Chapter 8

Comparing means of multiple groups - ANOVA

Contents

8.1 Introduction

In Chapter 3 the test of difference in mean of two groups was introduced

$$H_0: \quad \mu_1 - \mu_2 = \delta_0. \tag{8-1}$$

Often we are interested in testing if the mean of the two groups are different ($H_0 : \mu_1 = \mu_2$), against the alternative ($\mu_1 \neq \mu_2$). Often we will face a situation where we have data in multiple (more than two) groups leading to the natural extension of the two-sample situation to a multi-sample situation. The hypothesis of k groups having the same means can then be expressed as

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_k. \tag{8-2}$$

Or in words we have k groups (often referred to as treatments) and we want to test if they all have the same mean against the alternative that at least one group is different from the other groups. Note, that the hypothesis is not expressing any particular values for the means, but just that they are all the same.

The purpose of the data analysis in such a multi-group situation can be expressed as a two-fold purpose:

1. Answer the question: are the group means (significantly) different (hypothesis test)?
2. Tell the story about (or "quantify") the groups and their potential differences (estimates and confidence intervals)

The statistical analysis used for such an analysis is called one-way Analysis of Variance (ANOVA). Though there is an initial contradiction in the name, as ANOVA is used to compare the means of populations and not their variances, the name should not be met with confusion. An ANOVA expresses how different the means of k populations are by measuring how much of the variance in data is explained by grouping the observations (in other words: the variance explained by fitting a model with a mean for each population). If enough of the variation is explained, then a significant difference in population means can be concluded.

The one-way ANOVA is the natural multi-sample extension of the independent two-sample setup covered in Chapter 3. We will also present a natural multi-sample extension of the two paired-sample situation from Chapter 3. This generalization, where the k samples are somehow dependent, e.g. if the same individuals are used in each of the groups, is called two-way ANOVA.

8.2 One-way ANOVA

8.2.1 Data structure and model

As mentioned above we assume that we have data from k groups, also assume n_i repetitions in group (i), this imply that we can order data in a table like:

Tr_1	y_{11}	$\cdots$	y_{1,n_1}
$\vdots$	$\vdots$	$\ldots$	
Tr_k	$y_{k,1}$	$\cdots$	y_{k,n_k}

The total number of observations is $n = \sum_{i=1}^{k} n_i$, note that there does not have to be the same number of observations within each group (treatment).

As for the two-sample case in Chapter 3 there are some standard assumptions that are usually made in order for the methods to come to be 100% valid. In the case of one-way ANOVA, these assumptions are expressed by formulating a "model" much like how regression models in Chapters 5 and 6 are expressed

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2). \tag{8-3}$$

The model is expressing that the observations come from a normal distribution within each group, that each group (i) has a specific mean, and that the variance is the same (σ^2) for all groups. Further, we see explicitly that we have a number of observations (n_i) within each group $(j = 1, \ldots, n_i)$.

As noted above the relevant hypothesis to fulfil the first purpose of the analysis is that of equal group means (8-2). It turns out that a slight modification of (8-3) is convenient

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2). \tag{8-4}$$

Now, the situation is described with a μ that corresponds to the overall mean (across all groups), and then $\alpha_i = \mu_i - \mu$ is the difference between each group mean and the overall mean. The individual group mean is then $\mu_i = \mu + \alpha_i$, and the null hypothesis is expressed as

$$H_0 : \alpha_1 = \cdots = \alpha_k = 0, \tag{8-5}$$

with the alternative $H_1 : \alpha_i \neq 0$ for at least one i. The concept is illustrated in Figure 8.1 (for $k = 3$), the black dots are the measurements y_{ij}, the red line is the overall average, red dots are the average within each group, and the blue lines are the difference between group average and the overall average $(\hat{\alpha}_i)$.

Let's have a look at an example, before we discuss the analysis in further details.

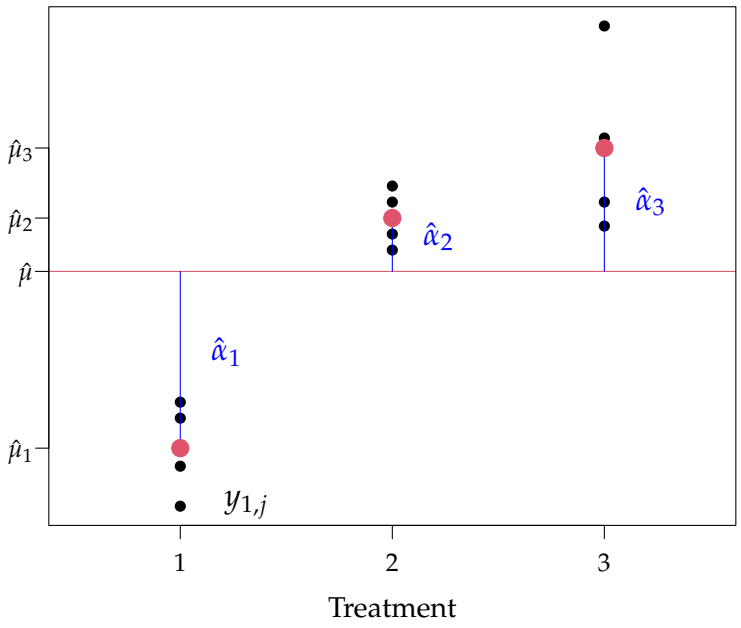


Figure 8.1: Conceptual plot for the ANOVA problem.

|||| Example 8.1 Basic example

The data used for Figure 8.1 is given by:

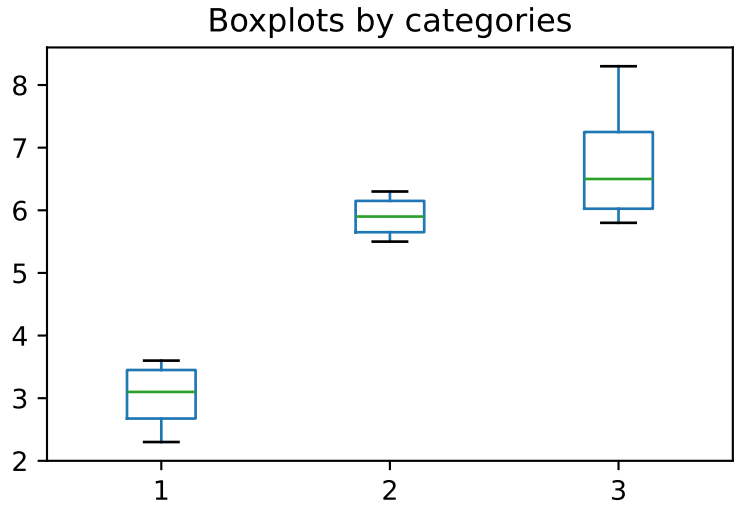
Group A	Group B	Group C
2.8	5.5	5.8
3.6	6.3	8.3
3.4	6.1	6.9
2.3	5.7	6.1

The question is of course: is there a difference in the means of the groups (A, B and C)? We start by having a look at the observations:

```
y = np.array([2.8, 3.6, 3.4, 2.3,
              5.5, 6.3, 6.1, 5.7,
              5.8, 8.3, 6.9, 6.1])
treatm = pd.Categorical([1, 1, 1, 1,
                         2, 2, 2, 2,
                         3, 3, 3, 3])

D = pd.DataFrame({'y': y, 'treatm': treatm})

D.boxplot(by='treatm', grid=False)
plt.title('Boxplots by categories')
plt.suptitle('') # Removing automatic titles
plt.xlabel('')
plt.show()
```



By using `pd.Categorical` the treatments are not considered as numerical values by Python, but rather as factors (or grouping variables), and we can get the boxplot of the within group variation. This plot gives information about the location of data and variance homogeneity (the model assumption), of course with only 4 observations in each group it is difficult to asses this assumption.

Now we can calculate the parameter estimates ($\hat{\mu}$ and $\hat{\alpha}_i$) by:

```
mu = np.mean(y)
muis = D.groupby('treatm',observed=True)['y'].mean()
alpha = muis - mu
print(mu)

5.233333333333333


print(muis)

treatm
1    3.025
2    5.900
3    6.775
Name: y, dtype: float64


print(alpha)

treatm
1   -2.208333
2    0.666667
3    1.541667
Name: y, dtype: float64
```

So our estimate of the overall mean is $\hat{\mu} = 5.23$, and the group levels (offsets from the overall sample mean) are $\hat{\alpha}_1 = -2.21$, $\hat{\alpha}_2 = 0.67$ and $\hat{\alpha}_3 = 1.54$. The question we need to answer is: how likely is it that the observed differences in group means are random variation? If this is very unlikely, then it can be concluded that at least one of them is significantly different from zero.

The shown use of the pandas function `groupby` function is a convenient way of finding the `mean` of `y` for each level of the factor `treatm`. By the way if the `mean` is substituted by any other function, e.g. the `variance`, we could similarly find the sample variance within each group (we will have a closer look at these later):

```
D.groupby('treatm',observed=True)['y'].var()

treatm
1    0.349167
2    0.133333
3    1.249167
Name: y, dtype: float64
```

8.2.2 Decomposition of variability, the ANOVA table

A characteristic of ANOVA in general and one-way ANOVA specifically is the fact that the overall variability (measured by the total variation) decomposes into interpretable components – it is these components which are used for hypothesis testing and more. For the one-way ANOVA presented in this section the total variation, that is, the variation calculated across all the data completely ignoring the fact that the data falls in different groups, can be decomposed into two components: a component expressing the group differences and a component expressing the (average) variation within the groups:

Theorem 8.2 Variability decomposition

The total sum of squares (SST) can be decomposed into sum of squared errors (SSE) and treatment sum of squares ($SS(Tr)$)

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2}_{\text{SS(Tr)}}, \tag{8-6}$$

where

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{k}\sum_{j=1}^{n_i} y_{ij}, \quad \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}. \tag{8-7}$$

Expressed in short form

$$SST = SS(Tr) + SSE. \tag{8-8}$$

Before we turn to the proof of the theorem, we will briefly discuss some interpretations and implications of this. First we look at each of the three terms separately.

The SST expresses the *total variation*. Let us compare with Equation (1-6) the formula for sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2. \tag{8-9}$$

We can see that if the sample variance formula is applied to the the y_{ij}s joined into a single sample (i.e. a single index counts through all the n observations),

then the sample variance is simply SST divided by $n-1$. The sample variance expresses then the *average variation per observation*. Therefore, we have

$$SST = (n-1) \cdot s_y^2, \tag{8-10}$$

where s_y^2 is the sample variance for all the y_{ij}s seen as a single sample (i.e. a sample from single population).

The group mean differences are quantified by the $SS(Tr)$ component, which can basically be seen directly from the definition, where the overall mean is subtracted from each group mean. As discussed above it can alternatively be expressed by deviations $\hat{\alpha}_i$

$$SS(Tr) = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{k} n_i \hat{\alpha}_i^2, \tag{8-11}$$

so $SS(Tr)$ is the sum of squared α_i's multiplied by the number of observations in group n_i.

|||| Remark 8.3

$SS(Tr)$ is also the key expression to get the idea of why we call the whole thing "analysis of variance": if we, for a second, assume that we have the same number of observations in each group: $n_1 = \ldots = n_k$, and let us call this common number m. Then we can express $SS(Tr)$ directly in terms of the variance of the k means

$$SS(Tr) = (k-1) \cdot m \cdot s_{\text{means}}^2, \tag{8-12}$$

where

$$s_{\text{means}}^2 = \frac{1}{k-1} \sum_{i=1}^{k} (\bar{y}_i - \bar{y})^2. \tag{8-13}$$

Let us emphasize that the formulas of this remark is not thought to be formulas that we use for practical purposes, but they are expressed to show explicitly that "$SS(Tr)$ quantifies the group differences by variation". Another way of thinking of $SS(Tr)$ is that it quantifies the "the variance explained by grouping the observations", i.e. the variance explained by fitting a model with a mean for each group.

Finally, SSE expresses the average variability within each group, as each individual observation y_{ij} is compared with the mean in the group to which it

belongs. In Figure 8.1 these are the differences between each of the black dots with the relevant read dot. Again we can link the formula given above to basic uses of the sample variance formula:

|||| Theorem 8.4 Within group variability

The sum of squared errors SSE divided by $n-k$, also called the residual mean square $MSE = SSE/(n-k)$ is the weighted average of the sample variances from each group

$$MSE = \frac{SSE}{n-k} = \frac{(n_1-1)s_1^2 + \cdots + (n_k-1)s_k^2}{n-k}, \tag{8-14}$$

where s_i^2 is the variance within the ith group

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2. \tag{8-15}$$

When $k = 2$, that is, we are in the two-sample case presented in Section 3.2, the result here is a copy of the pooled variance expression in Method 3.52

$$\text{For } k = 2: \; MSE = s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n-2}. \tag{8-16}$$

Intuitively, we would say that if some of the $\hat{\alpha}_i$'s are large (in absolute terms), then it is evidence against the null hypothesis of equal means. So a large $SS(Tr)$ value is evidence against the null hypothesis. It is also natural that "large" should be relative to some variation. SSE is the within group variation, and it also seems reasonable that if $\hat{\alpha}_i$ is large and the variation around $\hat{\mu}_i$ is small then this is evidence against the null hypothesis. It does therefore seem natural to compare $SS(Tr)$ and SSE, and we will get back to the question of exactly how to do this after the proof of Theorem 8.2:

Proof

Add and subtract $\bar{y}_i$ in SST to get

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i+\bar{y}_i-\bar{y})^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(y_{ij}-\bar{y}_i)^2+(\bar{y}_i-\bar{y})^2+2(y_{ij}-\bar{y}_i)(\bar{y}_i-\bar{y})\right] \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i-\bar{y})^2+2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)(\bar{y}_i-\bar{y}) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2+\sum_{i=1}^{k}n_i(\bar{y}_i-\bar{y})^2+2\sum_{i=1}^{k}(\bar{y}_i-\bar{y})\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i),\end{aligned} \tag{8-17}$$

now observe that $\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)=0$, and the proof is completed.

■

Example 8.5

We can now continue our example and calculate SST, SSE, and $SS(Tr)$:

```
muis = muis.values # Coverting to numpy array
alpha = muis - mu
SST = np.sum((y - mu)**2)
SSE = (np.sum((y[treatm == 1] - muis[0])**2) +
       np.sum((y[treatm == 2] - muis[1])**2) +
       np.sum((y[treatm == 3] - muis[2])**2))
SSTr = 4 * np.sum(alpha**2)
print(np.round([SST, SSE, SSTr],3))

[35.987  5.195 30.792]
```

For these data we have that $n_1 = n_2 = n_3 = 4$, so according to Theorem 8.2 we could also find SSE from the average of the variances within each group:

```
vars = D.groupby('treatm',observed=True)['y'].var()
print((12 - 3) * np.mean(vars))

5.195000000000002
```

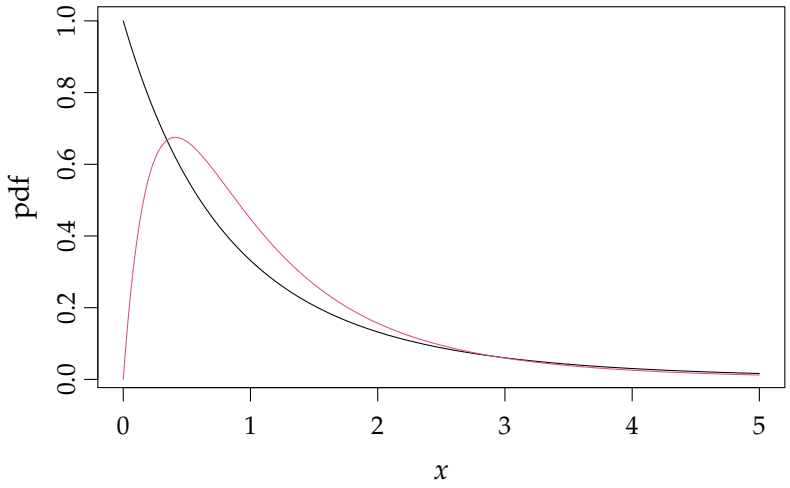


Figure 8.2: pdf of the F-distribution with 2 and 9 degrees of freedom (black line), and with 4 and 9 degrees of freedom (red line).

Now we have established that we should compare $SS(Tr)$ and SSE in some way, it should of course be quantified exactly in which way they should be compared. Now it turns out that the numbers $SS(Tr)/(k-1)$ and $SSE/(n-k)$ are both central estimators for σ^2, when the null hypothesis is true, and we can state the following theorem:

|||| Theorem 8.6

Under the null hypothesis

$$H_0: \quad \alpha_i = 0, \quad i = 1, 2, \ldots, k, \tag{8-18}$$

the test statistic

$$F = \frac{SS(Tr)/(k-1)}{SSE/(n-k)}, \tag{8-19}$$

follows an F-distribution with $k-1$ and $n-k$ degrees of freedom.

The F-distribution is generated by the ratio between independent χ^2 distributed random variables, and the shape is shown in Figure 8.2 for two particular choices of degrees of freedom.

As we have discussed before, the null hypothesis should be rejected if $SS(Tr)$ is large and SSE is small. This implies that we should reject the null hypothesis

when the test statistic (F) is large in the sense of Theorem 8.6 (compare with $F_{1-\alpha}$). The statistics are usually collected in an ANOVA table like this:

Source of variation	Degrees of freedom	Sums of squares	Mean sum of squares	Test-statistic F	p-value
Treatment	$k-1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{\text{obs}} = \frac{MS(Tr)}{MSE}$	$P(F > F_{\text{obs}})$
Residual	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	SST			

|||| Example 8.7

We can now continue with our example and find the F-statistic and the p-value:

```
F = (SSTr / (3 - 1)) / (SSE / (12 - 3))
pv = 1 - stats.f.cdf(F, 3 - 1, 12 - 3)
print(F, pv)

26.67228103946101 0.0001650052218172826
```

So we have a test statistic $F = 26.7$ and a p-value equal to 0.000165 and we reject the null hypothesis on e.g. level $\alpha = 0.05$. The calculations can of course also be done directly in Python, by:

```
fit = smf.ols('y ~ treatm', data=D).fit()
anova = sm.stats.anova_lm(fit)
print(anova)

             df      sum_sq     mean_sq          F    PR(>F)
treatm      2.0  30.791667  15.395833  26.672281  0.000165
Residual    9.0   5.195000   0.577222        NaN       NaN
```

Note, that in the direct Python calculation it is very important to include `treatm` as a factor (categorical), in order to get the correct analysis.

If we reject the null hypothesis, it implies that the observations can be finally described by the initial model re-stated here

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2), \tag{8-20}$$

and the estimate of the error variance σ^2 is $\hat{\sigma}^2 = SSE/(n-k) = MSE$.

|||| Remark 8.8 When multiple groups = 2 groups

When $k = 2$, that is, we are in the two-sample case studied in Chapter 3, we already saw above in Theorem 8.4 that $MSE = s_p^2$. Actually, this means that the analysis we get from a one-way ANOVA when we apply it for only $k = 2$ groups, which could be perfectly fine - nothing in the ANOVA approach really relies on k having to be larger than 2 - corresponds to the pooled t-test given as Method 3.53. More exact

$$\text{for } k = 2: \ F_{\text{obs}} = t_{\text{obs}}^2, \tag{8-21}$$

where t_{obs} is the pooled version coming from Methods 3.52 and 3.53. Thus the p-value obtained from the $k = 2$ group ANOVA equals exactly the p-value obtained from the pooled t-test given in Method 3.53.

8.2.3 Post hoc comparisons

If we reject the overall null hypothesis above, and hence conclude that $\alpha_i \neq 0$ for at least one i it makes sense to ask which of the treatments are actually different. That is, trying to meet the second of the two major purposes indicated in the beginning. This can be done by pairwise comparison of the treatments. We have already seen in Chapter 3, that such comparison could be based on the t-distribution. We can construct confidence interval with similar formulas except that we should use $MSE = SSE/(n-k)$ as the estimate of the error variance and hence also $n - k$ degrees of freedom in the t-distribution:

Method 8.9 Post hoc pairwise confidence intervals

A single pre-planned $(1-\alpha)\cdot 100\%$ confidence interval for the difference between treatment i and j is found as

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}, \tag{8-22}$$

where $t_{1-\alpha/2}$ is based on the t-distribution with $n-k$ degrees of freedom.

If all $M = k(k-1)/2$ combinations of pairwise confidence intervals are calculated using the formula M times, but each time with $\alpha_{\text{Bonferroni}} = \alpha/M$ (see Remark 8.14 below).

Similarly one could do pairwise hypothesis tests:

Method 8.10 Post hoc pairwise hypothesis tests

A single pre-planned level α hypothesis tests

$$H_0:\ \mu_i = \mu_j,\quad H_1:\ \mu_i \neq \mu_j, \tag{8-23}$$

is carried out by

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}}, \tag{8-24}$$

and

$$p\text{-value} = 2\cdot P(T > |t_{\text{obs}}|), \tag{8-25}$$

where the t-distribution with $n-k$ degrees of freedom is used.

If all $M = k(k-1)/2$ combinations of pairwise hypothesis tests are carried out use the approach M times but each time with test level $\alpha_{\text{Bonferroni}} = \alpha/M$ (see Remark 8.14 below).

Example 8.11

Returning to our small example we get the pairwise confidence intervals. If the comparison of A and B was specifically planned before the experiment was carried out, we would find the 95%-confidence interval as:

```
print(muis[0] - muis[1] + np.array([-1, 1]) *
       stats.t.ppf(1 - 0.05 / 2, 12 - 3) * np.sqrt(SSE / (12 - 3) *
(1/4 + 1/4)))

[-4.090 -1.660]
```

and we can hence also conclude that treatment A is different from B. The p-value supporting this claim is found as:

```
tobs = (muis[0] - muis[1]) / np.sqrt(SSE / (12 - 3) * (1/4 + 1/4))
print(2 * (1 - stats.t.cdf(np.abs(tobs), 12 - 3)))

0.0004613963065729365
```

If we do all three possible comparisons, $M = 3 \cdot 2/2 = 3$, and we will use an overall $\alpha = 0.05$, we do the above three times, but using each time $\alpha_{\text{Bonferroni}} = 0.05/3 = 0.016667$:

```
alpha_bonf = 0.05 / 3
# A-B
print(alpha[0] - alpha[1] + np.array([-1, 1]) *
      stats.t.ppf(1-alpha_bonf/2, 12 - 3) * np.sqrt(SSE/(12 - 3) * (1/4
+ 1/4)))

[-4.451 -1.299]


# A-C
print(alpha[0] - alpha[2] + np.array([-1, 1]) *
      stats.t.ppf(1-alpha_bonf/2, 12 - 3) * np.sqrt(SSE/(12 - 3) * (1/4
+ 1/4)))

[-5.326 -2.174]


# B-C
print(alpha[1] - alpha[2] + np.array([-1, 1]) *
      stats.t.ppf(1-alpha_bonf/2, 12 - 3) * np.sqrt(SSE/(12 - 3) * (1/4
+ 1/4)))

[-2.451  0.701]
```

and we conclude that treatment A is different from B and C, while we cannot reject that B and C are equal. The p-values for the last two comparisons could also be found, but we skip that now.

The so-called Bonferroni correction done above, when we do all possible post hoc comparisons, has the effect that it becomes more difficult (than without the correction) to claim that two treatments have different means.

|||| Example 8.12

The 0.05/3-critical value with 9 degrees of freedom is $t_{0.9917} = 2.933$ whereas the 0.05-critical value is $t_{0.975} = 2.262$:

```
print(stats.t.ppf(1 - alpha_bonf / 2, 12 - 3), stats.t.ppf(1 - 0.05 /
2, 12 - 3))

2.9333240883739897 2.2621571628540993
```

So two treatment means would be claimed different WITH the Bonferroni correction if they differ by more than half the width of the confidence interval

$$2.933 \cdot \sqrt{SSE/9 \cdot (1/4 + 1/4)} = 1.576 \tag{8-26}$$

whereas without the Bonferroni correction should only differ by more than

$$2.262 \cdot \sqrt{SSE/9 \cdot (1/4 + 1/4)} = 1.215 \tag{8-27}$$

to be claimed significantly different.

|||| Remark 8.13 Least Significant Difference (LSD) values

If there is the same number of observations in each treatment group $m = n_1 = \ldots = n_k$ the LSD value for a particular significance level

$$LSD_\alpha = t_{1-\alpha/2}\sqrt{2 \cdot MSE/m} \tag{8-28}$$

will have the same value for all the possible comparisons made.
The LSD value is particularly useful as a "measuring stick" with which we can go and compare all the observed means directly: the observed means with difference higher than the LSD are significantly different on the α-level. When used for all of the comparisons, as suggested, one should as level use the Bonferroni corrected version $LSD_{\alpha_{\text{Bonferroni}}}$ (see Remark 8.14 below for an elaborated explanation).

|||| Remark 8.14 Significance by chance in multiple testings!

Imagine that we performed an ANOVA in a situation with $k = 15$ groups. And then we do all the $M = 15 \cdot 14/2 = 105$ possible pairwise hypothesis tests. Assume for a moment that the overall null hypothesis is true, that is, there really are no mean differences between any of the 15 groups. And think about what would happen if we still performed all the 105 tests with $\alpha = 0.05$! How many significant results would we expect among the 105 hypothesis tests? The answer is that we expect $\alpha \cdot 105 = 0.05 \cdot 105 = 5.25$, that is, approximately 5 significant tests are expected. And what would the probability be of getting at least one significant test out of the 105? The answer to this question can be found using the binomial distribution

$$\begin{aligned} &P(\text{"At least one significant result in 105 independent tests"}) \\ &\qquad = 1 - 0.95^{105} \\ &\qquad = 0.9954. \end{aligned} \tag{8-29}$$

So whereas we, when performing a single test, have a probability of $\alpha = 0.05$ of getting a significant result, when we shouldn't, we now have an overall Type I error probability of seeing at least one significant result, when we shouldn't, of 0.9954! This is an extremely high (overall) Type 1 risk. This is also sometimes called the "family wise" Type 1 risk. In other words, we will basically always with $k = 15$ see at least one significant pairwise difference, if we use $\alpha = 0.05$. This is why we recommend to use a correction method when doing multiple testings like this. The Bonferroni correction approach is one out of several different possible approaches for this.

Using the Bonferroni corrected $\alpha_{\text{Bonferroni}} = 0.05/105$ in this case for each of the 105 tests would give the family wise Type 1 risk

$$\begin{aligned} &P(\text{"At least one significant result in 105 independent tests"}) \\ &\qquad = 1 - (1 - 0.05/105)^{105} \\ &\qquad = 0.049 \end{aligned} \tag{8-30}$$

8.2.4 Model control

The assumptions for the analysis we have applied in the one-way ANOVA model are more verbally expressed as:

1. The data comes from a normal distribution in each group
2. The variances from each group are the same

The homogeneous variances assumption can be controlled by simply looking at the distributions within each sample, most conveniently for this purpose by the group-wise box plot already used in the example above.

The normality within groups assumption could in principle also be investigated by looking at the distributions within each group - a direct generalization of what was suggested in Chapter 3 for the two-group setting. That is, one could do a q-q plot within each group. It is not uncommon though, that the amount of data within a single group is too limited for a meaningful q-q plot investigation. Indeed for the example here, we have only 4 observations in each group, and q-q plots for 4 observations do not make much sense.

There is an alternative, where the information from all the groups are pooled together to a single q-q plot. If we pool together the 12 residuals, that is, within group deviations, they should all follow the same zero-mean normal distribution, given by

$$\varepsilon_{ij} \sim N(0, \sigma^2). \tag{8-31}$$

|||| Method 8.15 Normality control in one-way ANOVA

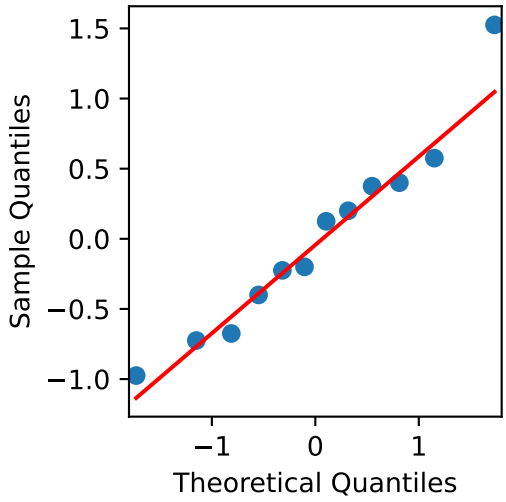
To control for the normality assumptions in one-way ANOVA we perform a q-q plot on the pooled set of n estimated residuals

$$e_{ij} = y_{ij} - \bar{y}_i, \; j = 1, \ldots, n_i, \; i = 1 \ldots, k. \tag{8-32}$$

|||| Example 8.16

For the basic example we get the normal q-q plot of the residuals by

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




```
print(residuals)

0    -0.975
1    -0.725
2    -0.675
3    -0.400
4    -0.225
5    -0.200
6     0.125
7     0.200
8     0.375
9     0.400
10    0.575
11    1.525
dtype: float64
```

8.2.5 A complete worked through example: plastic types for lamps

Example 8.17 Plastic types for lamps

On a lamp two plastic screens are to be mounted. It is essential that these plastic screens have a good impact strength. Therefore an experiment is carried out for 5 different types of plastic. 6 samples in each plastic type are tested. The strengths of

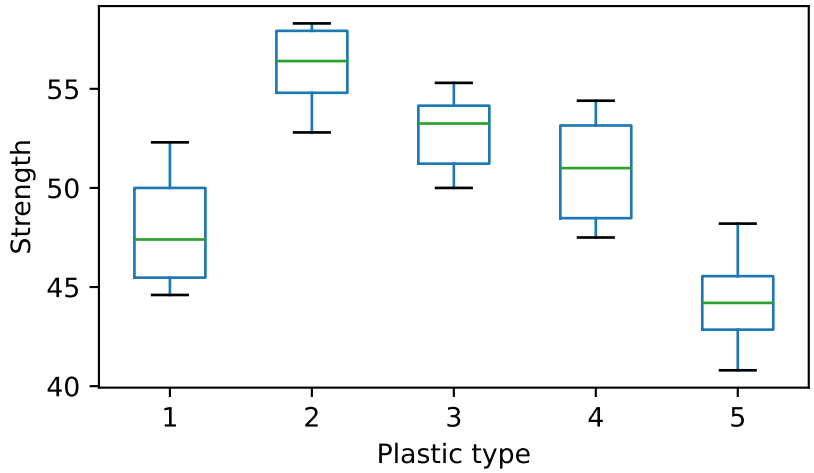
these items are determined. The following measurement data was found (strength in kJ/m^2):

Type of plastic				
I	II	III	IV	V
44.6	52.8	53.1	51.5	48.2
50.5	58.3	50.0	53.7	40.8
46.3	55.4	54.4	50.5	44.5
48.5	57.4	55.3	54.4	43.9
45.2	58.1	50.6	47.5	45.9
52.3	54.6	53.4	47.8	42.5

We run the following in Python:

```
D = pd.DataFrame({
       'strength': [44.6, 52.8, 53.1, 51.5, 48.2, 50.5, 58.3, 50.0,
53.7, 40.8,
                    46.3, 55.4, 54.4, 50.5, 44.5, 48.5, 57.4, 55.3,
54.4, 43.9,
                    45.2, 58.1, 50.6, 47.5, 45.9, 52.3, 54.6, 53.4,
47.8, 42.5],
       'plastictype': pd.Categorical(np.tile(np.arange(1, 6), 6))
})

D.boxplot(by='plastictype', grid=False)
plt.suptitle('') # Removing automatic titles
plt.title('')
plt.xlabel('Plastic type')
plt.ylabel('Strength')
plt.tight_layout()
plt.show()
```




```
fit = smf.ols('strength ~ plastictype', data=D).fit()
anova = sm.stats.anova_lm(fit)
print(anova)

               df   sum_sq   mean_sq          F        PR(>F)
plastictype   4.0  491.76  122.9400  18.233863  3.987701e-07
Residual     25.0  168.56    6.7424        NaN           NaN
```

The ANOVA results are more nicely put in a table here:

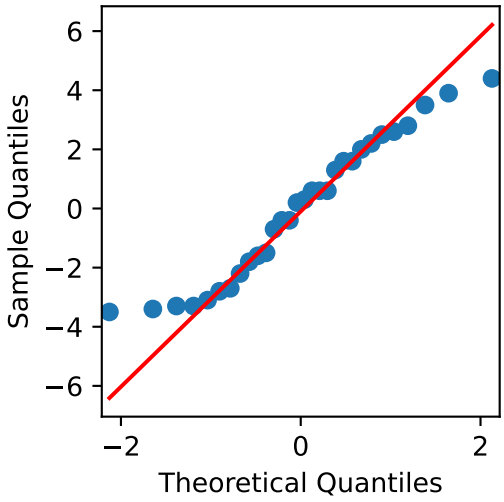
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Plastictype	4	491.76	122.94	18.23	$4 \cdot 10^{-7}$
Residuals	25	168.56	6.74		

From the box plot we see that there appears to be group mean differences and extremely low p-value in the ANOVA table confirms this: there is very strong evidence against the null hypothesis of the five means being the same

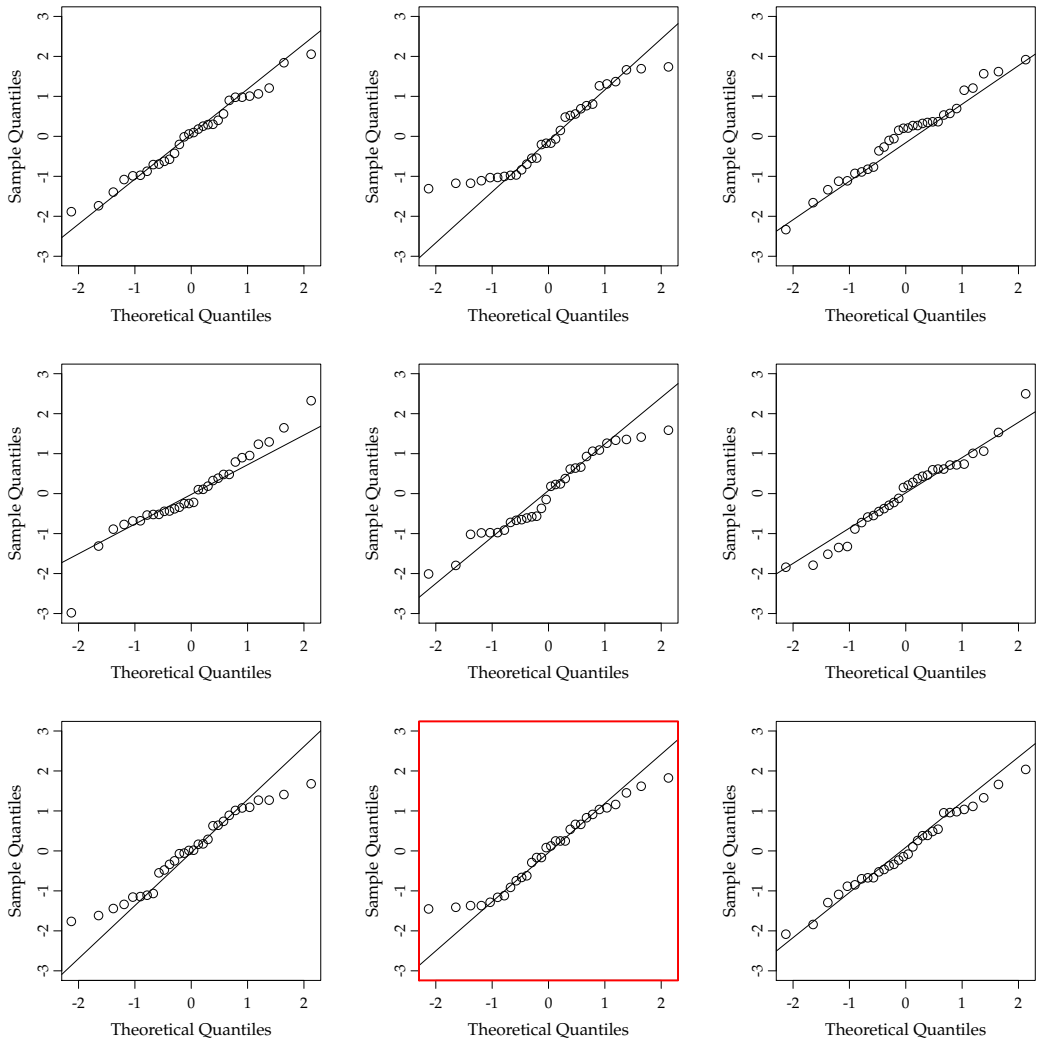
$$H_0 : \mu_1 = \cdots = \mu_5. \tag{8-33}$$

Model assumptions: the box plots do not indicate clear variance differences (although it can be a bit difficult to know exactly how different such patterns should be for it to be a problem. Statistical tests exist for such varicance comparisons, but they are not included here). Let us check for the normality by doing a normal q-q plot on the residuals:

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



Or using the idea of comparing with repeated plots on the standardized residuals: (See Section 3.1.8)



There appears to be no important deviation from normality.

To complete the story about (quantifying) the five plastic types, we first compute the five means:

```
print(D.groupby('plastictype',observed=True)['strength'].mean())

plastictype
1    47.9
2    56.1
3    52.8
4    50.9
5    44.3
Name: strength, dtype: float64
```

And then we want to construct the $M = 5 \cdot 4/2 = 10$ different confidence intervals

using Method 8.9. As all n_is equal 6 in this case, all 10 confidence intervals will have the same width, and we can use Remark 8.13 and compute the (half) width of the confidence intervals, the *LSD*-value. And since there are 10 multiple comparisons we will use $\alpha_{\text{Bonferroni}} = 0.05/10 = 0.005$:

```
LSD_0_005 = stats.t.ppf(1 - 0.005 / 2, 25) * np.sqrt(2 * 6.74 / 6)
print(LSD_0_005)

4.61387770149341
```

So Plastictypes are significantly different from each other if they differ by more than 4.61. A convenient way to collect the information about the 10 comparisons is by ordering the means from smallest to largest and then using the so-called compact letter display:

Plastictype	Mean	
5	44.3	a
1	47.9	ab
4	50.9	bc
3	52.8	cd
2	56.1	d

Plastic types with a mean difference less than the *LSD*-value, hence not significantly different share letters. Plastic types not sharing letters are significantly different. We can hence read off all the 10 comparisons from this table.

One could also add the compact letter information to the box plot for a nice plotting - it is allowed to be creative (while not changing the basic information and the results!) in order to communicate the results.

8.3 Two-way ANOVA

8.3.1 Data structure and model

The one-way ANOVA is the natural multi-sample extension of the independent two-sample situation covered in Section 3.2. The k samples are hence completely independent from each other, which e.g. in a clinical experiment would mean that different patients get different treatments – and hence each patient only tries a single treatment. Often this would be the only possible way to do a comparison of treatments.

However, sometimes it will be possible to apply multiple treatments to the same

patient (with some time in between). This could then lead to a multi-treatment setup, where the sample within each treatment consists of the same patients. This is the natural extension of the paired-design setup covered in Section 3.2.3, where we "pair" if there is exactly 2 treatments. With more than two treatments we use the phrase "block". A block would then be the patient in this case - and the same blocks then appear in all treatment samples. The "block" name comes from the historical background of these methods coming from agricultural field trials, where a block would be an actual piece of land within which all treatments are applied.

|||| Remark 8.18 Design: independent sampling or blocking?

For the project manager who is in charge of designing the study there will be a choice to make in cases where both approaches are practicable feasible: should the independent samples approach or the blocked approach be used? Should we use, say, 4 groups of 20 patients each, that is 80 patients all together, or should we use the same 20 patients in each of the four groups? The costs would probably be more or less the same. It sounds nice with 80 patients rather than 20? However, the answer is actually pretty clear if whatever we are going to measure will vary importantly from person to person. And most things in medical studies do vary a lot from person to person due to many things: gender, age, weight, BMI, or simply due to genetic differences that means that our bodies will respond differently to the medicine. Then the blocked design would definitely be the better choice! This is so, as we will see below, in the analysis of the blocked design the block-main-variability is accounted for and will not "blur" the treatment difference signal. In the independent design the person-to-person variability may be the main part of the "within group" variability used for the statistical analysis. Or differently put: in a block design each patient would act as his/her own control, the treatment comparison is carried out "within the block".

For the actual study design it would in both cases be recommended to randomize the allocation of patients as much as possible: In the independent design patients should be allocated to treatments by randomization. In the block design all patients receive all treatments but then one would randomize the order in which they receive the treatments. For this reason these two types of experimental designs are usually called the *Completely Randomized Design* and the *Randomized Block Design*.

We looked above in the one-way part at an example with 3 treatments with 4 observations for each. If the observations were on 4 different persons (and not 12) it would make sense and would be important to include this person

variability in the model. The resulting model becomes

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2), \tag{8-34}$$

so there is an overall mean μ, a treatment effect α_i and a block effect β_j and our usual random error term ε_{ij}.

The design is illustrated in the table below, so we have k treatments ($A_1, \ldots, A_k$) and l blocks ($B_1, \ldots, B_l$):

	B_1	$\ldots$	B_l
A_1	y_{11}	$\ldots$	$y_{1,l}$
$\vdots$	$\vdots$	$\ldots$	$\vdots$
A_k	$y_{k,1}$	$\ldots$	$y_{k,l}$

We can now find the parameters in the model above by

$$\hat{\mu} = \frac{1}{k \cdot l} \sum_{i=1}^{k} \sum_{j=1}^{l} y_{ij}, \tag{8-35}$$

$$\hat{\alpha}_i = \left(\frac{1}{l} \sum_{j=1}^{l} y_{ij} \right) - \hat{\mu}, \tag{8-36}$$

$$\hat{\beta}_j = \left(\frac{1}{k} \sum_{i=1}^{k} y_{ij} \right) - \hat{\mu}. \tag{8-37}$$

Or expressed more compact, with the definitions of the terms obvious from the above

$$\hat{\mu} = \bar{\bar{y}}, \tag{8-38}$$

$$\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{\bar{y}}, \tag{8-39}$$

$$\hat{\beta}_j = \bar{y}_{\cdot j} - \bar{\bar{y}}. \tag{8-40}$$

In a way, these means are the essential information in these data. All the rest we do is just all the statistics to distinguish signal from noise. It does not change the fact, that these means contain the core story. It also shows explicitly how we now compute means, not only across one way in the data table, but also across the other way. We compute means both row-wise and column-wise. Hence the name: two-way ANOVA.

Example 8.19

Lets assume that the data we used in the previous section was actually a result of a randomized block design and we could therefore write:

	Group A	Group B	Group C
Block 1	2.8	5.5	5.8
Block 2	3.6	6.3	8.3
Block 3	3.4	6.1	6.9
Block 4	2.3	5.7	6.1

In this case we should of course keep track of the blocks as well as the treatments:

```
y = np.array([2.8, 3.6, 3.4, 2.3,
              5.5, 6.3, 6.1, 5.7,
              5.8, 8.3, 6.9, 6.1])

treatm = pd.Categorical([1, 1, 1, 1,
                         2, 2, 2, 2,
                         3, 3, 3, 3])
block = pd.Categorical([1, 2, 3, 4,
                        1, 2, 3, 4,
                        1, 2, 3, 4])

D = pd.DataFrame({'y': y, 'treatm': treatm, 'block': block})
```

Now we can calculate the parameter estimates ($\hat{\mu}$ and $\hat{\alpha}_i$, and $\hat{\beta}_j$):

```
mu = np.mean(y)
alpha = D.groupby('treatm',observed=True)['y'].mean() - mu
beta = D.groupby('block',observed=True)['y'].mean() - mu
print(mu)

5.233333333333333


print(alpha)

treatm
1   -2.208333
2    0.666667
3    1.541667
Name: y, dtype: float64


print(beta)

block
1   -0.533333
2    0.833333
3    0.233333
4   -0.533333
Name: y, dtype: float64
```

so our estimates of the overall mean (μ) and α_i remain the same while the estimated block effects are $\hat{\beta}_1 = -0.53$, $\hat{\beta}_2 = 0.83$, $\hat{\beta}_3 = 0.23$ and $\hat{\beta}_4 = -0.53$.

8.3.2 Decomposition of variability and the ANOVA table

In the same way as we saw for the one-way ANOVA, there exists a decomposition of variation for the two-way ANOVA:

Theorem 8.20 Variation decomposition

The total sum of squares (SST) can be decomposed into sum of squared errors (SSE), treatment sum of squares ($SS(Tr)$), and a block sum of squares ($SS(Bl)$)

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{l}(y_{ij}-\hat{\mu})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{l}(y_{ij}-\hat{\alpha}_i-\hat{\beta}_j-\hat{\mu})^2}_{\text{SSE}} + \underbrace{l\cdot\sum_{i=1}^{k}\hat{\alpha}_i^2}_{\text{SS(Tr)}} + \underbrace{k\cdot\sum_{j=1}^{l}\hat{\beta}_j^2}_{\text{SS(Bl)}}$$

$$= \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{l}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{\bar{y}})^2}_{\text{SSE}} + \underbrace{l\cdot\sum_{i=1}^{k}(\bar{y}_{i\cdot}-\bar{\bar{y}})^2}_{\text{SS(Tr)}} + \underbrace{k\cdot\sum_{j=1}^{l}(\bar{y}_{\cdot j}-\bar{\bar{y}})^2}_{SS(Bl)}, \tag{8-41}$$

Expressed in short form

$$SST = SS(Tr) + SS(Bl) + SSE. \tag{8-42}$$

Note, how the SST and $SS(Tr)$ are found exactly as in the one-way ANOVA. If one ignores the block-way of the table, the two-way data has exactly the same structure as one-way data (with the same number of observations in each group). Further, note how $SS(Bl)$ corresponds to finding a "one-way $SS(Tr)$", but on the other way of the table (and ignoring the treatment-way of the data table). So from a computational point of view, finding these three, that is, finding SST, $SS(Tr)$ and $SS(Bl)$ is done by known one-way methodology. And then the last one, SSE, could then be found from the decomposition theorem as

$$SSE = SST - SS(Tr) - SS(Bl). \tag{8-43}$$

Example 8.21

Returning to the example we get (SST and $SS(Tr)$ remain unchanged):

```
beta = beta.values # Converting to numpy array
SSBl = 3 * np.sum(beta**2)
SSE = SST - SSTr - SSBl
print(np.round([SST, SSE, SSTr, SSBl],3))

[35.987  1.242 30.792  3.953]
```

Again, tests for treatment effects and block effects are done by comparing $SS(Tr)$ or $SS(Bl)$ with SSE:

Theorem 8.22

Under the null hypothesis

$$H_{0,Tr}: \quad \alpha_i = 0, \quad i = 1, 2, \ldots, k, \tag{8-44}$$

the test statistic

$$F_{Tr} = \frac{SS(Tr)/(k-1)}{SSE/((k-1)(l-1))}, \tag{8-45}$$

follows an F-distribution with $k-1$ and $(k-1)(l-1)$ degrees of freedom. Further, under the null hypothesis

$$H_{0,Bl}: \quad \beta_j = 0, \quad j = 1, 2, \ldots, l, \tag{8-46}$$

the test statistic

$$F_{Bl} = \frac{SS(Bl)/(l-1)}{SSE/((k-1)(l-1))}, \tag{8-47}$$

follows an F-distribution with $l-1$ and $(k-1)(l-1)$ degrees of freedom.

Example 8.23

Returning to our example we get:

```
# Test statistics
Ftr = SSTr / (3-1) / ( SSE / ((3-1) * (4-1)))
Fbl = SSBl / (4-1) / ( SSE / ((3-1) * (4-1)))
print(Ftr, Fbl)

74.39597315436248 6.367785234899335


# p-values
pv_tr = 1 - stats.f.cdf(Ftr, 3 - 1, (3 - 1) * (4 - 1))
pv_bl = 1 - stats.f.cdf(Fbl, 4 - 1, (3 - 1) * (4 - 1))
print(pv_tr, pv_bl)

5.823829718287765e-05 0.027048337827318747
```

or directly in Python:

```
fit = smf.ols('y ~ treatm + block', data=D).fit()
anova = sm.stats.anova_lm(fit)
print(anova)

            df     sum_sq    mean_sq          F    PR(>F)
treatm     2.0  30.791667  15.395833  74.395973  0.000058
block      3.0   3.953333   1.317778   6.367785  0.027048
Residual   6.0   1.241667   0.206944        NaN       NaN
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatm	2	30.79	15.40	74.40	0.0001
block	3	3.95	1.32	6.37	0.0270
Residuals	6	1.24	0.21		

we see that the block effect is actually significant on a 5% confidence level, and also that the p-value for the treatment effect is changed (the evidence against $H_{0,Tr}$ is stronger) when we accounted for the block effect.

The test statistics and p-values are often collected in an analysis of variance table as already shown above:

Source of variation	Degrees of freedom	Sums of squares	Mean sums of squares	Test statistic F	p-value
Treatment	$k-1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{Tr} = \frac{MS(Tr)}{MSE}$	$P(F > F_{Tr})$
Block	$l-1$	$SS(Bl)$	$MS(Bl) = \frac{SS(Bl)}{l-1}$	$F_{Bl} = \frac{MS(Bl)}{MSE}$	$P(F > F_{Bl})$
Residual	$(l-1)(k-1)$	SSE	$MSE = \frac{SSE}{(k-1)(l-1)}$		
Total	$n-1$	SST			

8.3.3 Post hoc comparisons

The post hoc investigation is done following the same approach and principles as for one-way ANOVA with the following differences:

1. Use the MSE and/or SSE from the two-way analysis instead of the MSE and/or SSE from the one-way analysis
2. Use $(l-1)(k-1)$ instead of $n-k$ as degrees of freedom and as denominator for SSE

With these changes the Method boxes 8.9 and 8.10 and the Remark 8.13 can be used for post hoc investigation of treatment differences in a two-way ANOVA.

|||| Example 8.24

Returning to our small example we now find the pairwise treatment confidence intervals within the two-way analysis. If the comparison of A and B was specifically planned before the experiment was carried out, we would find the 95%-confidence interval as:

```
print(muis[0] - muis[1] + np.array([-1,1]) *
      stats.t.ppf(0.975, (4-1)*(3-1)) * np.sqrt(SSE/((4-1) *
(3-1))*(1/4 + 1/4)))

[-3.662 -2.088]
```

and we can hence also conclude that treatment A is different from B. The p-value supporting this claim is found as:

```
tobs = (muis[0] - muis[1]) / np.sqrt(SSE/6 * (1/4 + 1/4))
print(2 * (1 - stats.t.cdf(abs(tobs), 6)))

0.0001094734143394227
```

If we do all three possible comparisons, $M = 3 \cdot 2/2 = 3$, and we will use an overall $\alpha = 0.05$, we do the above three times, but using each time $\alpha_{\text{Bonferroni}} = 0.05/3 = 0.017$:

```
alpha = alpha.values
alpha_bonf = 0.05 / 3
# A vs. B
print(alpha[0] - alpha[1] + np.array([-1, 1]) *
      stats.t.ppf(1 - alpha_bonf/2, 6) * np.sqrt(SSE/6 * (1/4 + 1/4)))


[-3.932 -1.818]

# A vs. C
print(alpha[0] - alpha[2] + np.array([-1, 1]) *
      stats.t.ppf(1 - alpha_bonf/2, 6) * np.sqrt(SSE/6 * (1/4 + 1/4)))


[-4.807 -2.693]

# B vs. C
print(alpha[1] - alpha[2] + np.array([-1, 1]) *
      stats.t.ppf(1 - alpha_bonf/2, 6) * np.sqrt(SSE/6 * (1/4 + 1/4)))


[-1.932  0.182]
```

and we conclude that treatment A is different from B and C, while we cannot reject that B and C are equal. The p-values for the last two comparisons could also be found, but we skip that.

8.3.4 Model control

Also model control runs almost exactly the same way for two-way ANOVA as for one-way:

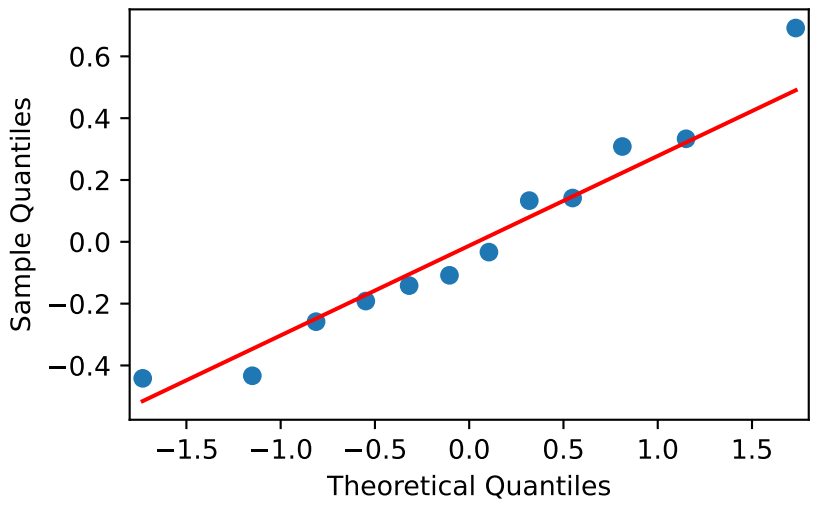
- Use a q-q plot on residuals to check for the normality assumption
- Check variance homegenity by categorized box plots

The only difference is that the box plotting to investigate variance homogeneity should be done on the residuals - NOT on the actual data. And that we can investigate both potential treatment heterogeneity as block heterogeneity.

Example 8.25

First the residual normality plot:

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

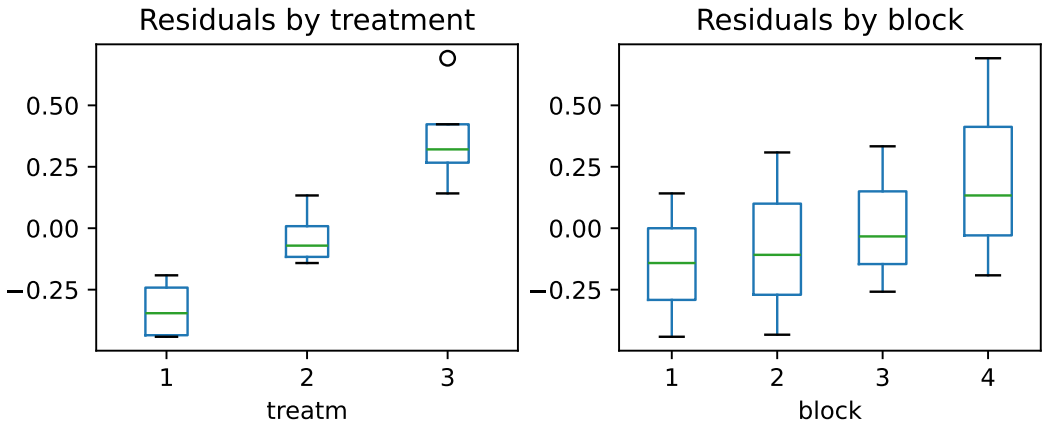



```
print(fit.resid.values)
```

```
[-0.442 -0.433 -0.258 -0.192 -0.142 -0.108 -0.033  0.133  0.142  0.308
  0.333  0.692]
```

Then the investigation of variance homogeneity:

```
D['residuals'] = fit.resid # Add residuals to DataFrame
fig, ax = plt.subplots(ncols=2)
D.boxplot(column='residuals', by='treatm', ax=ax[0], grid=False)
ax[0].set_title('Residuals by treatment')
D.boxplot(column='residuals', by='block', ax=ax[1], grid=False,)
ax[1].set_title('Residuals by block')
plt.suptitle('')
plt.tight_layout()
plt.show()
```



Actually, if we've had data with a higher number of observations for each block, we might have had a problem here as blocks 2 and 3 appears to be quite different on their variability, however since there are very few observations (3 in each block) it is not unlikely to get this difference in variance when there is no difference (but again: it is not very easy to know, exactly where the limit is between what is OK and what is not OK in a situation like this. It is important information to present and take into the evaluation of the results, and in the process of drawing conclusions).

8.3.5 A complete worked through example: Car tires

Example 8.26 Car tires

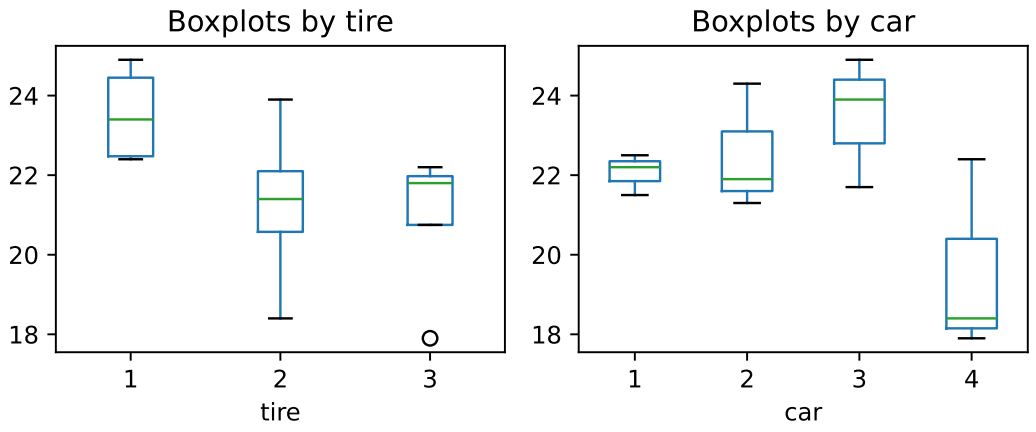
In a study of 3 different types of tires ("treatment") effect on the fuel economy, drives of 1000 km in 4 different cars ("blocks") were carried out. The results are listed in the following table in km/l.

	Car 1	Car 2	Car 3	Car 4	*Mean*
Tire 1	22.5	24.3	24.9	22.4	22.525
Tire 2	21.5	21.3	23.9	18.4	21.275
Tire 3	22.2	21.9	21.7	17.9	20.925
Mean	21.400	22.167	23.167	19.567	21.575

Let us analyse these data with a two-way ANOVA model, but first some explorative plotting:

```
# Collecting the data in a data frame
D = pd.DataFrame({
        'y': [22.5, 24.3, 24.9, 22.4,
              21.5, 21.3, 23.9, 18.4,
              22.2, 21.9, 21.7, 17.9],
        'car': pd.Categorical([1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4]),
        'tire': pd.Categorical([1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3]),
})

fig, ax = plt.subplots(ncols=2)
D.boxplot(column='y', by='tire', ax=ax[0],grid=False)
ax[0].set_title('Boxplots by tire')
D.boxplot(column='y', by='car', ax=ax[1],grid=False)
ax[1].set_title('Boxplots by car')
plt.suptitle('')
plt.tight_layout()
plt.show()
```



Then the actual two-way ANOVA:

```
fit = smf.ols('y ~ car + tire', data=D).fit()
anova = sm.stats.anova_lm(fit)
print(anova)
```

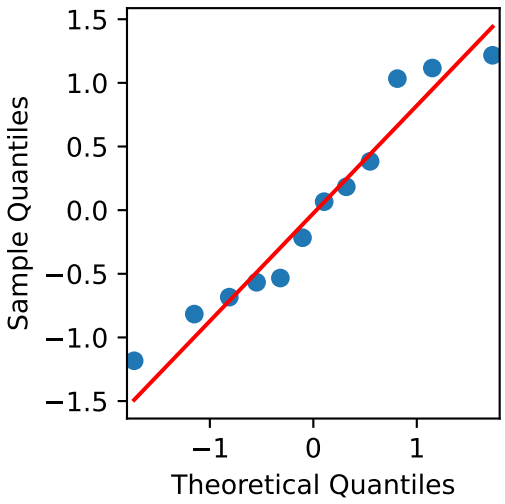
```
            df      sum_sq   mean_sq          F    PR(>F)
car        3.0  25.175833  8.391944  7.025814  0.021726
tire       2.0  15.926667  7.963333  6.666977  0.029888
Residual   6.0   7.166667  1.194444       NaN       NaN
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
car	3	25.18	8.39	7.03	0.0217
tire	2	15.93	7.96	6.67	0.0299
Residuals	6	7.17	1.19		

Conclusion: Tires (treatments) are significantly different and Cars (blocks) are significantly different.

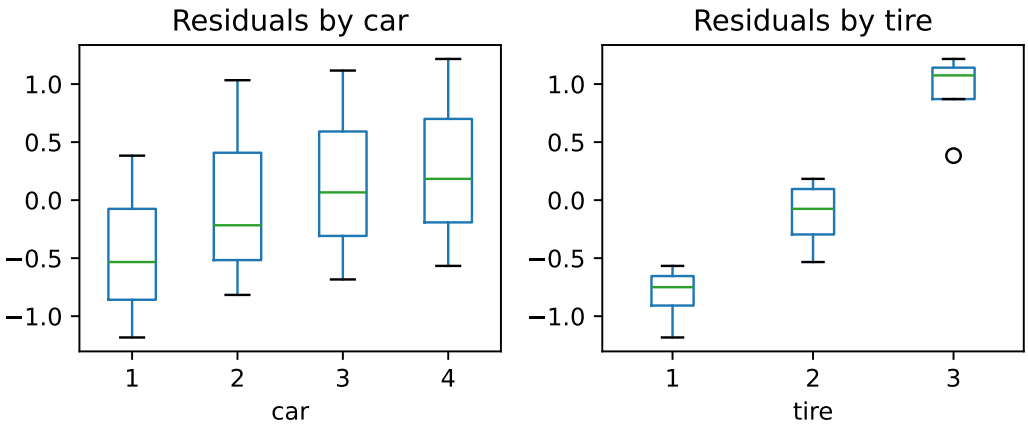
And the model control (for the conclusions to be validated). First the residual normality plot:

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



Then the investigation of variance homogeneity:

```
D['residuals'] = fit.resid # Add residuals to DataFrame
fig, ax = plt.subplots(ncols=2)
D.boxplot(column='residuals', by='car', ax=ax[0],grid=False)
ax[0].set_title('Residuals by car')
D.boxplot(column='residuals', by='tire', ax=ax[1],grid=False)
ax[1].set_title('Residuals by tire')
plt.suptitle('')
plt.tight_layout()
plt.show()
```



It seems like the variance for Car 2 and Car 3 is difference, however, as in the previous example, there are very few observations (only 3) for each car, hence this difference in variation is not unlikely if there is no difference. Thus we find that there do not see any important deviations from the model assumptions.

Finally, the post hoc analysis, first the treatment means:

```
print(D.groupby('tire',observed=True)['y'].mean())
```

```
tire
1    23.525
2    21.275
3    20.925
Name: y, dtype: float64
```

We can find the 0.05/3 (Bonferroni-corrected) *LSD*-value from the two-way version of Remark 8.13:

```
LSD_bonf = stats.t.ppf(1-0.05/6, 6) * np.sqrt(2*1.19/4)
print(LSD_bonf)

2.5358194018640283
```

So tires are significantly different from each other if they differ by more than 2.54. A convenient way to collect the information about the 3 comparisons is by ordering the means from smallest to largest and then using the so-called compact letter display:

Tire	Mean		
3	20.925	a	
2	21.275	a	b
1	23.525		b

There is no significant difference between mean of Tire 2 and 3, and no significant difference between mean of 2 and 1, but there is significant difference between mean of 1 and 3.

8.4 Perspective

We have already seen how the R-version of the ANOVA, both one-way and two-way, are carried out by the R-function `lm`. We also used `lm` for simple and multiple linear regression (MLR) analysis in Chapters 5 and 6. "lm" stands for "linear model", and in fact from a mathematical perspective all these models are what can be termed *linear models*, or sometimes *general linear models*. So differently put, the ANOVA models can in fact be expressed as multiple linear regression models, and the theory and matrix notation etc. from MLR can be used to also work with ANOVA models.

This becomes convenient to understand if one moves on to situations, models and statistical analysis going beyond the current course. An example of this would be situations where we have as well factors as quantitative (continuous) regression input in the same data set.

Important to know also is that the two basic ANOVA versions presented in this material is just the start to be able to handle more general situations. An example could be that, a two-way ANOVA could also occur in a different way than shown here: if we perform what would be a completely randomized study, that is, we have independent sampled groups, but with the groups being represented by a two-way treatment factor structure, say, factor A with 5 levels and factor B with 3 levels. Hence, we have all 15 groups consisting of all combina-

tions of the two treatments, but with several observations within each of the 15 groups. This would sometimes be called a two-way ANOVA with replications, whereas the randomized block setting covered above then would be the two-way ANOVA without replication (there is only and exactly one observation for each combination of treatment and block).

And then the next step could be even more than two treatment factors, and maybe such a multi-factorial setting could even be combined with blocking and maybe some quantitative x-input (then often called covariates) calling for extensions of all this.

Another important extension direction are situations with different levels of observations/variability: there could be hierarchical structures in the data, e.g. repeated measurement on an individual animal, but having also many animals in the study, and animals might come from different farms, that lies in different regions within different countries. This calls for so-called hierarchical models, multi-level models, variance components models or models, where both treatment factors and such hierarchical random effects are present – the so-called mixed models.

All of this and many other good things can be learned in statistics courses building further on the methods presented in this material!

Glossaries

Block [Blok] The block name comes from the historical background of agricultural field trials, where a block would be an actual piece of land within which all treatments are applied 26, 27

***P*-value** [*p*-værdi (for faktisk udfald af en teststørrelse)] 15, 31, 33

Acronyms

ANOVA Analysis of Variance 1, 2, 6, 11, 12, 17, 18, 21, 24, 26, 28, 29, 32, 34, 36, 39, *Glossary:* Analysis of Variance

cdf cumulated distribution function *Glossary:* cumulated distribution function

CI confidence interval 12, 14, 16, 23, 24, 32, *Glossary:* confidence interval

CLT Central Limit Theorem *Glossary:* Central Limit Theorem

IQR Inter Quartile Range *Glossary:* Inter Quartile Range

LSD Least Significant Difference *Glossary:* Least Significant Difference

pdf probability density function *Glossary:* probability density function