph (top right) is the **Normal Q-Q plot** (quantile to quantile plot) which plots the probability distribution of the standardised residuals with the standard normal probability distribution. If the distribution of the residuals was a perfect standard normal distribution then the Normal Q-Q plot would be a perfect straight line. Departures from the straight line indicate non-normality. Points 6, 11 and 13 lie a little off the straight line but this is nothing to worry about.

The third graph (bottom left) is **the Scale – Location plot** which shows a plot of the square root of the positive standardised residuals against the fitted values. Like the Residuals vs. Fitted plot it highlights any patterns the variance, each residual has been standardised by dividing through by its estimated variance. When fitting a regression model, the variance of the residuals can be smaller at the ends of the regression line. standardised residuals account for this and can be used to detect outliers and hetroscedasticty.

The fourth graph shows **Cook’s distance** which measure’s how influential a point is in the analysis. Cook's distance measures the effect of deleting a given observation on the statistical model. Data points with large residuals ([outliers](http://en.wikipedia.org/wiki/Outlier)) and/or high [leverage](http://en.wikipedia.org/wiki/Leverage_(statistics)) may often have a large Cook's distance value and are considered to merit closer examination in the analysis.

**Leverage points** are those observations made at extreme or outlying values of the independent variables such that the lack of neighboring observations means that the fitted regression model will pass close to that particular observation.

In our plot there seem to be no influential points that might be having a large effect on the parameter estimates.

**Reporting the Results**

We can report the results of the ANOVA as follows:

A one way analysis of variance indicated that the type of soil the crop was grown in significantly affected the crop yield, F(2, 27) = 4.245, MSE = 11.69, p = .025, =0.239. As shown in Table 1, Tukey’s HSD test indicated that the mean yield of crops grown on loam soil was significantly greater than the mean yield of crops grown on sand soil.

Table 1

|  |  |  |  |
| --- | --- | --- | --- |
| **Soil Type** | **Mean Yield** | **Standard deviation** | **n** |
| Sand | 9.9A | 3.92 | 10 |
| Clay | 11.5AB | 2.67 | 10 |
| Loam | 14.3B | 3.54 | 10 |

Note. Means sharing a letter in their superscript are not significantly

different at the .05 level according to a Tukey’s HSD test.

Exercise

A manufacturer of paper used for making grocery bags is interested in improving the strength of the product. It is thought that tensile strength is a function of the wood concentration in the pulp and that the range of hardwood concentrations of interest is between 5% and 20%. A team of engineers decides to investigate four levels of hardwood concentration: 5%, 10%, 15%, and 20%. They decide to make up six test specimens at each concentration level. All 24 specimens are tested for tensile strength and the data is available on Blackboard in the tensile.txt file.

1. What is the factor in this experiment?
2. What is the response variable in this experiment?
3. How many levels does the factor have?
4. State the null and alternative hypotheses associated with this experiment.
5. Carry out a one way analysis of variance on this data and state your conclusion. Use the 5% significance level and be sure to check that the assumptions of ANOVA are satisfied.
6. Write up the results presenting the basic descriptive statistics (exploratory data analysis), the results of the ANOVA and the pairwise comparisons, and a brief interpretation of the results.