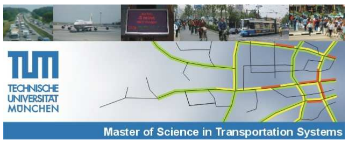
M.Sc. in ‚Transportation Systems'



Applied Statistics in Transport
ANOVA

Prof. Regine Gerike
Technische Universität München, mobil.TUM
regine.gerike@tum.de

Munich, 10/01/2011

Analysis of Variance

- Response variable: ratio/interval scaled
- Explanatory variables: categorical (ordinal, nominal, grouped interval data), are called factors
- Each factor has two or more levels
- For one single factor we can use the t-test

- ANOVA can be classified by the number of explanatory variables
- One-way ANOVA: one explanatory variable
- Two- and more way ANOVA: interaction effects can be included.

- We test whether the differences between the means of the different groups are high enough to conclude on differences in the populations; is the variation between the groups higher than within the groups?

ANOVA - Example

- We want to check whether the results of four different teaching methods differ significantly.
- The results (response, dependent variable) are tested with the number of points in the final test.
- The random sample contains 5 randomly drawn students from each group; the total sample size is 20.
- The following table shows the results of the final test:

	Teaching method			
	1	**2**	**3**	**4**
Scores	2	3	6	5
	1	4	8	5
	3	3	7	5
	3	5	6	3
	1	0	8	2
Total	10	15	35	20
Mean	2	3	7	4

ANOVA - Example

- Hypotheses:
- Null hypothesis: There are no differences in the results of the 4 teaching methods: H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_p$
- Alternative hypothesis: H_1: $\mu_i \neq \mu_{i'}$: Not all parameters μ are different but at least 2 arbitrary parameters μ_i and $\mu_{i'}$ differ.
- We accept H_1 if at least 2 teaching methods differ significantly in their results.

ANOVA versus t-tests

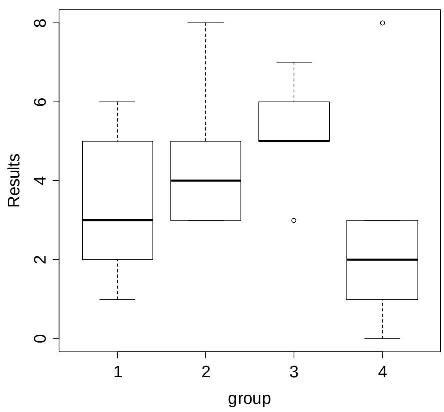
- In our example we could replace the ANOVA by $\binom{4}{2} = 6$ t-tests for independent samples, each of them comparing 2 samples.
- Disadvantage:
- In case of many t-tests, for some of them we will reject even though it is true (α-error), e.g. with a probability of 5%.
- For 6 tests we get on average 0.3 of such cases and not 0.05.
- This effect is called α-error-accumulation.
- Additionally: It is tedious to do those multiple tests.

- If we compare two samples: We get identical results with t-tests and ANOVA.

ANOVA - Example

- First step: Look at the data!

```
> a_vector<-c(2,3,6,5,  1,4,8,5,  3,3,7,5,  3,5,6,3,   1,0,8,2)
> group<-factor(rep(1:4,c(5,5,5,5)))
> plot(a_vector~group,ylab="Results",cex.lab=1.5,cex.axis=1.5)
```



ANOVA – Approach, Total Variation SSY

We calculate the total variation in the response variable (SSY) and

Partition it ('analyse it') into informative components:

Explained and unexplained variation.

Explained variation: treatment sum of squares (SSA)

Unexplained variation: error sum of squares (SSE, residual sum of squares)

First step: compute the total variation/the overall variance: $\hat{\sigma}^2_{total} = \frac{\sum_m (x_m - \bar{x})^2}{n-1}$

Sum of squares = SS: $\sum_m (x_m - \bar{x})^2$

Here SSY (total sum of squares), mean(a) gives 4.

```
> mean(a)
[1] 4
> #computing the total sum of squares SSY:
> (SSY<-((a-mean(a))^2))
     [,1] [,2] [,3] [,4]
[1,]    4    1    4    1
[2,]    9    0   16    1
[3,]    1    1    9    1
[4,]    1    1    4    1
[5,]    9   16   16    4
```

```
> a
     [,1] [,2] [,3] [,4]
[1,]    2    3    6    5
[2,]    1    4    8    5
[3,]    3    3    7    5
[4,]    3    5    6    3
[5,]    1    0    8    2
```

```
> sum(SSY)
[1] 100
> #d.f.=n*p-1=4*5-1
> (Var_total<-sum(SSY/19))
[1] 5.263158
```

$\rightarrow \hat{\sigma}^2_{total}$

ANOVA – Approach, Treatment Sum of Squares SSA

- Determine the part of variation that can be attributed to the four different teaching methods
- Question: How would the results look like if they were determined exclusively by the four different teaching methods?
- There would be no differences within the four teaching methods.
- Best guess for the effect of a teaching method: mean result.
- If the teaching methods were the only variance-generating source, all persons taught by one method should have the same results.

	Teaching method			
	1	**2**	**3**	**4**
Scores	2	3	6	5
	1	4	8	5
	3	3	7	5
	3	5	6	3
	1	0	8	2
Total	10	15	35	20
Mean	2	3	7	4

- This gives the following matrix:

```
> #treatment sum of squares
> (a_Treat_mean<-matrix(rep(cbind(mean(a[,1]),
+ mean(a[,2]),mean(a[,3]),mean(a[,4])),5),byrow=T,nrow=5))
     [,1] [,2] [,3] [,4]
[1,]    2    3    7    4
[2,]    2    3    7    4
[3,]    2    3    7    4
[4,]    2    3    7    4
[5,]    2    3    7    4
> mean(a_Treat_mean) #[1] 4
[1] 4
```

ANOVA – Approach, Treatment Sum of Squares SSA

```
> #treatment sum of squares
> (a_Treat_mean<-matrix(rep(cbind(mean(a[,1]),
+ mean(a[,2]),mean(a[,3]),mean(a[,4])),5),byrow=T,nrow=5))
     [,1] [,2] [,3] [,4]
[1,]    2    3    7    4
[2,]    2    3    7    4
[3,]    2    3    7    4
[4,]    2    3    7    4
[5,]    2    3    7    4
> mean(a_Treat_mean) #[1] 4
[1] 4
> #Squares deviations from the grand mean:
> (a_SSA<-((a_Treat_mean-mean(a_Treat_mean))^2))
     [,1] [,2] [,3] [,4]
[1,]    4    1    9    0
[2,]    4    1    9    0
[3,]    4    1    9    0
[4,]    4    1    9    0
[5,]    4    1    9    0
> (mean_diff_treat<-numeric(4))
[1] 0 0 0 0
> for (i in 1:4) {mean_diff_treat[i]<-(sum(a_SSA[,i]))}
> (a_SSA<-rbind(a_SSA,mean_diff_treat))
                [,1] [,2] [,3] [,4]
                   4    1    9    0
                   4    1    9    0
                   4    1    9    0
                   4    1    9    0
                   4    1    9    0
mean_diff_treat   20    5   45    0
> (a_SSA<-sum(a_SSA[6,]))
[1] 70
> #d.f._treat=p-1=4-1
> (Var_treat<-sum(a_SSA/(4-1)))
[1] 23.33333
```

Grand Mean

$\hat{\sigma}^2_{treat}$

Formula for computing the treatment variance:

$$SSA = \sum_i n * (\bar{A}_i - \bar{G})^2 = n * \sum_i (\bar{A}_i - \bar{G})^2$$

$$\hat{\sigma}^2_{treat} = \frac{SSA}{(p-1)} = \frac{n * \sum_i (\bar{A}_i - \bar{G})^2}{(p-1)}$$

ANOVA – Approach, Error Sum of Squares SSE

- Error variance: part of the overall variance that cannot be attributed to the teaching methods
- Independent from teaching methods, comes from other influences, such as motivation, intelligence
- Differences in results within one teaching method can be attributed to variables other than the teaching method
- We subtract the mean of each teaching method from the results for eliminating the influence of the teaching method:

```
> #Error sum of squares:
> a_SSE<-a
> for (i in 1:4) {a_SSE[,i]<-a[,i]-mean(a[,i])}
> (sum_diff_error<-numeric(4))
[1] 0 0 0 0
> for (i in 1:4) {sum_diff_error[i]<-(sum(a_SSE[,i]))}
> (a_SSE<-rbind(a_SSE,sum_diff_error))
               [,1] [,2] [,3] [,4]
                  0    0   -1    1
                 -1    1    1    1
                  1    0    0    1
                  1    2   -1   -1
                 -1   -3    1   -2
sum_diff_error    0    0    0    0
```

	Teaching method			
	1	**2**	**3**	**4**
Scores	2	3	6	5
	1	4	8	5
	3	3	7	5
	3	5	6	3
	1	0	8	2
Total	10	15	35	20
Mean	2	3	7	4

ANOVA – Approach, Error Sum of Squares SSE

```
> #Error sum of squares:
> a_SSE<-a
> for (i in 1:4) {a_SSE[,i]<-a[,i]-mean(a[,i])}
> (sum_diff_error<-numeric(4))
[1] 0 0 0 0
> for (i in 1:4) {sum_diff_error[i]<-(sum(a_SSE[,i]))}
> (a_SSE<-rbind(a_SSE,sum_diff_error))
               [,1] [,2] [,3] [,4]
                  0    0   -1    1
                 -1    1    1    1
                  1    0    0    1
                  1    2   -1   -1
                 -1   -3    1   -2
sum_diff_error    0    0    0    0
```

	Teaching method			
	1	**2**	**3**	**4**
Scores	2	3	6	5
	1	4	8	5
	3	3	7	5
	3	5	6	3
	1	0	8	2
Total	10	15	35	20
Mean	2	3	7	4

The squared differences between the mean of each teaching method and the individual results give the individual sum of squares (SSE1…4).

Divide those by the d.f. for getting the individual error variances:

```
> a_SSE_square<-a_SSE[1:5,]
> for (i in 1:4) {a_SSE_square[,i]<-a_SSE[1:5,i]^2}
> (sum_diff_error_squ<-numeric(4))
[1] 0 0 0 0
> for (i in 1:4) {sum_diff_error_squ[i]<-(sum(a_SSE_square[,i]))}
> (a_SSE_square<-rbind(a_SSE_square,sum_diff_error_squ))
                   [,1] [,2] [,3] [,4]
                      0    0    1    1
                      1    1    1    1
                      1    0    0    1
                      1    4    1    1
                      1    9    1    4
sum_diff_error_squ    4   14    4    8
> (SSE1<-a_SSE_square[6,1]) #similar for SSE2 to SSE4
sum_diff_error_squ
                 4
> #for computing the error variances we need SSE and the degrees of freedom:
> #the sum of differences within one group needs to be zero, so we have 4 d.f. per group.
> (Var_error1<-SSE1/4)  #similar for Var_error2 to Var_error4
sum_diff_error_squ
                 1
> #Total error variance = sum of error variances/d.f.
> (a_SSE<-sum(a_SSE_square[6,]))
[1] 30
> (Var_error<-(sum(a_SSE_square[6,])/(4+4+4+4)))
[1] 1.875
```

$\hat{\sigma}^2_{error}$

Formula for computing the error variance:

$SSE_i = \sum_i (x_{mi} - \bar{A}_i)^2$

$d.f._{\cdot error} = \sum_i d.f._{\cdot i} = \sum_i (n-1) = p*(n-1)$

$\hat{\sigma}^2_{error} = \frac{\sum_i SSE_i}{\sum_i d.f._{\cdot i}}$

ANOVA – Approach, Testing Significance

Null hypothesis: There are no differences in the results of the 4 teaching methods:
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_p$

Alternative hypothesis: H1: $\mu_i \neq \mu_{i'}$:
Not all parameters μ are different but at least 2 arbitrary parameters μ_i and $\mu_{i'}$ differ.

We accept H_1 if at least 2 teaching methods differ significantly in their results.

If H_0 is true, the treatment variance is equal to the error variance (there are no treatment effects); both can be used to estimate the population variance: $\hat{\sigma}^2_{treat} = \hat{\sigma}^2_{error}$.

We check the identity of the two variances with the F-test: $F = \hat{\sigma}^2_{treat} / \hat{\sigma}^2_{error}$

For our example: $F = \frac{23.33}{1.88} = 12.41$.

d.f. nominator = p-1, d.f.e denominator = p*(n-1)

qf(0.99,3,16)=5.292214

1-pf(12.444444,3,16)=0.0001875344

The empirical F-value is much higher than the critical value at the 1%-level. We can reject the null hypothesis: At least two of the teaching methods differ in their results at the 1%-level.

ANOVA – Approach, Summary, R

$SSA = 70$	$df_{treat} = 3$	$\hat{\sigma}^2_{treat} = 23.33$
$SSE = 30$	$df_{error} = 16$	$\hat{\sigma}^2_{error} = 1.88$
$SSY = 100$	$df_{total} = 19$	$\hat{\sigma}^2_{total} = 5.26$

With:

$SSY = SSA + SSE$, $df_{total} = df_{treat} + df_{error}$

The total sum of squares is the sum of the treatment sum of squares and the error sum of squares.

The degrees of freedom of the total variance are the sum of the d.f. of the treatment variance and of the d.f. of the error variance.

The variances are not additive.

```
> adfv
 [1] 2 1 3 3 1 3 4 3 5 0 6 8 7 6 8 5 5 5 3 2
> (TeachingMethod<-factor(rep(1:4,c(5,5,5,5))))
 [1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4
Levels: 1 2 3 4
> model_aov<-aov(adfv~TeachingMethod)
> summary(model_aov)
               Df Sum Sq Mean Sq F value    Pr(>F)    
TeachingMethod  3 70.000  23.333  12.444 0.0001875 ***
Residuals      16 30.000   1.875                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA – Individual Effects

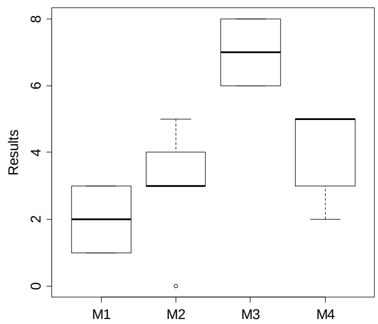
```
> summary.lm(aov(adfv~TeachingMethod))

Call:
aov(formula = adfv ~ TeachingMethod)

Residuals:
       Min         1Q     Median         3Q        Max 
-3.000e+00 -1.000e+00  2.317e-17  1.000e+00  2.000e+00 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       2.0000     0.6124   3.266  0.00485 ** 
TeachingMethod2   1.0000     0.8660   1.155  0.26517    
TeachingMethod3   5.0000     0.8660   5.774 2.85e-05 ***
TeachingMethod4   2.0000     0.8660   2.309  0.03460 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.369 on 16 degrees of freedom
Multiple R-squared:   0.7,	Adjusted R-squared: 0.6437 
F-statistic: 12.44 on 3 and 16 DF,  p-value: 0.0001875
```



ANOVA – Outlook

- Possible Reasons for a non-significant result of one-factorial ANOVA:
- Treatment has no influence (Var_treat is too small)
- Error variance is too big compared to treatment variance (Var_error too high)

We have no influence on the effect of the treatment, however we can reduce the error variance by:

1. Keeping other variables constant (age, sex,...), disadvantage: only statements for the analysed population possible (e.g. only for men of age 45-50)
2. Control variables: Analysis of Covariance, Crawley chapter 12
3. Vary variables systematically: multivariate ANOVA: error variance can be reduced, additionally we can analyse the influence of several variables + interactions among them, disadvantage: high sample size necessary: about 10 per cell
4. Probability for a significant result can be increased by higher sample sizes

ANOVA – Outlook

Important points for **survey design**, ask the following questions in advance:

- What is the dependent variable, how should it be measured?
- Which independent variables could potentially have an influence on this variable?
- Which factors with which levels should be included?
- What magnitude of ANOVA-effects do we expect?