Contents

1 ANOVA

ANOVA means ANalysis Of VAriance.
The ANOVA is a tool for studying the influence of one or more qualitative variables on the mean of a numerical variable in a population.

In ANOVA the response variable is numerical and the explanatory variables are categorical.

Example: A local school board is interested in comparing test scores on a standardized reading test for fourth–grade students in their district. They select a random sample of children from each of the local elementary schools. Since it is known that reading scores vary at that age from males to females, they use an analysis, which can tell them if there is a difference in those scores between the schools correcting for the influential factor gender.
Response variable: reading score (numerical)
Factor variables: gender and the elementary school the student is attending (factors are categorical).
Level of gender are male/female

One possible treatment would be "male from Rideau Park School"

We would analyze if the mean reading score depends on the categorical variables "Sex" and "School".

Definition:

- The *response variable*(dependent variable) is the variable of interest to be measured in the experiment.
- *Factors* are the variables whose effect on the response variable is studied in the experiment.
- *Factor levels* are the values of a factor in the experiment.
- *Treatments* are the possible factor level combinations in the experiment. (One factor level for each factor is combined with factor level from other factors)
- A *designed experiment* is an experiment in which the researcher chooses the treatments to be analyzed and the method for assigning individuals to treatments.

1.1 Completely randomized One-Way ANOVA

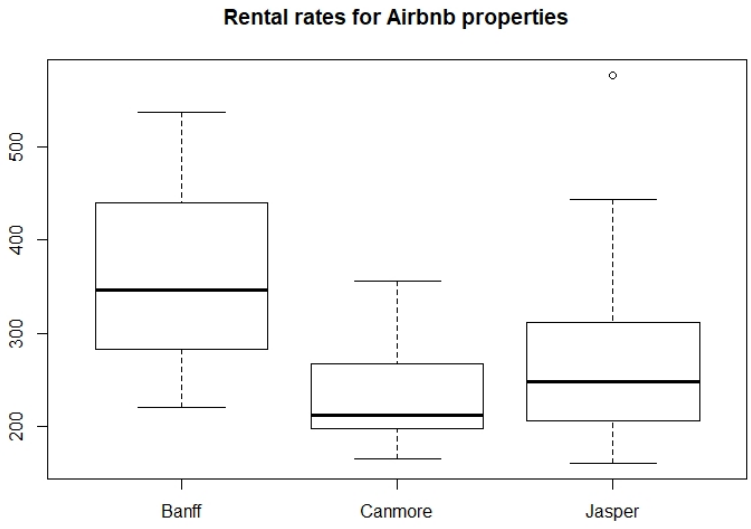
We will start with revisiting the One-Way-ANOVA model which you might have seen in the previous course.
In *Completely Randomized One-way ANOVA* "one-way" indicates that only one factor is considered and "completely randomized" indicates that the experimental units are assumed to be randomly assigned to the factor levels, or have been randomly sampled within each factor level.

Example 1
Where shall we stay for our trip to the mountains in Reading Week?
Is the mean fee charged for one night in an AirBnB property accommodating at least 6 adults differ for Jasper, Banff and Canmore?

On December 28, 2019, we selected random samples of properties which can host at least 6 adults and were available between February 15 and 22, 2020, from the AirBnB website for all three towns.

The data on rate per night in Canadian Dollars are:

Banff	Jasper	Canmore
283	192	165
417	161	129
220	302	209
345	206	184
307	216	215
537	443	167
732	206	213
347	279	202
440	312	156
226	676	198




```
     town  n  mean     sd median min max
1   Banff 10 365.5 114.16  346.0 220 537
2 Canmore 10 238.8  63.40  212.0 165 356
3  Jasper 10 289.4 129.88  247.5 161 576
```

Do the data provide sufficient evidence that the mean rates charged in that week are different in Canmore, Jasper, and Banff?

Example 2

In 1968 Dr. B. Spock was tried in United States district Court of Boston on charges of conspiring to violate the Selective Service Act by encouraging young men to resist being drafted into military service for Vietnam.

The defense in this case challenged the method by which jurors were selected, claiming that women were underrepresented. In fact, the Spock jury had no women.

In Boston the jurors are selected from a *venire* of 30 people, who were selected at random from the City Directory.

The Spock defense pointed to the venire of their trial which only included one woman, who was then released by the prosecution.

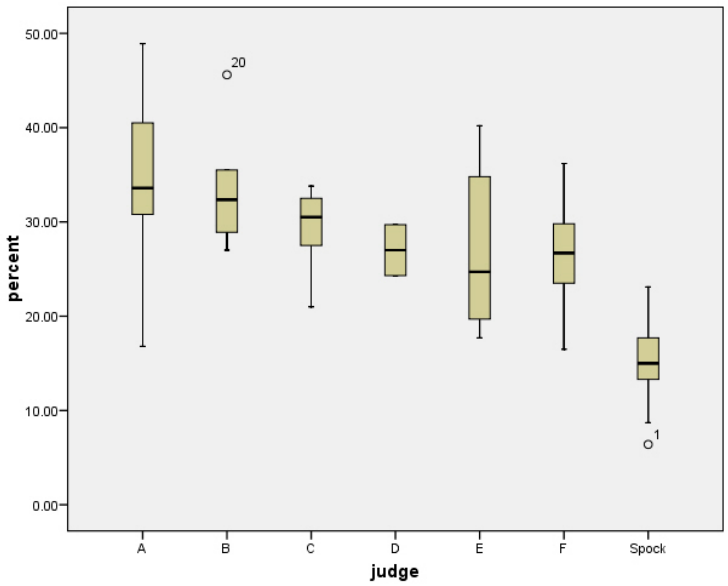
They argued that the judge in the trial had a history of venires in which women were systematically underrepresented, contrary to the law.

They compare this district judge's recent venires with the venires of six other Boston district judges.

The response variable: proportion of women included in venires
factor variable: district judge

The data:

judge	Statistic	value
A	Mean	34.1200
	Std. Deviation	11.94182
	Sample Size	5
B	Mean	33.6167
	Std. Deviation	6.58222
	Sample Size	6
C	Mean	29.1000
	Std. Deviation	4.59293
	Sample Size	9
D	Mean	27.0000
	Std. Deviation	3.81838
	Sample Size	2
E	Mean	26.9667
	Std. Deviation	9.01014
	Sample Size	6
F	Mean	26.8000
	Std. Deviation	5.96888
	Sample Size	9
Spock	Mean	14.6222
	Std. Deviation	5.03879
	Sample Size	9



The question to be explored, if the factor variable (the treatment) has an impact on the mean response variable, or is the mean of the response variable different for the different treatments. (Is the data indicating that the mean proportion of women in venires differs for the different judges).

The Model:
It is assumed that the population is normally distributed, but that the means of the experimental units depend on the treatment.
Let k be the number of different treatments (this would be 7 for the Spock data) and let
x_{ij} be the j-th measurement of the response variable for treatment i, (the proportion of women on the j-th venire for judge i) then assume that

$$x_{ij} = \mu_i + e_{ij}$$

where μ_i is the mean of the response variable for experimental units with treated with treatment i, and
e_{ij} is the error in this measurement, or the part in the measurement that can not be explained through the treatment.
It is assumed that the error is normally distributed with mean zero and standard deviation σ. ($e_{ij} \sim \mathcal{N}(0, \sigma)$)
This statement includes a very strong assumption: The standard deviation is the same for all treatments, which is similar to the assumption of equal standard deviation for the pooled t-test.

In this model the population is described through k potentially different means.
Since we want to see if the treatment impacts the mean of the response variable, the null hypothesis of the test of interest is

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ vs. H_a: at least one of the means is different from the others.

The arguments underlying an ANOVA
Every experiment results in data, that brings in a certain amount of variability (variance).
In an ANOVA the total variance is divided into portions that can be attributed to the different factors of interest.

The analysis of these portions will show the effect of the factors on the response.

Example:
Consider data from two different populations A and B:

```
Set A
     a a a                      b b b
---------------------------------------


Set B
     a b          a b           a b
---------------------------------------
```

The **total variance** in the two data sets is almost the same, but for set A the **variability within the groups** is much less than in set B.

Set A
Since the **variance in the groups** in relationship to the **total variance** is relatively small, the total variance can only be explained by the variance (difference) **between the groups**, so that one would conclude that the means for the groups must be different.

Set B
The variance of the groups is close to the total variance, so that the total variance is explained by this variance. One concludes that the variance **between** the groups can't be large, so that the means in the groups might not be different.

ANOVA is based on the comparison of the **variance of the sample means** with the **variance within the k samples**.
The calculation of the variances are all based on**sum of squares**.

The different sum of squares in a one way ANOVA:

TotalSS: The total variance in the experiment, is the variance of the combined k samples. It is based on the **total sum of squares**

$$TotalSS = \sum (x_{ij} - \bar{x})^2 = \sum x_{ij}^2 - \frac{(\sum x_{ij})^2}{n}$$

where $\bar{x}$ is the overall mean, from all k samples and $n = n_1 + n_2 + \ldots + n_k$.

In ANOVA we analyze the total variance. In one way ANOVA part of the variance can be explained through the use the different treatments and the leftover of the variance must then be due to error in the measurements (or other factors not included in the model).

SST: The variance explained through the treatment is based on the **sum of squares for treatment** (SST), it measures the variation among the k sample means (from one sample to the others):

$$SST = \sum n_i(\bar{x}_i - \bar{x})^2$$

SSE: The variance explained through the error in the measurements is based on the **sum of squares for error** (SSE), it is also used to estimate the variation within the k samples:

$$SSE = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)s_k^2$$

Assuming the standard deviations in all k populations are the same, than this is an estimate for the variation inside the populations which is the variance from the model, σ^2.

It is possible to proof algebraically, that always

$$TotalSS = SST + SSE$$

Therefore, it is only necessary to calculate two of the sum of squares and find the third one with this last identity.

Each of the sum of squares, when divided by its appropriate degrees of freedom, provides an estimate of the total variance in all measurements and the variances due to treatment and error in the experiment.

df_{Total}: Since $TotalSS$ involves n squares its degrees of freedom are $df = n - 1$.

df_T: Since SST involves k squares its degrees of freedom are $df = k - 1$.

df_E: Since SSE involves $(n_1 - 1) + (n_2 - 1) + \ldots + (n_k - 1) = n - k$ squares its degrees of freedom are $df = n - k$.

Observe that also for the degrees of freedom the identity holds that

$$df(TotalSS) = df(SST) + df(SSE)$$

Mean squares (MS) are calculated by dividing the sum of squares by their degrees of freedom $MS = SS/df$. All the results are displayed in an ANOVA table:

ANOVA Table for k independent Random Samples

Source	df	SS	MS	F
Treatments	$k-1$	SST	$MST = SST/(k-1)$	MST/MSE
Error	$n-k$	SSE	$MSE = SSE/(n-k)$	
Total	$n-1$	$TotalSS$		

With

$$F := \frac{MST}{MSE}$$

the variation due to the Error is compared with the variation due to the treatment. If the variation due to the treatment is much larger than the variation due to the Error (i.e. F is large), the data indicate that the treatment has an effect on the mean response.

Continue AirBnB example:
From the summary statistics obtain SSE and SST:

$$\begin{aligned} SSE &= (n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots + (n_k-1)s_k^2 \\ &= (10-1)114.16^2 + (10-1)63.40^2 + (10-1)129.88^2 \\ &= 305276 \end{aligned}$$

and

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_k\bar{x}_k}{n_1 + n_2 + \ldots + n_k} = \frac{8937}{30} = 297.9$$

That gives

$$\begin{aligned} SST &= \sum n_i(\bar{x}_i - \bar{x})^2 \\ &= 10(365.5-297.9)^2 + 10(238.8-297.9)^2 + 10(289.4-297.9)^2 \\ &= 81348 \end{aligned}$$

and

$$TotalSS = SST + SSE = 386624$$

The ANOVA table for the AirBnB data:

Source	*df*	*SS*	*MS*	*F*
town	2	81348	$MST = 40674$	3.6
Error	27	305276	$MSE = 11307$	
Total	29	386624		

As explained above, the F-score shall tell us if there is a significant difference between the mean rates charged in Jasper, Canmore, and Banff. In a next step we will discuss how large we should expect F to be if there is no difference. By comparing the calculated F against those values, we wil be able to find out, if these data provide evidence for a difference.

Continue Spock example:
From the summary statistics obtain SSE and SST:

$$\begin{aligned} SSE &= (n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots + (n_k-1)s_k^2 \\ &= (5-1)11.94^2 + (6-1)6.58^2 + (9-1)4.59^2 + (2-1)3.81^2 + \\ &\quad (6-1)9.01^2 + (9-1)5.97^2 + (9-1)5.04^2 \\ &= 1864.04 \end{aligned}$$

and

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_k\bar{x}_k}{n_1 + n_2 + \ldots + n_k} = \frac{1240.76}{46} = 26.97$$

That gives

$$\begin{aligned} SST &= \sum n_i(\bar{x}_i - \bar{x})^2 \\ &= 5(34.12-26.97)^2 + \ldots + 9(14.62-29.97)^2 \\ &= 1987.91 \end{aligned}$$

and

$$TotalSS = SST + SSE = 3891.95$$

The ANOVA table for the Spock data:

Source	*df*	SS	MS	F
judge	6	1987.91	$MST = 331.31$	6.932
Error	39	1864.04	$MSE = 47.79$	
Total	45	3891.95		

In a next step see how this information can help in deciding if the data provide sufficient evidence that not all the treatment means are the same.

1.2 The extra sum of square principle

Before continuing a second rationale behind the F-statistic shall be introduced. Another way to interpret the F-statistic is based on the following:
With the F statistic two standardized sum of squares are compared. These sums of squares can be interpreted as measuring the variance that remains unexplained through certain models.
So the F-statistic compares if a more complex model (full model) explains significantly more variance in the data than a simpler model (reduced model). If we find that the full model explains more of the variance, we reject the reduced model, and find the full model correct.
In one way ANOVA the F-statistic compares the following **models**

	Population	1	2	...	k
Reduced	Mean	μ	μ	...	μ
	standard deviation	σ	σ	...	σ
Full	Mean	μ_1	μ_2	...	μ_k
	standard deviation	σ	σ	...	σ

If the reduced model is rejected in favour of the full model find that the means are not all the same and the full model is a better description of the observed data.
Fitting the models:
The basic idea is to estimate the parameters in both models and to see whether the estimated model returns a good representation of the observed data.
Estimates:

	Population	1	2	...	k
Reduced	Mean	$\bar{x}$	$\bar{x}$	$\bar{x}$	$\bar{x}$
Full	Mean	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_k$

Now find the residual sum of squares based on these estimates. They measure how much of the variation remains unexplained through these models. The residual sum of squares for the full model is SSE (based on $\bar{x}_i$), and the residual sum of squares for the reduced model is Total SS (based on $\bar{x}$).

The F-statistic shows how much more variation is explained through the full model (after standardization) than through the reduced model,

$$F = \frac{(Total\ SS - SSE)/df_T}{SSE/df_E}$$

If F is large it indicates that the full model is significantly superior to the reduced model in describing the observed data, and it is decided that the full model should be adopted.

As mentioned above the goal is to use a 1-way ANOVA for the analysis of the means of different populations $\mu_1, \mu 2, \ldots, \mu_k$. In a next step we now want to use the information from the ANOVA table, to test hypotheses and give confidence intervals concerning those population means.

1.3 Testing for Equality of All Treatment Means

The hypotheses of interest are

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \quad \text{versus} \quad H_a : \text{at least one of the means differs from the others}$$

Use the following argument for developing the test:

- Remember that that the variances in the k populations are assumed to be all the same σ^2. The statistic
$$MSE = \frac{SSE}{n-k}$$
is a pooled estimate of σ^2 (a weighted average of all k sample variances), whether or not H_0 is true.

- If H_0 is true, then the **variation in the sample means**, measured by
$$MST = \frac{SSG}{k-1}$$
also provides an unbiased estimate of σ^2, this is derived from $\sigma_{\bar{x}}^2 = \sigma^2/n$.
However, if H_0 is false and the population means are not the same, then MST is larger than σ^2.

- The test statistic
$$F = \frac{MST}{MSE}$$
tends to be much larger than 1, if H_0 is false. Hence, H_0 can be rejected for large values of F.
What values of F have to be considered large, we learn from the distribution of F.

- If H_0 is true, the statistic
$$F = \frac{MST}{MSE}$$
has an F distribution with $df_1 = (k-1)$ and $df_2 = (n-k)$ degrees of freedom.
Upper tailed critical values of the F distribution can be found in Table VIII-XI

The ANOVA F-Test

1. The Hypotheses are

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \text{ versus}$$
$$H_a : \text{at least one of the means differs from the others}$$

2. Assumption: The population follows a normal distribution with means $\mu_1, \mu_2, \ldots, \mu_k$ and equal variance σ^2. The samples are independent random samples from each population.

3. Test statistic:
$$F_0 = \frac{MST}{MSE}$$
based on $df_1 = (k-1)$ and $df_2 = (n-k)$.

4. P-value: $P(F > F_0)$, where F follows an F-distribution with $df_1 = (k-1)$ and $df_2 = (n-k)$.

5. Decision: If P-value$\leq \alpha$, then reject H_0.
If P-value$> \alpha$, then do not reject H_0.

6. Put into context.

Continue AirBnB example:
The data on rental rates of AirBnB properties in Jasper, Banff, and Canmore resulted in the following ANOVA table

Source	df	SS	MS	F
town	2	81348	$MST = 40674$	3.6
Error	27	305276	$MSE = 11307$	
Total	29	386624		

Use this information now to test if the data provide sufficient evidence that the mean rates in the three towns are hot all the same during Reading Week in 2020.
Conduct an ANOVA F-test:

1. Hypotheses: Let μ_i=mean rate charged for one night in an AirBnB property suitable for at least 6 adults in Reading Week in town i.

$$H_0 : \mu_J = \mu_C = \mu_B \text{ versus}$$
$$H_a : \text{at least one of the means differs from the others}$$

choose $\alpha = 0.05$

2. The samples can be considered random samples. The box plots all look reasonable symmetric and do not provide strong evidence against the assumption of normality. The standard deviations are close enough to each other to be not of concern of violating the assumption of equal variances in the subgroups.

3. Test statistic: (from ANOVA table)

$$F_0 = 3.6 \quad \text{with} \quad df_n = 2, df_d = 27$$

4. P-value: The P-value is the uppertail area for $F_0 = 3.6$

   Use the table for the F-distribution, find $df_n = 2$ and $df_d = 27$.

   $F_0 = 3.6$ falls between critical values for $\alpha = 0.05$ and $\alpha = 0.025$, and we conclude that $0.025 <$P-value < 0.05

5. Decision: The P-value$< 0.005 < 0.05 = \alpha$, we reject H_0.

6. Put into context:
   At significance level 0.05 the data provide sufficient evidence that the mean rental rates for AirBnB properties for at least six guests in Banff, Jasper, and Canmore are not all the same in the 2020 Reading Week.

The question arising out of this answer is: Where are the (significant) differences?

Continue Spock example:
The data resulted in the following ANOVA table:

Source	*df*	SS	MS	F
judge	6	1987.91	$MST = 331.31$	6.932
Error	39	1864.04	$MSE = 47.79$	
Total	45	3891.95		

Conduct an ANOVA F-test:

1. Hypotheses: Let μ_i=mean proportion of women on a venire for judge i.

   $H_0 : \mu_A = \mu_B = \mu_C = \mu_D = \mu_E = \mu_F = \mu_{Spock}$ versus
   H_a : at least one of the means differs from the others

   choose $\alpha = 0.05$

2. According to the box plot the data can be assumed to be normal (all boxes are pretty symmetric). Only for one judge we find the standard deviation much higher than for the others, this could cause problems.

3. Test statistic:

   $$F_0 = 6.932 \quad \text{with} \quad df_n = 6, df_d = 39$$

4. P-value: The P-value is the uppertail area for $F_0 = 6.932$

   Use the table for the F-distribution, find $df_1 = 6$ and $df_2 = 30$ (largest df in the table smaller than 39).

   6.932 is larger than the largest value in the block of values for this particular choice of df, which is 3.95. Therefore the P-value is smaller than 0.005, the upper tail area for 3.95.

   State: P-value< 0.005.

5. Decision: The P-value$< 0.005 < 0.05 = \alpha$, we reject H_0.

6. Put into context:
   At significance level 0.05 the data provide sufficient evidence that at least for one judge the mean proportion of women on venires is different from the mean proportion of the other judges.

The test above indicates that the mean proportion of women are not the same for all judges. The next task would be to find out where out where the differences can be found.

Example 3
Sevoflurane and Desflurane Concentration in Post-Anesthesia Care Units (PACUs).

According to Wikipedia desflurane is a highly fluorinated methyl ethyl ether used for maintenance of general anesthesia, and sevoflurane is a sweet-smelling, nonflammable, highly fluorinated methyl isopropyl ether used as an inhalational anaesthetic for induction and maintenance of general anesthesia. Both are commonly used volatile anesthetic agents used during surgeries. Both are usually considered safe for anesthesia but are recognized as occupational hazard for health professionals. A study aimed to document and investigate differences in desflurane and sevoflurane concentrations in the breath of nurses in pediatric PACUs and adult PACUs. The study also included a control group of residents, who were not exposed to the agents in the previous 3 month.
For all participants measurements indicating the concentration of desflurane and sevoflurane were taken in the morning. Nurses were about to start a day shift (after they had worked the day before the same shift).
Do the data indicate that there are differences in the mean concentration of desflurane concentration in the breath between the three groups (pediatric, adult, control)?

Group	n	mean	st.dev.
Pediatric	10	1.74	0.417
Adult	10	1.77	0.416
Control	10	1.21	0.213

From the summary statistics obtain SSE and SST:

$$\begin{aligned} SSE &= (n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_3-1)s_3^2 \\ &= (10-1)0.417^2 + (10-1)0.416^2 + (10-1)0.213^2 \\ &= 3.5308 \end{aligned}$$

and

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1+n_2+n_3} = \frac{47.2}{30} = 1.573$$

That gives

$$\begin{aligned} SST &= \sum n_i(\bar{x}_i - \bar{x})^2 \\ &= 10(1.74-1.573)^2 + 10(1.77-1.573)^2 + 10(1.21-1.573)^2 \\ &= 1.98467 \end{aligned}$$

and

$$TotalSS = SST + SSE = 5.51547$$

Producing the following ANOVA table for desflurane

Source	*df*	*SS*	*MS*	*F*
Group	2	1.985	$MST = 0.9923$	7.588
Error	27	3.531	$MSE = 0.1308$	
Total	29	5.515		

Conducting an ANOVA F-test

1. Hypotheses: Let μ_0 = mean desflurane concentration in the control group,
   μ_1 = mean desflurane concentration in nurses working in a pediatric PACU, and
   μ_3 = mean desflurane concentration in nurses working in an adult PACU.

   $$H_0 : \mu_0 = \mu_1 = \mu_3 \text{ versus}$$
   $$H_a : \text{at least one of the means is different}$$

   choose $\alpha = 0.05$

2. Random sample? Is the assumption of a completely randomized ANOVA met? (Hint: No.)

   The standard deviations are not too different, so the assumption of equal standard deviations is not badly violated.

   The error has to be normal, later residual analysis will be discussed illustrating how to check this assumption.

3. Test statistic:
   $$F_0 = 7.588 \quad \text{with} \quad df_n = 2, df_d = 27$$

4. P-value: The P-value is the uppertail area for $F_0 = 7.588$.

   Use the table for the F-distribution, find $df_1 = 2$ and $df_2 = 27$.
   It shows that for 6.49 the uppertail area is 0.005. This is already the largest value for these degrees of freedom.

   Since $7.588 > 6.49$, its uppertail area must be smaller than 0.005, therefore

   P-value < 0.005.

5. Since the P-value is smaller than $\alpha = 0.05$, reject H_0.

6. At significance value of 5% the data provide sufficient evidence that the mean concentrations of desflurane are not the same for the three groups.

   Next question: Where are the differences?

For Sevoflurane the ANOVA table is given by

Source	*df*	*SS*	*MS*	*F*	Pr(>F)
Group	2	0.0507	$MST = 0.02533$	2.118	0.14
Error	27	0.3230	$MSE = 0.01196$		
Total	29	5.515			

Pr(>F) is giving the P-value. What would be the conclusion for the mean Sevoflurane concentration measured for the 3 groups?

In the next section we show how pairwise comparisons (through confidence intervals and tests) are conducted in ANOVA.

1.4 Estimating and Testing Differences in Treatment Means

Is the ANOVA F-test significant, one can conclude at the given significance level that not all population means are the same, but through this test we can not determine where the differences occur. In order to locate the differences pairwise comparisons should be conducted, i.e. look at parameters
$\mu_1 - \mu_2, \ldots, \mu_1 - \mu_k$
$\mu_2 - \mu_3, \ldots, \mu_2 - \mu_k$
$\ldots$
$\mu_{k-1} - \mu_k$

There is a total of $k(k-1)/2$ parameters for pairwise comparisons of k means.
The comparisons can be done with the help of t-tests or t-CIs.

For some studies the pairwise comparisons are not the main focus, like in the Spock example; here it is more relevant to compare μ_{Spock} with the average of all the other means

$$\bar{\mu} = \frac{\mu_A + \mu_B + \mu_C + \mu_D + \mu_E + \mu_F}{6}$$

The parameter of interest in that case would be

$$\frac{1}{6}\mu_A + \frac{1}{6}\mu_B + \frac{1}{6}\mu_C + \frac{1}{6}\mu_D + \frac{1}{6}\mu_E + \frac{1}{6}\mu_F - \mu_{Spock}$$

For the study on the concentration of volatile anesthetic agents in the breath in a control group versus nurses working in pediatric and adult PACUs one might be really interested in comparing the two groups of nurses with the control group reflected in

$$\frac{1}{2}(\mu_P + \mu_A) - \mu_C = \frac{1}{2}\mu_P + \frac{1}{2}\mu_A - \mu_C$$

When comparing the rates payable for staying one night in an AirBnB we might be more interested in comparing the prices between options for Banff National Park (Banff and Canmore) versus Jasper National Park (Jasper), so we compare the mean rate paid in Jasper, μ_J, with the average of the mean rates paid in Banff and in Canmore, $1/2(\mu_B + \mu_C)$. resulting in the contrast

$$\gamma = \mu_J - \frac{1}{2}(\mu_B + \mu_C) = 1 \cdot \mu_J + (-\frac{1}{2}) \cdot \mu_B + (-\frac{1}{2}) \cdot \mu_C$$

In order to describe this family of parameters of interest we look at *contrasts*.

Contrasts:
Linear combinations of the group means have the form

$$\gamma = C_1\mu_1 + C_2\mu_2 + \ldots + C_k\mu_k$$

If the coefficients add to zero ($C_1 + C_2 + \ldots + C_k = 0$) the linear combination is called a contrast.

Example 4
For pairwise comparisons the linear combinations is

$$\gamma = \mu_1 - \mu_2 = 1\mu_1 + (-1)\mu_2 + 0\mu_3 + \ldots + 0\mu_k$$

with $C_1 = 1, C_2 = -1, C_3 = \ldots = C_k = 0$. Since these total to zero a pairwise comparison leads to a contrast.

The linear combination for the Spock example is

$$\gamma = \frac{1}{6}\mu_A + \frac{1}{6}\mu_B + \frac{1}{6}\mu_C + \frac{1}{6}\mu_D + \frac{1}{6}\mu_E + \frac{1}{6}\mu_F - \mu_{Spock},$$

which gives $C_A = C_B = \ldots = C_F = 1/6$ and $C_{Spock} = -1$, with a total $C_A + \ldots + C_F + C_{Spock} = 0$, so it is a contrast.

For the anesthesia example it would be

$$\gamma = \frac{1}{2}\mu_P + \frac{1}{2}\mu_A - \mu_C$$

with $C_P = C_A = 1/2$ and $C_C = -1$, with a total $C_P + C_A + C_C = 0$, so it is a contrast.

The AirBnB example results in the contrast

$$\gamma = \mu_J - \frac{1}{2}(\mu_B + \mu_C) = 1 \cdot \mu_J + (-\frac{1}{2}) \cdot \mu_B + (-\frac{1}{2}) \cdot \mu_C$$

A contrast γ becomes the unknown population parameter. How can this parameter be estimated from the available data?

The sample contrast estimating $\gamma = C_1\mu_1 + C_2\mu_2 + \ldots + C_k\mu_k$ is

$$c = C_1\bar{x}_1 + C_2\bar{x}_2 + \ldots + C_k\bar{x}_k$$

To find a confidence interval for γ, we study the distribution of c. c is based on a random sample therefore assumes random values, it is a random variable and therefore is described by a distribution. For the distribution of c one finds

$$\mu_c = \gamma, SE(c) = s_p\sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \ldots + \frac{C_k^2}{n_k}}$$

where $s_p = \sqrt{MSE}$
So c is an unbiased estimator for γ (why?).
From this we conclude

$$t = \frac{c - \gamma}{SE(c)}$$

has a t-distribution with $df = n - k$.
From this we derive the

$(1 - \alpha)100\%$ **Confidence Interval for** γ

$$c \pm t^{df}_{1-\alpha/2} SE(c)$$

$t^{df}_{1-\alpha/2}$ is the $1 - \alpha/2$ percentile of the t-distribution with $df = n - k$.

t-test for contrasts

1. Hypotheses:
   $H_0 : \gamma = 0$ versus $H_a : \gamma \neq 0$
   $H_0 : \gamma \leq 0$ versus $H_a : \gamma > 0$
   $H_0 : \gamma \geq 0$ versus $H_a : \gamma < 0$

2. Assumption: Random samples, the response variable is normally distributed, and the standard deviation is the same for all treatments.

3. Test Statistic:

$$t_0 = \frac{c}{SE(c)}, \quad df = n - k$$

4. P-value:

Hypotheses	P-value
two-tailed	$2P(t > abs(t_0))$
upper-tailed	$P(t > t_0)$
lower-tailed	$P(t < t_0)$

5. Decision:
   P-value$\leq \alpha$, then reject H_0 P-value$> \alpha$, then do not reject H_0

Application: A confidence interval for comparing two means within the ANOVA model is then: $(1 - \alpha)100\%$ Confidence interval for $\mu_i - \mu_j$:

$$(\bar{x}_i - \bar{x}_j) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

with $s_p^2 = MSE$ and $t_{\alpha/2}$ the $(1 - \alpha/2)$ percentile of the t-distribution with $df = n - k$.

Continue AirBnB Example: Find confidence intervals for all pairwise comparisons:

1. $\gamma = \mu_J - \mu_C = 1 \cdot \mu_J + (-1) \cdot \mu_C + 0 \cdot \mu_B : \quad c = 1 \cdot \bar{x}_J + (-1) \cdot \bar{x}_C + 0 \cdot \bar{x}_B = \bar{x}_J - \bar{x}_C$
2. $\gamma = \mu_J - \mu_B = 1 \cdot \mu_J + 0 \cdot \mu_C + (-1) \cdot \mu_B : \quad c = 1 \cdot \bar{x}_J + 0 \cdot \bar{x}_C + (-1) \cdot \bar{x}_B = \bar{x}_J - \bar{x}_B$
3. $\gamma = \mu_C - \mu_B = 0 \cdot \mu_J + 1 \cdot \mu_C + (-1) \cdot \mu_B : \quad c = 0 \cdot \bar{x}_J + 1 \cdot \bar{x}_C + (-1) \cdot \bar{x}_B = \bar{x}_C - \bar{x}_B$

Since the sample sizes are all the same $SE(c)$ are the same for all three confidence intervals

$$SE(c) = s_p\sqrt{\frac{(1)^2}{10} + \frac{(-1)^2}{10}}, \text{with} \quad s_p = \sqrt{MSE} = \sqrt{11307} = 106.33$$

such that $SE(c) = 47.55$, using the t-critical value (for a 95% confidence interval) $t_{0.025}^{27} = 2.052$ (from the t-table), one finds that the margin of error for all confidence intervals is $2.052(47.55) = 97.58$
The confidence intervals are therefore:

1. $(289.4 - 238.8) \pm 97.58 : \quad c = 50.6 : [-46.98, 148.18]$
2. $(289.4 - 365.5) \pm 97.58 : \quad c = -76.11 : [-173.68, 21.48]$
3. $(238.8 - 365.5) \pm 97.58 : \quad c = -126.7 : [-224.28, -29.12]$

From 1. we are 95% confident that the difference in the mean rate to be paid in Jasper versus Canmore falls between \$-47 and \$148. Since 0 falls within the confidence interval the data do not provide sufficient evidence that the mean rates are different.
2. DIY
From 3. we are 95% confident that the difference in the mean rate to be paid in Banff versus Canmore falls between \$29 and \$224. Since 0 falls outside the confidence interval the data provide sufficient evidence that the mean rates in Banff are somewhere between \$29 and \$224 higher than in Canmore.

When comparing the rates in the two national parks we use the contrast

$$\gamma = \mu_J - \frac{1}{2}(\mu_B + \mu_C) = 1 \cdot \mu_J + (-\frac{1}{2}) \cdot \mu_B + (-\frac{1}{2}) \cdot \mu_C$$

Do the data provide sufficient evidence that the rates charged in or just outside Banff National Park are higher than the rates charged in Japer National Park?

1. Hypotheses:
   $H_0 : \gamma \geq 0$ versus $H_a : \gamma < 0$

   $\alpha = 0.05$

2. Assumption: Random samples, the response variable is normally distributed, and the standard deviation is the same for all treatments.

   According to the introduction of the example the samples are random samples, we will see later how to check the other assumptions.

3. Test Statistic:
   Use data with R:

```
> levels(airbnb$town)
[1] "Banff"   "Canmore" "Jasper"

> contrast(ls,contr)
 contrast         estimate   SE df t.ratio p.value
 c(-0.5, -0.5, 1)    -12.8 41.2 27 -0.310  0.7592
```

   From this output relevant for the test statistic:

$$c = -12.8,\ SE(c) = 41.2, t_0 = -0.310, df = 27$$

4. P-value:
   We are conducting a lower tail test, so

   P-value $= P(t < -0.310) = 0.7592/2 = 0.3796$

   Be careful with the P-value from the output, it gives the two-tailed, therefore we have to divide it by 2 to get the lower tailed. (Draw diagram)

5. Decision:
   P-value$> 0.05 = \alpha$, do not reject H_0 at significance level of 5%

6. Context:
   At significance level of 5% the data do not provide sufficient evidence that the mean rate charged for AirBnB properties suitable for at least 6 people in Reading Week is significantly higher in Banff National Park than in Jasper National Park.

Continue Spock Example: We obtain 95% confidence intervals for $\mu_A - \mu_{Spock}$, $\mu_B - \mu_{Spock}$, $\mu_C - \mu_{Spock}$, $\mu_D - \mu_{Spock}$,$\mu_E - \mu_{Spock}$, and $\mu_E - \mu_{Spock}$, in order to figure out if the mean on μ_{Spock} is significant different from all the other means.

$\mu_A - \mu_{Spock}$:
For $\gamma = \mu_A - \mu_{Spock}$ $C_A = 1, C_{Spock} = -1$ and all others are 0.

$$c = \bar{x}_A - \bar{x}_{Spock},\ SE(c) = s_p\sqrt{\frac{(1)^2}{n_A} + \frac{(-1)^2}{n_{Spock}}} = s_p\sqrt{\frac{1}{5} + \frac{1}{9}}$$

$t^{39}_{0.975} = 2.023$, so we get

$$\begin{aligned}&(34.12 - 14.62) \pm 2.023 \cdot 6.89\sqrt{\tfrac{1}{5} + \tfrac{1}{9}}\\ &19.5 \pm 7.775\\ &(11.73\ ;\ 27.27)\end{aligned}$$

in the same way we get

	95% CI
$\mu_B - \mu_{Spock}$	(11.5848 26.4152)
$\mu_C - \mu_{Spock}$	(7.8476 21.1124)
$\mu_D - \mu_{Spock}$	(1.3815 23.3785)
$\mu_E - \mu_{Spock}$	(4.9348 19.7652)
$\mu_F - \mu_{Spock}$	(5.5476 18.8124)

None of the confidence intervals captures zero, we are 95% confident that the mean proportion of women on venires for any judge and Spock's judge are different (95% confident for each interval).
We are looking at 6 confidence interval every single CI has a 95% chance of capturing the parameter of interest. But there is also the chance that any of them does not capture the parameter we are looking for. And when looking at the whole set of CIs, it is likely that we miss at least one of the parameters with the six intervals. This is an undesirable effect of conducting many different comparisons.
The other CI we are interested in is a 95%CI for $\gamma = \frac{1}{6}\mu_A + \frac{1}{6}\mu_B + \frac{1}{6}\mu_C + \frac{1}{6}\mu_D + \frac{1}{6}\mu_E + \frac{1}{6}\mu_F - \mu_{Spock}$:

$$(29.6 - 14.62) \pm 2.042 \cdot 6.89\sqrt{\frac{1}{6^2 \cdot 5} + \frac{1}{6^2 \cdot 6} + \frac{1}{6^2 \cdot 9} + \frac{1}{6^2 \cdot 2} + \frac{1}{6^2 \cdot 6} + \frac{1}{6^2 \cdot 9} + \frac{(-1)^2}{9}}$$

$$14.98 \pm 2.042 \cdot 6.89 \cdot 0.382 = [9.61, 20.35]$$

Since zero is not included with the interval we are 95% confident, that the mean proportion of women on venires from Spock's judge is significant lower than the mean for all the other 6 judges.

Continue Desflurane Example:
Test if the mean desflurane concentration is in the morning higher for nurses working in PACUs than in the control group.
Choose $\gamma = \frac{1}{2}(\mu_P + \mu_A) - \mu_C$.
t-test for contrasts

1. Hypotheses: $H_0 : \gamma \leq 0$ versus $H_a : \gamma > 0$

2. Assumption: Addressed these when doing the ANOVA

3. Test Statistic:
   $c = \frac{1}{2}(1.74 + 1.77) - 1.21 = 0.545, s_p = \sqrt{MSE} = 0.362,$

$$SE(c) = 0.362\sqrt{\frac{0.5^2}{10} + \frac{0.5^2}{10} + \frac{(-1)^2}{10}} = 0.1402$$

   and

$$t_0 = \frac{0.545}{0.1402} = 3.628, \quad df = 27$$

4. P-value: Upper tailed test ($H_a : \gamma > 0$), P-value = $P(t > 3.628)$ In table IV use df=27, and find that P-value<0.005.

5. Decision: P-value< $\alpha = 0.05$, reject H_0

6. At significance level of 5% the data provide sufficient evidence that the mean concentration of desflurane is higher for nurses working in PACUs then in the control group.

1.5 Multiple Comparisons

As mentioned above, when simultaneously conducting **several** comparisons (confidence intervals, tests) at significance level of α, the probability that one of the results in the overall statement is wrong can be much higher than α. The more comparisons are made the more likely it becomes that at least one decision is wrong, i.e. the overall error probability increases with every extra comparison included with the overall statement.
This effect is problematic, because we already have to accept that in inferential statistics we might make mistakes. This was found acceptable as long as the probability for making an error was known and could be chosen (α). This control is lost when making multiple comparisons without taking the described effect into account. In order to counteract the effect different strategies have been developed to control the overall or experiment wise (versus comparison wise) error rate in multiple comparisons.

Different multiple comparison methods

1. Tukey-HSD (honest significant difference) procedure:
   Can only be used when it is required to investigate ALL pairwise comparisons of several means. Since taking advantage of this special situation the CIs are narrower than for the other methods, when it can be used.

2. Bonferroni method:
   Can be used in very general situation where we want to control the overall error rate of a multiple comparison, it usually low power.

   For the Bonferroni method, only α for the t-CI has to be adjusted: When doing m pairwise comparisons, then choose $\alpha^* = \alpha/m$ for each comparison.

3. Scheffé method:
   Can be used also for multiple contrasts, but will result in wider CIs than Tukey if applied to pairwise comparisons.

Conducting a multiple comparison using Bonferroni's method

1. Choose the acceptable experiment wise error rate, α.
2. Find m= the number of comparisons you want to make.
3. Calculate $\alpha^* = \alpha/m$ the comparison wise error rate
4. Rank the sample means from the samples of those populations you want to compare from smallest to largest
5. Determine the margin of error (ME) for *each* comparison ($\mu_i - \mu_j$)

$$ME_{ij} = t^{n-k}_{\alpha^*/2}\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

6. Place a bar under those pairs of treatments that differ less than the margin of error. A pair of treatments not connected by a bar implies a difference in the population means.

The confidence level associated with all statements at once is at least $(1 - \alpha)$.

Continue AirBnB example:

1. $\alpha = 0.05$.
2. $m = 3(3-1)/2 = 3$ is the number of comparisons ($mu_J - \mu_B$, $\mu_J - \mu_C, \mu_B - mu_C$).
3. $\alpha^* = \alpha/3 = 0.01666$ the comparison wise error rate.
4.

$\bar{x}_C$	$< \bar{x}_J$	$< \bar{x}_B$
238.8	289.4	365.5

5. The margin of error (ME) is the same for all three comparisons because the sample sizes are all the same.
$$ME = t^{27}_{0.008}(106.33)\sqrt{\frac{1}{10} + \frac{1}{10}} = 131.77$$
with $t^{27}_{0.008} \approx 2.771$ from table IV.
6. $\bar{x}_J - \bar{x}_C = 50.6 < ME$, underline $\bar{x}_J$ and $\bar{x}_C$
$\bar{x}_B - \bar{x}_C = 126.7 < ME$, continue line to $\bar{x}_B$
$\bar{x}_B - \bar{x}_J = 76.1 < ME$

$\bar{x}_C$	$< \bar{x}_J$	$< \bar{x}_B$
238.8	289.4	365.5

7. At experimentwise error rate of 5% the data do not indicate any significant pairwise differences between the mean rates for AirBnB properties housing at least 6 people in Reading Week 2002

   This is odd, since the F-test said there is at least one difference, but this method is no sensitive enough to detect where the difference is.

Example 5
Spock's example using the Bonferroni method

1. $\alpha = 5\%$
2. $m = 7(6)/2 = 21$
3. $\alpha^* = 0.05/21 = 0.00238$
4. the ranked means are

$$14.62, 26.80, 26.97, 27.00, 29.10, 33.62, 34.12$$

5. For the percentile find $\alpha^*/2 \approx 0.001$ for $df = 39$ use $t^{39}_{0.001} = 3.31$ (from online calculator - the best we can get from the table is $t^{39}_{0.005} = 2.708$)

   This gives the following margin of errors for the different comparisons

		Judge/sample size						
		Spock's	A	B	C	D	E	F
		9	5	6	9	2	6	9
Spock's	9		13.05	12.33	11.03	18.29	12.33	11.03
A	5			14.17	13.05	19.57	14.17	13.05
B	6				12.33	19.11	13.51	12.33
C	9					18.29	12.33	11.03
D	2						19.11	18.29
E	6							12.33
F	9							

6. If we put the bars on, we get

```
 14.62    26.80    26.97    27.00     29.10    33.62     34.12
Spock's     F        E        D         C        B         A
-------                    -------
         ---------------------------------------------------
```

In conclusion we find at 95% confidence that the mean proportion of women on venires is significantly different for Spock's judge and judges besides F, E, C, B and A. The mean proportion of women on venires for judges F-A do not differ significantly.
The mean proportion of women on venires from Spock's judge and judge D do not differ significantly. But judge D is special because only two of his venires were included in the study resulting in a very large standard error, which in conclusion is not significantly different from any other judge.

Example 6
Compare the mean desflurane concentration for the three groups.

1. $\alpha = 0.05$.
2. $m = 3(3-1)/2 = 3$ is the number of comparisons ($mu_0 - \mu_1$, $\mu_0 - \mu_3, \mu_1 - mu_3$).
3. $\alpha^* = \alpha/3 = 0.01666$ the comparison wise error rate.
4.

$\bar{x}_0$	$< \bar{x}_1$	$< \bar{x}_3$
1.21	1.74	1.77

5. The margin of error (ME) is the same for all three comparisons because the sample sizes are all the same.
$$ME = t^{27}_{0.008}(0.362)\sqrt{\frac{1}{10} + \frac{1}{10}} = 0.4486$$
with $t^{27}_{0.008} \approx 2.771$ from table IV.
6. $\bar{x}_3 - \bar{x}_0 = 0.56 > ME$
$\bar{x}_3 - \bar{x}_1 = 0.03 < ME$, underline $\bar{x}_1$ and $\bar{x}_3$
$\bar{x}_1 - \bar{x}_0 = 0.53 > ME$

$\bar{x}_0$	$< \bar{x}_1$	$< \bar{x}_3$
1.21	1.74	1.77
	———	

7. At experimentwise error rate of 5% the data provide sufficient evidence that the mean concentration of desflurane in the breath in the control group is significant lower than for both pediatric and adult PACU nurses in the morning after a shift on the day before. The analysis did not show a significant difference when comparing the mean concentration for pediatric and adult PACU nurses.

1.6 Residual Analysis - Checking the Assumptions

The assumptions for all the tests in ANOVA can be written as:

1. Samples are random samples
2. The model is an appropriate description of the population

To understand what needs to be checked reconsider the one-way ANOVA model:

$$x_{ij} = \mu_i + e_{ij}, \quad e_{ij} \sim \mathcal{N}(0, \sigma)$$

The model implies that observations can be viewed as treatment mean plus some error, where the error comes from a normal distribution with mean 0 and common standard deviation, σ.

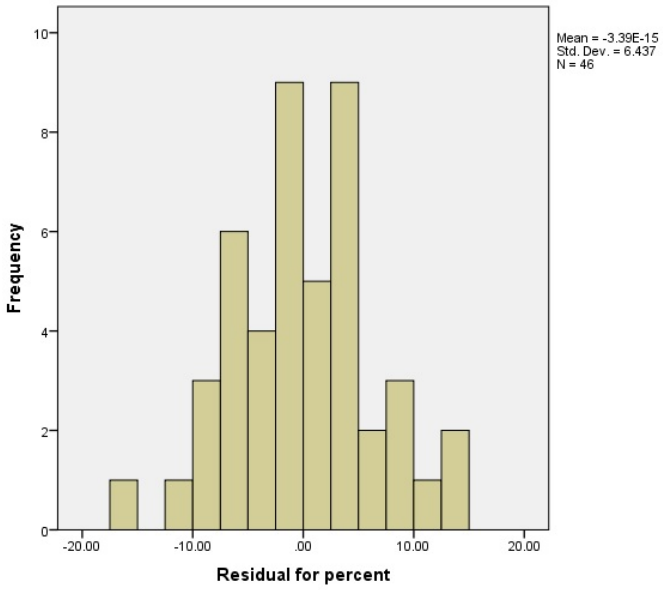
In order to check this assumption we focus on the error, which we will check for normality and homoscedasticity.
To do this we first need to extract the error from the measurements which leads us to the residuals. For one way ANOVA they are:

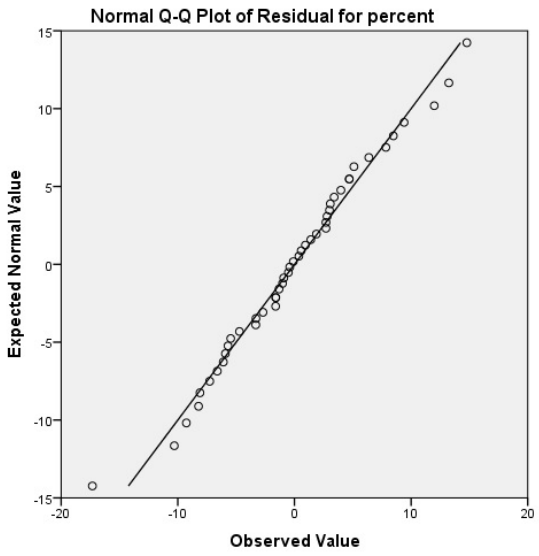
$$r_{ij} = x_{ij} - \bar{x}_i,$$

the difference between the measurement and the mean of all measurements taken for the same treatment. This results in a residual for each measurement in the sample which are interpreted as measurements for the error.
To check the assumptions we check,

1. if is reasonable to assume that the residuals are from a normal distribution (normality), and
2. if it is reasonable to assume that the standard deviations are the same for all treatments (homoscedasticity)

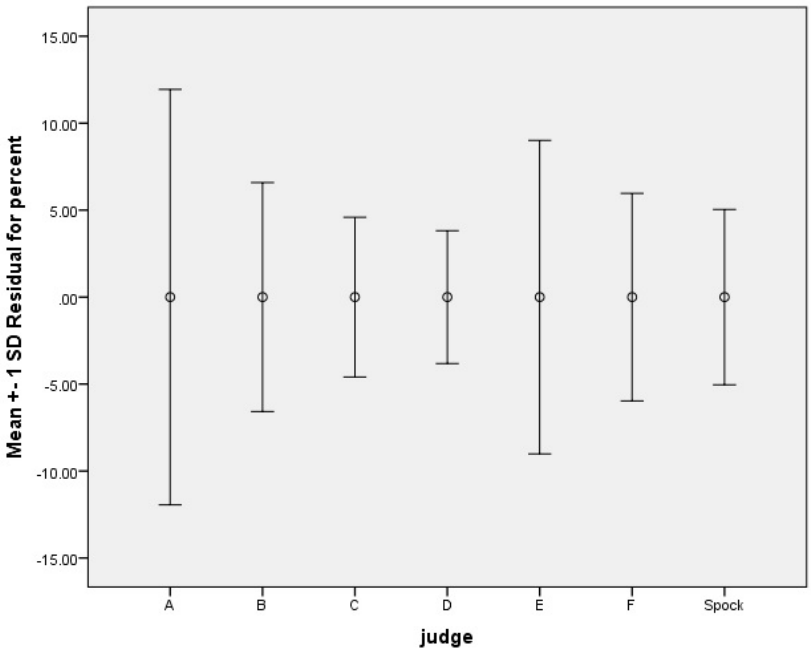
To check normality obtain histogram and QQ-plot for the residuals





Both graphs do not imply a strong deviation from normality, so we are not concerned about the normality assumption being violated.

To check homoscedasticity we plot the residuals versus the treatment (judges):



The standard deviation for judge A seems quite a bit larger than for judge D and is beyond of being acceptable.

In summary: We have no information if the venires were chosen at random, and this would be a problem for the ANOVA being appropriate. We also remain concerned that the homoscedasticity assumption might be violated. This would result in inappropriate for σ, and we should for example redo the confidence intervals for the pairwise comparisons of means using 2-sample t-confidence intervals.

2 Extra ANOVA Example

2.1 The example

A researcher wishes to try three different techniques to lower blood pressure of individuals diagnosed with high blood pressure. The subjects are randomly assigned to the three groups; the first takes medication, the second exercises and the third follows a certain diet.
After 4 weeks the reduction in each person's blood pressure is recorded.
The data

Medication	Exercise	Diet
10	6	5
12	8	9
9	3	12
15	0	8
13	2	4
$\bar{x}_1 = 11.8$	$\bar{x}_2 = 3.8$	$\bar{x}_3 = 7.6$
$s_1^2 = 5.7$	$s_2^2 = 10.2$	$s_3^2 = 10.3$

2.2 F-test

First find out if the treatment has an effect on the reduction in blood pressure. This calls for an ANOVA F-test.
Let μ_1= mean reduction in blood pressure for patients with high blood pressure treated with medication, μ_2, μ_3 similar.

- Hypotheses:

  $H_0 : \mu_1 = \mu_2 = \mu_3$ versus H_a : at least one mean is different

  Choose $\alpha = 0.05$

- Assumptions:
  The samples seem to be random samples. We do not know though how the patients were recruited. The standard deviations are close enough to each other to be of no concern.

  The samples are too small to check for normality. We will do a residual analysis later.

- Test statistic:

$$\bar{x}_{GM} = (10 + 6 + 5 + \ldots + 4)/(5 + 5 + 5) = 7.73$$
$$SSG = 5(11.8 - 7.73)^2 + 5(3.8 - 7.73)^2 + 5(7.6 - 7.73)^2 = 160.13$$
$$MSG = 160.13/(3 - 1) = 80.07$$
$$SSE = (5 - 1)5.7 + (5 - 1)10.2 + (5 - 1)10.3 = 104.8$$
$$MSE = 104.8/(15 - 3) = 8.73$$
$$F_0 = 80.07/8.73 = 9.17$$

  Therefore $F_0 = 9.17$ with $df_n = 2$ and df_d=12

- P-value:
  The test statistic is larger then the critical value for 0.005 (8.51). Therefore the P-value< 0.005
- Decision:
  At $\alpha = 0.05$ reject H_0, since P-value$< \alpha$.
- Context:
  At significance level of 5% the data provide sufficient evidence that the mean reduction in blood pressure is not the same for the three different treatment alternatives.

Now we know that the treatment has an effect on the reduction of the blood pressure. Which of the treatment is the best? Are there significant differences?

2.3 Multiple comparison

We will use the Bonferroni method.

1. $\alpha = 0.05$
2. m=3(2)/2=3
3. $\alpha^* = 0.05/3 = 0.0166$
4. The ranked means are

Exercise	Diet	Medication
3.8	7.6	11.8

5. For the percentile use $\alpha^*/2 = 0.008333 \approx 0.01$, $df_E = 12$, use $t^{12}_{0.01} = 2.681$. Since the sample sizes are the same for all three groups the Margin of Error is the same for all comparisons

$$ME = 2.681\sqrt{8.73}\sqrt{\frac{2}{5}} = 5.001$$

6. Result:

Exercise	Diet	Medication
3.8	7.6	11.8

   ———————

   ———————

   At experiment wise error rate of 5% the data provide sufficient evidence that the mean reduction in blood pressure for patients taking medication is significant higher than for patients who exercise. No significant difference could be found in the mean reduction for patients who used a certain diet and the other treatment groups.

2.4 Contrast

Another question we might have if the medication results in a significant higher mean reduction in blood pressure than the conservative treatments (diet, exercise).

- To answer the question we could use the following contrast:

$$\gamma = 2\mu_1 - (\mu_2 + \mu_3)$$

- A point estimate is:

$$c = 2\bar{x}_1 - (\bar{x}_2 + \bar{x}_3) = 12.2$$

We will conduct a t-test:

1. $H_0 : \gamma \leq 0$ versus $H_a : \gamma > 0$
   $\alpha = 0.05$
2. Assumptions: above
3. Test statistic:
$$t_0 = \frac{c}{SE(c)} = \frac{12.2}{3.237} = 3.769, \quad df = n - k = 12$$
   $SE(c) = \sqrt{8.73}\sqrt{2^2/5 + (-1)^2/5 + (-1)^2/5} = 3.237$
4. P-value: The test statistic t_0 is larger than 3.055, the 0.005 critical value with 12 df, there fore P-value<0.005.
5. Decision: P-value< α therefore reject H_0.
6. Context:
   At significance level of 5% the data provide sufficient evidence that medication results in a higher mean reduction in blood pressure than the more conservative treatments.