MSc Applied Information and Data Science

HSLU Lucerne University of Applied Sciences and Arts

Data Analytics for Data Scientists

Design of Experiments (DoE)

Lecture 07 Addendum: Introduction to Analysis of Variance (ANOVA)

2025

Prof. Dr. Jürg Schwarz

Program: ~ 17:10 until 17:55

~ 17:00	**Begin of the lesson**

	Lecture: Jürg Schwarz
	◦ One-way ANOVA / Two-way ANOVA
	  ◦ An introductory example
	  ◦ Key steps & Designs of ANOVA
	  ◦ Main effects / Interaction
	◦ Running ANOVA with R
	  ◦ Details
	◦ Overview over Statistical Hypothesis Tests
	  ◦ Dependence on level of measurement

17:55	**End of the lesson**

Detailed table of contents → Slide 29

One-way ANOVA: An introductory example

Research in the Field of Human Resource Management

Survey of assistant's salaries

Table: Salary by level of experience [a]

	Level of Experience			
	1	**2**	**3**	**All**
All	36.-	38.-	42.-	39.-

[a] The table displays rounded mean values in CHF/h.

Grand Mean

Data

Sample of n = 96 assistants

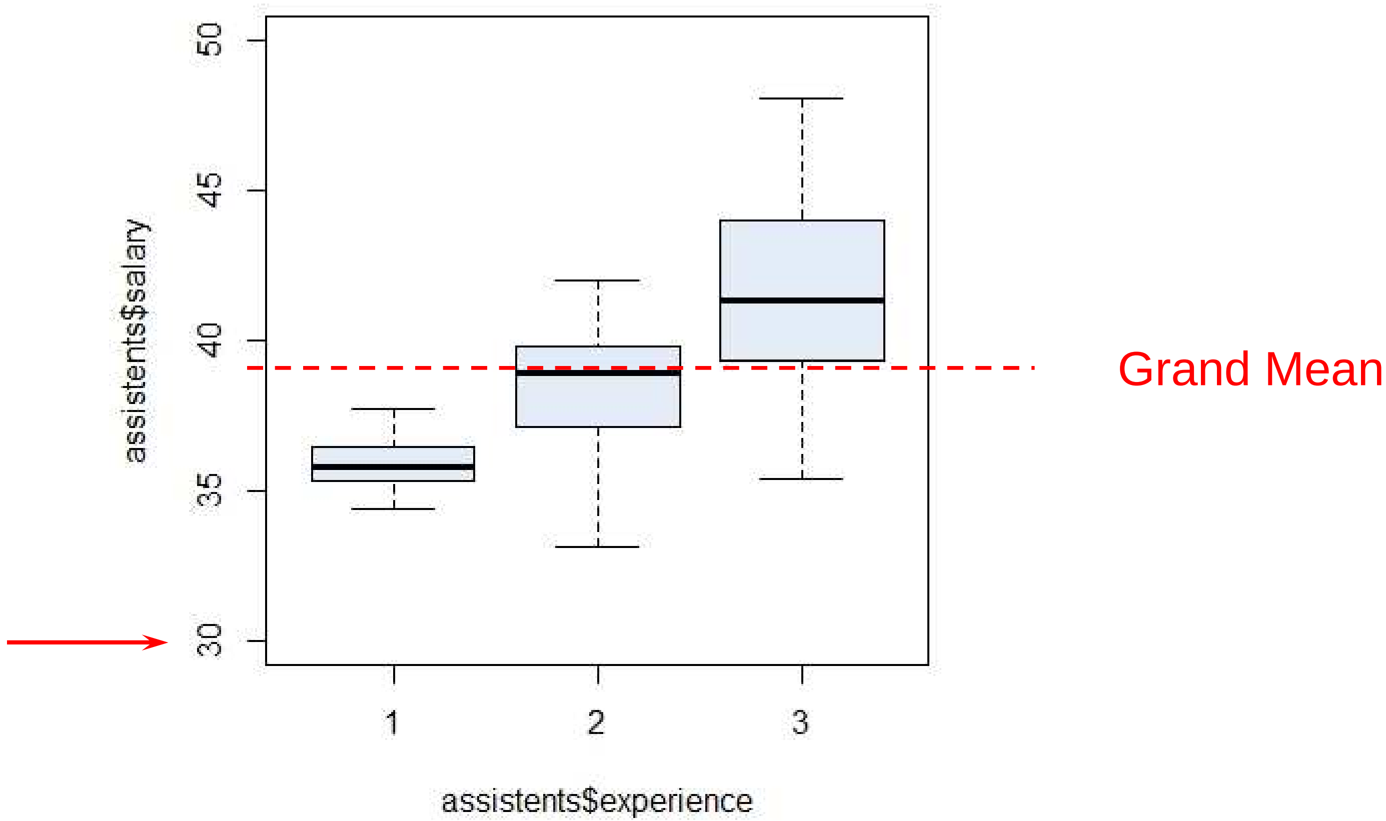
Among other variables: work experience (3 levels), salary (in CHF/h)

Typical questions

Does work experience have an impact on salary?

Are there salary differences depending on work experience?

Boxplot



According to the boxplot, salary (variable *salary*) differs more or less significantly with respect to work experience (variable *experience*).

Questions

Question in everyday language

Does work experience have an impact on salary?

Research question

What is the impact of work experience on salary?

What kind of model is suitable?

Is analysis of variance the right model?

Statistical question

Forming hypothesis

H_0: "No model" (= No overall model and no significant factors)

H_A: "Model" (= Overall model and significant factors)

Can H_0 be rejected?

Answer

Linear model with *salary* as the dependent variable (y_{gk} = salary of assistant k in group g)

$$y_{gk} = \bar{y} + \alpha_g + \varepsilon_{gk}$$

$\bar{y} =$ grand mean

$\alpha_g =$ effect of group g

$\varepsilon_{gk} =$ random term

* / ** See more on → Slide 16

How to run a One-way ANOVA in R

Scales

Dependent variable: metric

Independent variable(s): categorical (called "factor")

R script: ##### LECTURE 07

```
> oneway.test(salary ~ factor(experience), var.equal = TRUE, data = as ...)

  One-way analysis of means

data:  salary and factor(experience)
F = 46.483, num df = 2, denom df = 93, p-value = 1.013e-14
```

The overall model is significant: $F(2, 93) = 46.483$, $p = .000$*

```
> fit  <-  aov(salary ~ factor(experience), data = assistents)
> summary(fit, intercept = TRUE)
                   Df Sum Sq Mean Sq  F value   Pr(>F)
(Intercept)         1 143337  143337 25997.83  < 2e-16 ***
factor(experience)  2    513     256    46.48 1.01e-14 ***
Residuals          93    513       6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

There is a main effect of experience (levels 1, 2, 3) on salary, $F(2, 93) = 46.48$, $p = .000$. The value of adjusted $R^2 = .4892$** indicates that 48.9% of the variance in salary **around the grand mean** is uncovered by the model (here by *experience* only).

Key Steps in Analysis of Variance

1. **Design of experiment**
   - ANOVA is typically used for the analysis of results from experiments
     → Design of experiments (DoE) and ANOVA are related to each other*
2. **Calculating differences and sum of squares**
   - Differences between group means, individual values and grand mean are squared and summed up. This leads to the fundamental equation of ANOVA.
   - Test statistics for hypothesis test are calculated from the means of the sums of squares.
3. **Verification of the model**
   - Is the model as a whole significant (F test)?
   - Are the factors significant? Is there a significant interaction? (F tests)
   - How much variance does the analysis of variance explain?
     → Adjusted R squared / Partial Eta squared
4. **Considering other aspects** (post hoc tests)
5. **Testing of assumptions** (homogeneity of variance, etc.)
6. **Interpretation of the model and reporting** (profile plots)

*R. A. Fisher: Statistical methods for research workers (1925) / The Design of Experiments (1937)

Designs of ANOVA

Cross Section

Number of Variables			
Dependent Metric	Independent Categorical **Factor**	Independent Metric **Covariate**	
1	1	-	**One-way ANOVA** Analysis of variance with one factor
1	2 or more	-	**Multi-factorial ANOVA** Analysis of variance with two or more factors
2 or more	2 or more	-	**MANOVA** Multivariate Analysis of variance
1	1	1 or more	**One-way ANCOVA** Analysis of covariance with one factor
1	2 or more	1 or more	**Multi-factorial ANCOVA** Analysis of covariance with two or more factors
2 or more	2 or more	1 or more	**MANCOVA** Multivariate Analysis of covariance

Longitudinal

Number of Variables			
Dependent Metric	Independent Categorical **Factor**	Independent Metric **Covariate**	
1	2 or more*	-	**Repeated Measures ANOVA** Analysis of variance with repeated measurement
1	2 or more*	1 or more	**Repeated Measures ANCOVA** Analysis of covariance with repeated measurement

*The factor for the longitudinal dimension must always be introduced into the model.

Two-Way ANOVA

Research in human resource management: Survey of assistant's salaries

Table: Salary by level of experience and position [a]

		Level of Experience			
		1	**2**	**3**	**All**
Position	**Office**	35.-	37.-	39.-	37.-
	Lecture	37.-	40.-	44.-	40.-
All		36.-	38.-	42.-	39.-

[a] The table displays rounded mean values in CHF/h.

Now two factors are in the design

- Work experience (Level of experience 1-3): Variable *experience*
- Work position (Position in the office or in lectures): Variable *position*

Typical questions

Do experience and work position have an effect on salary? (→ main effects)
Is there an interaction between experience and work position? (→ interaction)

Main effects

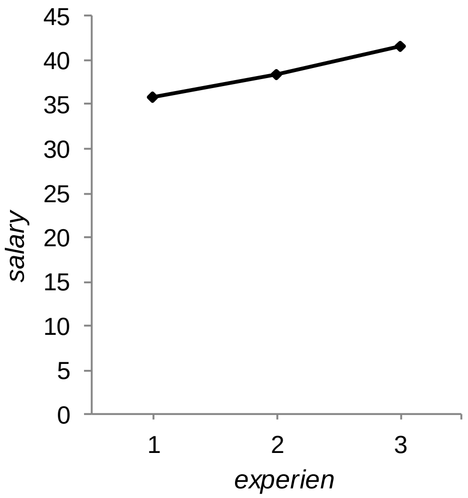
The direct effect of an independent variable on the dependent variable is called main effect.

In the example:

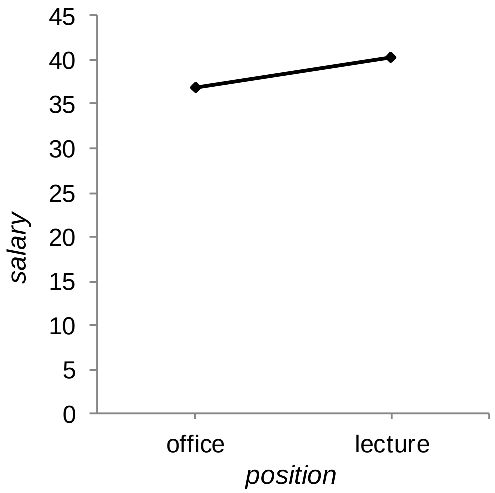
- The main effect of *experience* shows that the assistant's salary depends on their level of work experience.
- The main effect of *position* shows that the assistant's salary depends on whether they work in the office or in lectures.

Profile plots are used as visualization:

Main effect *experience*



Main effect *position*



If the profile plot shows a (nearly) horizontal line,
the main effect in question is probably not significant.

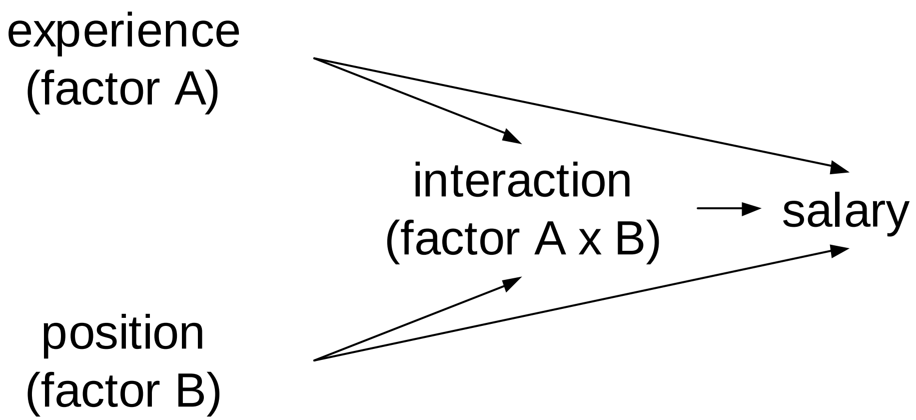
Interaction effect

An interaction between experience and position means there is dependency between the two variables.

The independent variables have a complex influence on the dependent variable.

The factors do not just function additively but act together in a different manner.

An interaction means that the effect of one factor depends on the value of another factor.



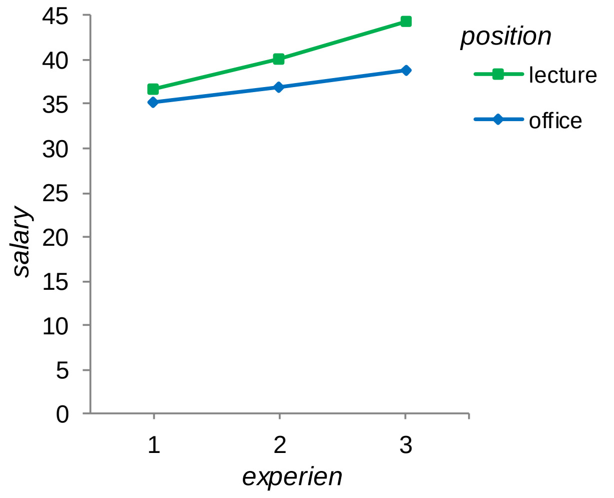
Interaction effects

In the example: The interaction between *experience* and *position* means that ...

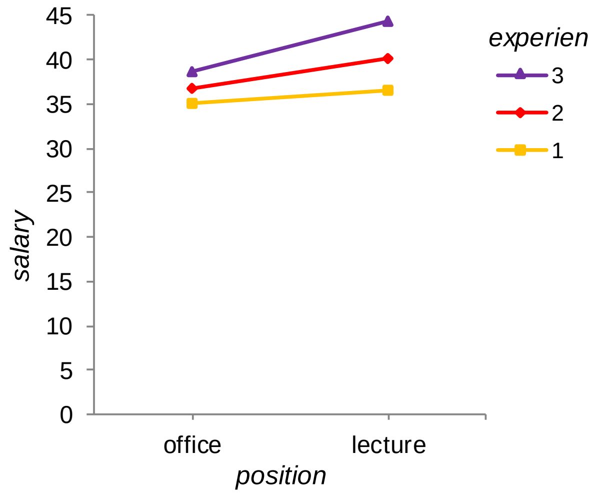
- the effect of work experience on salary is not the same for assistants who work in the office and for assistants who work in lectures.
- the difference in salary between assistants working in lectures and assistants working in the office depends on the level of experience.

Profile plots:

Separate lines for *position*



Separate lines for *experience*



If there is no interaction, the lines are parallel.

If there is an interaction, the lines are not parallel.

The more the lines deviate from being parallel, the more likely is an interaction.

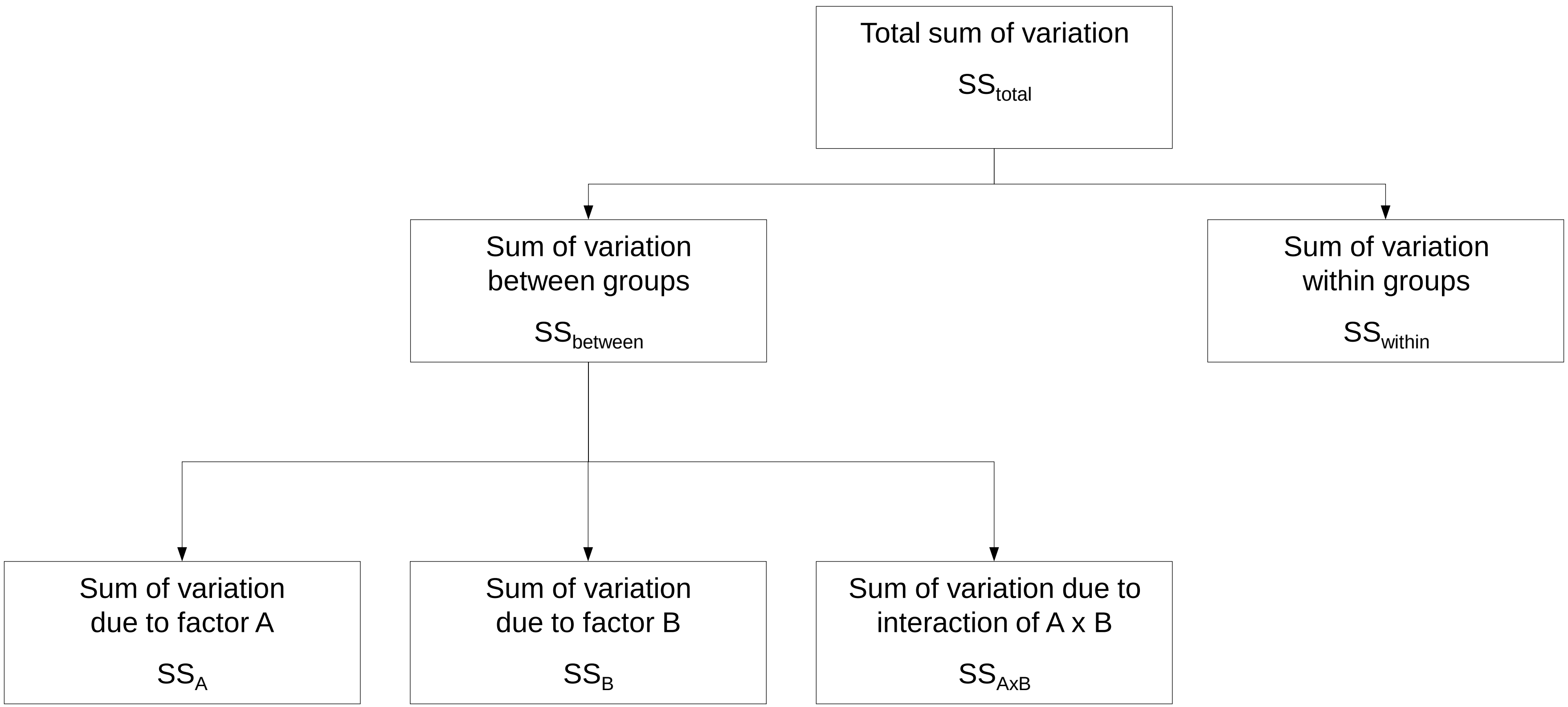
Sum of Squares (with interaction)

Given $SS_{total} = SS_{between} + SS_{within}$

With $SS_{between} = SS_{Experience} + SS_{Position} + SS_{Experience\ x\ Position}$

Follows $SS_{total} = (SS_{Experience} + SS_{Position} + SS_{Experience\ x\ Position}) + SS_{within}$

Where $SS_{Experience\ x\ Position}$ is the interaction of both factors **simultaneously**



Prerequisites of ANOVA

ANOVA is relatively robust against violations of prerequisites.

1. Sampling

   Randomly assigning participants to the treatment or control groups / No self-selection

   A well designed study avoids violation of this assumption.

2. Distribution of residuals

   Residuals (= error) are **normally** distributed

   Correction → e.g. logarithmizing

   Topic is not discussed any further in this module

3. Homogeneity of variances

   Residuals (= error) have **constant variance**

   Correction → weighting of variances

   Topic is not discussed any further in this module

4. Balanced design

   Same sample size in all groups

   Unbalanced designs can be corrected by "Type III" of the sum of squares.
   R-code: `summary(fit, intercept = TRUE, type="III")`

Running ANOVA with R

Back to One-way ANOVA

Output as on → Slide 6

* / ** See more on → Slide 16

How to run in R

Scales

Dependent variable: metric
Independent variable(s): categorical (called "factor")

R script: ###### LECTURE 07

```
> oneway.test(salary ~ factor(experience), var.equal = TRUE, data = as ...)

  One-way analysis of means

data:  salary and factor(experience)
F = 46.483, num df = 2, denom df = 93, p-value = 1.013e-14
```

The overall model is significant: $F(2, 93) = 46.483$, $p = .000$*

```
> fit  <-  aov(salary ~ factor(experience), data = assistents)
> summary(fit, intercept = TRUE)
                   Df Sum Sq Mean Sq  F value   Pr(>F)
(Intercept)         1 143337  143337 25997.83  < 2e-16 ***
factor(experience)  2    513     256    46.48 1.01e-14 ***
Residuals          93    513       6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a main effect of experience (levels 1, 2, 3) on salary, $F(2, 93) = 46.48$, $p = .000$.
The value of adjusted $R^2 = .4892$** indicates that 48.9% of the variance in salary **around the grand mean** is uncovered by the model (here by *experience* only).

*Overall model

In principle, the hypothesis structure requires two-stage testing:

H_0: "No model" (= No overall model and no significant factors)

H_A: "Model" (= Overall model and significant factors)

The factors may be considered only if the "overall model" is significant.

This can be tested with `oneway.test()`
In this module, the "overall model" is significant in each case → `oneway.test()` is dropped

**Adjusted R^2

The coefficient of determination R^2 is a measure of the goodness of fit of the model in terms of: How much of the variance in the data can be explained by the model.

This can be estimated with `lm()`

```
> fit_lm <- lm(salary ~ factor(experience), data = assistents)
> summary(fit_lm, intercept = TRUE)
                 :
Multiple R-squared:  0.4999, Adjusted R-squared:  0.4892
F-statistic: 46.48 on 2 and 93 DF,  p-value: 1.013e-14
```

In this module, this measure is not important → `lm()` is dropped

Multiple testing

Hypothesis structure of ANOVA

$H_0: \bar{y}_1 = \bar{y}_2 = \bar{y}_3$

$H_A: \bar{y}_i \neq \bar{y}_j$ for at least one pair ij

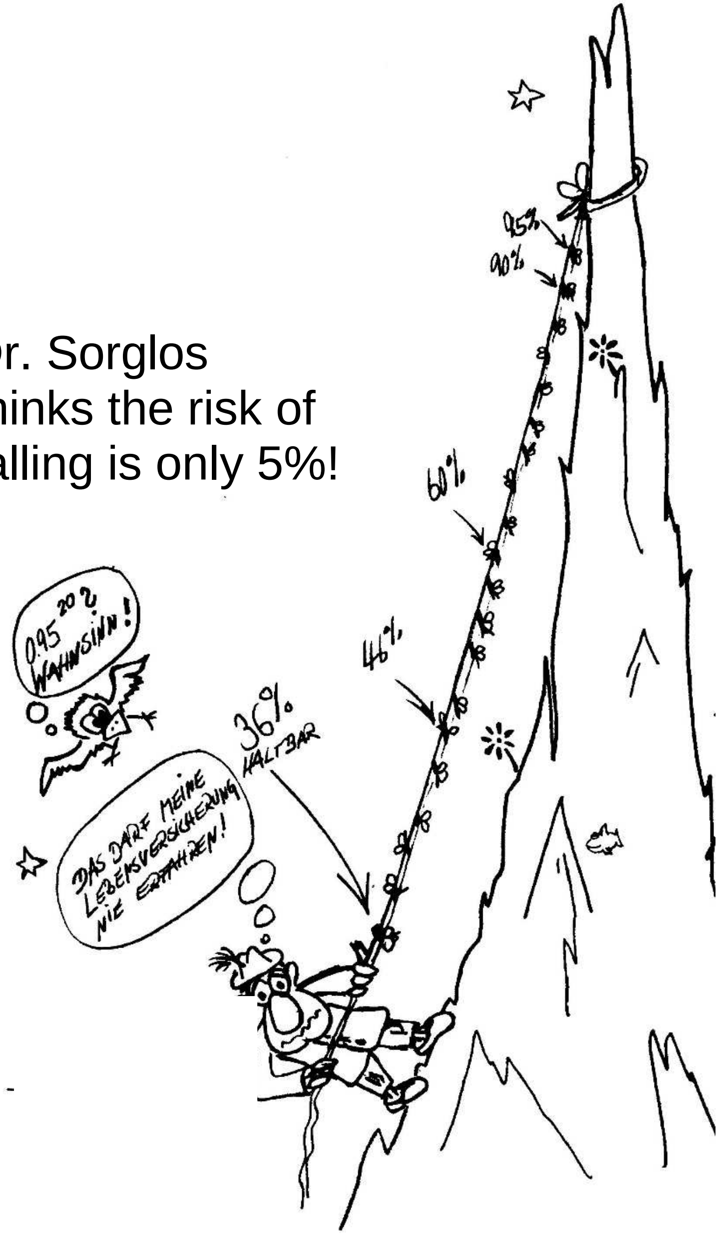
If H_0 is rejected, the group means differ

But, which of the groups differ?

Why not simply compare means pairwise?

<u>Example</u>: In the case of a rope with 20 knots,
each knot has α = 5% as the probability of failure.
All knots together have a probability of failure of
$1 - (1 - 0.05)^{20} = 0.64$.
The risk of a deadly fall therefore is 64%!
In order to keep this risk at the desired 5% level,
each knot may not exceed the probability of failure
of $\alpha_B = \alpha/\text{number of knots} = 5\%/20 = 0.25\%$.

Dr. Sorglos
thinks the risk of
falling is only 5%!



Cartoon: Dubben, H.-H. (2006). Der Hund, der Eier legt: Erkennen von Fehlinformation ...

Multiple testing → Post hoc tests

There are different methods to compare groups in pairs.
All methods are similar, however, in that they solve the problem of multiple testing.

<u>Example Bonferroni correction</u>:
If k means are pairwise tested, it becomes necessary to conduct $n = k \cdot (k - 1)/2$ tests.
To keep significance level $\alpha = 5\%$ for the entire test, each pairwise test must be conducted using significance level $\alpha = 5\%/n$.

```
> pairwise.t.test(assistents$salary, assistents$experience, p.adj = "bonf")

  Pairwise comparisons using t tests with pooled SD

data:  assistents$salary and assistents$experience

  1       2
2 0.00012 -
3 3.8e-15 2.3e-06

P value adjustment method: bonferroni
```

Example: Groups 1 and 2 have a significant difference: p = .000 (↔ .00012 rounded)

Compare with run `p.adj = "none"` → p = .000 (↔ .000039 rounded)

Effect size

Run ANOVA to obtain «partial eta squared»

```
library(effectsize)
eta_squared(fit)$Eta2
```

```
...
[1] 0.4999084     ≈ 0.5
```

Partial eta squared η_p^2 relates the variance explained by one factor to the variance not explained by other factors in the model.

Effect size f for the one-way analysis of variance according to Cohen (1992), calculated from η_p^2

$$f = \sqrt{\frac{\eta_p^2}{1-\eta_p^2}} = \sqrt{\frac{0.5}{1-0.5}} = \sqrt{\frac{0.5}{0.5}} = 1.00$$

	Small	Medium	Large
Effect size *f*	0.10	0.25	0.40

There is a main effect of experience (levels 1, 2, 3) on salary, $F(2, 93) = 46.48$, $p = .000$.
Effect size is $f = 1.00$.

The effect size is large (Cohen 1992).

Table from the article *A Power Primer* by Cohen (1992)

Test	ES index	Effect size: Small	Medium	Large
1. m_A vs. m_B for independent means	$d = \frac{m_A - m_B}{\sigma}$	.20	.50	.80
2. Significance of product–moment r	r	.10	.30	.50
3. r_A vs. r_B for independent rs	$q = z_A - z_B$ where z = Fisher's z	.10	.30	.50
4. $P = .5$ and the sign test	$g = P - .50$	.05	.15	.25
5. P_A vs. P_B for independent proportions	$h = \phi_A - \phi_B$ where ϕ = arcsine transformation	.20	.50	.80
6. Chi-square for goodness of fit and contingency	$w = \sqrt{\sum_{i=1}^{k} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$	.10	.30	.50
7. One-way analysis of variance	$f = \frac{\sigma_m}{\sigma}$	.10	.25	.40
8. Multiple and multiple partial correlation	$f^2 = \frac{R^2}{1 - R^2}$	.02	.15	.35

In our case the equation is different because R does not provide σ_m and σ

Running Two-way ANOVA with R

Only main effects, no interaction

```
> fit  <-  aov(salary ~ factor(experience) + factor(position),   data = assist ...)
> summary(fit, intercept = TRUE)
                   Df Sum Sq Mean Sq F value Pr(>F)    
(Intercept)         1 143337  143337 56615.1 <2e-16 ***
factor(experience)  2    513     256   101.2 <2e-16 ***
factor(position)    1    280     280   110.5 <2e-16 ***
Residuals          92    233       3                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a main effect of experience (levels 1, 2, 3) on salary, $F(2, 92) = 101.2$, $p = .000$.
There is a main effect of position (levels 1, 2) on salary, $F(1, 92) = 110.5$, $p = .000$.

experience (1, 2, 3) and position (1, 2) have a significant effect on salary.

With interaction

```
factor()  +  factor()→ without interaction
factor()  *  factor()→ with interaction
```

```
> fit  <-  aov(salary ~ factor(experience) * factor(position),   data = …)
> summary(fit, intercept = TRUE)
                                     Df Sum Sq Mean Sq  F value  Pr(>F)
(Intercept)                           1 143337  143337 78925.76 < 2e-16 ***
factor(experience)                    2    513     256   141.12 < 2e-16 ***
factor(position)                      1    280     280   154.08 < 2e-16 ***
factor(experience):factor(position)   2     69      35    19.13 1.2e-07 ***
Residuals                            90    163       2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

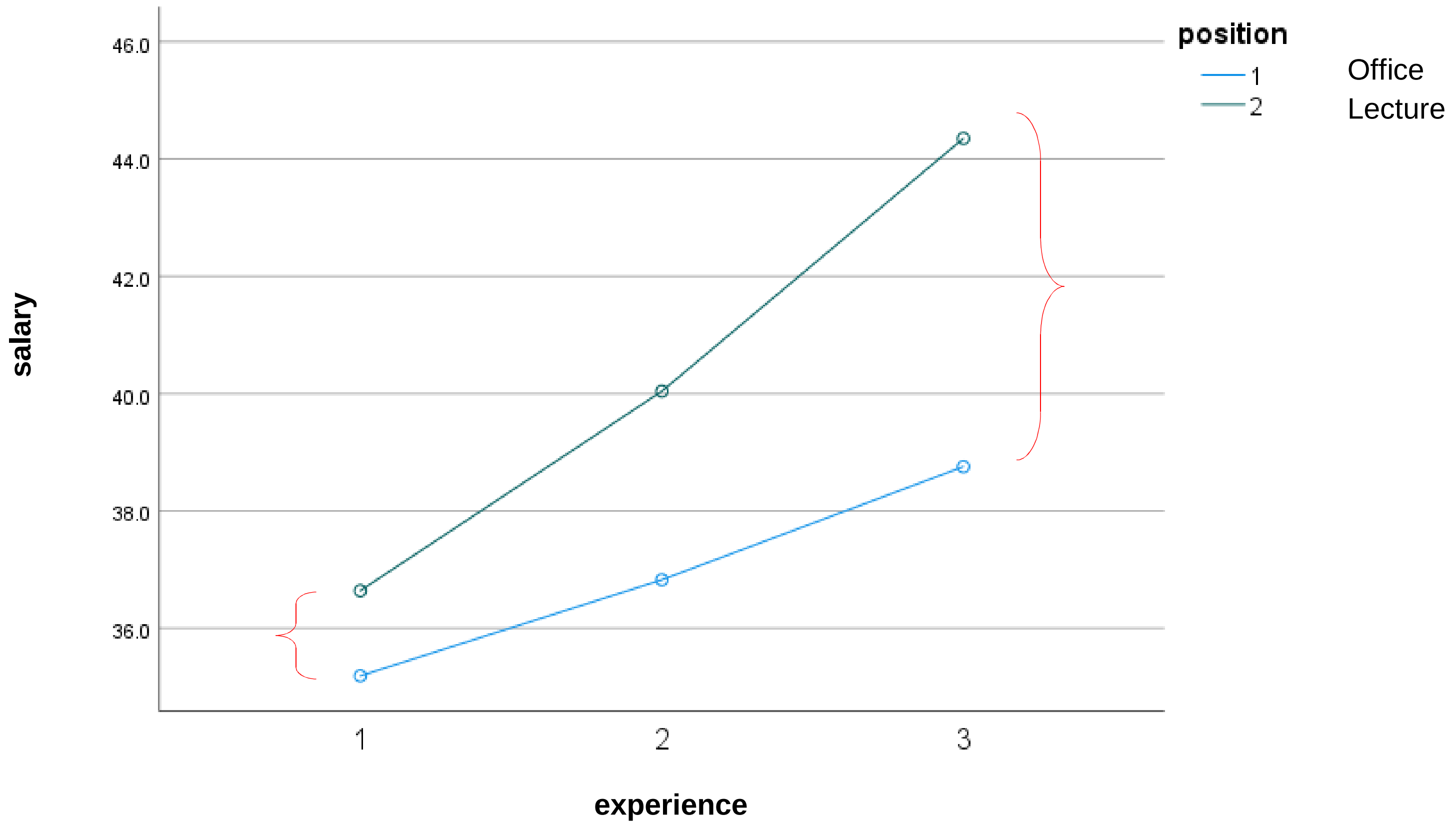
There is a main effect of experience (levels 1, 2, 3) on salary, $F(2, 90) = 141.12$, $p = .000$.
There is a main effect of position (levels 1, 2) on salary, $F(1, 90) = 154.08$, $p = .000$.

There is a significant interaction, $F(2, 90) = 19.13$, $p = .000$.

Interpretation of the interaction

Do different levels of experience influence the impact of different levels of position differently?

Yes, if experience has values 2 or 3 then the influence of position is raised.

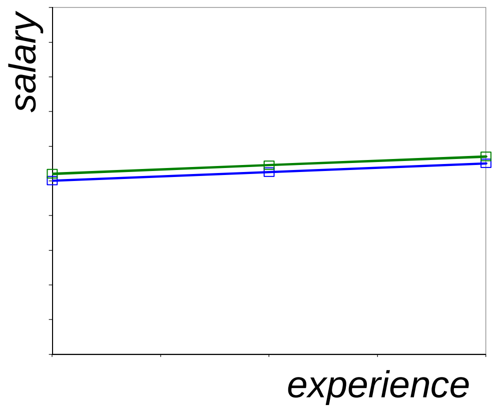


Simplified: Lines are not parallel

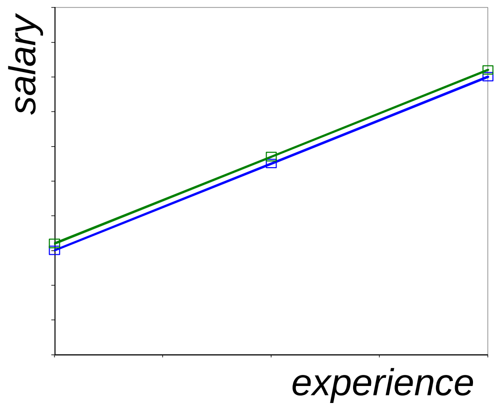
Interpretation: Experience is more important in lectures than in offices.

Examples of interactions

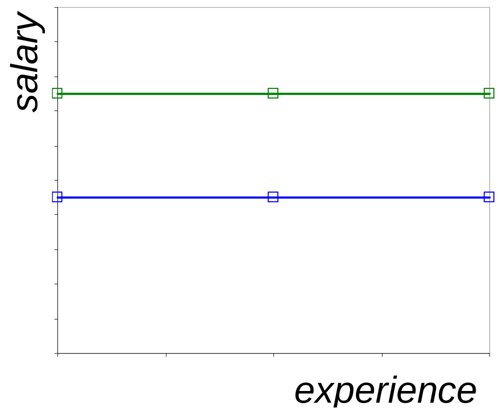
- ✗ Main effect of *experience*
- ✗ Main effect of *position*
- ✗ Interaction



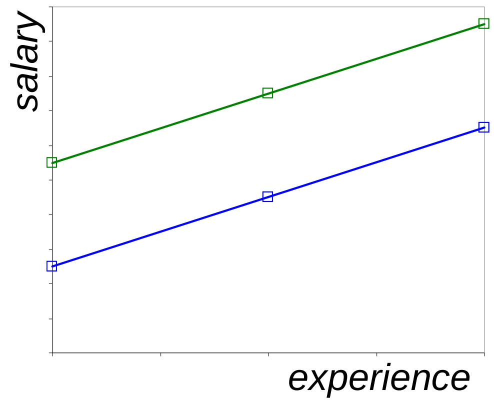
- ✓ Main effect of *experience*
- ✗ Main effect of *position*
- ✗ Interaction



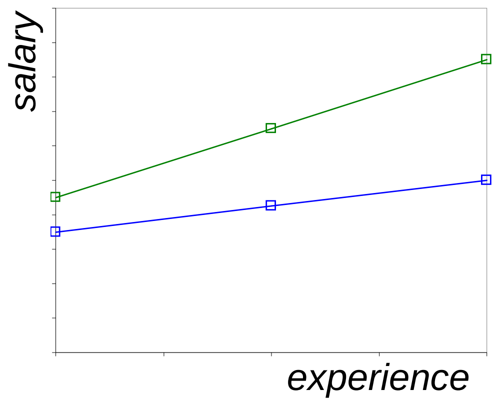
- ✗ Main effect of *experience*
- ✓ Main effect of *position*
- ✗ Interaction



- ✓ Main effect of *experience*
- ✓ Main effect of *position*
- ✗ Interaction



- ✓ Main effect of *experience*
- ✓ Main effect of *position*
- ✓ Interaction



- ✗ Main effect of *experience*
- ✗ Main effect of *position*
- ✓ Interaction

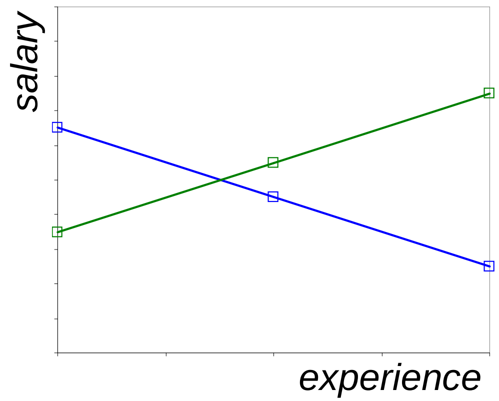


Table of contents

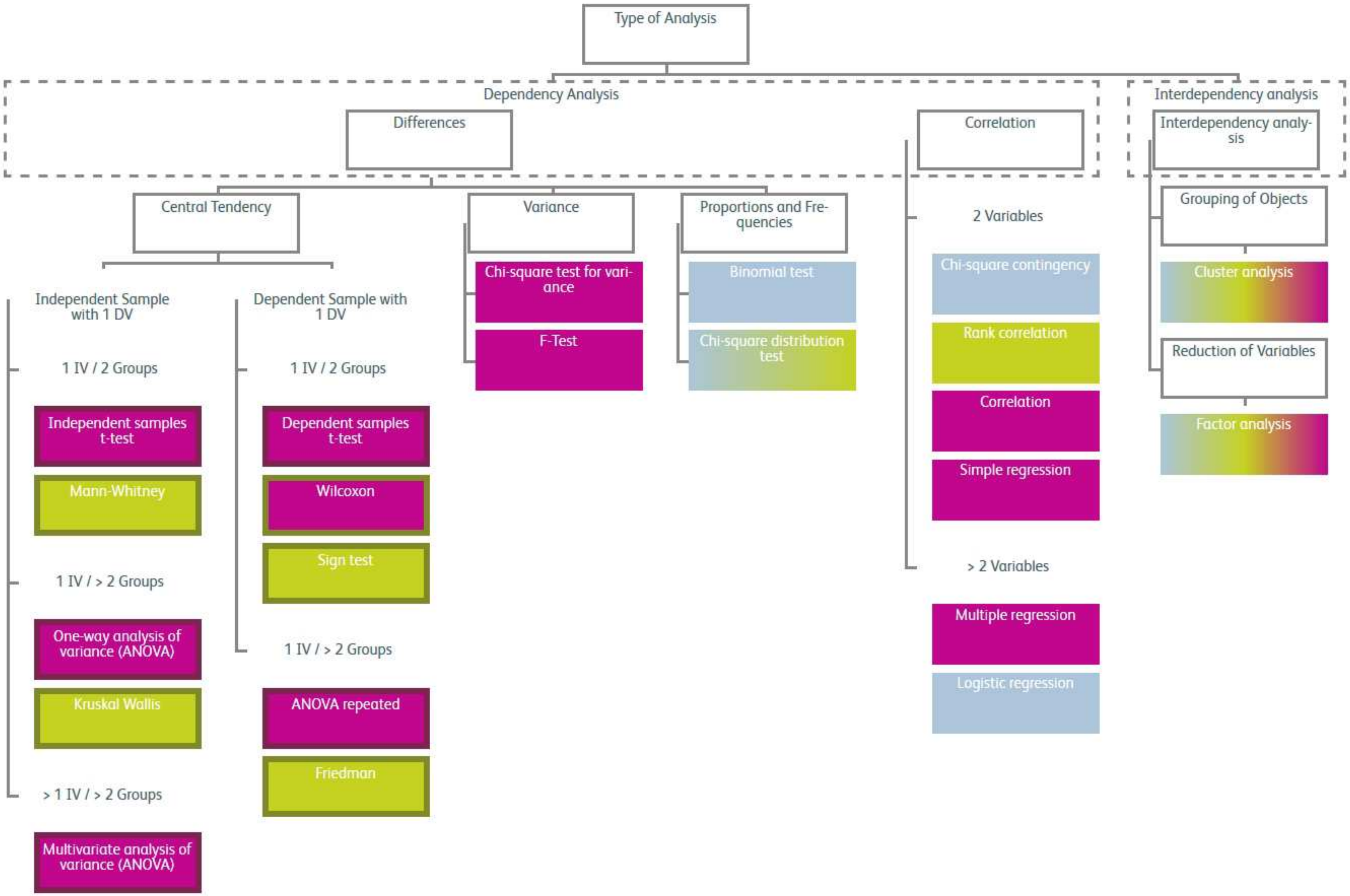
Appendix

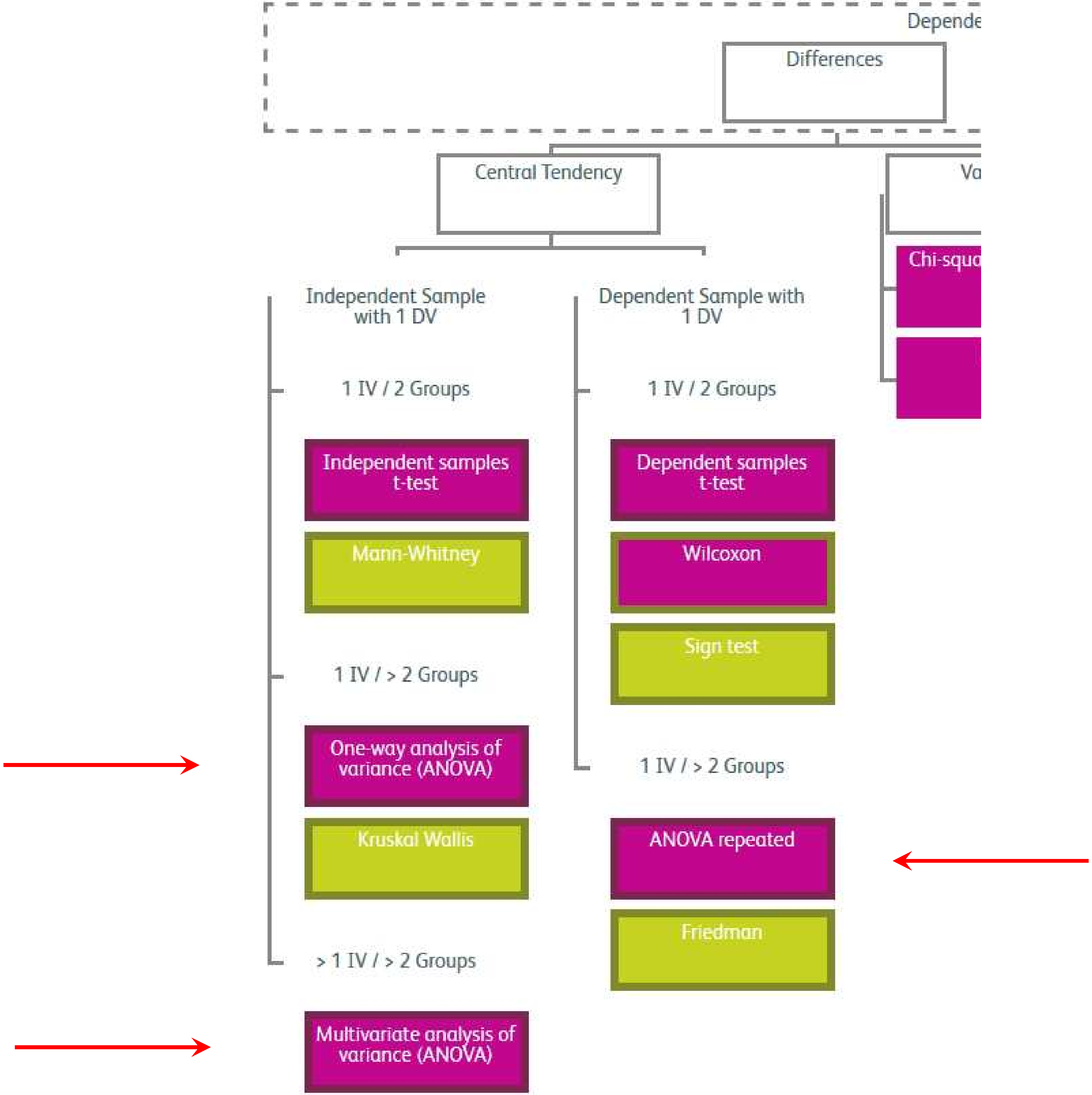
Overview over Statistical Hypothesis Tests

Choosing the type of analysis depending on level of measurement

Dependent Variable (DV)	Independent Variable (IV): Categorical (Nominal & Ordinal)	Independent Variable (IV): Metric (Interval & Ratio)
Metric (Interval & Ratio)	**Analysis of Variance (ANOVA)** IV mostly categorical (factors) <u>In addition</u>: Introduce metric IV as covariate(s) (AN**C**OVA)	**Regression Analysis** IV mostly metric <u>In addition</u>: Introduce categorical IV as dummy variable(s)
Categorical (Nominal & Ordinal)	**Chi-Square Test** No distinction between DV and IV <u>In addition</u>: Introduce layer variable to separate subgroups	**Logistic Regression Analysis** IV mostly metric <u>In addition</u>: Introduce categorical IV as dummy variable(s)

Decision tree → www.empirical-methods.hslu.ch/home-english



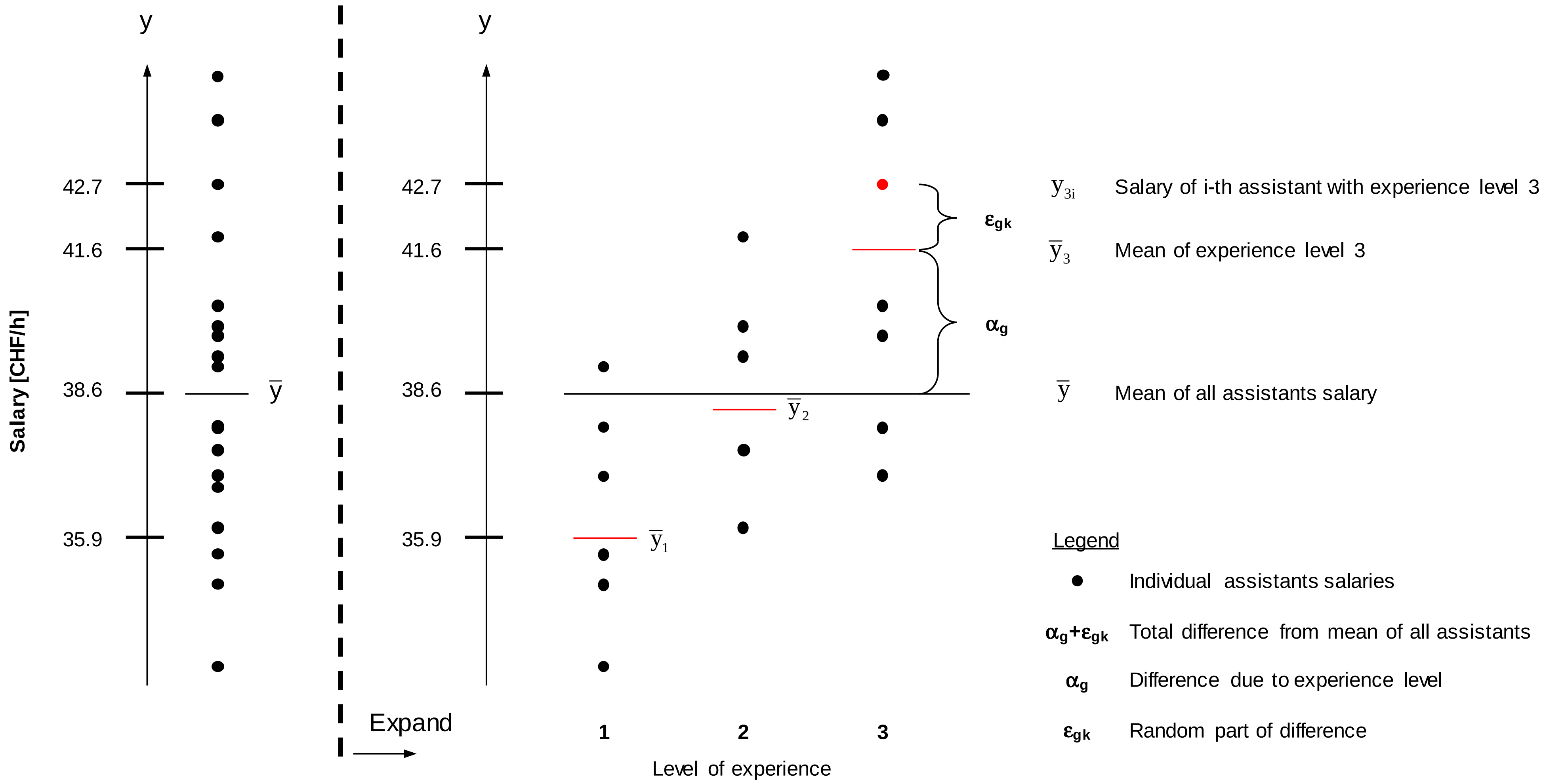

Depende
Differences
Central Tendency
Va
Chi-squa
Independent Sample with 1 DV
1 IV / 2 Groups
Independent samples t-test
Mann-Whitney
1 IV / > 2 Groups
One-way analysis of variance (ANOVA)
Kruskal Wallis
> 1 IV / > 2 Groups
Multivariate analysis of variance (ANOVA)
Dependent Sample with 1 DV
1 IV / 2 Groups
Dependent samples t-test
Wilcoxon
Sign test
1 IV / > 2 Groups
ANOVA repeated
Friedman

Guess: What if $\bar{y}_1 \approx \bar{y}_2 \approx \bar{y}_3$?

Sum of Squares

Step by step

Survey on assistant's salary: Salaries differ by level of experience.



Basic idea of ANOVA

If $\bar{y}_1 \approx \bar{y}_2 \approx \bar{y}_3$, then $SS_{between} << SS_{within}$

Total sum of squared differences $\mathbf{SS_{total}}$ is split into two parts
(SS is short for Sum of Squares)

- $\mathbf{SS_{between}}$ Sum of squared differences <u>due to groups</u> ("between groups", treatments)
  (here: between levels of experience)
- $\mathbf{SS_{within}}$ Sum of squared differences <u>due to randomness</u> ("within groups", also SS_{error})
  (here: within each experience group)

Fundamental equation of ANOVA:

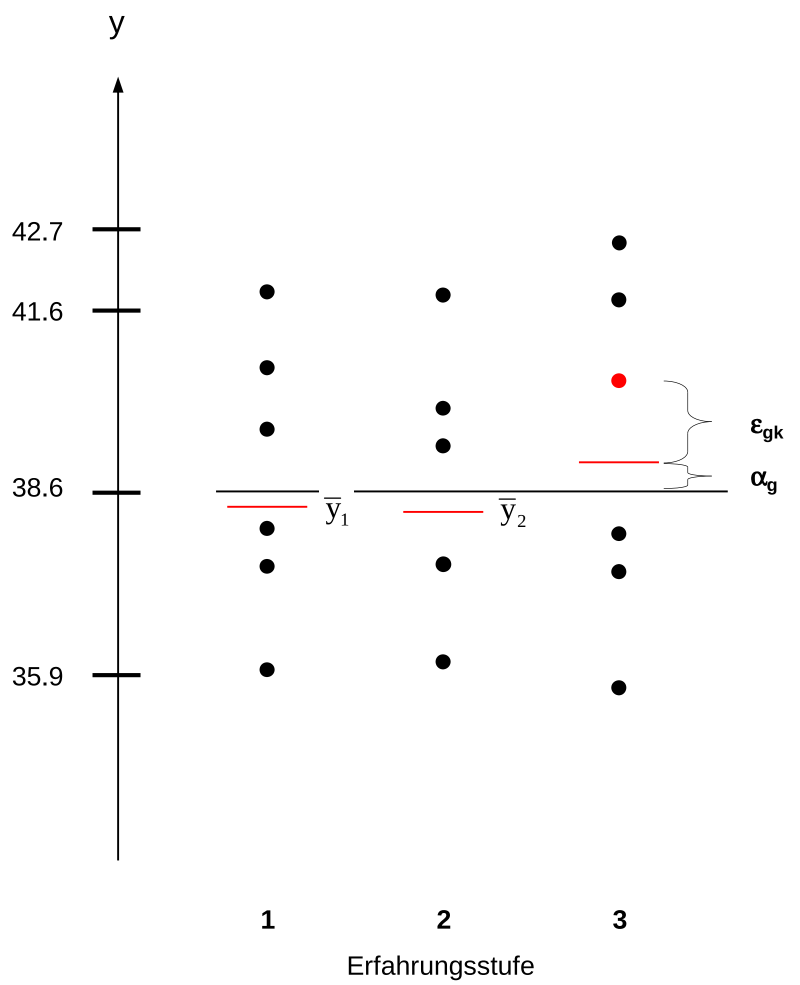
$$\underbrace{\sum_{g=1}^{G}\sum_{k=1}^{K_g}(y_{gk}-\bar{y})^2}_{SS_{total}} = \underbrace{\sum_{g=1}^{G}K_g(\bar{y}_g-\bar{y})^2}_{SS_{between}} + \underbrace{\sum_{g=1}^{G}\sum_{k=1}^{K_g}(y_{gk}-\bar{y}_g)^2}_{SS_{within}}$$

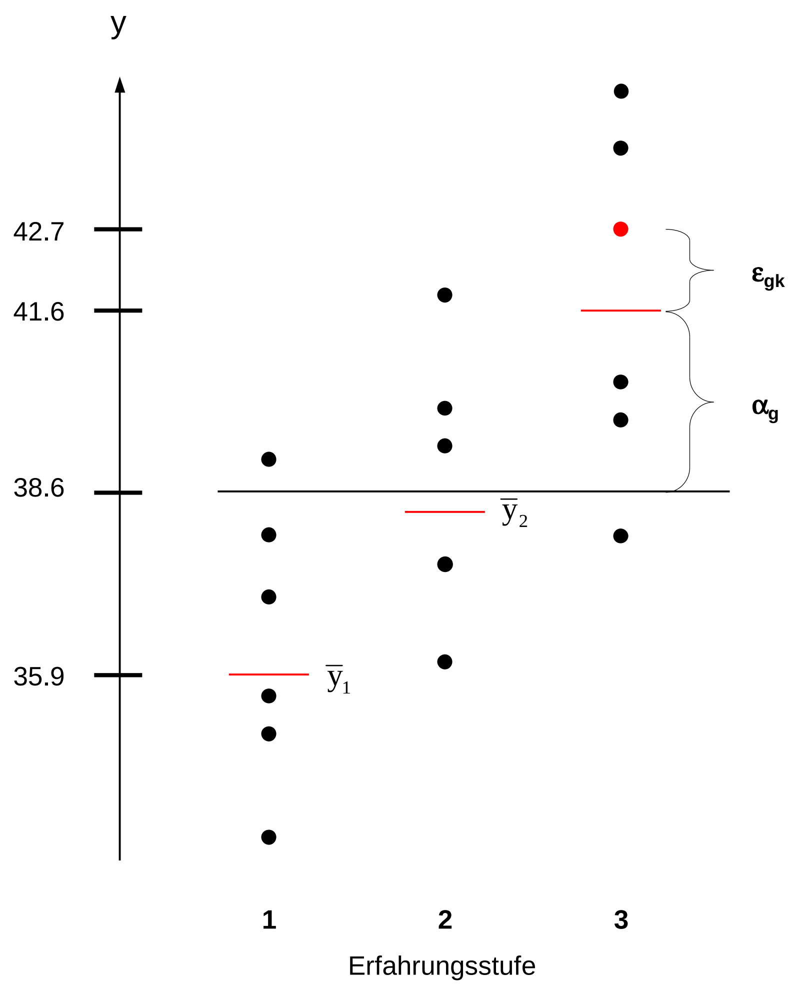
g: index for groups from 1 to G (here: G = 3 levels of experience)

k: index for individuals within each group from 1 to K_g
(here: $K_1 = K_2 = K_3 = 32$, $K_{total} = K_1 + K_2 + K_3 = 96$ assistants)

Comparison

If $\bar{y}_1 \approx \bar{y}_2 \approx \bar{y}_3$, then $SS_{between} << SS_{within}$





$$F = \frac{SS_{between}}{SS_{within}} \quad \text{"small"}$$

$$F = \frac{SS_{between}}{SS_{within}} \quad \text{"large"}$$

If $\bar{y}_1 \approx \bar{y}_2 \approx \bar{y}_3$, then $MS_{between} << MS_{within}$

Hypothesis testing for the model

Test statistic F for significance testing is computed by relation of means of sum of squares (MS = Mean Squares)

$$MS_{total} = \frac{SS_{total}}{K_{total} - 1}$$ Mean of SS_{total}

$$MS_{between} = \frac{SS_{between}}{G - 1}$$ Mean of $SS_{between}$

$$MS_{within} = \frac{SS_{within}}{K_{total} - G}$$ Mean of SS_{within}

MS_{xyz} is proportional to SS_{xyz}

Therefore the comparison

$F = \frac{SS_{between}}{SS_{within}}$ "small" vs. $F = \frac{SS_{between}}{SS_{within}}$ "large"

keeps its meaning.

Calculating test statistic F and hypothesis testing for the global model

$$F = \frac{MS_{between}}{MS_{within}}$$

F follows an F distribution with (G – 1) and (K_{total} – G) degrees of freedom

The F test verifies the hypothesis that the group means are equal:

H_0: $\bar{y}_1 = \bar{y}_2 = \bar{y}_3$

H_A: $\bar{y}_i \neq \bar{y}_j$ for at least one pair ij