

Wissenschaftliche Methodik I

Varianzanalyse I

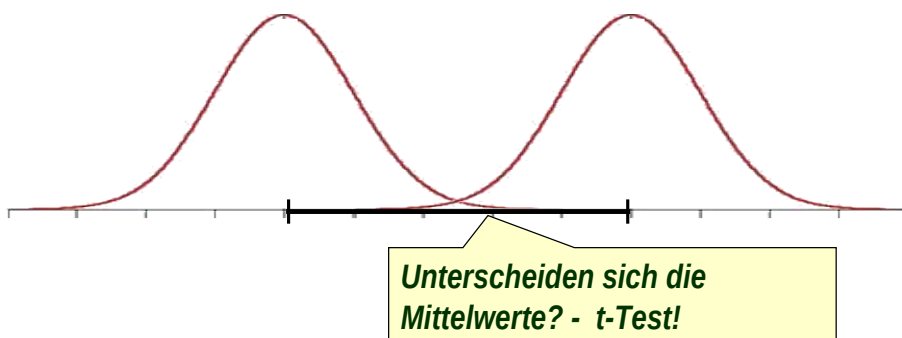
Beispieldatensätze zu dieser VL (enthalten im Workspace "ANOVA_1.Rdata") :

- *points.csv*
- *d1.csv* und *d2.csv*
- *bZIP.csv* und *bZIP_2.csv*
- *Repmes.csv*

Verwendete Pakete:

- *nortest*

Verschiedenheit von Mittelwerten



- Nullhypothese: $H_0: \mu_1 = \mu_2$
- Alternativhypothese: $H_1: \mu_1 \neq \mu_2$

T-Statistik:

- Metrische Daten: t-Test

- Zumindest Stichproben-Mittelwerte sind annähernd normalverteilt
- Varianzhomogenität

Prüfgröße

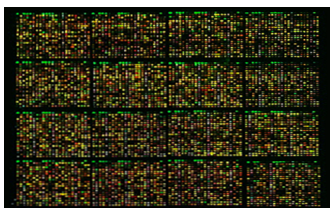
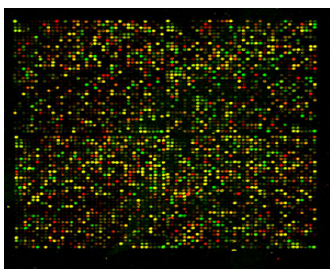
$$T = \frac{\bar{X} - \bar{Y}}{S \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Mittelwertdifferenz

Standardabweichung

Maß für
Probenumfang

Multiples Testen:



- Gene Array: Vergleich zweier Zellkulturen

- 24.000 Gene

- Für Irrtumswahrscheinlichkeit 5%:

- Error = $(1 - 0.95) * 24.000 = 1200$
Falsch-Positive

- 2 Vergleiche unter 2 Bedingungen

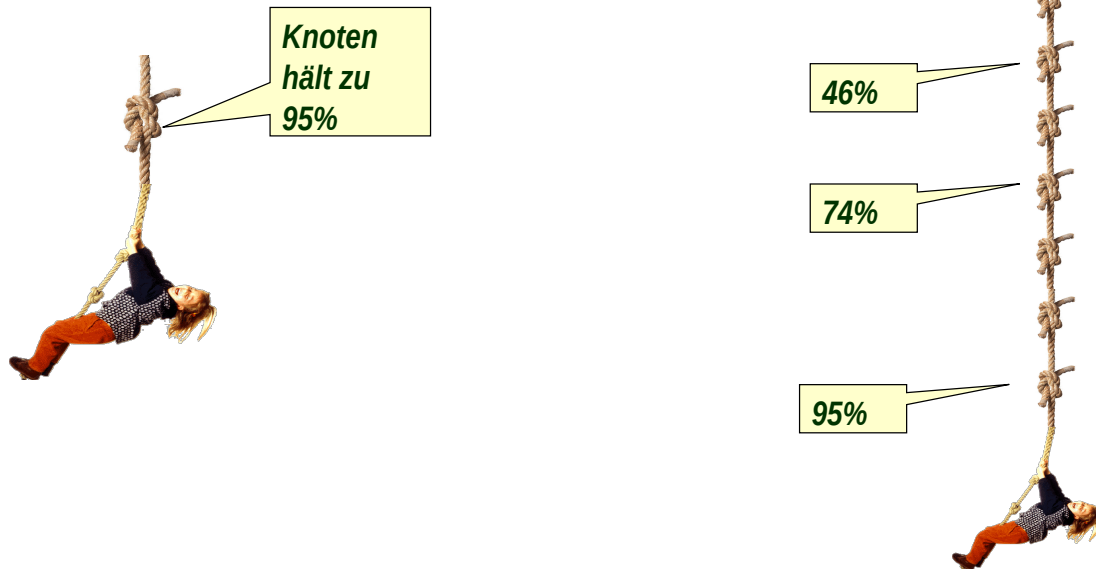
- Error = $(1 - 0.95^2) * 24.000 = 2340$

- 4 Vergleiche unter 4 Bedingungen:

- Error = $(1 - 0.95^4) * 24.000 \approx 4452$

19%

Kumulierte Fehlerraten:



Kumulation des α -Fehlers:

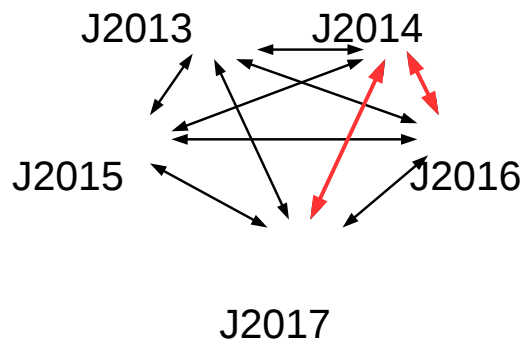
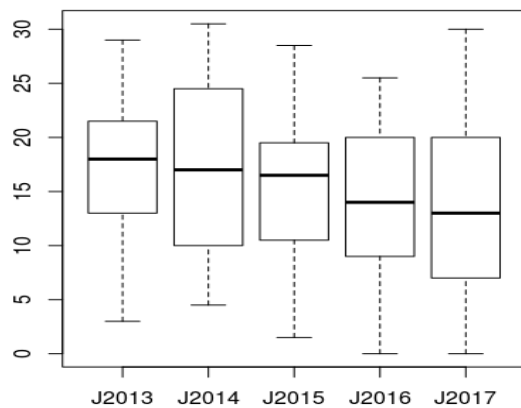
Gruppen	Vergleiche	Kumulierter α -Fehler
3	3	0.143
4	6	0.264
5	10	0.401
6	15	0.537
7	21	0.659
8	28	0.762
9	36	0.842
10	45	0.901
11	55	0.940
12	66	0.966
13	78	0.982
14	91	0.991
15	105	0.995

$$n * (n-1)/2$$

Ab 6 Gruppen ist das Ergebnis purer Zufall!

Kumulation des α -Fehlers:

Erreichte Punkte in "Pflanzliche Systeme"



Bonferroni-Korrektur:

Das α -Fehler-Niveau wird für jeden einzelnen Test soweit herabgesetzt, dass der kumulierte Fehler nur noch 0.05 beträgt:

$$\alpha_{adj} = \frac{\alpha}{N_{Tests}}$$

Beispiel: 5 Gruppen \Rightarrow 10 Tests $\Rightarrow \alpha_{adj} = 0.05 / 10 = 0.005$

Nachteil:

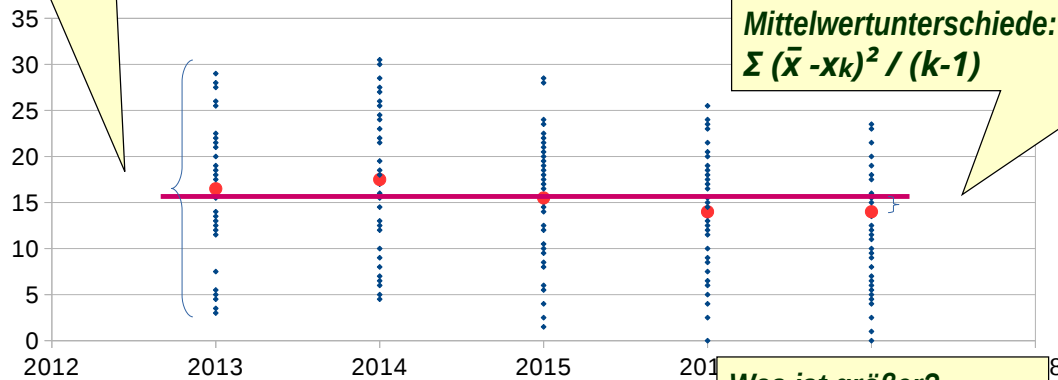
- \Rightarrow sehr niedriges Alpha-Niveau bei den einzelnen Tests
- \Rightarrow geringe Auflösung ("Power"), d.h. großer β -Fehler

Alternative: Berechnung einer Varianzanalyse

Vergleich mehrerer Mittelwerte:

Streuung in der Gruppe:
 $\sum (\bar{x} - x_i)^2 / (N-k)$

Punktzahl 1. Klausur "Pflanzliche Systeme"



Mittelwertunterschiede:
 $\sum (\bar{x} - \bar{x}_k)^2 / (k-1)$

Was ist größer?

Theorie der ANOVA:

- Wie groß ist die gesamte Streuung im Datensatz?
 - Summe der Abweichungsquadrate (SS) aller Einzelwerte vom Gesamtmittelwert
- Welcher Teil der Streuung ist erklärbar? *among group variation*
 - Welcher Teil der Streuung geht auf den Gruppeneffekt zurück?
 - SS der Gruppenmittelwerte vom Gesamtmittelwert!
- Welcher Teil der Streuung ist zufällig? *within group variation*
 - Wie groß ist die Streuung innerhalb der Gruppen?
 - SS der Einzelwerte vom Gruppenmittelwert!

Theorie der ANOVA: **F-Statistik**

- Verhältnis *among / within group variation*:

– Der **F-Wert**

$$F_{\text{Stichprobe}} = \frac{S_2^2}{S_1^2} = \frac{\frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2}{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2}$$

Berechnung mit dem Statistikpaket "R"

```
> summary(aov(Punkte~Jahrgang, data=points))
      Df Sum Sq Mean Sq F value Pr(>F)
Jahrgang    4    447   111.64    2.161  0.074 .
Residuals  243  12554    51.66
---
Signif. Codes:  ***: p < 0.001; **: p < 0.01; *: p < 0.05
```

Es gibt keinen signifikanten Effekt des Jahrgangs

Einfaktorielle ANOVA:

```
> ps.aov = aov(Punkte~Jahrgang, data=points)
> summary(ps.aov)
      Df Sum Sq Mean Sq F value Pr(>F)
Jahrgang    4    447   111.64    2.161  0.074 .
Residuals  243  12554    51.66
---
Signif. Codes:  ***: p < 0.001; **: p < 0.01; *: p < 0.05
```

Abhängige Variable

Faktor

- Einfluss einer unabhängigen Variablen (Faktor) mit k verschiedenen Ausprägungen (level) auf eine abhängige Variable

- Nullhypothese:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k = \mu_{\text{total}}$$

- Alternativhypothese:

$$H_1: \mu_x \neq \mu_{\text{total}}$$

Einfaktorielle ANOVA:

„in Abhängigkeit von“

```
> ps.aov = aov(Punkte~Jahrgang, data=points)
> summary(ps.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Jahrgang	4	447	111.64	2.161	0.074 .
Residuals	243	12554	51.66		

Signif. Codes: ***: p < 0.001; **: p < 0.01; *:p < 0.05

```
>summary(aov(Punkte~Jahrgang, data=points))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Jahrgang	4	447	111.64	2.161	0.074 .
Residuals	243	12554	51.66		

Signif. Codes: ***: p < 0.001; **: p < 0.01; *:p < 0.05

ANOVA: Voraussetzungen

Nicht so leicht zu messen! Aber: wenn die Werte normal verteilt sind, sind es die Residuen auch!

- **Normalverteilung**

- Nicht notwendig die Werte, aber die Residuen innerhalb jeder Gruppe, müssen normal verteilt sein
- Andernfalls wäre der Gruppenmittelwert kein guter Schätzer für die Gruppeneigenschaft!

- Bei $n < 50$: Shapiro-Wilk Test: `shapiro.test(subset(datcol, datgroup=="group1"))`
Bei $50 < n < 100$: Kolmogorov-Smirnov mit Lilliefors-Korrektur:
`library("nortest")`
`lillie.test(subset(datcol, datgroup=="group1"))`

ANOVA: Voraussetzungen

- Normalverteilung

- Nicht notwendig die Werte, aber die Residuen innerhalb jeder Gruppe, müssen normal verteilt sein
- Andernfalls wäre der Gruppenmittelwert kein guter Schätzer für die Gruppeneigenschaft!

Normierung von Daten durch Logarithmieren:

```
> Log = log10(data$value)
> data2 = data.frame(data, Log)
```

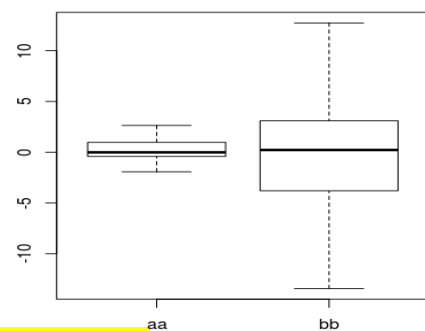
ANOVA: Voraussetzungen

- Normalverteilung

```
> summary(aov(x~t3, data=d1))
      Df Sum Sq Mean Sq F value Pr(>F)
t3      1    3.2   3.169    0.145  0.705
Residuals 98 2149.0  21.929
```

- Varianzhomogenität

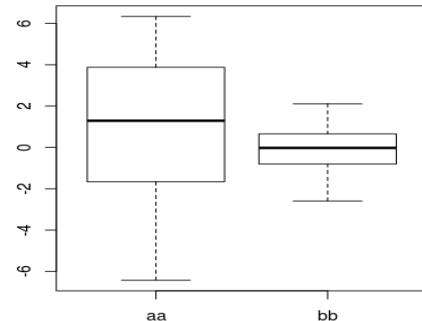
- Weniger problematisch bei gleichen Probenumfängen
- Größerer β -Fehler, wenn geringe Varianz in "kleiner" Gruppe



ANOVA: Voraussetzungen

- Normalverteilung

```
> summary(aov(y~t3, data=d2))
              Df Sum Sq Mean Sq F value Pr(>F)
t3              1   19.7   19.706    4.569  0.035 *
Residuals     98  422.6    4.313
```



- Varianzhomogenität

- Weniger problematisch bei gleichen Probenumfängen
- Größerer β -Fehler, wenn geringe Varianz in "kleiner" Gruppe
- Größerer α -Fehler, wenn große Varianz in "kleiner" Gruppe

Alternative zur ANOVA: non-parametrische Tests

- Wenn keine Normalverteilung besteht und nicht durch Normierung erzeugt werden kann,

Male: $1+2+6+8+9 = 26$

- Wenn Varianzen stark inhomogen oder Gruppen sehr unterschiedlich groß sind:

Female: $3+4+5+7+10 = 29$

factor	variable	rank
male	12,6	1
male	12,7	2
female	13,5	3
female	13,9	4
female	15,2	5
male	17	6
female	27	7
male	28	8
male	42	9
female	45	10

- Rangsummen-Tests

- Mann-Whitney U Test bzw. Wilcoxon W Test für 2 Gruppen
- Kruskal-Wallis für > 2 Gruppen

Alternative zur ANOVA: non-parametrische Tests

$$U1 = n1 * n2 + (n1 (n1+1))/2 - R1$$

$$= 25 + 15 - 26 = 14$$

$$U2 = n1 * n2 + (n2 (n2+1))/2 - R2$$

$$= 25 + 15 - 29 = 11$$

$$Z = (U_{\min} - \hat{u}) / SE_U$$

$$\text{Male: } 1+2+6+8+9 = 26$$

$$\text{Female: } 3+4+5+7+10 = 29$$

factor	variable	rank
male	12,6	1
male	12,7	2
female	13,5	3
female	13,9	4
female	15,2	5
male	17	6
female	27	7
male	28	8
male	42	9
female	45	10

```
> wilcox.test(y~t3, data=d2)
```

Wilcoxon rank sum test with continuity correction

data: y by t3

W = 1111, p-value = 0.1685

alternative hypothesis: true location shift is not equal to 0

Alternative zur ANOVA: non-parametrische Tests

```
> wilcox.test(y~t3, data=d2)
```

Wilcoxon rank sum test with continuity correction

data: y by t3

W = 1111, p-value = 0.1685

alternative hypothesis: true location shift is not equal to 0

```
> kruskal.test(y~t3, data = d2)
```

Kruskal-Wallis rank sum test

data: y by t3

Kruskal-Wallis chi-squared = 1.9075, df = 1, p-value = 0.1672

Varianzanalyse: Ergebnis

- H_0 : alle Mittelwerte sind gleich
- H_1 : Mittelwerte sind nicht alle gleich
- Die ANOVA sagt nicht, welche Mittelwerte verschieden sind!
 - Geplante Kontraste
 - a priori Hypothesen nötig: Experimentdesign!
 - Post Hoc Tests
 - Keine a priori Hypothesen; multiple Tests

Geplante Kontraste: Beispiel

qPCR für 4 bZIP Transkriptionsfaktoren

- \bar{X}_i : mittlere ΔC_t Werte aus 10 Wiederholungen
- Vermutung: \bar{X}_3 und \bar{X}_4 unterscheiden sich

Kontraste:

$$\widehat{C} = w_1 \bar{X}_1 + w_2 \bar{X}_2 + w_3 \bar{X}_3 + w_4 \bar{X}_4$$

Nur 3; 4:

$$\widehat{C} = 0 * \bar{X}_1 + 0 * \bar{X}_2 + 1 * \bar{X}_3 + (-1) * \bar{X}_4$$

$$\sum w_i = 0$$

t-Statistik:

$$t = \frac{\widehat{C}}{\sqrt{MS_w \left[\sum_{i=1}^k \frac{w_i^2}{n_i} \right]}}$$

Beispiel bZIP: Durchführung in R

1) Analyse der Faktor-Stufen:

```
> levels(bZIP$bzip)
[1] "a" "b" "c" "d"
```

```
> levels(bZIP_2$bzip)
NULL
```

```
> levels(as.factor(bZIP_2$bzip))
[1] "1" "2" "3" "4"
```

2) Analyse der Kontrast-Tabelle des Datensatzes

```
> contrasts(bZIP$bzip)
  b c d
a 0 0 0
b 1 0 0
c 0 1 0
d 0 0 1
```

Vergleich "b" mit Mittelwert

... "c" mit Mittelwert

...und "d" mit Mittelwert

Beispiel bZIP: Durchführung in R

3) Ändern der Kontrast-Tabelle: nur "c" mit "d" vergleichen

```
> contrasts(bZIP$bzip) <- c(0,0,-1,1)
```

```
> contrasts(bZIP$bzip)
  [,1] [,2] [,3]
a    0 -0.8535534 -0.1464466
b    0  0.4225890 -0.7559223
c   -1  0.2154822  0.4511845
d    1  0.2154822  0.4511845
```

Von R berechnete Coeffizienten
für die "Gesamt-ANOVA";
für uns nicht relevant

4) Durchführung der ANOVA:

```
> model <- aov(dCT ~ bzip, data= bZIP)
```

Beispiel bZIP: Durchführung in R

5) "gesplittete" Anzeige:

...den 1. Vergleich anzeigen

```
> summary.aov(model, split=list(bzip=list("c vs.d"=1)))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
bzip	3	282.6	94.20	11.50	1.96e-05	***
bzip: c vs.d	1	224.4	224.45	27.39	7.35e-06	***
Residuals	36	295.0	8.19			

6) Zurücksetzen der Kontrast-Tabelle:

```
> contrasts(bZIP$bzip) <- contr.treatment
```

```
> contrasts(bZIP$bzip)
```

	2	3	4
a	0	0	0
b	1	0	0
c	0	1	0
d	0	0	1

Geplante Kontraste: Beispiel 2

Geplanter Kontrast: \bar{X}_4 unterscheidet sich vom Mittelwert der anderen

$$\widehat{C} = 1 \cdot \bar{X}_1 + 1 \cdot \bar{X}_2 + 1 \cdot \bar{X}_3 + (-3) \cdot \bar{X}_4$$

$$\sum w_i = 0$$

Parallele Vergleiche von Faktorstufen:

$$\widehat{C}_a = 1 \cdot \bar{X}_1 + -1 \cdot \bar{X}_2 + 0 \cdot \bar{X}_3 + 0 \cdot \bar{X}_4$$

$$\widehat{C}_a = 0 \cdot \bar{X}_1 + 0 \cdot \bar{X}_2 + 1 \cdot \bar{X}_3 + -1 \cdot \bar{X}_4$$

$$\sum w_{1i} \cdot w_{2i} = 0$$

Beachte: Geplante Kontraste müssen orthogonal sein, d.h. unabhängig.

1/2 & 3/4: $\sum w_{1i} \cdot w_{2i} = (+1)(0) + (-1)(0) + (0)(+1) + (0)(-1) = 0$

1/2 & 2/3: $\sum w_{1i} \cdot w_{2i} = (+1)(0) + (-1)(+1) + (0)(-1) + (0)(0) = -1$

Varianzanalyse: Post Hoc Tests

- $H_1: \mu_x \neq \mu_{\text{total}}$ » mindestens zwei μ verschieden!

- Anzahl möglicher Vergleiche: $C = k(k - 1) / 2$

multiples Testen!

- Post Hoc Tests berücksichtigen "**error inflation**"

- Tests mit Annahme von Varianz-Homogenität:

- Least Squared Difference (LSD)

Keine Korrektur α Fehler

- Duncan

- Ryan-Einot-Gabriel-Welsh

- **Tukey HSD**

Zu konservativ!

- Scheffe

- Bonferroni

Power

ANOVA mit Messwiederholung

- Stichproben sind abhängig:

- z.B. dieselbe Person vor / nach Behandlung
- z.B. dieselbe Stadt in verschiedenen Jahren

- Da die Proben nicht unabhängig sind, können sie nicht als zufällige Auswahl aus der Grundgesamtheit gelten!

- Vorteil: Individualeffekte können aus der Berechnung der Gruppenvarianzen ausgeschlossen werden!

- Die Rest-Varianz nimmt ab – aber auch DF!

ANOVA mit Messwiederholung

- Beispiel: **Konzentrationstest (Werte von 0 bis 100)**
 - dieselben Personen morgens, mittags und abends getestet
- Untersucht wird der Effekt der Tageszeit

```
> summary(aov(alert~daytime, data=repmes))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
daytime	2	42.5	21.23	0.406	0.67
Residuals	27	1412.5	52.31		

- Als repeated measure ANOVA:

```
> summary(aov(alert~daytime+Error(person/daytime), data=repmes))
```

Error: person

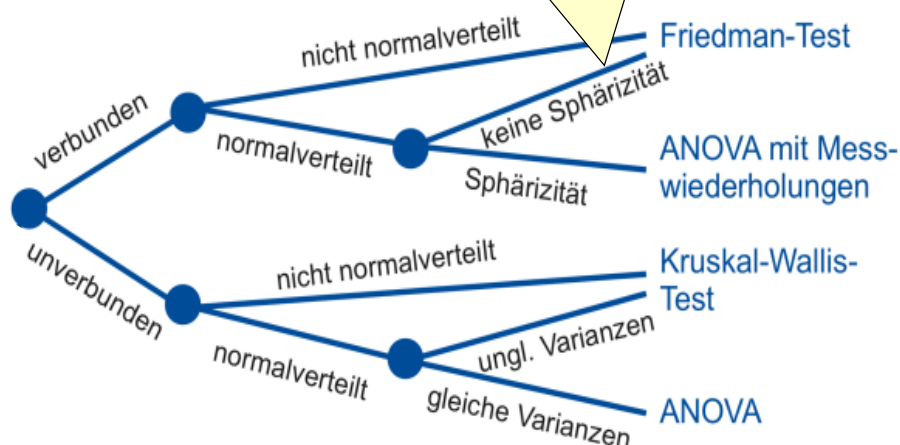
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	9	1388	154.3		

Error: person:daytime

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
daytime	2	42.47	21.233	15.79	0.000109 ***
Residuals	18	24.20	1.344		

Strategie-Schema:

Varianzhomogenität: bei unabhängigen Merkmalen
Sphärizität: bei verbundenen Merkmalen



Friedman Test für verbundene Datensätze

- Aufmerksamkeit als repeated measure ANOVA:

```
> summary(aov(alert~daytime+Error(person/daytime), data=repmes))
```

```
Error: person
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	9	1388	154.3		

```
Error: person:daytime
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
daytime	2	42.47	21.233	15.79	0.000109 ***
Residuals	18	24.20	1.344		

Achtung! Der Friedman-Test kann nicht mit Wiederholungen von "daytime" innerhalb von "person" umgehen: wenn mehrmals gemessen wurde, müssen Mittelwerte verwendet werden!

- Als Rangsummen-Test nach Friedman:

```
> friedman.test(alert~ daytime | person, data = repmes)
```

```
Friedman rank sum test
```

```
data: alert and daytime and person
```

```
Friedman chi-squared = 12, df = 2, p-value = 0.002479
```