



TECHNISCHE UNIVERSITÄT  
IN DER KULTURHAUPTSTADT EUROPAS  
CHEMNITZ

Professur Psychologie digitaler Lernmedien

Institut für Medienforschung

Philosophische Fakultät



Statistik I

# Einfaktorielle Varianzanalyse

Squid Game (2021). Netflix.

# Überblick

- Einführung
- Alphafehler-Kumulierung
- Grundprinzip der Varianzanalyse
- Empirischer  $F$ -Wert
- Quadratsummen innerhalb und zwischen den Zellen
- Zähler- und Nennerfreiheitsgrade
- Inferenzstatistische Entscheidung und Ergebnisdarstellung
- Post-hoc-Analysen
- Inferenzstatistische Voraussetzungen

# Einführung

(z. B. Rasch, Frieese, Hofmann & Naumann, 2021)

- **t-Test:** Inferenzstatistischer Vergleich zwischen zwei Mittelwerten
- **Varianzanalyse (engl. analysis of variance, ANOVA):** Statistisches Verfahren zum simultanen Vergleich mehrerer Mittelwerte
- **Einfaktorielle Varianzanalyse:** Varianzanalyse zu einem einfaktoriellen Versuchsdesign
- **Warum nicht mehrere Mittelwertvergleiche mittels mehrerer t-Tests?**
  - Verringerte Teststärke
  - Alphafehler-Kumulierung

# Alphafehler-Kumulierung

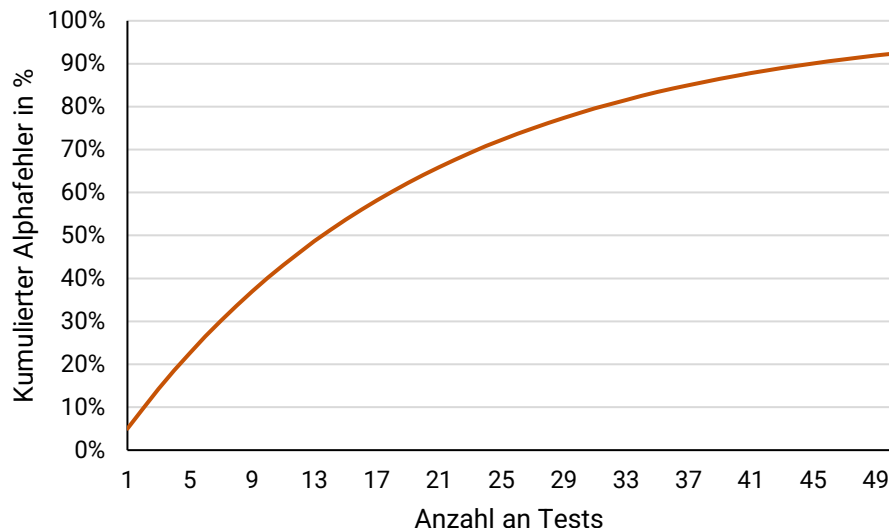
- **Alphafehler-Kumulierung:** Tritt prinzipiell bei mehreren inferenzstatistischen Tests (bzw. mehreren Hypothesen) auf
- **Quasi „Mehrfachwürfeln“:** Chance steigt an, mindestens einmal das Signifikanzniveau zu überwinden
- **Anders formuliert:** Wahrscheinlichkeit, dass ich mindestens ein Muster ( $H_1$ ) sehe, obwohl kein (oder ein anderes) Muster ( $H_0$ ) vorliegt, steigt an, je mehr inferenzstatistische Tests ich durchführe
- **Berechnung des kumulierten Alphafehlers:**

$$\pi = 1 - (1 - \alpha)^m$$

$\alpha$  = Signifikanzniveau, z. B. 5%  
 $m$  = Anzahl an Tests

# Alphafehler-Kumulierung

- **Beispiel:** Berechnung des kumulierten Alphafehlers für drei  $t$ -Tests, die mit einem Signifikanzniveau von 5% getestet werden:
- $\pi = 1 - (1 - 0.05)^3 \approx 0.143$
- Der kumulierte Alphafehler liegt somit bei 14.3% und nicht bei 5%
- Allgemein gilt für ein Signifikanzniveau von 5%:



# Alphafehler-Kumulierung

Wie hoch ist der kumulierte Alphafehler für fünf  $t$ -Tests, die mit einem Signifikanzniveau von 1% getestet werden?

- A: 4.5 %
- B: 5.0 %
- C: 1.0 %
- D: 4.9 %
- E: 5.1 %

# Grundprinzip der Varianzanalyse (z. B. Rasch, Frieese, Hofmann & Naumann, 2021)

- **Grundprinzip:** Zerlegung der Gesamtvarianz aller Messwerte in systematische Varianz und Residualvarianz (Fehlervarianz)
- Bzw. Zerlegung der totalen Quadratsumme ( $QS_{\text{Total}}$ ) in die Quadratsumme zwischen den Zellen ( $QS_{\text{Zwischen}}$ ) und die Quadratsumme innerhalb der Zellen ( $QS_{\text{Innerhalb}}$ )

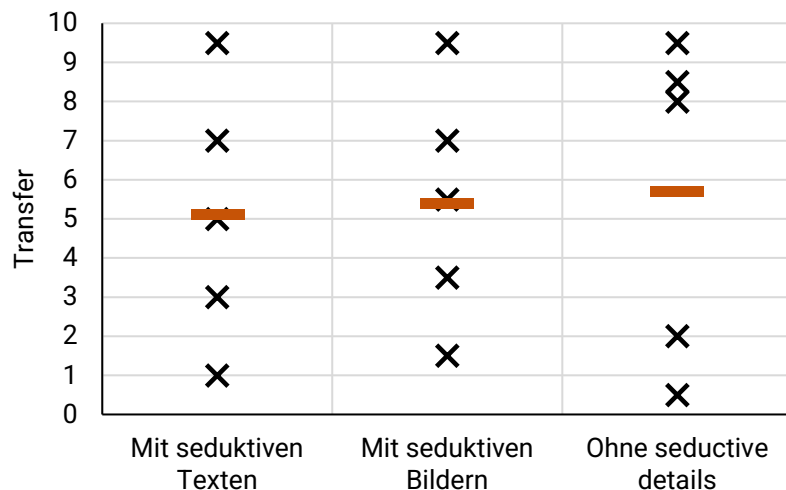


# Systematische Varianz vs. Residualvarianz (z. B. Rasch, Frieese, Hofmann & Naumann, 2021)

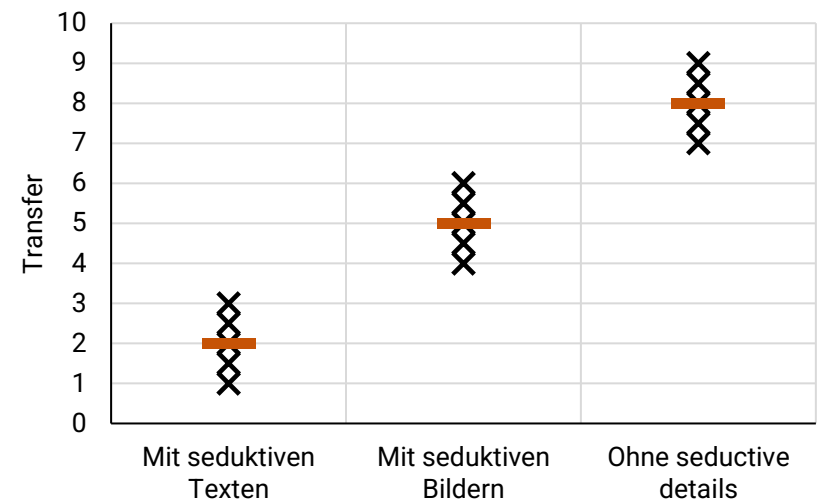
- **Systematische Varianz:** Varianz, die auf den Einfluss der experimentellen Faktoren zurückzuführen ist
  - Durch die Varianz zwischen den Bedingungen geschätzt (vgl.  $QS_{\text{Zwischen}}$ )
- **Residualvarianz:** Varianz, deren Ursachen nicht durch das Experiment erfassbar sind
  - Durch die durchschnittliche Varianz innerhalb der Bedingungen geschätzt (vgl.  $QS_{\text{Innerhalb}}$ )

# Beispiele für das Verhältnis zwischen systematischer Varianz und Residualvarianz

- **Beispiel:** Fiktive Ergebnisse zweier Studien zu seductive details
- Jeweils mit einem einfaktoriellen, dreifachgestuften Versuchsdesign (d. h. drei Versuchsbedingungen)
- Rohdaten als schwarze Kreuze, Mittelwerte als orangene Linien



Systematische Varianz gering  
Residualvarianz hoch



Systematische Varianz hoch  
Residualvarianz gering

# Beispiel zur Berechnung einer Varianzanalyse

- **Beispiel:** Fiktive Rohdaten zu einer Studie zum seductive details (vgl. rechte Abbildung auf der vorherigen Seite)

Mit seduktiven Texten

VPN	Transfer
1	2.0
2	1.5
3	1.0
4	2.5
5	3.0
<i>M</i>	2.0

Mit seduktiven Bildern

VPN	Transfer
6	5.0
7	5.5
8	4.5
9	6.0
10	4.0
<i>M</i>	5.0

Ohne seductive details

VPN	Transfer
11	8.5
12	8.0
13	7.0
14	9.0
15	7.5
<i>M</i>	8.0

# Berechnung des empirischen $F$ -Wertes

- Inferenzstatistische Überprüfung der Varianzverhältnisse mit Hilfe des  $F$ -Wertes
- **Formel zur Berechnung des empirischen  $F$ -Wertes:**

$$F = \frac{\frac{QS_{Zwischen}}{df_Z}}{\frac{QS_{Innerhalb}}{df_N}}$$

$F$	= Empirischer $F$ -Wert
$QS_{Zwischen}$	= Quadratsumme zwischen den Zellen
$QS_{Innerhalb}$	= Quadratsumme innerhalb der Zellen
$df_Z$	= Zählerfreiheitsgrade
$df_N$	= Nennerfreiheitsgrade

- $F$ -Werte und  $t$ -Werte lassen sich ineinander überführen
- Es gilt:  $F = t^2$

# Quadratsumme innerhalb der Zellen

- Formel zur Berechnung der Quadratsumme innerhalb der Zellen:

$$QS_{Innerhalb} = \sum_{i=1}^p \sum_{m=1}^n (x_{mi} - \bar{A}_i)^2$$

$x_{mi}$  = Wert der Person m in der Gruppe i  
 $\bar{A}_i$  = Mittelwert der Gruppe i

- Berechnung für das Beispiel:  $QS_{Innerhalb} = (2.0 - 2.0)^2 + (1.5 - 2.0)^2 + (1.0 - 2.0)^2 + (2.5 - 2.0)^2 + (3.0 - 2.0)^2 + \dots + (7.5 - 8.0)^2 = (2.5 + 2.5 + 2.5) = 7.5$

Mit seduktiven Texten	
1	2.0
2	1.5
3	1.0
4	2.5
5	3.0
M	2.0

Mit seduktiven Bildern	
6	5.0
7	5.5
8	4.5
9	6.0
10	4.0
M	5.0

Ohne seductive details	
11	8.5
12	8.0
13	7.0
14	9.0
15	7.5
M	8.0

# Quadratsumme zwischen den Zellen

- Formel zur Berechnung der Quadratsumme zwischen den Zellen:

$$QS_{Zwischen} = n \cdot \sum_{i=1}^p (\bar{A}_i - \bar{G})^2$$

$n$  = Anzahl an Versuchspersonen in einer Gruppe  
 $\bar{A}_i$  = Mittelwert der Gruppe  $i$   
 $\bar{G}$  = Gesamtmittelwert

- Für das Beispiel gilt:

- Anzahl an Versuchspersonen in einer Gruppe: 5
- Gruppenmittelwerte: 2.0 (mit seduktiven Texten), 5.0 (mit seduktiven Bildern) und 8.0 (ohne seductive details)
- Gesamtmittelwert: 5.0

- Berechnung:

$$QS_{Zwischen} = 5 \cdot [(2.0 - 5.0)^2 + (5.0 - 5.0)^2 + (8.0 - 5.0)^2] = 5 \cdot [9 + 0 + 9] = 90$$

# Freiheitsgrade

(Rasch, Frieese, Hofmann & Naumann, 2021)

- **Freiheitsgrade:** Legen die Genauigkeit von Populationsschätzern und damit die Form von Verteilungen fest, die auf Schätzern basieren wie z. B. die  $F$ -Verteilung
- **Zahl der Freiheitsgrade:** Gibt an, wie viele Werte theoretisch frei variieren können, wenn das Ergebnis bereits feststeht
- **Unterscheidung Zählerfreiheitsgrade und Nennerfreiheitsgrade**

# Zählerfreiheitsgrade

- **Definition:** Anzahl der bei der Berechnung eines Kennwerts frei variierbaren Werte im Zähler (Numerator)
- Hypothesenfreiheitsgrade  $df_h$  bzw.  $df_z$  bzw.  $df_{\text{treat}}$  bzw.  $df_{\text{Zwischen}}$
- Formel für eine einfaktorielle, univariate Varianzanalyse:
- $df_z = p - 1$   

$p = \text{Anzahl an Versuchsgruppen}$
- **Beispiel:** Bei drei Versuchsgruppen ist  $df_z = 3 - 1 = 2$

# Nennerfreiheitsgrade

- **Definition:** Anzahl der bei der Berechnung eines Kennwerts frei variierbaren Werte im Nenner (Denominator)
- Fehlerfreiheitsgrade  $df_e$  bzw.  $df_N$  bzw.  $df_{\text{error}}$  bzw.  $df_{\text{Fehler}}$  bzw.  $df_{\text{Innerhalb}}$
- $df_N = N - p$  bzw.  $p \cdot (n - 1)$

$N$  = Gesamter Stichprobenumfang

$p$  = Anzahl an Versuchsgruppen

$n$  = Anzahl an Versuchspersonen in einer Gruppe

- **Beispiel:** Bei drei Gruppen mit jeweils fünf Personen gilt:
- $df_N = 15 - 3 = 12$  bzw.  $3 \cdot (5 - 1) = 12$

# Zähler- und Nennerfreiheitsgrade

In einem Experiment mit einem einfaktoriellen, vierfachgestuften, univariaten Versuchsdesign werden pro Versuchsbedingung 20 Probanden getestet. Wie hoch sind die Zähler- und Nennerfreiheitsgrade?

- A:  $df_z = 1$  und  $df_N = 80$
- B:  $df_z = 1$  und  $df_N = 76$
- C:  $df_z = 3$  und  $df_N = 80$
- D:  $df_z = 3$  und  $df_N = 76$
- E:  $df_z = 2$  und  $df_N = 76$

# Beispiel zur Ermittlung des empirischen und kritischen $F$ -Wertes

- Berechnung des empirischen  $F$ -Wertes:

- $$F = \frac{\frac{QS_{Zwischen}}{df_Z}}{\frac{QS_{Innerhalb}}{df_N}}$$
 Einsetzen ergibt: 
$$F = \frac{\frac{90}{2}}{\frac{7.5}{12}} = 72$$

- Ermittlung des kritischen  $F$ -Wertes für  $df_Z = 2$  und  $df_N = 12$  und  $\alpha = .05$ :  $F_{\text{krit}} \approx 3.89$

# Inferenzstatistische Entscheidung und Ergebnisdarstellung

- Zwei Möglichkeiten bei der inferenzstatistischen Entscheidung
  - Nicht signifikant ( $F_{\text{emp}} < F_{\text{krit}}$ ):  $H_0$  wird vorläufig beibehalten
  - Signifikant ( $F_{\text{emp}} \geq F_{\text{krit}}$ ):  $H_0$  wird zugunsten der  $H_1$  verworfen
- **Beispiel:** Da  $F_{\text{emp}} = 72 \geq F_{\text{krit}} \approx 3.89$  wird  $H_0$  zugunsten der  $H_1$  verworfen, d. h. das Ergebnis ist signifikant
- **Inferenzstatistische Ergebnisdarstellung**
  - **Signifikantes Ergebnis:** Angabe des empirischen  $F$ -Wertes (inklusive  $df_Z$  und  $df_N$ ) und  $p$ -Wertes sowie der Effektstärke
  - **Nicht signifikantes Ergebnis:** Zusätzlich Angabe der Teststärke
- **Beispiel:**  $F(2,12) = 72, p < .001, \eta_p^2 = .92$

# Post-hoc-Analysen

(z. B. Rasch, Frieese, Hofmann & Naumann, 2021)

- **Varianzanalyse testet unspezifisch:** Analyse von Mittelwertsunterschieden ohne Aussage, welche Mittelwerte sich (signifikant) voneinander unterscheiden und welche nicht
- **Mehrere *t*-Tests** zur spezifischen Testung denkbar, dagegen sprechen aber Teststärkeverlust und Alphafehler-Kumulierung
- **Post-hoc-Verfahren:** Stattdessen Einsatz sog. Post-hoc-Verfahren, um ausgewählte Mittelwerte spezifisch gegeneinander zu testen
- **Tukey HSD-Test:** Post-hoc-Verfahren für den paarweisen Vergleich der Gruppenmittelwerte
- **HSD:** Berechnung über die kleinste noch signifikante Differenz zwischen zwei Gruppenmittelwerten (engl. honest significant difference, HSD)

# Tukey HSD-Test

Die HSD beträgt für das vorherige Beispiel 1.32. Die Mittelwerte für die drei Versuchsbedingungen lagen bei 2.0, 5.0 und 8.0. Welche Aussagen sind zutreffend?

- A: Die drei Mittelwerte unterscheiden sich nicht signifikant voneinander.
- B: Nur der niedrigste Mittelwert (2.0) und der höchste Mittelwert (8.0) unterscheiden sich signifikant voneinander.
- C: Alle drei Mittelwerte unterscheiden sich signifikant voneinander.
- D: Nur der mittlere Mittelwert (5.0) unterscheidet sich signifikant von den anderen beiden Mittelwerten.
- E: 42.

# Inferenzstatistische Voraussetzungen (z. B. Rasch, Friese, Hofmann & Naumann, 2021)

- **Unabhängigkeit der Messwerte in den einzelnen Bedingungen:** Gilt nur für Versuchsdesigns ohne Messwiederholung
- **Intervallskalenniveau der abhängigen Variable:** Überprüfung aufwändig und schwierig
- **Normalverteilung der abhängigen Variable in der Population** (getrennt für jede Versuchsbedingung oder auf Basis der Residuen): Überprüfung z. B. mittels Shapiro-Wilk-Test
- **Varianzhomogenität** als Gleichheit der Populationsvarianzen, aus denen die Stichproben stammen: Überprüfung mittels Levene-Test
- **Überprüfung der vier Voraussetzungen in der Forschungspraxis** (leider) eher unüblich (vgl. *t*-Tests)

# Inferenzstatistische Voraussetzungen (z. B. Rasch, Frieese, Hofmann & Naumann, 2021)

- **Annahmeverletzungen:** Bei Verletzungen einer der genannten Voraussetzungen kann statt des  $F$ -Tests (parametrisches Verfahren) u. a. ein nonparametrisches Verfahren eingesetzt werden (vgl.  $t$ -Test)
- **Robustheit:** Allerdings reagiert der  $F$ -Test unter folgenden Bedingungen relativ robust gegenüber Verletzungen der Voraussetzungen
  - Ungefähr gleichgroße Stichproben der Gruppen
  - Hinreichend große Stichproben
- **In der Praxis:** Nonparametrische Verfahren selten genutzt

# Beispiele für Varianzanalysen und $F$ -Tests in Fachzeitschriften

Descriptive statistics for pretest and posttest scores and mental effort invested in learning from the animations are shown in Table 2. On the tests, missing answers were scored as errors. The significance level for the comparisons was set at 0.05, and eta-squared is reported as a measure of effect size, with 0.01 indicating a small, 0.06 a moderate and 0.14 a large effect. The conditions did not differ in pretest scores,  $F(3,157) = 1.21$ ,  $MSE = 1.02$ ,  $p = 0.31$ ,  $\eta_p^2 = 0.02$ .

A 2-by-2 ANOVA revealed a main effect of Pausing on posttest scores: participants studying animations with pauses ( $M = 6.46$ ,  $sd = 1.79$ ) performed significantly better on the posttest than participants studying animations without pauses ( $M = 5.80$ ,  $sd = 2.16$ ),  $F(1,157) = 4.39$ ,  $p = 0.04$ ,  $MSE = 3.92$ ,  $\eta_p^2 = 0.03$ . There was no significant main effect of Cueing,  $F(1,157) = 0.49$ ,  $p = 0.48$ ,  $\eta_p^2 = 0.003$  nor an interaction effect between Cueing and Pausing,  $F(1,157) = 2.85$ ,  $p = 0.09$ ,  $\eta_p^2 = 0.02$ .

A 2-by-2 ANOVA on mental effort invested during animation study showed a significant main effect of Cueing: learning from animations with cues ( $M = 2.19$ ,  $sd = 1.05$ ) required significantly less investment of mental effort than learning from animations without cues ( $M = 2.73$ ,  $sd = 1.70$ ),  $F(1,157) = 5.67$ ,  $p = 0.02$ ,  $MSE = 2.01$ ,  $\eta_p^2 = 0.03$ . There was no significant main effect of Pausing  $F(1,157) = 2.33$ ,  $p = 0.13$ ,  $\eta_p^2 = 0.01$ , nor a significant interaction effect  $F(1,157) = 0.33$ ,  $p = 0.56$ ,  $\eta_p^2 = 0.002$ .

Quelle: Spanjers, van Gog, Wouters und van Merriënboer (2012)

**3.2.1.1. Sensitivity.** We calculated mean hit rates (“yes” answers to identical clockworks) and mean false alarm rates (“yes” answers to differing clockworks) for each participant in each condition of the two-factorial design. Next, we calculated the sensitivity measure  $d'$ . Data as displayed in Table 3 were submitted to a 2 (“presentation mode”; between-subjects)  $\times$  2 (“target clock functionality”; within-subjects) mixed-factor ANOVA. As hypothesized, participants’ performance was significantly higher with the Vex presentation mode,  $F(1,75) = 38.95$ ,  $p < .001$ ,  $\eta_C^2 = .29$ , suggesting that the Vex mode facilitated visual search. However, the significant interaction of the factors presentation mode and target clock functionality showed that there are substantial differences within each

Quelle: Huff, Bauhoff und Schwan (2012)

A repeated measures ANOVA with study condition (simultaneous vs. sequential) manipulated within-subjects on hits-fa in proportional scores showed a significant effect of study condition  $F(1,24) = 6.27$ ,  $p < .02$ ,  $MSE = .029$ , as recognition accuracy was .51 under simultaneous, while it was .39 under sequential study condition. As for Experiment 1, correlational analysis between accuracy and digit and fluency scores did not reveal any significant correlations indicating that performance was not affected by individual differences in working memory processes.

Quelle: Mammarella, Fairfield und Di Domenico (2013)

# Zusammenfassung

- **Varianzanalyse:** Statistisches Verfahren zum simultanen Vergleich mehrerer Mittelwerte
- **Grundprinzip:** Zerlegung der Gesamtvarianz in systematische Varianz und Residualvarianz
- **Inferenzstatistische Überprüfung** der Varianzverhältnisse mit dem kritischen und empirischen  $F$ -Wert
- **Post-hoc-Verfahren** wie der Tukey HSD-Test zur spezifischen Testung ausgewählter Mittelwerte
- **Voraussetzung von Varianzanalysen:** Intervallskalenniveau, Normalverteilung, Varianzhomogenität und ggf. Unabhängigkeit der Messwerte

$$F = \frac{\frac{QS_{Zwischen}}{df_Z}}{\frac{QS_{Innerhalb}}{df_N}}$$

# Prüfungsliteratur

- Rasch, B., Friese, M., Hofmann, W., & Naumann, E. (2021). *Quantitative Methoden 2: Einführung in die Statistik für Psychologie, Sozial- & Erziehungswissenschaften* (5. Aufl.). Heidelberg: Springer.
  - Einfaktorielle Varianzanalyse (S. 1–36)

# Weiterführende Literatur

- Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin: Springer.
  - Einfaktorielle Versuchspläne (S. 205–220)
  - Kontraste und Mehrfachvergleiche für einfaktorielle Versuchspläne (S. 221–235)
- Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (5. Aufl.). Weinheim: Beltz.
  - Einfaktorielle Varianzanalyse (S. 392–429)
- Leonhart, R. (2022). *Lehrbuch Statistik. Einstieg und Vertiefung* (5. Auflage). Bern: Huber.
  - Einfaktorielle Varianzanalyse mit festen Effekten (S. 383–422)
- Sedlmeier, P., & Renkewitz, F. (2018). *Forschungsmethoden und Statistik: Ein Lehrbuch für Psychologen und Sozialwissenschaftler* (3. Aufl.). München: Pearson.
  - Der *F*-Test in der einfaktoriellen Varianzanalyse (S. 429–462)