Chapter 8

Comparing means of multiple groups - ANOVA (solutions to exercise)

Contents

Import Python packages

```python
# Import all needed python packages
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import scipy.stats as stats
import statsmodels.api as sm
import statsmodels.formula.api as smf
import statsmodels.stats.power as smp
import statsmodels.stats.proportion as smprop
```

8.1 Environment action plans

Exercise 8.1 Environment action plans

To investigate the effect of two recent national Danish aquatic environment action plans the concentration of nitrogen (measured in g/m^3) have been measured in a particular river just before the national action plans were enforced (1998 and 2003) and in 2011. Each measurement is repeated 6 times during a short stretch of river. The result is shown in the following table:

	N_{1998}	N_{2003}	N_{2011}
	5.01	5.59	3.02
	6.23	5.13	4.76
	5.98	5.33	3.46
	5.31	4.65	4.12
	5.13	5.52	4.51
	5.65	4.92	4.42
Row means	5.5517	5.1900	4.0483

Further, the total variation in the data is $SST = 11.4944$.

You got the following output from Python corresponding to a one-way analysis of variance (where most of the information, however, is replaced by the letters A-E as well as U and V):

```
fit = smf.ols('nitrogen ~ year', data=D).fit()
print(sm.stats.anova_lm(fit))
```

```
          Df  SumSq  MeanSq  Fvalue    Pr(>F)
Year       A    B       C       U         V
Residuals  D  4.1060    E
```

a) What numbers did the letters A-D substitute?

Solution

One should check the structure of the oneway ANOVA table, so A and D are the degrees of freedom:

A $= k - 1 = 3 - 1 = 2$
D $= n - k = 18 - 3 = 15$

From the table we see $SSE = 4.1060$. B is the treatment sum-of-squares:

$$B = SS(Tr) = SST - SSE = 11.4944 - 4.1060 = 7.3884,$$

And finally, C is the $MS(Tr)$-value

$$C = MS(Tr) = SS(Tr)/2 = 7.3884/2 = 3.6942.$$

b) If you use the significance level $\alpha = 0.05$, what critical value should be used for the hypothesis test carried out in the analysis (and in the table illustrated with the figures U and V)?

Solution

The test carried out in the ANOVA table is an F-test (Theorem 8.6), with the test statistic:

$F = \frac{SS(tr)/(k-1)}{SSE/(n-k)}$

which follows the F-distribution with degrees of freedom: $\nu_1 = k - 1 = 2$ and $\nu_2 = n - k = 15$.

When using the significance level $\alpha = 0.05$, the critical value that we need is:

$$F_{0.05}(2, 15)$$

We can find the value of $F_{0.05}(2, 15)$ using Python:

```
print(stats.f.ppf(0.95, 2, 15))

3.6823203436732412
```

Hence we see that

$$F_{0.05}(2,15) = 3.682$$

c) Can you with these data demonstrate statistically significant differences in the mean nitrogen concentrations from year to year (at significance level $\alpha = 0.05$)?

Solution

the results of the F-test can be found in the ANOVA table.

U is the F-statistic

$$F = \frac{C}{E} = \frac{(11.4944 - 4.1060)/2}{4.1060/15} = 13.496,$$

which we already see is much larger than the critical value 3.682 - hence we can already conclude that we will reject the null hypothesis of "no difference" and therefore conclude that there *is* a significant difference in the mean nitrogen concentrations, at significance level $\alpha = 0.05$.

V is the p-value (using the F(2,15)-distribution)

$$P(F > 13.496) = 0.00044,$$

Which we can also compute using Python:

```python
print(1-stats.f.cdf(13.496, 2, 15))
```

```
0.00044354250479539115
```

We can also base our conclusion on the p-value, which we see is (much) smaller than 0.05 (and therefore conclude that there *is* a significant difference in the mean nitrogen concentrations, at significance level $\alpha = 0.05$).

d) Compute the 90% confidence interval for the mean difference between year 2011 and year 1998.

Solution

Since we are only computing *one* confidence interval we do not need to make a Bonferroni-correction of the significance level and we use $\alpha = 0.10$ to get the 90% confidence interval.

We use the formula for a single pre-planned pairwise post hoc confidence interval:

$$4.0483 - 5.5517 \pm t_{1-\alpha/2}(15)\sqrt{MSE \cdot (1/6 + 1/6)}$$

To calculate this we need the quantile $t_{0.95}$ from a t-distribution with $n - k = 15$ degrees of freedom:

```
print(stats.t.ppf(0.95, df=15))

1.7530503556925547
```

Hence the confidence interval becomes:

$$-1.5034 \pm 1.753\sqrt{4.1060/15 \cdot (1/6 + 1/6)} = [-2.252; -0.7545]$$

8.2 Environment action plans (part 2)

|||| Exercise 8.2 Environment action plans (part 2)

This exercise is using the same data as the previous exercise, but let us repeat the description here. To investigate the effect of two recent national Danish aquatic environment action plans the concentration of nitrogen (measured in g/m^3) have been measured in a particular river just before the national action plans were enforced (1998 and 2003) and in 2011. Each measurement is repeated 6 times during a short stretch of river. The result is shown in the following table, where we have now added also the variance computed within each group.

	N_{1998}	N_{2003}	N_{2011}
	5.01	5.59	3.02
	6.23	5.13	4.76
	5.98	5.33	3.46
	5.31	4.65	4.12
	5.13	5.52	4.51
	5.65	4.92	4.42
Row means	5.5517	5.1900	4.0483
Row variances	0.2365767	0.1313200	0.4532967

Further, the overall mean of the data is $\bar{y} = 4.930$.

a) Compute the three sums of squares (SST, $SS(Tr)$ and SSE) using the three means and three variances, and the overall mean (show the formulas explicitly).

|||| Solution

The treatment sum-of-squares $SS(Tr)$ can be computed from the three means as

$$\begin{aligned} SS(Tr) &= \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 \\ &= 6 \cdot (5.5517 - 4.930)^2 + 6 \cdot (5.1900 - 4.930)^2 + 6 \cdot (4.0483 - 4.930)^2 \\ &= 7.388. \end{aligned}$$

The residual error sum-of-squares *SSE* is defined by by

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

alternatively written as:

$$SSE = \sum_{i=1}^{k} (n_i - 1) s_i^2$$

and now we can insert the values

$$\begin{aligned} SSE &= 5s_1^2 + 5s_2^2 + 5s_3^2 = 5 \cdot 0.2366 + 5 \cdot 0.1313 + 5 \cdot 0.4533 \\ &= 4.106 \end{aligned}$$

Finally, we compute *SST*:

$$SST = SS(Tr) + SSE = 7.3884 + 4.106 = 11.49$$

The raw data can be read into Python by the following code:

```
nitrogen = np.array([
    5.01, 5.59, 3.02,
    6.23, 5.13, 4.76,
    5.98, 5.33, 3.46,
    5.31, 4.65, 4.12,
    5.13, 5.52, 4.51,
    5.65, 4.92, 4.42
])
year = pd.Categorical(np.tile(["1998", "2003", "2011"], 6))
D = pd.DataFrame({"nitrogen": nitrogen, "year": year})
```

b) Find the *SST*-value using Python's sample variance function "`var(ddof=1)`".

|||| Solution

From eqn. (8-10) we see that SST can be calculates from teh sample variance of the data:

$$SST = (n-1)s_y^2$$

```
print(17 * D["nitrogen"].var(ddof=1))

11.494400000000002
```

c) Run the ANOVA in Python and produce the ANOVA table in Python.

|||| Solution

It may be done as follows:

```
fit = smf.ols("nitrogen ~ year", data=D).fit()
anova = sm.stats.anova_lm(fit)
print(anova)

            df    sum_sq   mean_sq          F    PR(>F)
year       2.0  7.388433  3.694217  13.495787  0.000444
Residual  15.0  4.105967  0.273731        NaN       NaN
```

d) Do a complete post hoc analysis, where all the 3 years are compared pairwise.

|||| Solution

We want to make three pairwise comparisons: compare 1998 with 2003, 1998 with 2011 and 2003 with 2011 (number of comparisons $M = 3$). Since there is an equal number of observations in all three groups ($n_i = 6$), we can compare the groups by computing the "Least Siginifcant Difference" (the LSD-value, see eqn. (8-28)). And since there are 3 multiple comparisons we will use $\alpha_{\text{Bonferroni}} = 0.05/3 = 0.01667$

$$LSD_{0.01667} = t_{1-0.01667/2} \cdot \sqrt{2 \cdot 0.2737/6} = 0.8136.$$

For this we need the quantile $t_{1-0.01667/2}$ from a t-distribution with $n-k=15$ degrees of freedom:

```
print(stats.t.ppf(1-0.01667/2, 15))

2.6936395462653473
```

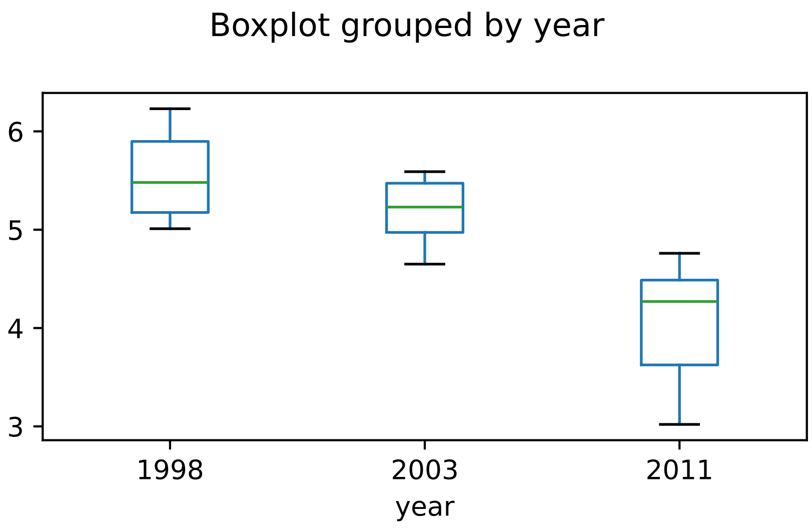
which gives us:

$$LSD = 0.8136$$

Now we look at the differences in mean nitrogen concentration and see if this difference is larger than LSD. Wee conclude that there is a significant difference between year 1998 and 2011 $(5.5517-4.0483=1.5034)$, there is a significant difference between year 2003 and 2011 $(5.1900-4.0483=1.1417)$, but there is no significant difference between year 1998 and 2003 $(5.5517-5.1900=0.3617)$.

The differences could also be visualized in the following plot:

```
D.boxplot(column='nitrogen', by='year', grid=False)
plt.title("")
plt.tight_layout()
plt.show()
```



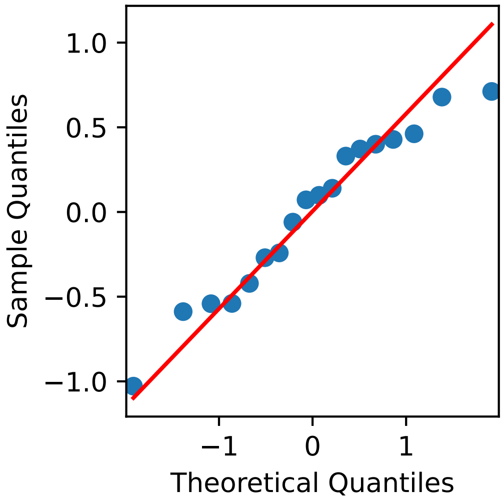
e) Use Python to do model validation by residual analysis.

|||| Solution

the model assumes that residuals follow a normal distribution with equal residual variance in all groups. The box plot from the question above does not indicate clear variance differences, so it seems the assumption about equal residual variance in each group is not violated.

We also check is the residuals seem to follow a normal distribution by doing a normal q-q plot on the residuals:

```
sm.qqplot(fit.resid.values, line='q',a=1/2)
plt.tight_layout()
plt.show()
```



There appears to be no important deviation from normality - the points do not deviate too much from the straight line.

Ideally, we woul also want to check if the residuals are *independent*.

8.3 Plastic film

|||| Exercise 8.3 Plastic film

A company is starting a production of a new type of patch. For the product a thin plastic film is to be used. Samples of products were received from 5 possible suppliers. Each sample consisted of 20 measurements of the film thickness and the following data were found:

	Average film thickness $\bar{x}$ in μm	Sample standard deviation s in μm
Supplier 1	31.4	1.9
Supplier 2	30.6	1.6
Supplier 3	30.5	2.2
Supplier 4	31.3	1.8
Supplier 5	29.2	2.2

From the usual calculations for a one-way analysis of variance the following is obtained:

Source	Degrees of freedom	Sums of Squares
Supplier	4	$SS(Tr) = 62$
Error	95	$SSE = 362.71$

a) Is there a significant ($\alpha = 5\%$) difference between the mean film thicknesses for the suppliers?

|||| Solution

The F-test statistics for one-way ANOVA is

$$F_{\text{obs}} = \frac{MS(Tr)}{MSE} = \frac{SS(Tr)/(k-1)}{SSE/(n-k)} = \frac{62/4}{362.71/95} = 4.06,$$

and the relevant critical value is $F_{0.05}(4,95)$ to be found in Python by: `stats.f.ppf(0.95, 4, 95) = 2.467`. So the answer is:

Yes, the null hypothesis is rejected, since $Fobs = 4.06$ is larger than the critical value 2.47.

Or we could find the p-value:

```
print(1 - stats.f.cdf(4.06, 4, 95))
```

```
0.004405521419133418
```

and conclude that it is this is so small, we have strong evidence against the null hypothesis.

b) Compute a 95% confidence interval for the difference in mean film thicknesses of Supplier 1 and Supplier 4 (considered as a “single pre-planned” comparison).

|||| Solution

The "ANOVA post hoc" confidence interval is to be used

$$31.4 - 31.3 \pm t_{0.975}(95)\sqrt{MSE\left(\frac{1}{n_1}+\frac{1}{n_2}\right)},$$

where $MSE = \frac{SSE}{n-k} = \frac{362.71}{95}$, so since $t_{0.975}(95)$ - to be found in Python as: `stats.t.ppf(0.975, 95)=1.985` it becomes

$$0.1 \pm 1.985\sqrt{\frac{362.71}{95}\left(\frac{1}{20}+\frac{1}{20}\right)} = [-1.127; 1.327]$$

8.4 Brass alloys

|||| Exercise 8.4 Brass alloys

When brass is used in a production, the modulus of elasticity, E, of the material is often important for the functionality. The modulus of elasticity for 6 different brass alloys are measured. 5 samples from each alloy are tested. The results are shown in the table below where the measured modulus of elasticity is given in GPa:

Brass alloys					
M1	M2	M3	M4	M5	M6
82.5	82.7	92.2	96.5	88.9	75.6
83.7	81.9	106.8	93.8	89.2	78.1
80.9	78.9	104.6	92.1	94.2	92.2
95.2	83.6	94.5	87.4	91.4	87.3
80.8	78.6	100.7	89.6	90.1	83.8

An analysis of variance is carried out using Python:

```
fit = sm.ols('elasmodul ~ alloy', data=D).fit()
print(sm.stats.anova_lm(fit))
```

and the following output is obtained: (however some of the values have been substituted by the symbols A, B, and C)

```
            df    sum_sq  mean_sq        F      PR(>F)
alloy       A    1192.51  238.501  9.9446  3.007e-05
Residuals   B       C      23.983
```

a) What are the values of A, B, and C?

|||| Solution

The A and B are the degrees of freedom, which in the oneway ANOVA is $k-1$ and $n-k$, where $k=6$ is the number of groups and $n=30$ is the number of observations. C can be found by

$$C = SSE = MSE \cdot (n-k) = 23.983 \cdot 24 = 575.59$$

So the answer is:

$A = 5$, $B = 24$ and $C = 575.59$.

b) The assumptions for using the one-way analysis of variance is (choose the answer that lists all the assumptions and that NOT lists any unnecessary assumptions):

1) The data must be normally and independently distributed within each group and the variances within each group should not differ significantly from each other

2) The data must be normally and independently distributed within each group

3) The data must be normally and independently distributed and have approximately the same mean and variance within each group

4) The data should not bee too large or too small

5) The data must be normally and independently distributed within each group and have approximately the same IQR-value in each group

|||| Solution

It is difficult to make a lot of arguments here, but simply emphasize that only in Answer 1 all assumptions needed, and no unnecessary assumptions, are listed.

c) Compute a 95% confidence interval for the single pre-planned difference between brass alloy 1 and 2.

|||| Solution

A pre-planned post hoc 95% confidence interval between two groups in a one-way ANOVA is

$$\bar{y}_1 - \bar{y}_2 \pm t_{0.975}\sqrt{MSE\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

where $t_{0.975}$ is the 0.975th quantile in a t-distribution with $n - k = 24$ degrees of freedom (`stats.t.ppf(0.975, 24) = 2.064`) So we have to compute the means of the M1 and M2 groups:

$$\bar{y}_1 = 84.62, \quad \bar{y}_2 = 81.14,$$

($\bar{y}_1 - \bar{y}_2 = 3.48$) and then plug in $MSE = 23.983$ and $n_1 = n_2 = 5$. So the answer is:

$$3.48 \pm 2.064\sqrt{23.983\left(\frac{2}{5}\right)} = [-2.91, 9.87].$$

8.5 Plastic tubes

Exercise 8.5 Plastic tubes

Some plastic tubes for which the tensile strength is essential are to be produced. Hence, sample tube items are produced and tested, where the tensile strength is determined. Two different granules and four possible suppliers are used in the trial. The measurement results (in MPa) from the trial are listed in the table below:

	Granule	
	g1	g2
Supplier a	34.2	33.1
Supplier b	34.8	31.2
Supplier c	31.3	30.2
Supplier d	31.9	31.6

The following is run in Python:

```
D = pd.DataFrame({
    "strength": [34.2,34.8,31.3,31.9,33.1,31.2,30.2,31.6],
    "supplier": pd.Categorical(["a","b","c","d","a","b","c","d"]),
    "granule": pd.Categorical([1,1,1,1,2,2,2,2])
})
fit = smf.ols("strength ~ supplier + granule", data=D).fit()
print(sm.stats.anova_lm(fit))
```

with the following result:

```
          D     sum_sq   mean_sq         F    PR(>F)
supplier  3.0  10.03375  3.344583  3.253749  0.179225
granule   1.0   4.65125  4.651250  4.524929  0.123339
Residual  3.0   3.08375  1.027917       NaN       NaN
```

a) Which distribution has been used to find the p-value 0.1792?

|||| Solution

The p-value is from the F-test from a two-way ANOVA using the $F(3,3)$-distribution.

Hence the correct answer is:

The F-distribution with the degrees of freedom $\nu_1 = 3$ and $\nu_2 = 3$.

b) What is the most correct conclusion based on the analysis among the following options (use $\alpha = 0.05$)?

1) A significant difference has been found between the variances from the analysis of variance

2) A significant difference has been found between the means for the 2 granules but not for the 4 suppliers

3) No significant difference has been found between the means for neither the 4 suppliers nor the 2 granules

4) A significant difference has been found between the means for as well the 4 suppliers as the 2 granules

5) A significant difference has been found between the means for the 4 suppliers but not for the 2 granules

|||| Solution

Since both of the p-values are larger than 0.05 none of the two usual hypothesis tests (of no group difference) are significant. So the correct answer is:

3) No significant difference has been found between the means for neither the 4 suppliers nor the 2 granules

8.6 Joining methods

|||| Exercise 8.6 Joining methods

To compare alternative joining methods and materials a series of experiments are now performed where three different joining methods and four different choices of materials are compared.

Data from the experiment are shown in the table below:

	Material				Row
Joining methods	1	2	3	4	average
A	242	214	254	248	239.50
B	248	214	248	247	239.25
C	236	211	245	243	233.75
Column average	242	213	249	246	

In a Python-run for two-way analysis of variance:

```
D = pd.DataFrame({
    "Strength": [242,214,254,248,248,214,248,247,236,211,245,243],
    "Joiningmethod": pd.Categorical(["A","A","A","A",
                                     "B","B","B","B",
                                     "C","C","C","C"]),
    "Material": pd.Categorical([1,2,3,4,1,2,3,4,1,2,3,4])
})
fit = smf.ols("Strength ~ Joiningmethod + Material", data=D).fit()
print(sm.stats.anova_lm(fit))
```

the following output is generated (where some of the values are replaced by the symbols A, B, C, D, E and F):

```
                          sum_sq mean_sq  F_val    PR(>F)
Joiningmethod          A    84.5       B  C        0.05041 .
Material               D       E  825.00  F        1.637e-05
Residuals              6    49.5    8.25
```

a) What are the values for A, B and C?

|||| Solution

A is the degrees of freedom for `Joiningmethod`, wich is the number of groups/levels minus 1: A=2 (and then actually the question can allready be answered). BUT also:

$$\text{B} = MS(\text{Joiningmethod}) = 84.5/2 = 84.5/2 = 42.25,$$

and

$$\text{C} = \frac{MS(\text{Joiningmethod})}{MSE} = \frac{42.25}{8.25} = 5.12.$$

So the answer becomes:

A=2, B=42.25 and C=5.12.

b) What are the conclusions concerning the importance of the two factors in the experiment (using the usual level $\alpha = 5\%$)?

|||| Solution

We can read off the answer from the two p-values given in the output - one of them is below α (Material p-value) and one is NOT (Joiningmethod p-value).

So the answer is:

Significant differences between materials can be concluded, but not between joining methods.

c) Do post hoc analysis for as well the Materials as Joining methods (Confidence intervals for pairwise differences and/or hypothesis tests for those differences).

|||| Solution

First we find the treatment and block means (and we print the ANOVA table):

```
D = pd.DataFrame({
    "Strength": [242,214,254,248,248,214,248,247,236,211,245,243],
    "Joiningmethod": pd.Categorical(["A","A","A","A",
                                     "B","B","B","B",
                                     "C","C","C","C"]),
    "Material": pd.Categorical([1,2,3,4,1,2,3,4,1,2,3,4])
})

fit = smf.ols("Strength ~ Joiningmethod + Material", data=D).fit()
print(sm.stats.anova_lm(fit))
```

```
                df  sum_sq  mean_sq           F    PR(>F)
Joiningmethod  2.0    84.5    42.25    5.121212  0.050408
Material       3.0  2475.0   825.00  100.000000  0.000016
Residual       6.0    49.5     8.25         NaN       NaN
```

```
print(D.groupby("Joiningmethod")["Strength"].mean())
```

```
Joiningmethod
A    239.50
B    239.25
C    233.75
Name: Strength, dtype: float64

<string>:1: FutureWarning: The default of observed=False is deprecated and will be chan
```

```
print(D.groupby("Material")["Strength"].mean())
```

```
Material
1    242.0
2    213.0
3    249.0
4    246.0
Name: Strength, dtype: float64
```

We can find the 0.05/3 (Bonferroni-corrected) *LSD*-value from the two-way version of Remark Remark **??** (see Section **??**) for the comparison of the 3 `Joiningmethods`:

```
LSD_bonf = stats.t.ppf(1 - 0.05 / 6, 6) * np.sqrt(2 * 8.25 / 4)
print(LSD_bonf)

6.676852987425712
```

We see that none of the three Joining methods are different from each other (although close), which matches fine with the p-value just above 0.05.

And we then do the same for the 4 Materials (that is, 6 pairwise comparisons): We can find the 0.05/6 (Bonferroni-corrected) LSD-value from the two-way version of Remark 8.13 for the comparison of the 4 Materials:

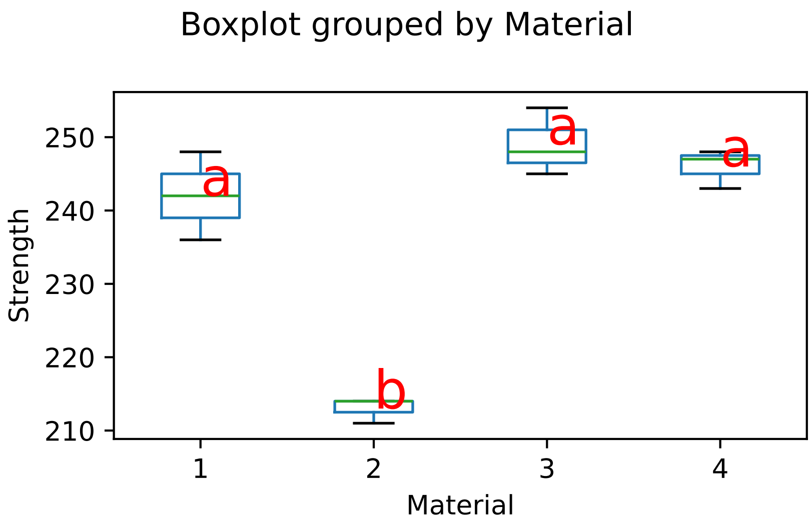
```
LSD_bonf = stats.t.ppf(1 - 0.05 / 12, 6) * np.sqrt(2 * 8.25 / 3)
print(LSD_bonf)
```

9.059516029948194

Solution

So we see that Material 2 is significantly smaller than each of the other three but none of these 3 are significantly different from each other:

```
D.boxplot(column=’Strength’, by=’Material’, grid=False)
plt.title("")
plt.ylabel(’Strength’)
letters = [’a’, ’b’, ’a’, ’a’]
# Add text on top of each boxplot
for i, letter in enumerate(letters):
    plt.text(i+1, D.groupby(’Material’)[’Strength’].mean().values[i],
             letter, fontsize=20, color=’red’)
plt.tight_layout()
plt.show()
```



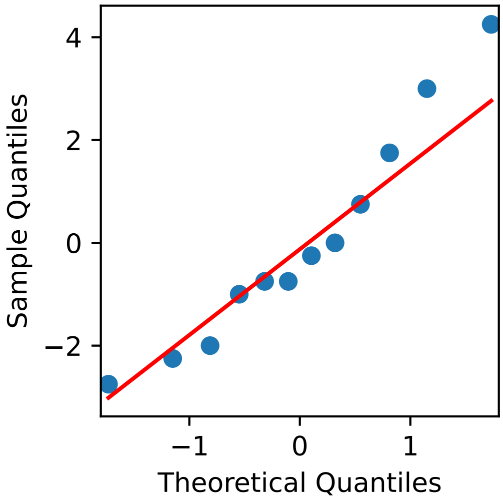
d) Do residual analysis to check for the assumptions of the model:

1. Normality
2. Variance homogeneity

|||| Solution

First the residual normality plot:

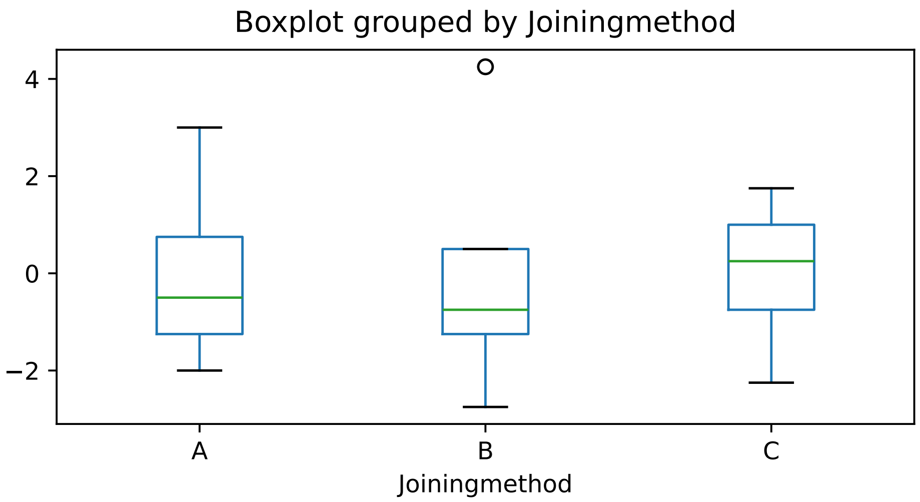
```
sm.qqplot(fit.resid.values, line="q",a=1/2)
plt.tight_layout()
plt.show()
```



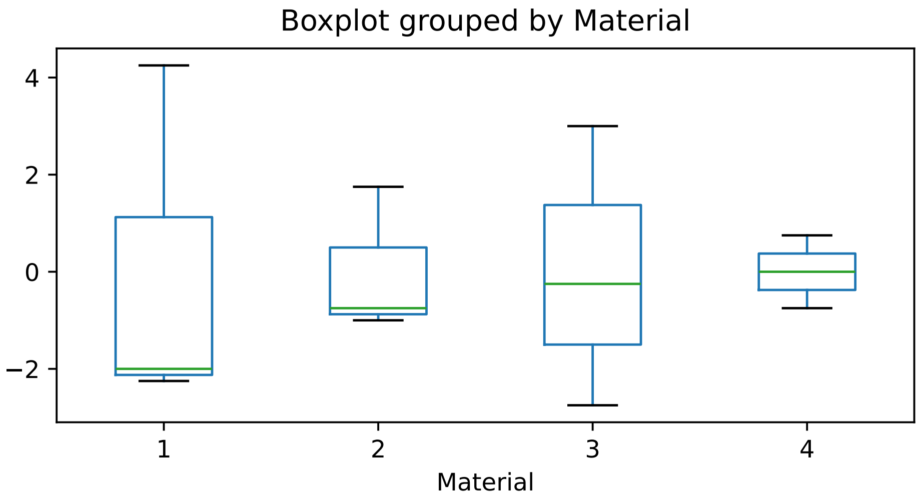
|||| Solution

Then the investigation of variance homogeneity:

```
D['residuals'] = fit.resid.values # add residuals to the data frame
D.boxplot(column='residuals', by='Joiningmethod', grid=False)
plt.title('')
plt.show()
```



```
D.boxplot(column='residuals', by='Material', grid=False)
plt.title('')
plt.show()
```



There may some indications of lower variability within Materials 2 and 4 compared to 1 and 3 (We do not, however, have the methodology (i.e. a test of difference in variance) in the course to deal with this).

8.7 Remoulade

Exercise 8.7 Remoulade

A supermarket has just opened a delicacy department wanting to make its own homemade "remoulade" (a Danish delicacy consisting of a certain mixture of pickles and dressing). In order to find the best recipe a taste experiment was conducted. 4 different kinds of dressing and 3 different types of pickles were used in the test. Taste evaluation of the individual "remoulade" versions were carried out on a continuous scale from 0 to 5.

The following measurement data were found:

	Dressing type				Row
Pickles type	A	B	C	D	average
I	4.0	3.0	3.8	2.4	3.30
II	4.3	3.1	3.3	1.9	3.15
III	3.9	2.3	3.0	2.4	2.90
Column average	4.06	2.80	3.36	2.23	

In a Python-run for twoway ANOVA:

```
fit = smf.ols('Taste ~ Pickles + Dressing', data=D).fit()
print(sm.stats.anova_lm(fit))
```

```
                 sum_sq    mean_sq      F_val      PR(>F)
Pickles.     A  0.326667  0.163333        E      0.287133
Dressing     B  5.536667  1.845556        F      0.002273
Residual     C       D    0.105556
```

a) What are the values of A, B, and C?

Solution

From the general definition of the two-way ANOVA table (see page **??** of Chapter **??**) the degrees of freedom are $k-1$, $l-1$ and $(k-1)(l-1)$, where $k=3$ is the number of rows, $l=4$ is the number of columns.

So the answer is:

A=2, B=3 and C=6.

b) What are the values of D, E, and F?

|||| Solution

E and F are the observed F-statistics, which are:

$$F_{\text{obs,Pickles}} = \frac{MS_{\text{Pickles}}}{MSE} = \frac{0.16333}{0.10556} = 1.547$$
$$F_{\text{obs,Dressing}} = \frac{MS_{\text{Dressing}}}{MSE} = \frac{1.84556}{0.10556} = 17.48$$

Actually, only one answer option has these two values. The D= SSE could be found from the total sum of squares

$$SST = \sum_{i=1}^{3}\sum_{j=1}^{4}(y_{ij} - 2.23)^2,$$

and then

$$D = SSE = SST - 0.3267 - 5.5367.$$

Or more easily using that the degrees of freedom ($= (r-1)(b-1) = 6$) and then

$$D = SSE = 6 \cdot MSE = 6 \cdot 0.10556 = 0.633.$$

In any case, the answer is:

$D = 0.633$, $E = 1.55$ and $F = 17.48$

c) With a test level of $\alpha = 5\%$ the conclusion of the analysis, what is the conclusion of the tests?

|||| Solution

We look at the P-values in the ANOVA table, and observe that the Dressing p-value is BELOW 0.05 and the Pickles p-value is ABOVE 0.05, and hence the answer is:

Only the choice of the dressing type has a significant influence on the taste.

8.8 Transport times

|||| Exercise 8.8 Transport times

In a study the transport delivery times for three transport firms are compared, and the study also involves the size of the transported item. For delivery times in days, the following data found:

	The size of the item			Row
	Small	Intermediate	Large	average
Firm A	1.4	2.5	2.1	2.00
Firm B	0.8	1.8	1.9	1.50
Firm C	1.6	2.0	2.4	2.00
Coulumn average	1.27	2.10	2.13	

In Python was run:

```
fit = smf.ols('Time ~ Firm + Itemsize', data=D).fit()
print(sm.stats.anova_lm(fit))
```

and the following output was obtained: (wherein some of the values, however, has been replaced by the symbols A, B, C and D)

```
             df   sum_sq  mean_sq      F      PR(>F)
Firm        2.0      A        B      4.2857  0.10124
Itemsize    2.0  1.44667     C         D     0.01929
Residual    4.0  0.23333  0.05833
```

a) What is A, B, C and D?

|||| Solution

We have a two-way ANOVA situation. The definition of the terms in the given ANOVA table can all be found in Section **??**, as:

$$C = \frac{SS(Bl)}{DF(Bl)} = \frac{1.44667}{2} = 0.723,$$

$$D = \frac{MS(Bl)}{MSE} = \frac{C}{MSE} = \frac{0.723}{0.05833} = 12.4.$$

And since B is given easily by A:

$$B = \frac{SS(Tr)}{DF(Tr)} = \frac{A}{2},$$

where $DF(Tr)$ denotes the degrees of freedom for the treatment Firm. We just to find $A = SS(Tr)$. We could use the defining formula, and that the overall average is $\bar{y}_{\cdot\cdot} = 5.5/3 = 1.83$

$$A = SS(Tr) = 3 \cdot (2 - 1.83)^2 + 3 \cdot (1.5 - 1.83)^2 + 3 \cdot (2 - 1.83)^2 = 0.5,$$

but more easily we could find B from the F-value as

$$B = 4.2857 \cdot 0.05833 = 0.25,$$

and then

$$A = 2, B = 0.5.$$

Hence the correct answer is:

A = 0.5, B = 0.25, C =0.723 and D = 12.4.

b) What is the conclusion of the analysis (with a significance level of 5%)?

|||| Solution

We look at the two p-values, and see that the Itemsize p-value is less than 5% ("groups significantly different") and that the Firm p-value is NOT ("groups NOT significantly different") and hence the correct answer is:

Only the size of the item has a significant influence on the delivery time.