Chapter 8

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

Contents

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 mean	5.5517	5.1900	4.0483

Further, the total variation in the data is $SST = 11.4944$. You got the following output from R 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.olm('N ~ 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?

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)?

c) Can you with these data demonstrate statistically significant (at significance level $\alpha = 0.05$) differences in N-mean values from year to year (both conclusion and argument must be valid)?

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

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

The data can be read into Python and the means and variances computed by the following in Python:

```
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))
df = pd.DataFrame({"nitrogen": nitrogen, "year": year})
print(df.groupby("year")["nitrogen"].mean())

<string>:1: FutureWarning: The default of observed=False is deprecated and will be
year
1998    5.551667
2003    5.190000
```

```
2011    4.048333
Name: nitrogen, dtype: float64

print(df.groupby("year")["nitrogen"].var())

year
1998    0.236577
2003    0.131320
2011    0.453297
Name: nitrogen, dtype: float64

print(df["nitrogen"].mean())

4.930000000000001
```

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).

b) Find the SST-value in Python using the sample variance function `var`.

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

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

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

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 (both conclusion and argument must be correct)?

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).

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

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

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

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

```
              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?

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

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

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?

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

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?

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

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).

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

   1. Normality
   2. Variance homogeneity

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))
```

```
            D     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   0.633333  0.105556
```

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

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

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

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?

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