

# Statistik

## 05. ANOVA

Roland Schäfer

Institut für Germanistische Sprachwissenschaft  
Friedrich-Schiller-Universität Jena

stets aktuelle Fassungen: <https://github.com/rsling/VL-Statistik>

1

## ANOVA

- Überblick
- Graphische Einführung

- Einfaktorielle ANOVA
- Zweifaktorielle ANOVA

2

Nächste Woche | Überblick

ANOVA

- Vergleiche von Mittelwerten zwischen mehr als zwei Gruppen
- Mittelwertvergleiche mit mehreren Unabhängigen
- Warum kann man über Varianzen Mittelwerte vergleichen?

- Gravetter & Wallnau (2007)
- Bortz & Schuster (2010)
- indirekt: Maxwell & Delaney (2004)

- Einschränkung beim t-test: immer nur 2 Gruppen
- t-Test bei mehr als 2 Gruppen: komplizierte paarweise Vergleiche
- stattdessen ANOVA: ANalysis Of VAriance
- Vergleich von Varianzen zwischen beliebigen Gruppen
- Schluss auf Mittelwerte nur indirekt über die Varianzen
- bei zwei Gruppen: Konvergenz von t-Test und ANOVA

- ANOVA vergleicht immer **mehrere Gruppen**
- Gruppen bei der einfaktoriellen ANOVA = den Ausprägungen **einer unabhängigen Variable** (z. B. Text-Register)
- diese Variablen heißen hier **Faktoren**.
- Einfluss der Faktoren auf **eine abhängige** (z. B. Satzlänge, Lesezeit)
- bei mehreren Faktoren (z. B. Text-Register und Jahrhundert): **mehrfaktorielle ANOVA**.

# Idee bei ANOVA (z. B. drei Gruppen)

- NULL:  $\bar{x}_1 = \bar{x}_2 = \bar{x}_3$
- aber: Es gibt keinen “Differenzwert” für drei Mittel (also sowas wie den t-Wert).
- daher Varianzvergleich
- F-Wert (Verteilung unter NULL bekannt) als Test-Statistik

$$F = \frac{\text{Varianz zwischen Stichprobenmitteln}}{\text{Varianz in den Stichproben}} = \frac{\text{Varianz zwischen Stichprobenmitteln}}{\text{Varianz per Zufall}}$$

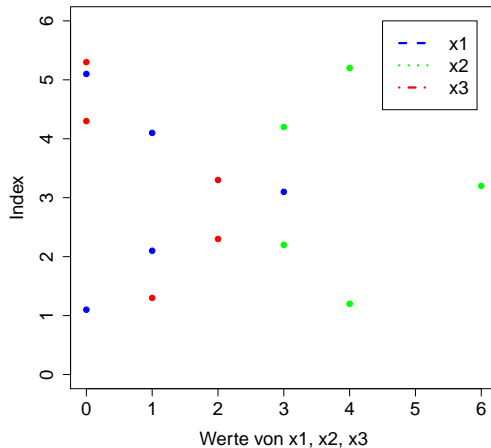


# Drei Stichproben

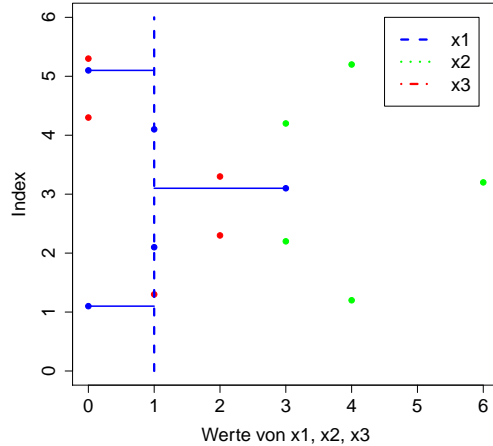
$$x_1 = [0, 1, 3, 1, 0]$$

$$x_2 = [4, 3, 6, 3, 4]$$

$$x_3 = [1, 2, 2, 0, 0]$$

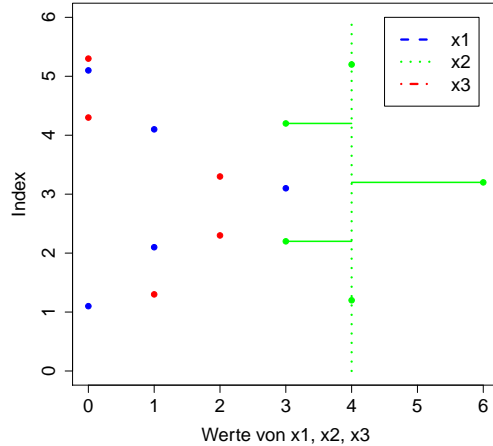


# Komponenten der Varianz von $x_1$



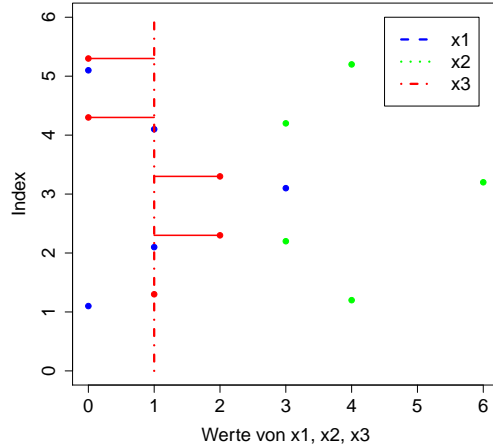
$$s^2(x_1) = 1.5$$

# Komponenten der Varianz von $x_2$



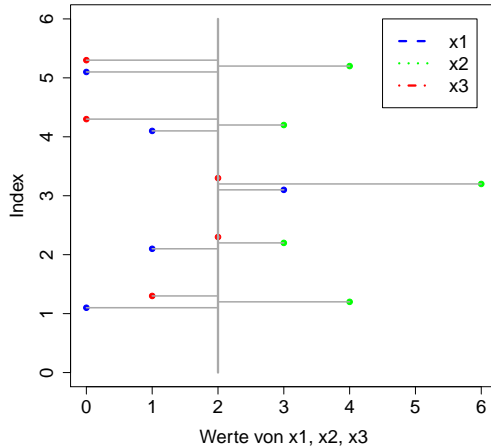
$$s^2(x_2) = 1.5$$

# Komponenten der Varianz von $x_3$



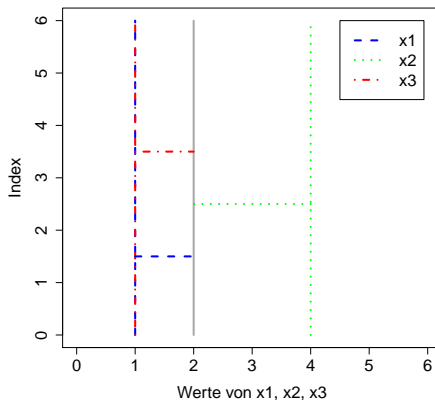
$$s^2(x_3) = 1$$

# Varianz in der zusammengefassten Stichprobe X



$$s^2(X) = 3.29$$

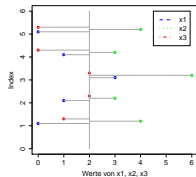
# Varianz zwischen den drei Gruppen



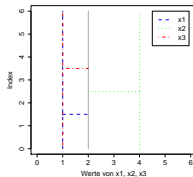
$$s^2([\bar{x}_1, \bar{x}_2, \bar{x}_3]) = 1.33$$

Achtung: Bei unterschiedlichen Stichprobengrößen  
ist das nicht ganz so einfach!

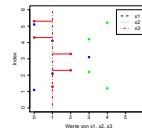
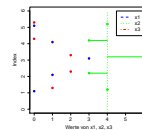
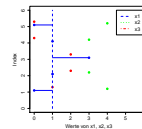
# Es gilt bezüglich der Varianzen



=



+

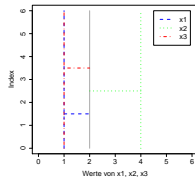


$$s^2(X) = s^2(\bar{x}_1, \bar{x}_2, \bar{x}_3) + s^2(x_1) + s^2(x_2) + s^2(x_3)$$

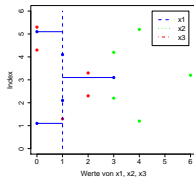
Wenn man den Abstand zwischen den Mitteln verschiebt,  
**muss** die Gesamtvarianz größer werden!

# Graphische Verdeutlichung des F-Werts

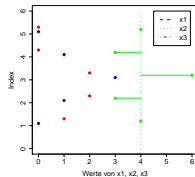
$$F = \frac{\text{Varianz zwischen Stichprobenmitteln}}{\text{Varianz in den Stichproben}} = \frac{\text{Varianz zwischen Stichprobenmitteln}}{\text{Varianz per Zufall}}$$



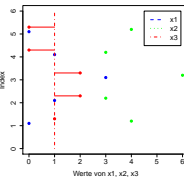
$F =$



+



+



Wenn man den Abstand zwischen den Mitteln verschiebt,  
**muss** die Gesamtvarianz größer werden!



# Wie funktioniert der F-Wert

- $F = \frac{\text{Varianz zwischen Stichprobenmitteln}}{\text{Varianz in den Stichproben}}$
- Warum?
- $F = \frac{\text{Unterschied durch Effekt} + \text{Unterschiede durch restliche Varianz}}{\text{Unterschied durch restliche Varianz}}$
- Unter Annahme der NULL gibt es keinen Effekt, ...
- also *Unterschied durch Effekt = 0*
- dann:  $F = \frac{0 + \text{Unterschiede durch restliche Varianz}}{\text{Unterschied durch restliche Varianz}} = 1$

- Anzahl der Gruppen  $x_i$ :  $k$
- Größe der Gruppen:  $n_i$
- Größe der Gesamtstichprobe  $X$ :  $N$
- Summen der Gruppen:  $T_i$
- Gesamtsumme:  $G$
- Mittel (anders als G&W):  $\bar{x}_i, \bar{X}$
- Summe der Quadrate (=Zähler der Varianz):  $SQ(x_i), SQ(X)$

Zur Erinnerung:  $s^2(x) = \frac{\sum(x-\bar{x})}{n-1} = \frac{SQ(x)}{df(x)}$

# Varianz ist Varianz beim F-Wert

$$F = \frac{\text{Varianz zwischen den Gruppen}}{\text{Varianz in den Gruppen}} = \frac{s_{\text{zwischen}}^2}{s_{\text{in}}^2} = \frac{\frac{SQ_{\text{zwischen}}}{df_{\text{zwischen}}}}{\frac{SQ_{\text{in}}}{df_{\text{in}}}}$$

denn

$$s^2(x) = \frac{SQ(x)}{df(x)}$$

Am einfachsten unter Beachtung von:

$$SQ_{gesamt} = SQ_{zwischen} + SQ_{in}$$

$$\text{Es gilt: } SQ_{gesamt} = SQ(X) = \sum (X - \bar{X})^2$$

$$\text{Außerdem: } SQ_{in} = \sum SQ(x_j)$$

$$\text{Damit: } SQ_{zwischen} = SQ_{gesamt} - SQ_{in}$$

$SQ_{zwischen}$  kann man auch direkt ausrechnen:

$$SQ_{zwischen} = \sum_i \left( \frac{T_i^2}{n_i} \right) - \frac{G^2}{N}$$

$$x_1 = [0, 1, 3, 1, 0]$$

$$x_2 = [4, 3, 6, 3, 4]$$

$$x_3 = [1, 2, 2, 0, 0]$$

Bitte alle  $SQ$  ausrechnen, inkl.  $SQ_{zwischen}$  direkt.

Tipp: Sie brauchen als Vorwissen **nur** den Stoff der ersten Statistik-Sitzung:

- arithmetisches Mittel
- $SQ$

Es gilt auch hier, ähnlich wie bei den SQ:

$$df_{gesamt} = df_{zwischen} + df_{in}$$

$$df_{gesamt} = N - 1$$

$$df_{zwischen} = k - 1$$

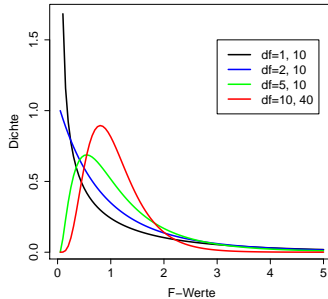
$$df_{in} = \sum_{i=1}^k (n_i - 1) = (N - 1) - (k - 1)$$

$$F = \frac{s_{zwischen}^2}{s_{in}^2} = \frac{\frac{SQ_{zwischen}}{df_{zwischen}}}{\frac{SQ_{in}}{df_{in}}}$$

Bitte ausrechnen für o. g. Beispiel.



## F-Verteilung:



In R für  $df_{\text{zwischen}} = 2$  und  $df_{\text{in}} = 12$  bei  $\text{SIG}=0.05$ :  
> `qf(0.95, 2, 12)`  $\Rightarrow 3.885294$

$$\eta^2 = \frac{SQ_{zwischen}}{SQ_{gesamt}}$$

(wieder ein  $r^2$ -Maß)

- Problem: Welche Gruppen unterscheiden sich denn nun?
- Lösung: Post(-Hoc)-Tests, z. B. Scheffé-Test:
  - ▶ paarweise ANOVA
  - ▶ aber:  $k$  wird gesetzt wie bei ursprünglicher ANOVA
  - ▶ dadurch Vermeidung kumulierten Alpha-Fehlers (Vorteil ggü. paarweisen t-Tests)
  - ▶ weiterer Vorteil: paarweise Post-Tests nur erforderlich, wenn Omnibus-ANOVA bereits Signifikanz gezeigt hat
  - ▶ und: Generalisierbarkeit zu mehrfaktorieller ANOVA (geht mit t-Test nicht)

Bitte ausrechnen für die oben gerechnete ANOVA.

# Wozu mehrfaktorielle Designs

Oft vermutet man den Einfluss **mehrerer Unabhängiger** auf eine Abhängige.  
Beispiel: Satzlängen

		Textsorte		
		Fiktion	Zeitung	Wissenschaft
Jahrhundert	19	$X_{11}$	$X_{12}$	$X_{13}$
	20	$X_{21}$	$X_{22}$	$X_{23}$

Hier also:  $2 \cdot 3 = 6$  Gruppen

# Ablauf der zweifaktoriellen ANOVA

- 1 erste ANOVA zwischen Zeilen
- 2 zweite ANOVA zwischen Spalten
- 3 dritte ANOVA für **Interaktionen** zwischen Zeilen und Spalten
- 4 Interaktion: Ungleichverteilung in Gruppen, die nicht durch die Spalten- und Zeileneffekte erklärt werden kann
- 5 Alle drei ANOVAs sind **unabhängig** voneinander!

- **Gesamtvarianz** = Varianz zwischen Gruppen + Varianz in den Gruppen
- **Varianz zwischen den Gruppen** =  
Haupt-Faktoren-Varianz + **Interaktions-Varianz**
- **Haupt-Faktoren-Varianz** =  
Varianz zwischen Faktor A-Gruppen +  
Varianz zwischen Faktor B-Gruppen

# Schritt 1(1): SQ/df zwischen den Gruppen

Jede Zelle der Tabelle ist eine Gruppe.

$$SQ_{\text{zwischen}} = \sum_i \left( \frac{T_i^2}{n_i} \right) - \frac{G^2}{N}$$
$$df_{\text{zwischen}} = k - 1 \quad (k = \text{Anzahl der Zellen/Gruppen})$$

Beachte: Keine Änderung verglichen mit einfaktorieller ANOVA!

## Schritt 1(2): SQ/df in den Gruppen

Jede Zelle der Tabelle ist eine Gruppe.

$$\begin{aligned}SQ_{in} &= \sum SQ(x_i) \\df_{in} &= \sum df(x_i)\end{aligned}$$

Beachte: Keine Änderung verglichen mit einfaktorieller ANOVA!



## Schritt 2(2): SQ/df für Gruppe A

Berechnung nach dem Schema für Zwischen-Gruppen-Varianz

		Textsorte			
		Fiktion	Zeitung	Wissenschaft	
Jahrhundert	19	$X_{11}$	$X_{12}$	$X_{13}$	$A_1$
	20	$X_{21}$	$X_{22}$	$X_{23}$	$A_2$

Auch hier keine wesentliche Änderung:

$$SQ_A = \sum_i \left( \frac{T_{A_i}^2}{n_{A_i}} \right) - \frac{G^2}{N}$$

$$df_A = k_A - 1 \quad (k_A = \text{Anzahl der Zeilen})$$

## Schritt 2(2): SQ/df für Gruppe A

Berechnung nach dem Schema für Zwischen-Gruppen-Varianz

		Textsorte		
		Fiktion	Zeitung	Wissenschaft
Jahrhundert	19	$X_{11}$	$X_{21}$	$X_{31}$
	20	$X_{12}$	$X_{22}$	$X_{32}$
		$B_1$	$B_2$	$B_3$

Auch hier keine Änderung:

$$SQ_B = \sum_i \left( \frac{T_{B_i}^2}{n_{B_i}} \right) - \frac{G^2}{N}$$

$$df_B = k_B - 1 \quad (k_B = \text{hier Anzahl der Spalten})$$

## Schritt 2(3): SQ/df für Interaktion $A \times B$

Die Varianz, die auf Kosten der Interaktion geht, ist  
die Zwischen-Gruppen-Varianz ohne die Einzelfaktor-Varianz.

$$\begin{aligned} SQ_{A \times B} &= SQ_{\text{zwischen}} - SQ_A - SQ_B \\ df_{A \times B} &= df_{\text{zwischen}} - df_A - df_B \end{aligned}$$

# Alle drei F-Werte ausrechnen

Die zweifaktorielle ANOVA  
erfordert wie gesagt drei Einzel-ANOVAs.

$$F_A = \frac{\frac{SQ_A}{df_A}}{\frac{SQ_{zwischen}}{df_{zwischen}}} = \frac{s_A^2}{s_{zwischen}^2}$$

$$F_B = \frac{\frac{SQ_B}{df_B}}{\frac{SQ_{zwischen}}{df_{zwischen}}} = \frac{s_B^2}{s_{zwischen}^2}$$

$$F_{A \times B} = \frac{\frac{SQ_{A \times B}}{df_{A \times B}}}{\frac{SQ_{zwischen}}{df_{zwischen}}} = \frac{s_{A \times B}^2}{s_{zwischen}^2}$$

Entsprechend sind **drei**  $\eta^2$  auszurechnen:

$$\eta_A^2 = \frac{SQ_A}{SQ_{gesamt} - SQ_B - SQ_{A \times B}}$$

$$\eta_B^2 = \frac{SQ_B}{SQ_{gesamt} - SQ_A - SQ_{A \times B}}$$

$$\eta_{A \times B}^2 = \frac{SQ_{A \times B}}{SQ_{gesamt} - SQ_A - SQ_B}$$

Wir fragen jeweils, welchen Anteil an der Varianz, die die anderen beiden Faktoren **nicht** erklären, der jeweilige dritte Faktor hat.

# Das jetzt alles zusammen

Bitte vollständige zweifaktorielle ANOVA  
bei  $\text{SIG}=0.05$  und  $\text{SIG}=0.01$  rechnen:

	<b>B1</b>	<b>B2</b>	<b>B3</b>
<b>A1</b>	1, 3, 1, 4	4, 3, 3, 6	8, 6, 8, 10
<b>A2</b>	8, 6, 6, 8	1, 6, 8, 1	1, 4, 1, 4

Nächste Woche | Überblick

- 1 Inferenz
- 2 Deskriptive Statistik
- 3 Nichtparametrische Verfahren
- 4 z-Test und t-Test
- 5 ANOVA
- 6 Freiheitsgrade und Effektstärken
- 7 Power und Severity
- 8 Lineare Modelle
- 9 Generalisierte Lineare Modelle
- 10 Gemischte Modelle



- Bortz, Jürgen & Christof Schuster. 2010. *Statistik für Human- und Sozialwissenschaftler*. 7. Aufl. Berlin: Springer.
- Gravetter, Frederick J. & Larry B. Wallnau. 2007. *Statistics for the Behavioral Sciences*. 7. Aufl. Belmont: Thomson.
- Maxwell, Scott E. & Harold D. Delaney. 2004. *Designing experiments and analyzing data: a model comparison perspective*. Mahwa, New Jersey, London: Taylor & Francis.

## Kontakt

Prof. Dr. Roland Schäfer  
Institut für Germanistische Sprachwissenschaft  
Friedrich-Schiller-Universität Jena  
Fürstengraben 30  
07743 Jena

<https://rolandschaefer.net>  
[roland.schaefer@uni-jena.de](mailto:roland.schaefer@uni-jena.de)

## Creative Commons BY-SA-3.0-DE

Dieses Werk ist unter einer Creative Commons Lizenz vom Typ *Namensnennung - Weitergabe unter gleichen Bedingungen 3.0 Deutschland* zugänglich. Um eine Kopie dieser Lizenz einzusehen, konsultieren Sie

<http://creativecommons.org/licenses/by-sa/3.0/de/> oder wenden Sie sich brieflich an Creative Commons, Postfach 1866, Mountain View, California, 94042, USA.