Unterschiede in zentraler Tendenz in einer oder zwei Gruppen: Gliederung

1 Ein-Stichproben-Tests

(z-Test, t-Test, weitere Tests)

2 Parametrische Tests für Mittelwertsunterschiede zweier Gruppen

(Unabhängigkeit und Abhängigkeit von Gruppen, t-Test für unabhängige und abhängige Gruppen, Welch-Test)

3 Prüfung der Voraussetzungen

(Robustheit, Varianzhomogenität, Normalverteilung)

4 Nicht-parametrische Tests für Unterschiede zweier Gruppen in der zentralen Tendenz

(Mann-Whitney U-Test, Vorzeichen-Test, Wilcoxon-Test)

Unterschiede in zentraler Tendenz in einer oder zwei Gruppen: Literatur

Einführende Literatur

📖 Bortz, J. & Schuster, Ch. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Auflage). Berlin: Springer. [Kap. 8]

📖 Diehl, J. M. & Arbinger, R. (2001). *Einführung in die Inferenzstatistik* (3. Auflage). Eschborn bei Frankfurt: Klotz Verlag. [Kap. 3-7, 15, 23, 25]

📖 Eid, M., Gollwitzer, M. & Schmitt, M. (2010). *Statistik und Forschungsmethoden*. Weinheim: Beltz. [Kap. 10-12]

Weiterführende Literatur

📖 Wilcox, R. R. (2010). *Fundamentals of modern statistical methods: Substantially improving power and accuracy* (2nd ed.). New York: Springer.

Statistische Tests

- Wir haben die Logik des Hypothesentests am Beispiel des Ein-Stichproben z-Tests kennengelernt. Er ermöglichte unter bestimmten Voraussetzungen die Prüfung der Fragestellung, ob ein Mittelwert einer Variablen in der Population einen bestimmten festen Wert aufweist.
- Wir werden nun eine Vielzahl weiterer statistischer Tests kennenlernen, die prinzipiell ebenfalls der obigen Logik folgen. Welcher Test am besten geeignet ist, hängt von einer Reihe von Faktoren ab, darunter …
  - der Art der Fragestellung/Hypothese,
  - dem gewählten Untersuchungsdesign,
  - dem Skalenniveau der beteiligten Variablen,
  - den Verteilungseigenschaften der beteiligten Variablen im Abgleich mit den diesbezüglichen Anforderungen der Testverfahren,
  - weiteren statistischen Überlegungen, etwa zur Power der Testverfahren.
- Im Folgenden werden wir nun zunächst alternative Verfahren betrachten, bei der es ebenfalls um die Hypothese geht, dass ein Mittelwert einen bestimmten Wert aufweist.

Ein-Stichproben z-Test

- Bezeichnung: Ein-Stichproben z-Test (Ein-Stichproben Gauß-Test).
- Einsatzbereich: Prüfung der Hypothese, dass ein Mittelwert einer intervallskalierten Variablen X in der Population einen bestimmten festen Wert μ_0 aufweist.
- Hypothesen: zweiseitig: H_0: $\mu = \mu_0$ und H_1: $\mu \neq \mu_0$ oder entsprechend einseitig.
- Voraussetzungen: Normalverteilte Werte von X (in der Population), Standardabweichung von X (in der Population) σ bekannt. (Einfache Zufallsstichprobe mit relativ zur Stichprobe großen Population.)
- Vorgehen: Bestimmung der Prüfgröße

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Die Zusätze in Klammern werden im Folgenden nicht mehr aufgeführt.

- Entscheidung: Zurückweisung der H_0, falls $|z| > z_{crit}$ mit $z_{crit} = z_{1-\alpha/2}$ bei zweiseitiger Prüfung und $z_{crit} = z_{1-\alpha}$ bei einseitiger Prüfung (und korrekter Richtung).

- Da in der Regel die Standardabweichung des Merkmals X in der Population σ unbekannt ist, wird dieser Test nur sehr selten angewandt. Wir wenden uns nun einem Test zu, der diese Annahme nicht benötigt.

Ein-Stichproben t-Test

- Bezeichnung: Ein-Stichproben t-Test.
- Einsatzbereich: Prüfung der Hypothese, dass ein Mittelwert einer intervallskalierten Variablen X in der Population einen bestimmten festen Wert μ_0 aufweist.
- Hypothesen: zweiseitig: H_0: $\mu = \mu_0$ und H_1: $\mu \neq \mu_0$ oder entsprechend einseitig.
- Voraussetzungen: Normalverteilte Werte von X.
- Vorgehen: Bestimmung der Prüfgröße

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \quad \text{mit} \quad df = n - 1$$

- Entscheidung: Zurückweisung der H_0, falls $|t| > t_{crit}$ mit $t_{crit} = t_{n-1;1-\alpha/2}$ bei zweiseitiger Prüfung und $t_{crit} = t_{n-1;1-\alpha}$ bei einseitiger Prüfung (und korrekter Richtung).

Ein-Stichproben t-Test

- Beispiel (ähnlich Bortz & Schuster, 2010, S. 119): Es wird untersucht, ob der Schlaf-Wach-Rhythmus mit einer Periodendauer von 24 Stunden (circadiane Rhythmik) erhalten bleibt, wenn die äußeren Zeitgeber fehlen. Dazu leben 7 Vpn von der Außenwelt (Uhrzeit, sozialen Kontakten, Tageslicht) abgeschnitten 10 Tage in einem Labor. Gemessen wird die Zeit zwischen dem Zubettgehen am vorletzten und am letzten Tag in Minuten. Die Hypothese lautet, dass sich der Wach-Schlaf Rhythmus unter diesen freilaufenden Bedingungen ändert.

- $H_0: \mu = 1440$ (Der Wach-Schlaf-Rhythmus beträgt unter freilaufenden Bedingungen 24 Stunden $= 24 \cdot 60 = 1440$ Minuten)

  $H_1: \mu \neq 1440$ (Der Wach-Schlaf-Rhythmus beträgt … nicht 1440 Minuten)

Vp	X
1	1452
2	1438
3	1524
4	1479
5	1554
6	1461
7	1476

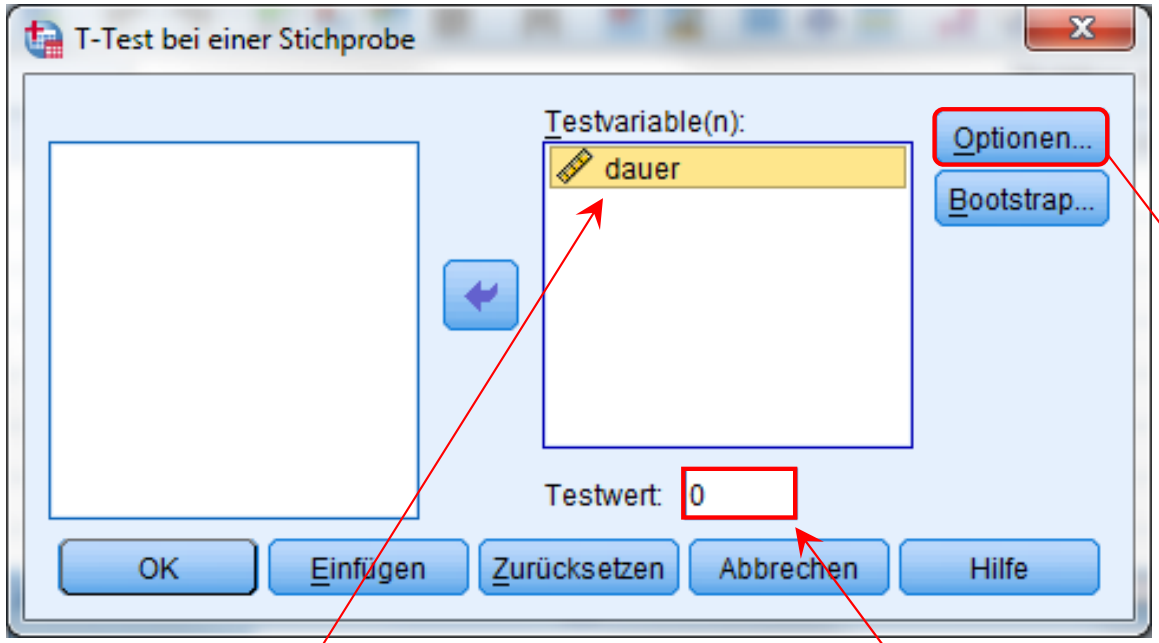
- Es werden Mittelwert und Standardabweichung von X bestimmt: $\bar{x} = 1483.43$ und $s = 41.34$. Einsetzen in die Prüfgröße ergibt:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1483.43 - 1440}{41.34/\sqrt{7}} = 2.78$$

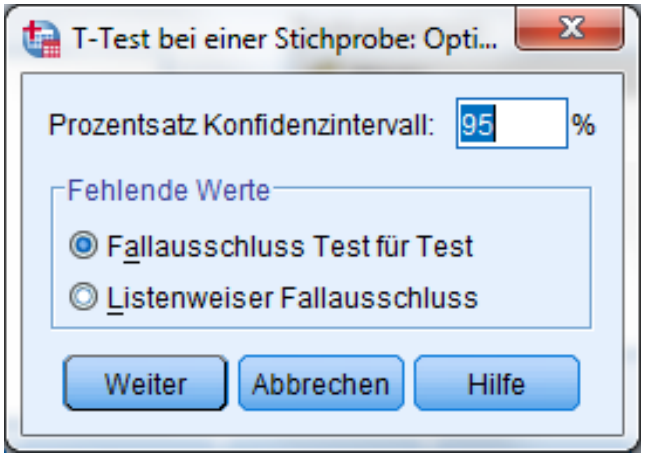
- Der tabellierten t-Verteilung entnehmen wir für $\alpha = 0.05$ und $df = n - 1 = 6$ den Wert $t_{crit} = t_{6;0.975} = 2.447$. Da $|t| > t_{crit}$ weisen wir die Nullhypothese zurück und schließen, dass sich die Periodendauer statistisch signifikant verlängert hat.

Ein-Stichproben t-Test und Konfidenzintervall in SPSS

- Der Ein-Stichproben t-Test sowie das entsprechende Konfidenzintervall finden sich in SPSS unter `Analysieren/Mittelwerte_vergleichen/T-Test_bei_einer_Stichprobe…`.



Unter `(Optionen)` kann für die Berechnung des Konfidenzintervalls der Konfidenzkoeffizient abweichend von 95% ($\alpha = 0.05$) spezifiziert werden.

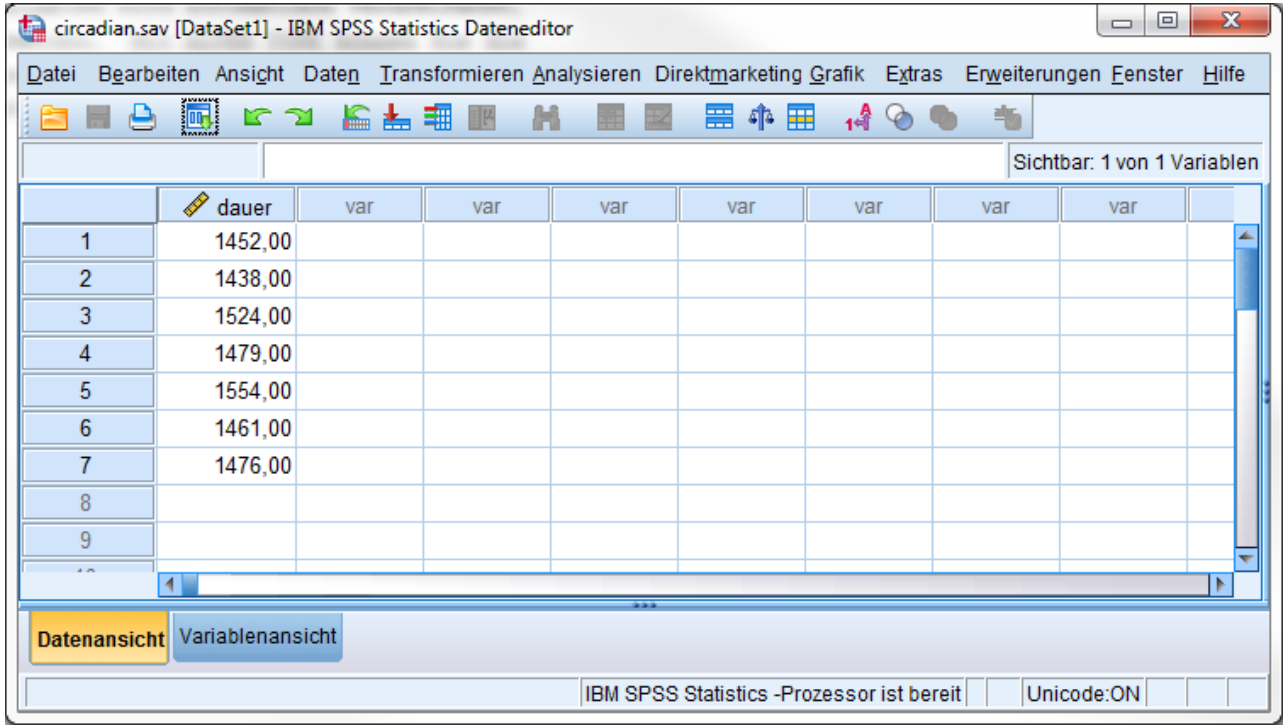


Hier ist die Variable X anzugeben. Werden mehrere Variablen spezifiziert, so wird mit allen getrennt derselbe Test ausgeführt.

Hier ist der Wert μ_0 anzugeben. Im Beispiel ist der voreingestellte Wert von `0` auf `1440` zu ändern.

Ein-Stichproben t-Test und Konfidenzintervall in SPSS

- In unserem Beispiel werden die Daten der $n = 7$ Vpn wie üblich in SPSS eingegeben:



- Im Ergebnis-Viewer werden dann zunächst tabellarisch die folgenden Statistiken ausgegeben:

Statistik bei einer Stichprobe

	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
dauer	7	1483,4286	41,34351	15,62638

n $\quad$ $\bar{x}$ $\quad$ s $\quad$ $s_{\bar{x}} = s / \sqrt{n}$

Ein-Stichproben t-Test und Konfidenzintervall in SPSS

Hier werden von SPSS die Grenzen des 95%-Konfidenzintervalls (KI) um die mittlere Differenz statt um $\bar{x}$ angegeben. Um das in der Regel gewünschte KI $\bar{x} \pm t_{n-1;1-\alpha/2} \cdot s/\sqrt{n}$ zu erhalten, müssen wir μ_0 addieren und erhalten dann das KI $[1440 + 5.19, 1440 + 81.66] = [1445.19, 1521.66]$. Man kann also auch an diesem KI sehen, das der Mittelwert (bei $\alpha = 0.05$) signifikant von μ_0 abweicht, da $\mu_0 = 1440$ nicht im KI liegt.

Test bei einer Sichprobe

Testwert = 1440

	T	df	Sig. (2-seitig)	Mittlere Differenz	95% Konfidenzintervall der Differenz	
					Untere	Obere
dauer	2,779	6	,032	43,42857	5,1922	81,6649

Prüfgröße t mit Freiheitsgraden df. Hier: $t(6) = 2.78$

SPSS gibt unter [Sig. (2-seitig)] statt dem p-Wert den Wert $2 \cdot p$ an, so dass man bei zweiseitigem Testen unmittelbar das gesetzte α mit diesem Wert vergleichen kann. Hier ist dieser Wert 0.032 kleiner als $\alpha = 0.05$ und damit wird die H_0 zurückgewiesen (nicht aber bei $\alpha = 0.01$).

Bei einseitiger Testung wäre α mit dem halbierten Wert zu vergleichen; hier wäre also zu prüfen, ob α kleiner als $0.032/2 = 0.016$ ist.

Ein-Stichproben Tests

- Neben den vorgestellten z- und t-Test gibt es noch weitere Ein-Stichproben Tests, die prüfen, ob ein Maß der zentralen Tendenz einen bestimmten Wert in der Population annimmt. Diese kommen mit noch schwächeren Voraussetzungen aus:

Bezeichnung	Voraussetzungen	zentrale Tendenz	nähere Beschreibung z.B. in
z-Test für eine Stich-probe; Ein-Stichproben Gauß-Test	X metrisch und normal-verteilt, σ bekannt	Mittelwert	✓
t-Test für eine Stichprobe	X metrisch und normalverteilt	Mittelwert	✓
Wilcoxon-Vorzeichen-Rangtest	X metrisch und symmetrisch verteilt	Median	Eid, Gollwitzer & Schmitt (2010, S. 280ff)
Vorzeichentest	X ordinalskaliert	Median	Eid, Gollwitzer & Schmitt (2010, S. 278ff)

- Daneben gibt es auch Ein-Stichproben Tests, die für andere univariate Kennwerte prüfen, ob sie in der Population einen bestimmten festen Wert aufweisen (z.B. die Varianz oder Häufig-keiten); vgl. z.B. Eid, Gollwitzer & Schmitt (2010, Kap. 10.3 - 10.5).

Unterschiede in zentraler Tendenz in einer oder zwei Gruppen: Gliederung

1 Ein-Stichproben-Tests

(z-Test, t-Test, weitere Tests)

2 Parametrische Tests für Mittelwertsunterschiede zweier Gruppen

(Unabhängigkeit und Abhängigkeit von Gruppen, t-Test für unabhängige und abhängige Gruppen, Welch-Test)

3 Prüfung der Voraussetzungen

(Robustheit, Varianzhomogenität, Normalverteilung)

4 Nicht-parametrische Tests für Unterschiede zweier Gruppen in der zentralen Tendenz

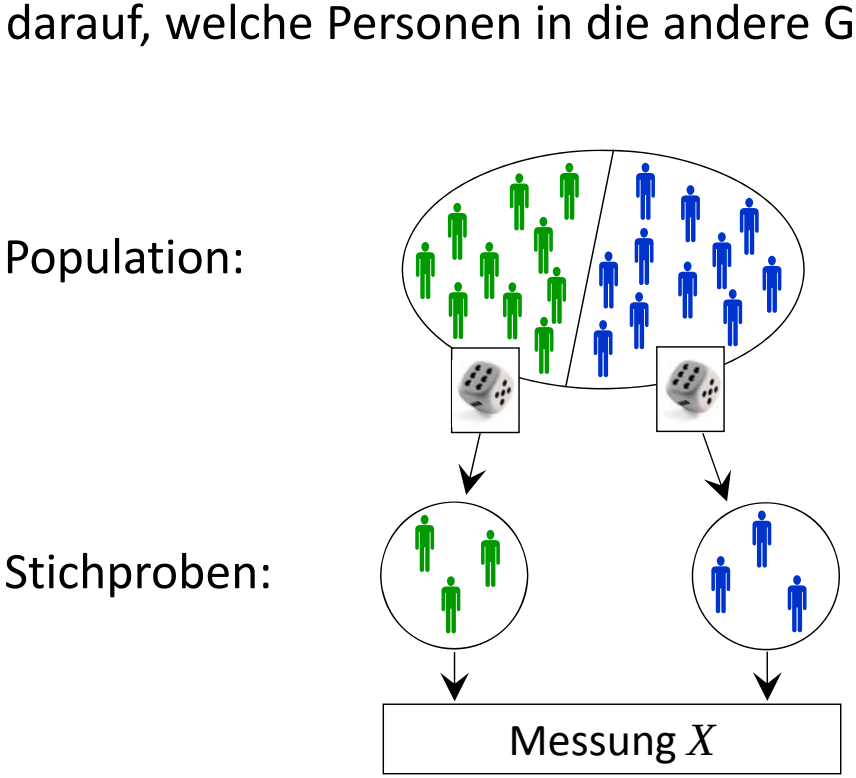
(Mann-Whitney U-Test, Vorzeichen-Test, Wilcoxon-Test)

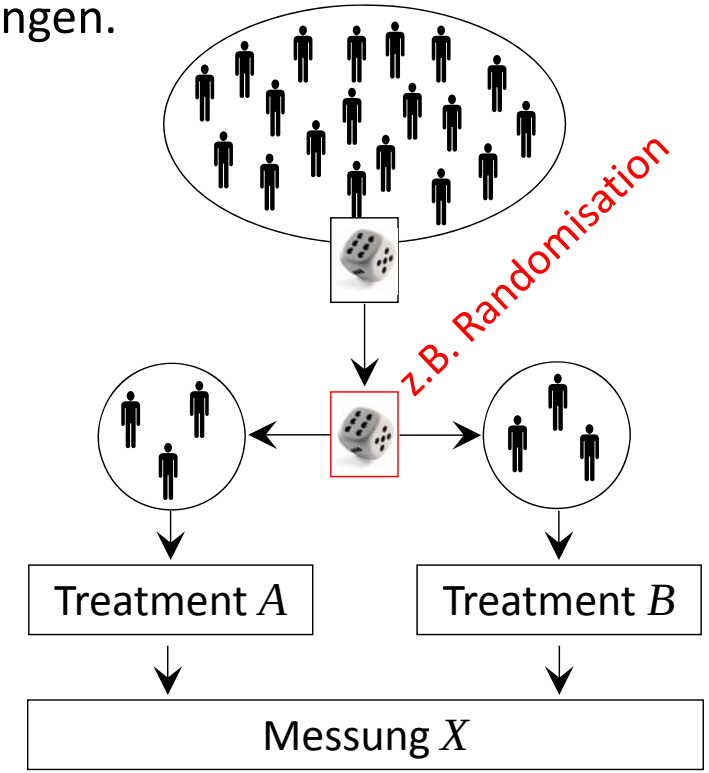
Zwei-Stichproben Tests

- In den Ein-Stichproben-Tests wurden Hypothesen geprüft, die sich auf den Unterschied eines Stichprobenmittelwertes (oder einer anderen Statistik) und eines fixen Wertes beziehen.
- Häufiger finden sich aber Unterschiedshypothesen, die einen Unterschied zwischen den Mittelwerten zweier Gruppen (Teilpopulationen) in einer Variable X prüfen wollen (Zwei-Stichproben-Tests).
- Beispiele für entsprechende wissenschaftliche Hypothesen sind:
  - Therapie A ist wirksam bei Panikattacken.
  - Lehrmethode A ist effektiver als Lehrmethode B.
  - Männer sind kreativer als Frauen.
  - Bei Paaren, in denen beide berufstätig sind, arbeiten Männer weniger im Haushalt.
  - Durch ein Raucherentwöhnungstraining reduziert sich die durchschnittliche Zahl an gerauchten Zigaretten pro Tag.
- Die beiden Gruppen (Messungen) können z.B. durch unterschiedliche Versuchsbedingungen (z.B. Experimental- vs. Kontrollgruppe), Merkmale der Personen (z.B. Geschlecht) oder unterschiedliche Zeitpunkte (z.B. vor und nach einer Therapie) zustande kommen.

Zwei-Stichproben Tests: (Un-) Abhängigkeit der Gruppen

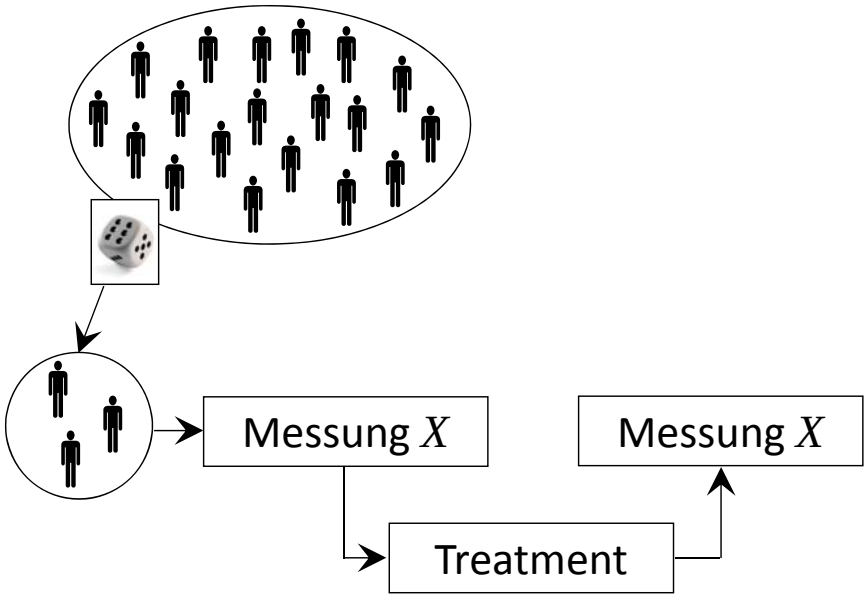
- Für die Testung ist wichtig zu wissen, ob die beiden Gruppen (Messungen, Stichproben), in denen die gleiche Variable X erhoben wurde, unabhängig oder abhängig sind.
- Unabhängige Gruppen liegen vor, wenn (a) unabhängig voneinander zufällig Gruppen aus Teilpopulationen gezogen werden oder (b) eine Zufallsstichprobe gezogen wird und die Personen (idealerweise zufällig: Randomisation) auf die Versuchsbedingungen zugewiesen werden. Die Auswahl der Personen in der einen Gruppe hat dabei also keinen Einfluss darauf, welche Personen in die andere Gruppe gelangen.

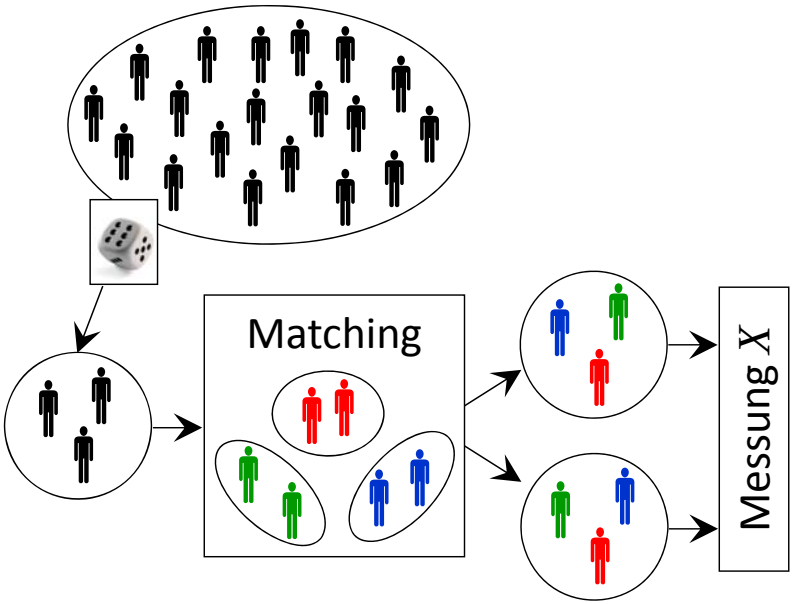




Zwei-Stichproben Tests: (Un-) Abhängigkeit der Gruppen

- Abhängige Gruppen (gebundene oder gepaarte Gruppen/Messungen/Stichproben) liegen vor, wenn …
  - mehrfache Messungen in einer Stichprobe vorliegen (repeated measures)
  - ein Matching (Parallelisierung) der Personen hinsichtlich einer oder mehrerer Drittvariablen stattgefunden hat
  - wenn natürliche Paare von Personen zufällig (z.B. eineiige Zwillinge) oder systematisch (z.B. Männer vs. Frauen aus Ehepaaren) auf die Gruppen verteilt werden





Zwei-Stichproben Tests: (Un-) Abhängigkeit der Gruppen

- Bei bestimmten Untersuchungsfragen kann man als Forscher entscheiden, ob man in der Studie einen Versuchsplan mit unabhängigen Gruppen (between-subject designs) oder mit Messwiederholung (within-subject designs) realisieren will.
- Beispiel: Man will untersuchen, ob Vpn sinnlose Silben besser erinnern können, wenn sie dafür einen finanziellen Anreiz erhalten.
- Einige Vor- und Nachteile von Messwiederholungsdesigns gegenüber unabhängigen Gruppen:
  - \+ Interindividuelle Unterschiede können nicht mehr mit der UV konfundiert sein, was die interne Validität der Studie erhöht.
  - \+ Interindividuelle Unterschiede können bei der statistischen Analyse als Fehlerquelle herausgezogen werden; dies erhöht die Power.
  - \+ Es besteht ein geringerer Vpn-Bedarf.
  - \- Es können z.B. Übungs- oder Ermüdungseffekte auftreten, so dass die Treatments nicht mehr unter vergleichbaren Bedingungen untersucht werden (daher: Ausbalancieren).
  - \- Es kann drop-out geben, d.h. Vpn stehen zu späteren Zeitpunkten nicht mehr zur Verfügung.

t-Test für unabhängige Gruppen

- Beispiel: Es interessiert die Frage, ob es Personen gelingt, ihre Antworten in Persönlichkeitstests im Sinne einer positiven Darstellung zu verzerren (z.B. als Bewerber auf eine Stelle). Dazu wird 18 studentischen Vpn der Persönlichkeitstest NEO-FFI zur Beantwortung vorgelegt, wobei die Vpn per Zufall entweder der Gruppe zugewiesen wurden, die den Test wahrheitsgemäß beantworten („Standard-Instruktion“) bzw. sich positiv darstellen soll („Faking good“-Instruktion).
- Im Folgenden werden nur die Punktwerte der Skala Gewissenhaftigkeit als dem NEO-FFI betrachtet. Es wird vermutet, dass es Personen gelingt, sich positiver (gewissenhafter) darzustellen.

Gemeinsame Instruktion: „Sie werden auf den folgenden Seiten eine Reihe von Aussagen über bestimmte Verhaltensweisen, Einstellungen und Interessen finden. Sie können diesen Aussagen mehr oder weniger zustimmen oder sie ablehnen.“

Gruppe 1 (Standard-Instruktion): „Es gibt keine richtigen und falschen Antworten. Antworten Sie bitte so, wie es nach ihrer persönlichen Ansicht auf Sie zutrifft.“

Gruppe 2 (Faking good-Instruktion): „Bitte stellen Sie sich vor, sie würden sich auf eine für sie attraktive Stelle bewerben und sie wollten daher ein besonders günstiges Bild von sich abgeben. Beantworten Sie bitte die Fragen so, wie Sie meinen, dass es für eine Einstellung auf diese Stelle möglichst günstig ist.“

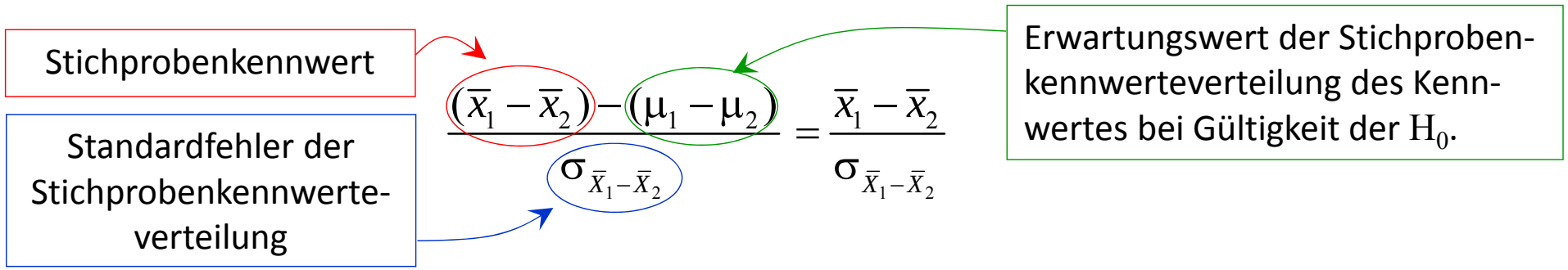
Gr. 1	Gr. 2
46	38
29	45
41	36
29	41
31	40
44	32
37	45
27	46
28	
33	

$n_1 = 10$	$n_2 = 8$
$\overline{x}_1 = 34.50$	$\overline{x}_2 = 40.38$
$s_1 = 7.03$	$s_2 = 4.93$
$s_1^2 = 49.39$	$s_2^2 = 24.27$

t-Test für unabhängige Gruppen

- Bezeichnung: t-Test für unabhängige Gruppen.
- Einsatzbereich: Prüfung der Hypothese, dass sich die Mittelwerte von zwei unabhängigen Gruppen in einer Variablen X unterscheiden.
- Hypothesen: zweiseitig: H_0: $\mu_1 = \mu_2$ und H_1: $\mu_1 \neq \mu_2$ oder entsprechend einseitig.
- Voraussetzungen: In beiden Teilpopulationen (Gruppen) muss gelten: Das metrische Merkmal X muss normalverteilt sein und die Varianzen müssen gleich sein (Homoskedastizität).
- Vorgehen: Ist die H_0 korrekt, so entstammen die beiden Mittelwerte $\bar{x}_1$ und $\bar{x}_2$ jeweils einer Population mit gleichem Mittelwert $\mu_1 = \mu_2$ oder $\mu_1 - \mu_2 = 0$.

  Um zu statistischen Entscheidung über die Gültigkeit der Hypothese auf der Basis der beobachteten Mittelwertsdifferenz zu kommen, konstruieren wir nach der üblichen Logik eine Prüfgröße:



t-Test für unabhängige Gruppen

- Wie lässt sich sich nun aber der Standardfehler der Stichprobenkennwerteverteilung der Mittelwertsdifferenzen $\sigma_{\bar{X}_1-\bar{X}_2}$ bestimmen?
- Für die Varianzen der Stichprobenkennwerteverteilungen (=Varianzfehler=quadrierter Standardfehler) der beiden Mittelwerte gilt bei Normalverteilung in beiden Gruppen:

$$\sigma^2_{\bar{X}_1} = \frac{\sigma_1^2}{n_1} \quad \text{und} \quad \sigma^2_{\bar{X}_2} = \frac{\sigma_2^2}{n_2}$$

- Bildet man nun die Stichprobenkennwerteverteilung der Differenzen der Mittelwerte, so gilt für deren Varianzfehler bei Unabhängigkeit von $\bar{X}_1$ und $\bar{X}_2$ (vgl. Rechenregel für Varianzen):

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}$$

Anwendung der Rechenregel: $\mathrm{Var}(c\cdot X + d\cdot Y) = c^2\cdot \mathrm{Var}(X) + d^2\cdot \mathrm{Var}(Y)$ falls X und Y unabhängig (mit $c = 1$ und $d = -1$):

- Setzen wir die Varianzfehler ein und vereinfachen unter Berücksichtigung der Annahme der Homoskedastizität $\sigma_1 = \sigma_2 := \sigma$, so resultiert

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

t-Test für unabhängige Gruppen

- Da aber die Varianz von X in der Population σ^2 unbekannt ist, müssen wir sie auf der Basis der Varianzen in den Stichproben s_1^2 und s_2^2 schätzen. Im Falle gleicher Stichprobenumfänge in beiden Gruppen $n_1 = n_2$ gilt:

$$\hat{\sigma}^2 = s_g^2 = \frac{s_1^2 + s_2^2}{2}$$

- Im allgemeineren Fall ungleicher Stichprobenumfänge $n_1 \neq n_2$ gilt für die gepoolte Varianz

$$\hat{\sigma}^2 = s_g^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

- Setzen wir in die Prüfgröße ein, die in diesem Fall wieder einer t-Verteilung folgt, so erhalten wir allgemein

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\hat{\sigma}^2 \cdot (1/n_1 + 1/n_2)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_g^2 \cdot (1/n_1 + 1/n_2)}}$$

$$= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad \text{mit} \quad df = n_1 + n_2 - 2$$

t-Test für unabhängige Gruppen

- Entscheidung: Zurückweisung der H_0, falls $|t| > t_{crit}$ mit $t_{crit} = t_{n_1+n_2-2;1-\alpha/2}$ bei zweiseitiger Prüfung und $t_{crit} = t_{n_1+n_2-2;1-\alpha}$ bei einseitiger Prüfung (und korrekter Richtung).

- Beispiel: Die Forschungshypothese „Es wird vermutet, dass es Personen gelingt, sich positiver (gewissenhafter) darzustellen.“ soll zweiseitig bei $\alpha = 0.05$ getestet werden:

  $H_0: \mu_1 = \mu_2, H_1: \mu_1 \neq \mu_2$

- Die Berechnung der Prüfgröße ergibt

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)\cdot s_1^2 + (n_2-1)\cdot s_2^2}{n_1+n_2-2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{34.50 - 40.38}{\sqrt{\frac{(10-1)\cdot 49.39 + (8-1)\cdot 24.27}{10+8-2} \cdot \left(\frac{1}{10} + \frac{1}{8}\right)}} = -2.00$$

Standard	Faking good
Gr. 1	Gr. 2
46	38
29	45
41	36
29	41
31	40
44	32
37	45
27	46
28	
33	

$n_1 = 10$	$n_2 = 8$
$\bar{x}_1 = 34.50$	$\bar{x}_2 = 40.38$
$s_1 = 7.03$	$s_2 = 4.93$
$s_1^2 = 49.39$	$s_2^2 = 24.27$

t-Test für unabhängige Gruppen

- Der tabellierten t-Verteilung entnehmen wir für $\alpha = 0.05$ und $df = n_1 + n_2 - 2 = 10 + 8 - 2 = 16$ sowie zweiseitiger Testung den Wert $t_{crit} = t_{16;0.975} = 2.120$.

  Da $|t| = 2.00 < t_{crit}$ behalten wir die Nullhypothese bei und schließen, dass die Selbstbeschreibungen der Gewissenhaftigkeit unter der Faking-Good Bedingung nicht statistisch signifikant positiver oder negativer ausfällt als in der Kontrollbedingung.

- Hätten wir die gerichtet formulierte Hypothese einseitig getestet, so hätten wir zunächst festgestellt, dass der Mittelwertsunterschied in die erwartete Richtung geht und dann statistisch getestet.

  Wir hätten dann bei $\alpha = 0.05$ einen kritischen Wert von $t_{crit} = t_{16;0.95} = 1.746$ erhalten.

  Da nun $|t| = 2.00 > t_{crit}$ ist, wären wir abweichend zu der Entscheidung gelangt, die Nullhypothese zurückzuweisen.

- Bisher haben wir (wie bei der Ein-Stichproben Testung) nicht geprüft, ob die Voraussetzungen der Varianzhomogenität und der Normalverteilung gegeben sind. Wir kommen darauf später zurück. Jetzt werden wir eine Modifikation des t-Tests für unabhängige Gruppen betrachten, die die Homoskedastizität nicht erfordert.

Welch-Test für unabhängige Gruppen

- Bezeichnung: t-Test für unabhängige Gruppen mit Welch-Korrektur; Welch-Test.
- Einsatzbereich, Hypothesen und Entscheidung wie bei t-test für unabhängige Gruppen.
- Voraussetzungen: In beiden Teilpopulationen (Gruppen) muss das metrische Merkmal X normalverteilt sein.
- Vorgehen: Das Aufgeben der Annahme homogener Varianzen führt zur Veränderung der Schätzung des Varianzfehlers der Differenzen (und damit auch der Prüfgröße) sowie der Freiheitsgrade wie folgt:

$$\hat{\sigma}^2_{\bar{X}_1-\bar{X}_2} = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2} = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{X}_1-\bar{X}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \text{mit} \quad df_{korr} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Welch-Test für unabhängige Gruppen

- Beispiel: Die Forschungshypothese „Es wird vermutet, dass es Personen gelingt, sich positiver (gewissenhafter) darzustellen.“ soll wieder zweiseitig bei $\alpha = 0.05$ getestet werden.

	Standard	Faking good
	$n_1 = 10$	$n_2 = 8$
	$\bar{x}_1 = 34.50$	$\bar{x}_2 = 40.38$
	$s_1 = 7.03$	$s_2 = 4.93$
	$s_1^2 = 49.39$	$s_2^2 = 24.27$

- Die Berechnung der modifizierten Prüfgröße und *df* ergibt:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{34.50 - 40.38}{\sqrt{\frac{49.39}{10} + \frac{24.27}{8}}} = -2.08$$

$$df_{korr} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{49.39}{10} + \frac{24.27}{8}\right)^2}{\frac{\left(\frac{49.39}{10}\right)^2}{10 - 1} + \frac{\left(\frac{24.27}{8}\right)^2}{8 - 1}} = 15.79$$

Konservativ bedeutet, zugunsten der H_0: ein kleinerer *df*-Wert führt zu einem größeren kritischen Wert und damit weniger eher zu einem statistisch signifikanten Resultat.

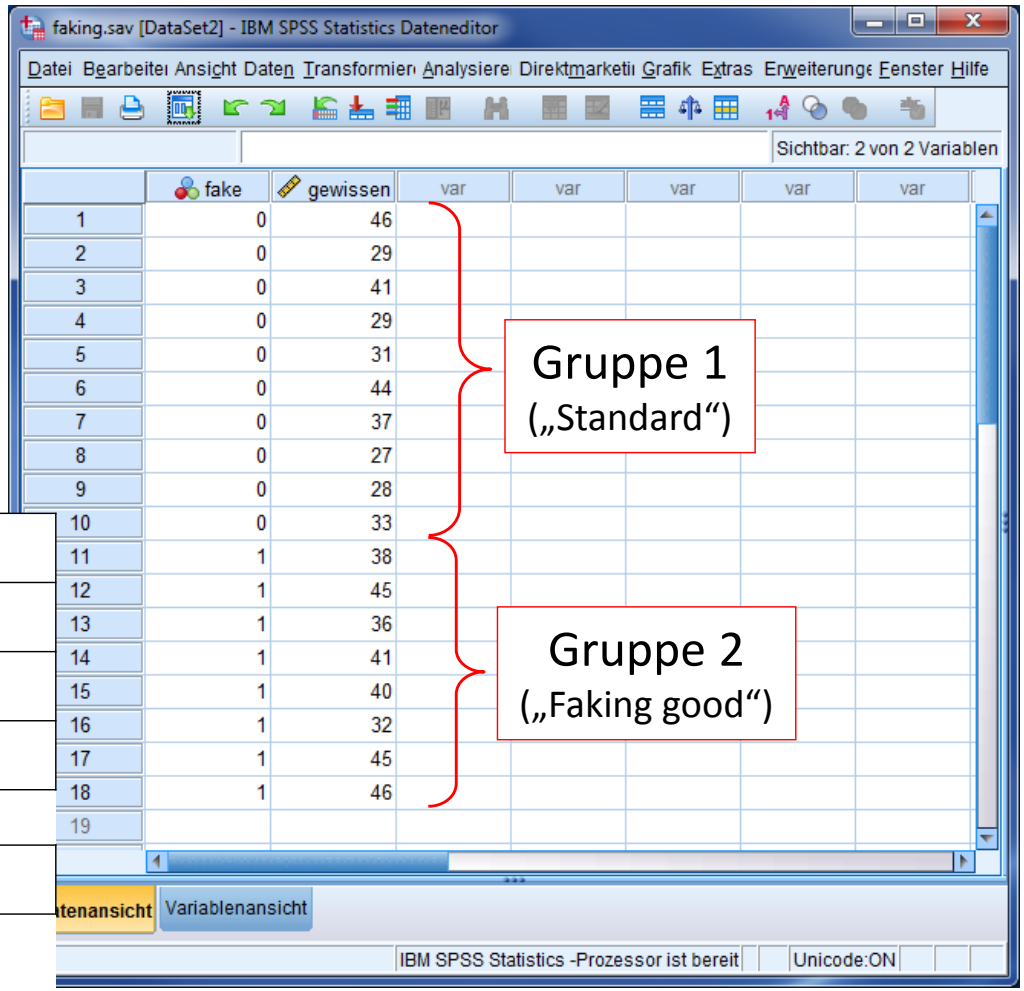
- Bei Aufsuchen des kritischen Wertes in der Tabelle wählen wir konservativ den Freiheitsgrad 15 und erhalten $t_{crit} = t_{15;0.975} = 2.131$. Da $|t| = 2.08 < t_{crit}$ behalten wir auch hier die H_0 bei.

t- und Welch-Test für unabhängige Gruppen in SPSS

- Den t-Test für unabhängige Stichproben und den Welch-Test erhält man unter: `Analysieren/Mittelwerte_vergleichen/T-Test_bei_unabhängigen_Stichproben….`

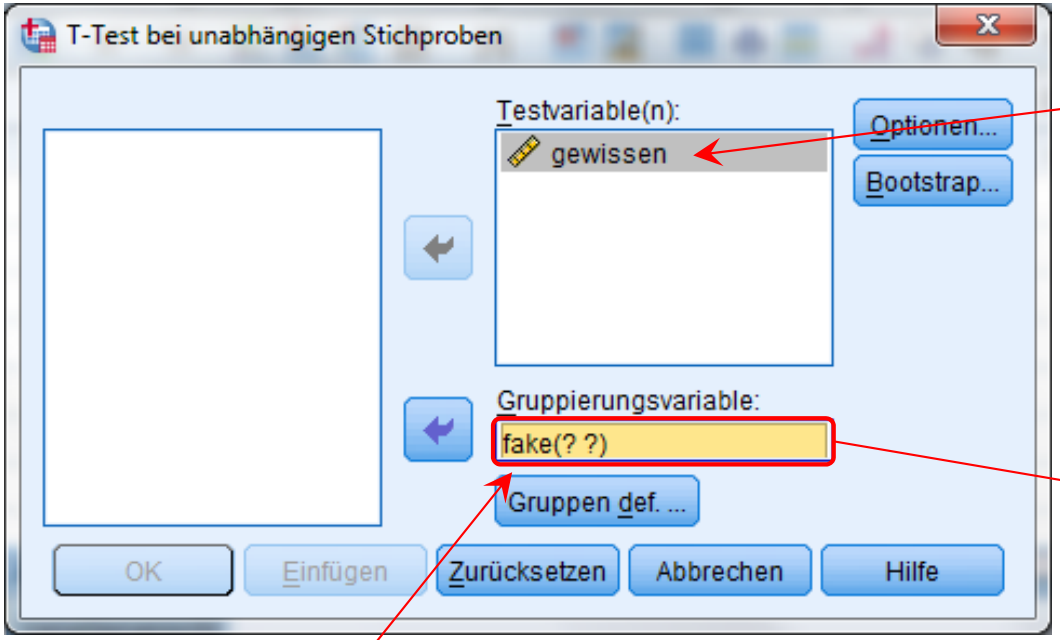
- In diesem Fall sind die Daten in SPSS so einzugeben, dass die Gruppenzugehörigkeit in einer Variable codiert wird (hier: FAKE) und die abhängige Variable X in einer zweiten Variablen (hier: GEWISSEN).

- Sinnvoll ist zudem, in der Variablenansicht für die Gruppierungs-Variable FAKE Wertelabels zu vergeben, hier z.B. 0=„Standard“ und 1=„Faking good“

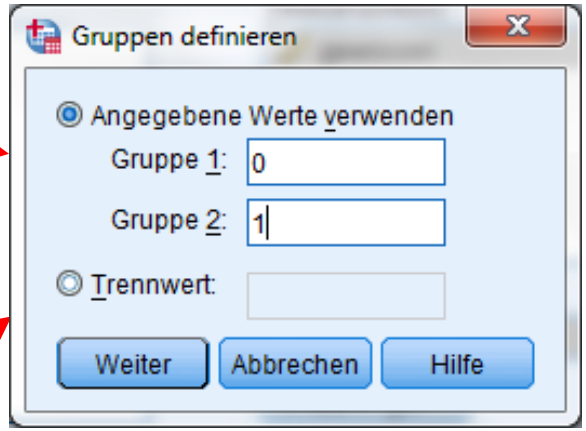


Gr. 1	Gr. 2
46	38
29	45
41	36
…	…
27	46
28	
33	

t- und Welch-Test für unabhängige Gruppen in SPSS



Hier wird die abhängige Variable angegeben. Werden mehrere Variablen spezifiziert, bestimmt SPSS für jede separat den t- & Welch-Test.



Hier wird die unabhängige Variable (Gruppierungsvariable) angegeben, in der die Gruppenzugehörigkeit codiert ist. Zusätzlich sind nach dem Drücken des Buttons (`Gruppen_def._…`) die beiden Werte anzugeben, die die beiden Gruppen definieren (also z.B. `0` vs. `1` bei der Variable FAKE für die Standard- und die Faking-good-Bedingung).

t- und Welch-Test für unabhängige Gruppen in SPSS

T-Test

Gruppenstatistiken

	fake	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
gewissen	0 Standard	10	34,50	7,028	2,222
	1 Faking good	8	40,38	4,926	1,742

Zunächst werden die geläufigen deskriptiven Statistiken Stichprobenumfang, Mittelwert und Standardabweichung sowie der Standardfehler der Mittelwertes der abhängigen Variablen (hier: GEWISSEN) für beide Gruppen getrennt (hier: „Standard“ vs. „Faking-good“ Bedingung) angegeben.

Man erkennt, dass deskriptiv der Mittelwertsunterschied in die in der Forschungshypothese postulierte Richtung geht.

t- und Welch-Test für unabhängige Gruppen in SPSS

Test bei unabhängigen Stichproben

		Levene-Test der Varianzgleichheit		T-Test für die Mittelwertgleichheit						
									95% Konfidenzintervall der Differenz	
		F	Signifikanz	T	df	Sig. (2-seitig)	Mittlere Differenz	Standardfehler der Differenz	Untere	Obere
gewissen	Varianzen sind gleich	2,392	,142	-1,999	16	,063	-5,875	2,939	-12,106	,356
	Varianzen sind nicht gleich			-2,081	15,79	,054	-5,875	2,824	-11,867	,117

In der oberen Zeile ("Varianzen sind gleich") steht die Prüfgröße t [T] mit den Freiheitsgraden df [df] und dem p-Wert [Sig. (2-seitig)] für die zweiseitige Testung des t-Tests für unabhängige Gruppen.

Im Beispiel führt der t-Test bei $\alpha = 0.05$ zu einem statistisch insignifikanten Ergebnis (da $.063$ größer als $.05$ ist).

In der unteren Zeile ("Varianzen sind nicht gleich") erhalten wir die analogen Ausgaben für den Welch-Test.

Beide Tests führen hier zu der gleichen statistischen Entscheidung.

In dem grün gerahmten Bereich wird ein Test zur Prüfung der Voraussetzung der Varianzhomogenität ausgegeben, auf den wir später zurückkommen. Auf der Basis des Ergebnisses dieses Tests kann entschieden werden, ob der t- oder Welch-Test verwendet werden sollte.

Parametrische Tests für Mittelwertsunterschiede zweier unabh. Gruppen

- Neben den vorgestellten t- und Welch-Tests gibt es noch weitere Tests für die Prüfung auf Unterschiedlichkeit zweier Maße der zentralen Tendenz in unabhängigen Gruppen.
- Von denen, die eine metrische abhängige Variable und normalverteilte Werte voraussetzen, ist noch der z-Test zu ergänzen, der aber selten eingesetzt wird, da er bekannte Populationsvarianzen voraussetzt.
- Auf Tests, die die Normalverteilungsannahme nicht benötigen, kommen wir später zurück.

Bezeichnung	Voraussetzungen	zentrale Tendenz	nähere Beschreibung z.B. in
z-Test für unabhängige Stichproben, Zwei-Stichproben Gauß-Test	X metrisch, in beiden Gruppen Normalverteilung und bekannte Varianzen σ^2	Mittel-werte	Diehl & Arbinger (2001, Kap. 5.4); Eid, Gollwitzer & Schmitt (2010, S. 305ff)
t-Test für unabhängige Gruppen	X metrisch, in beiden Gruppen Normalverteilung und homogene Varianzen	Mittel-werte	✓
Welch-Test	X metrisch, in beiden Gruppen Normalverteilung	Mittel-werte	✓

t-Test für abhängige Gruppen

- Beispiel (aus Howell, 2010, nach Daten von Everitt, 1994): Bei 15 Frauen (Mädchen) mit der Diagnose Anorexia nervosa (Magersucht) wurde vor und nach einer familientherapeutischen Behandlung das Körpergewicht in kg erfasst. Die Hypothese lautet, dass die Behandlung in dem Sinne wirksam ist, dass eine Gewichtszunahme zu verzeichnen ist.

- Folgende deskriptive Statistiken können berechnet werden:

$n = 15$	
$\bar{x}_1 = 41.39$	$\bar{x}_2 = 44.90$
$s_1 = 2.59$	$s_2 = 4.38$

Vp-Nr.	vorher	nachher
1	41.9	47.6
2	41.7	47.2
3	43.0	45.7
4	41.2	45.9
5	43.4	50.2
6	39.8	38.4
7	38.4	38.4
8	47.1	50.8
9	36.7	47.4
10	40.3	37.6
11	40.8	38.9
12	41.0	47.8
13	38.8	45.3
14	41.8	46.3
15	44.9	46.0

t-Test für abhängige Gruppen

- Bezeichnung: t-Test für abhängige (verbundene) Gruppen.
- Einsatzbereich: Prüfung der Hypothese, dass sich die Mittelwerte von zwei abhängigen Gruppen in einer Variablen X unterscheiden.
- Hypothesen: zweiseitig: H_0: $\mu_1 = \mu_2$ und H_1: $\mu_1 \neq \mu_2$ oder entsprechend einseitig.
- Voraussetzungen: Normalverteilte Differenzwerte $X_1 – X_2$.
- Vorgehen: Durch die paarweise Zuordnung der Messwerte lassen sich die Differenzen der n Messwertpaare $d_i = x_{i1} - x_{i2}$ bilden sowie deren Mittelwert $\overline{x}_d$ und Standardabweichung s_d.

Für die Prüfgröße können wir wieder schreiben:

$$\frac{\overline{x}_d - (\mu_1 - \mu_2)}{\sigma_{\overline{X}_d}} = \frac{\overline{x}_d}{\sigma_{\overline{X}_d}}$$

Für den Standardfehler resultiert

$$\hat{\sigma}_{\overline{X}_d} = \frac{s_d}{\sqrt{n}}$$

Gr. 1	Gr. 2	
x_{11}	x_{12}	$d_1 = x_{11} - x_{12}$
x_{21}	x_{22}	$d_2 = x_{21} - x_{22}$
…	…	…
x_{n1}	x_{n2}	$d_n = x_{n1} - x_{n2}$
$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_d = \overline{x}_1 - \overline{x}_2, s_d$

t-Test für abhängige Gruppen

- Einsetzen führt zu einer *t*-verteilten Prüfgröße:

$$t = \frac{\bar{x}_d}{\hat{\sigma}_{\bar{X}_d}} = \frac{\bar{x}_d}{s_d / \sqrt{n}} \quad \text{mit} \quad df = n - 1$$

- Entscheidung: Zurückweisung der H_0, falls $|t| > t_{crit}$ mit $t_{crit} = t_{n-1,1-\alpha/2}$ bei zweiseitiger Prüfung und $t_{crit} = t_{n-1,1-\alpha}$ bei einseitiger Prüfung (und korrekter Richtung).
- Der t-Test für abhängige Gruppen entspricht dem Ein-Stichproben t-Test für die Differenzwerte d_i.

- Weitere Tests für die Prüfung auf Unterschiedlichkeit zweier Mittelwerte in abhängigen Gruppen unter Verwendung der Normalverteilungsannahme sind nicht gebräuchlich. Auf Tests, die die Normalverteilungsannahme nicht benötigen, kommen wir später zurück.

t-Test für abhängige Gruppen

- Beispiel: Die gerichtete Hypothese, dass die Behandlung zu einer Gewichtszunahme führt, soll zweiseitig bei $\alpha = 0.05$ getestet werden.

  Es resultiert: $\bar{x}_d = -3.51 = \bar{x}_1 - \bar{x}_2 = 41.39 - 44.90$
  und $s_d = 3.82$

  Einsetzen erbringt die Prüfgröße:

$$t = \frac{\bar{x}_d}{s_d / \sqrt{n}} = \frac{-3.51}{3.82 / \sqrt{15}} = -3.56$$

- Der tabellierten t-Verteilung entnehmen wir für $\alpha = 0.05$ und $df = n - 1 = 14$ sowie zweiseitiger Testung den Wert $t_{crit} = t_{14;0.975} = 2.145.$

  Da $|t| = 3.56 > t_{crit}$ weisen wir die Nullhypothese zurück und schließen, dass das Gewicht der Frauen nach der Therapie statistisch signifikant höher ist als vorher.

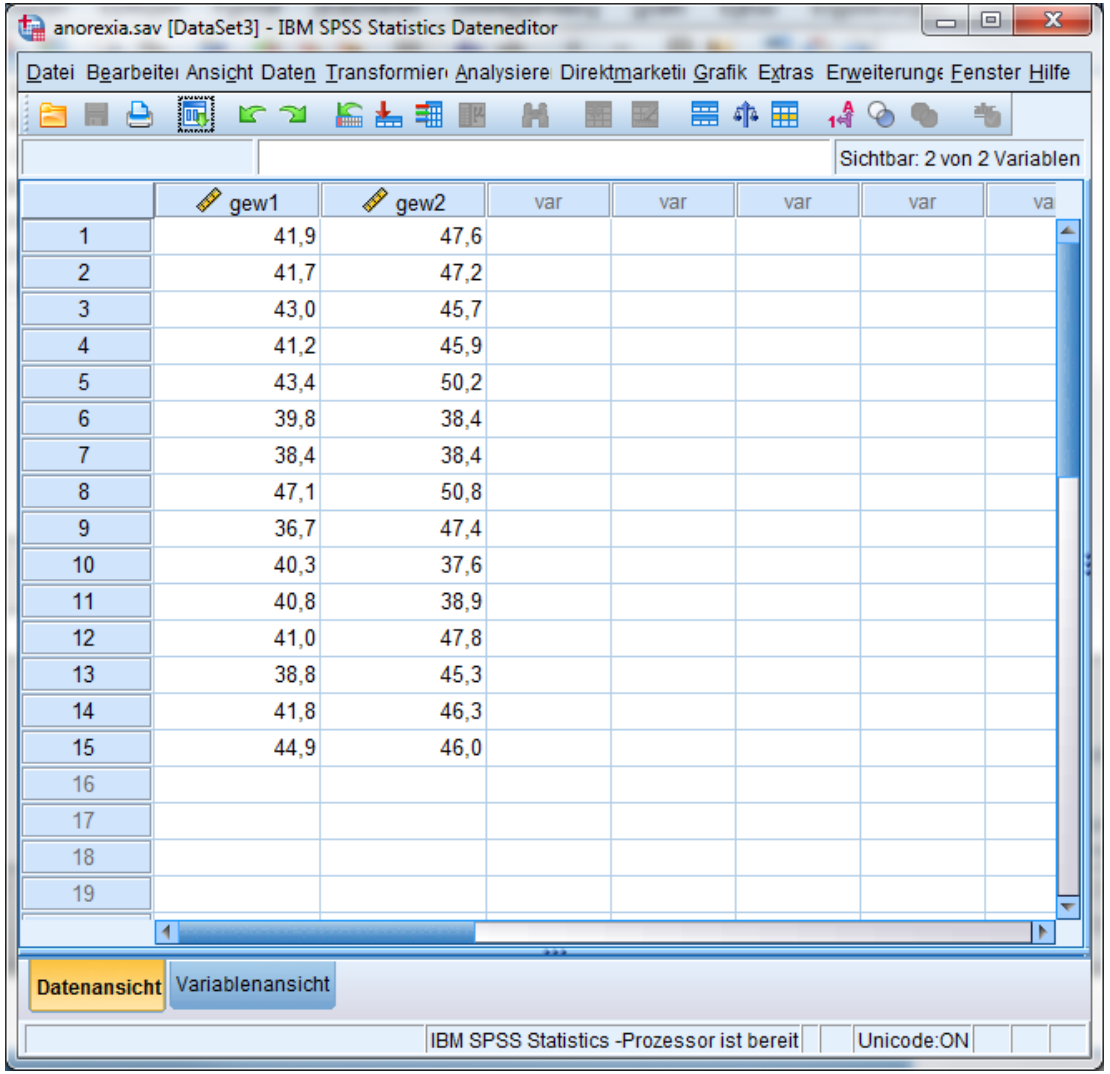
vorher	nachher	d_i
41.9	47.6	-5.7
41.7	47.2	-5.5
43.0	45.7	-2.7
41.2	45.9	-4.7
43.4	50.2	-6.8
39.8	38.4	1.4
38.4	38.4	0.0
47.1	50.8	-3.7
36.7	47.4	-10.7
40.3	37.6	2.7
40.8	38.9	1.9
41.0	47.8	-6.8
38.8	45.3	-6.5
41.8	46.3	-4.5
44.9	46.0	-1.1

$\bar{x}_1 = 41.39$	$\bar{x}_2 = 44.90$	$\bar{x}_d = -3.51$
$n = 15$		$s_d = 3.82$

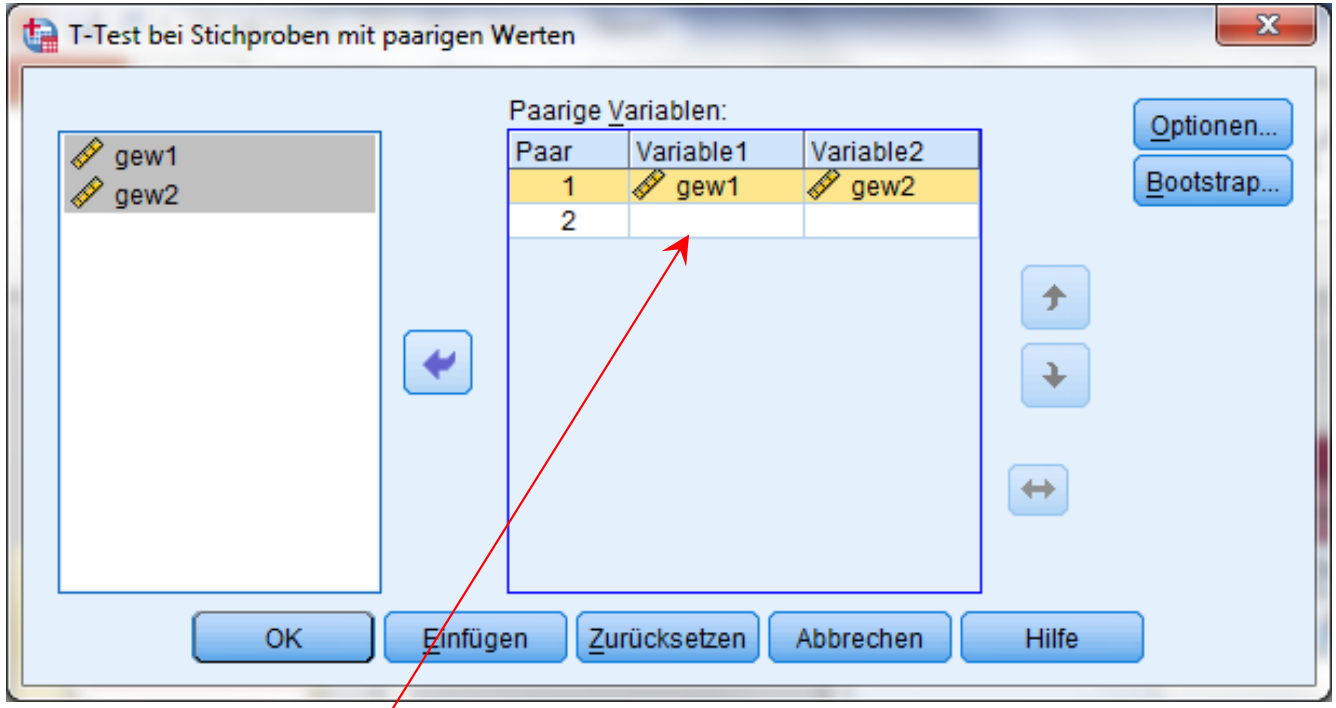
t-Test für abhängige Gruppen in SPSS

- Den t-Test für abhängige Stichproben erhält man unter: `Analysieren/Mittelwerte vergleichen/T-Test_bei_verbundenen_Stichproben`….

- In diesem Fall sind die Daten in SPSS so einzugeben, dass die zusammengehörigen Messwertpaare immer in zwei Variablen in einer Zeile stehen (hier GEW1 und GEW2).



t-Test für abhängige Gruppen in SPSS



Unter "Gepaarte Variablen" werden die beiden Variablen angegeben, deren Mittelwerte gegeneinander getestet werden sollen. Wenn mehrere Variablenpaare spezifiziert werden (in Paar 2 usw.), bestimmt SPSS für jedes Paar separat einen t-Test für abhängige Stichproben.

t-Test für abhängige Gruppen in SPSS

T-Test

Statistik bei gepaarten Stichproben

		Mittelwert	N	Standardabweichung	Standardfehler des Mittelwertes
Paaren 1	gew1	41,387	15	2,5909	,6690
	gew2	44,900	15	4,3826	1,1316

Zunächst werden die geläufigen deskriptiven Statistiken Stichprobenumfang, Mittelwert und Standardabweichung sowie der Standardfehler der Mittelwertes der abhängigen Variablen (hier: GEW1 und GEW2) für beide Gruppen getrennt angegeben.
Man erkennt, dass deskriptiv der Mittelwertsunterschied in die in der Forschungshypothese postulierte Richtung geht.

Korrelationen bei gepaarten Stichproben

		N	Korrelation	Signifikanz
Paaren 1	gew1 & gew2	15	,499	,058

Zusätzlich wird die lineare Produkt-Moment Korrelation der beiden Variablen mit dem (noch nicht behandelten) Test auf statistische Signifikanz ausgegeben.

t-Test für abhängige Gruppen in SPSS

Test bei gepaarten Stichproben

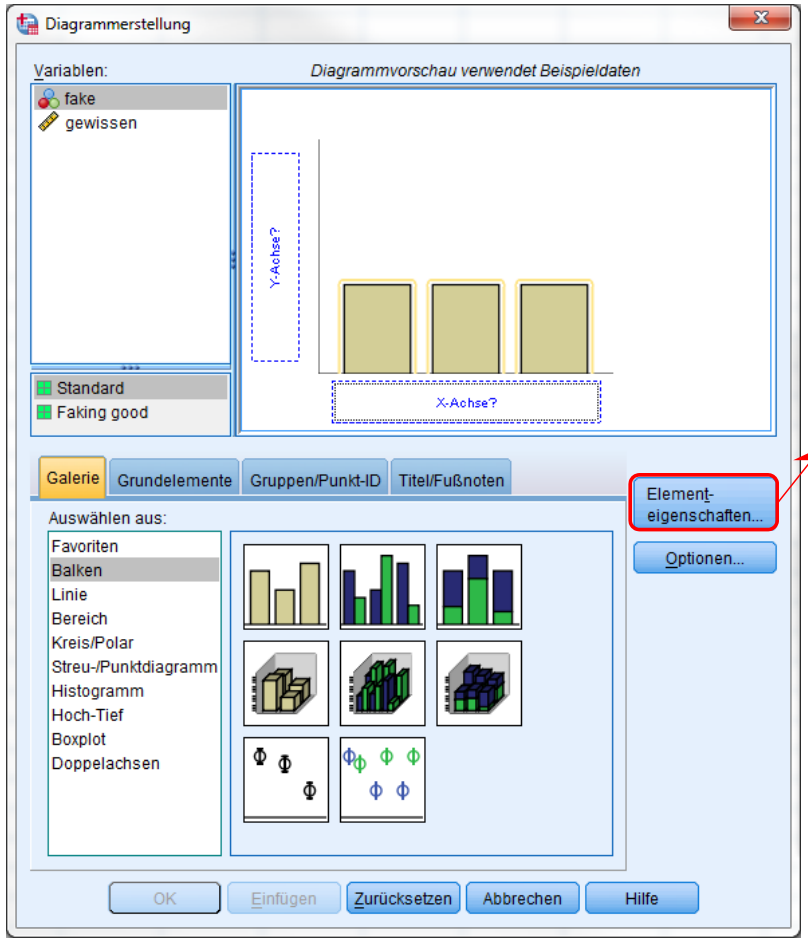
		Gepaarte Differenzen							
					95% Konfidenzintervall der Differenz				
		Mittelwert	Standardabweichung	Standardfehler des Mittelwertes	Untere	Obere	T	df	Sig. (2-seitig)
Paaren 1	gew1 - gew2	-3,5133	3,8191	,9861	-5,6283	-1,3984	-3,563	14	,003

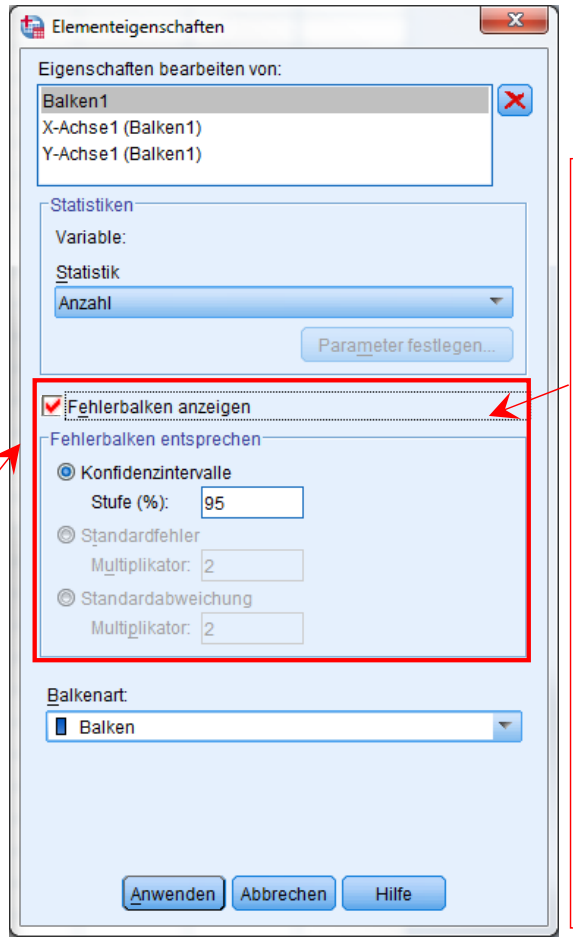
Hier steht die Prüfgröße t [T] mit den Freiheitsgraden df [df] und dem p-Wert [Sig. (2-seitig)] für die zweiseitige Testung des t-Tests für abhängige Stichproben.

Im Beispiel führt der t-Test bei $\alpha = 0.05$ zu einem statistisch signifikanten Ergebnis (da $.003$ kleiner als $.05$ ist).

Mittelwertsdarstellungen in SPSS

- Bei zwei (oder mehreren) Gruppen lassen sich in SPSS die Mittelwerte als Histogramme (oder Polygonzüge) darstellen und zusätzlich mit „Fehlerbalken“ versehen. Bei unabhängigen Gruppen wählt man unter `Grafik/Diagrammerstellung…` aus der ⌈Galerie⌉ die Option „Balken“ aus und zieht die UV (hier: FAKE) auf die Abszisse und die AV (hier: GEWISSEN) auf die Ordinate des Diagramms.



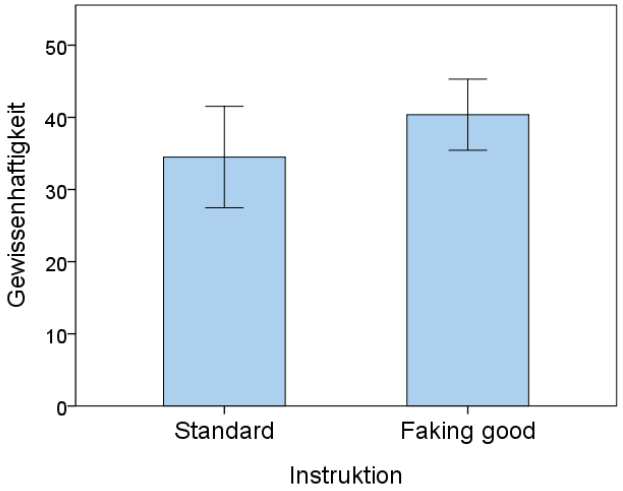


Unter [Elementeigenschaften] kann man nach Aktivieren der Option „Fehlerbalken anzeigen“ zwischen drei verschiedenen Varianten von Fehlerbalken wählen. Bei Wahl einer der beiden unteren Optionen wird man im Regelfall den „Multiplikator“ auf 1 setzen.

Mittelwertsdarstellungen in SPSS

- Im Beispiel resultieren dann (nach Bearbeitung) die untenstehenden Grafiken, wenn als Fehlerbalken die Standardabweichungen (unten links), die Standardfehler (Mitte) bzw. die Konfidenzintervalle (rechts) um die Mittelwerte der Gruppen abgetragen werden. Es ist also jeweils darauf zu achten, um welche Fehlerbalken es sich jeweils handelt.

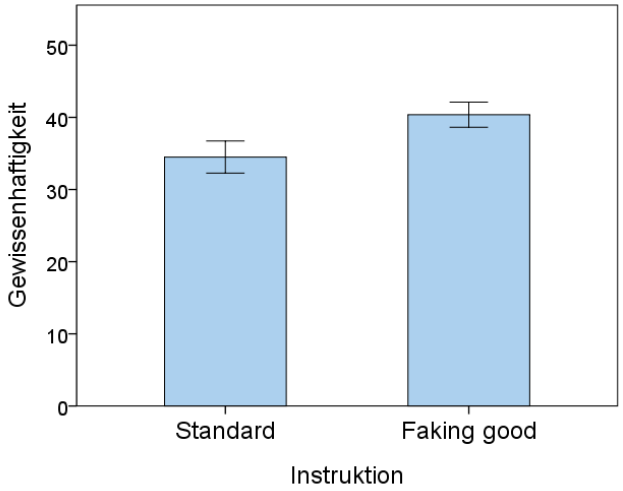
Standard	Faking good
$\overline{x}_1 = 34.50$	$\overline{x}_2 = 40.38$
$s_1 = 7.03$	$s_2 = 4.93$
$s_{\overline{X}_1} = 2.22$	$s_{\overline{X}_2} = 1.74$



$$\overline{x} \pm s$$

$$\overline{x}_1 \pm s_1 = 34.50 \pm 7.03$$

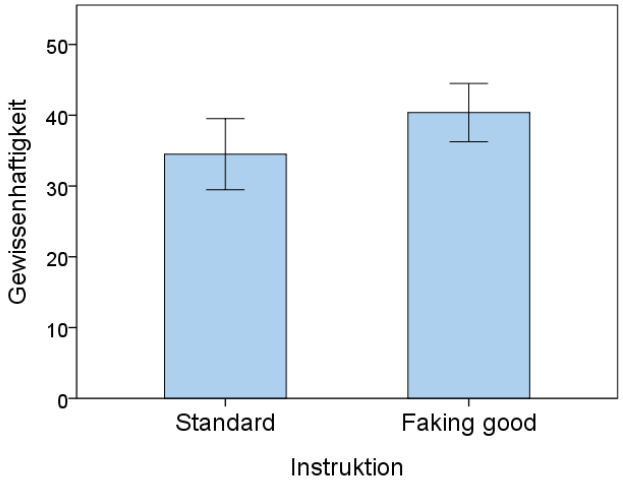
$$\overline{x}_2 \pm s_2 = 40.38 \pm 4.93$$



$$\overline{x} \pm s_{\overline{X}}$$

$$\overline{x}_1 \pm s_{\overline{X}_1} = 34.50 \pm 2.22$$

$$\overline{x}_2 \pm s_{\overline{X}_2} = 40.38 \pm 1.74$$



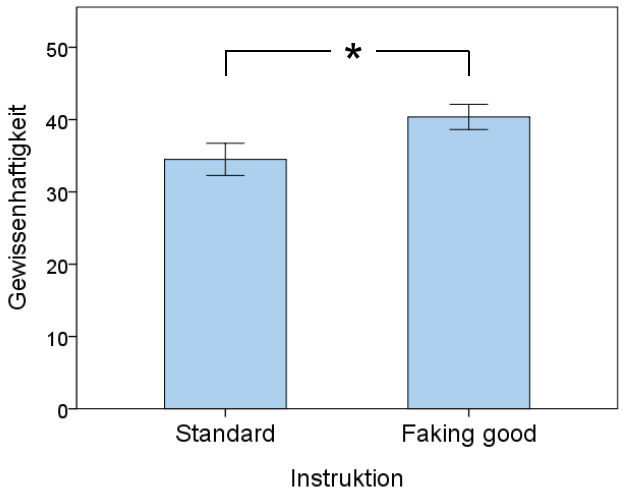
$$\overline{x} \pm s_{\overline{X}} \cdot t_{n-1;1-\alpha/2}$$

$$\overline{x}_1 \pm s_{\overline{X}_1} \cdot t_{9;0.975} = 34.50 \pm 5.03$$

$$\overline{x}_2 \pm s_{\overline{X}_2} \cdot t_{7;0.975} = 40.38 \pm 4.12$$

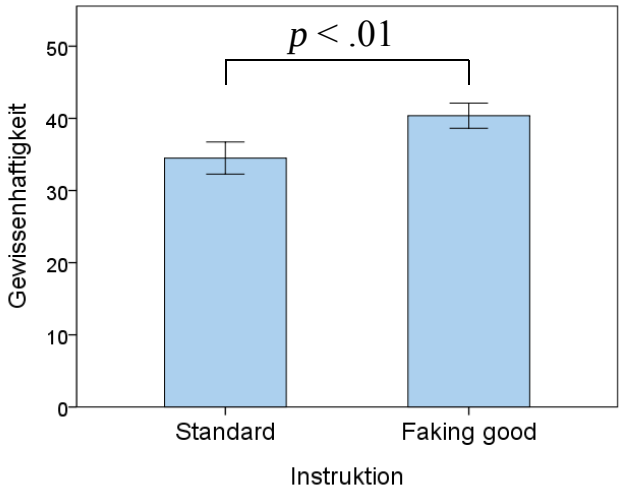
Mittelwertsdarstellungen in SPSS

- Manchmal wird zusätzlich dargestellt, ob sich die Mittelwerte (nicht) statistisch signifikant unterscheiden, z.B. wie folgt:



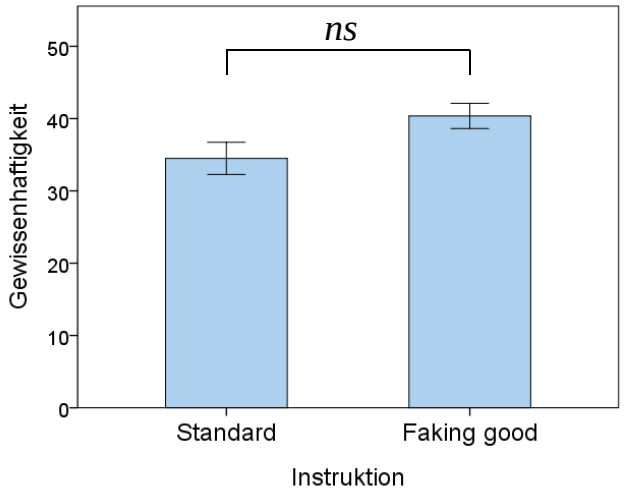
$\overline{x} \pm s_{\overline{x}}$

Die beiden Mittelwerte unterscheiden sich statistisch signifikant (üblicherweise bei $\alpha = .05$)



$\overline{x} \pm s_{\overline{x}}$

Die beiden Mittelwerte unterscheiden sich statistisch signifikant (bei $\alpha = .01$)



$\overline{x} \pm s_{\overline{x}}$

Die beiden Mittelwerte unterscheiden sich nicht statistisch signifikant (üblicherweise bei $\alpha = .05$)

Unterschiede in zentraler Tendenz in einer oder zwei Gruppen: Gliederung

1 Ein-Stichproben-Tests

(z-Test, t-Test, weitere Tests)

2 Parametrische Tests für Mittelwertsunterschiede zweier Gruppen

(Unabhängigkeit und Abhängigkeit von Gruppen, t-Test für unabhängige und abhängige Gruppen, Welch-Test)

3 Prüfung der Voraussetzungen

(Robustheit, Varianzhomogenität, Normalverteilung)

4 Nicht-parametrische Tests für Unterschiede zweier Gruppen in der zentralen Tendenz

(Mann-Whitney U-Test, Vorzeichen-Test, Wilcoxon-Test)

Prüfung von Voraussetzungen

- Alle statistischen Tests haben bestimmte Voraussetzungen (Annahmen). Darunter sind bestimmte bezüglich des Zustandekommens der Daten (z.B. einfache Zufallsstichprobe, Abhängigkeit bzw. Unabhängigkeit der Gruppen) und bezüglich der Verteilung der betrachteten Variablen (z.B. Normalverteilung, Varianzhomogenität).
- Bestimmte Voraussetzungen lassen sich relativ einfach prüfen (z.B. die nach dem Vorliegen abhängiger vs. unabhängiger Gruppen), bei anderen ist es aufwendiger und manche lassen sich nicht adäquat prüfen.
- Eine Voraussetzung, die in der psychologischen Forschung fast nie erfüllt ist, ist das Vorliegen einer einfachen Zufallsstichprobe. Vielmehr liegen meist überhaupt gar keine Zufallsstichproben vor, sondern sog. Gelegenheits- (anfallende, ad hoc-) Stichproben.
- Für Gelegenheitsstichproben existiert keine statistische Theorie. Es wird im Regelfall so getan, als läge eine Zufallsstichprobe vor. Dabei kann passieren, dass es bei der Ziehung derartiger Stichproben zu systematischen Verzerrungen (bias) kommt. (Dies kann zudem auch bei Zufallsstichproben, z.B. durch non-response, passieren.)

Prüfung von Voraussetzungen

- Neben Voraussetzungen bezüglich des Zustandekommens der Daten können auch Verteilungsvoraussetzungen verletzt sein. Die Voraussetzung der Normalverteilung kann z.B. dadurch verletzt sein, dass das Merkmal in der Population nicht normalverteilt ist (z.B. Reaktionszeiten, die sich linkssteil verteilen) oder sich Abweichungen aus irgend einem Grund in der Stichprobe ergeben (z.B. Messfehler, die zu Ausreißern führen).

- Die Prüfung von Verteilungsannahmen, die sich immer auf die Population beziehen, erfolgt ebenfalls mittels statistischer Tests, bei denen man sich fehlerhaft entscheiden kann (also α oder β-Fehler begehen kann).

Prüfung von Voraussetzungen

- Tests reagieren unterschiedlich stark auf Verletzungen der Voraussetzungen. Man sagt auch, sie sind unterschiedlich robust (oder empfindlich) gegenüber diesen Verletzungen. Ein Test ist umso robuster, je weniger sich durch die Verletzungen die Stichprobenkennwerteverteilung und deren Parameter ändern.

- Bei empfindlichen Tests können das faktische α und das nominelle α stark auseinanderfallen. Dies kann z.B. dazu führen, dass die Wahrscheinlichkeit eines faktischen α-Fehlers viel größer ist als das gesetzte, nominelle α-Niveau (von z.B. 0.05). Außerdem kann sich auch die Power der Tests verändern (d.h. größer oder kleiner werden).

- Bei der Robustheit spielt eine Rolle, ob die Verfahren in bestimmter Weise reagieren, z.B. so, dass das faktische α kleiner ist als das nominelle, was zu konservativen Entscheidungen (d.h. eher zugunsten der H_0) führt. Wird durch Verletzungen die H_1 begünstigt, spricht man von progressiven Entscheidungen (oder inflationiertem Risiko für α-Fehler).

Prüfung von Voraussetzungen

- Wie robust ein Testverfahren ist, kann z.B. mittels Computersimulationen untersucht werden. Dazu erzeugt man künstlich Daten mit bestimmten Verteilungseigenschaften, z.B. für die Prüfung der Robustheit des t-Tests für unabhängige Gruppen zwei normalverteilte Populationen mit gleichem μ und σ, aus denen wiederholt Stichproben der Größe $n_1 = n_2 = 100$ gezogen werden. [Anmerkung: Zieht man z.B. 1000 mal solche Stichproben und führt den t-Test immer zweiseitig mit $\alpha = 0.05$ durch, so sollte in 50 Tests (5%) ein statistisch signifikantes Ergebnis resultieren.]
- Dann verändert man die Daten so, dass eine oder beide Voraussetzungen in einem bestimmten Umfang nicht mehr erfüllt sind (z.B. indem man die Varianz in einer der Populationen erhöht) und zieht wieder viele Male die Stichproben und führt den Test durch (z.B. immer zweiseitig mit $\alpha = 0.05$). Resultiert dann in 5% der Ziehungen ein statistisch signifikantes Ergebnis, so ist der Test gegenüber den untersuchten Verletzungen unter den betrachteten Bedingungen robust. Ist er z.B. nur in 2% der Fälle signifikant, so reagiert der Test konservativ auf derartige Voraussetzungsverletzungen.
- Die Ergebnisse solcher Robustheitsuntersuchungen sind häufig komplex. Die Robustheit hängt meist von verschiedenen Bedingungen ab (n, Ausmaß der Verstöße, α, Gerichtetheit etc.) und davon, ab Verstöße gegen mehrere Voraussetzungen simultan vorliegen.

Prüfung von Voraussetzungen

- Einige Befunde zur Robustheit des Ein-Stichproben t-Tests und t-Test für abhängige Gruppen (vgl. Diehl & Arbinger, 2001, S. 97ff):
  - Bei Verletzung der Normalverteilungsannahme reagiert der Test meist progressiv und dies umso mehr, je stärker die Abweichungen sind. Besonders problematisch ist es, wenn die Verteilung nicht symmetrisch ist.
  - Die Robustheit nimmt mit wachsendem n zu. (Große Stichproben „härten“ den Test gegen die Verletzung der Normalverteilungsannahme.)
  - Bei zweiseitiger Prüfung sind die Tests robuster bei 5% als bei kleineren α-Werten.

Prüfung von Voraussetzungen

- Befunde zur Robustheit des t-Tests für unabhängige Gruppen (vgl. Diehl & Arbinger, 2001, S. 145ff):
  - Bei gleichem n in beiden Gruppen und normalverteilten Populationen ist der t-Test robust gegenüber heterogenen Populationsvarianzen.
  - Bei ungleichem n hingegen reagiert der t-Test empfindlich auf das Vorliegen von Varianzheterogenität.
  - Bei gleichem n und Varianzhomogenität ist der Test robust gegenüber Verletzungen der Normalverteilungsannahmen im Sinne mäßig schiefer Verteilungen. Symmetrische nichtnormalverteilte Verteilungen sind relativ unproblematisch.
  - Heterogene Varianzen und schiefe Verteilungen führen (insbesondere bei kleinen Stichproben) zu teilweise extremer Nichtrobustheit des Tests.

Prüfung von Voraussetzungen: Levene-Test

- Vorbemerkung: Da bei vielen Voraussetzungsprüfungen fehlende Abweichungen bzw. Gleichheit von Verteilungen (z.B. Normalverteilung) oder Parametern (z.B. Varianzen) postuliert wird, stecken die „Wunschhypothesen“ meist in der H_0. Dann ist es sinnvoll, zur Verringerung des β-Fehler Risikos ein größeres α zu wählen, üblich sind 0.10 oder 0.20.

- Bezeichnung: Varianzhomogenitätstest von Levene, Levene-Test.

- Einsatzbereich: Prüfung der Homogenität der Varianzen in zwei unabhängigen Gruppen (auch verallgemeinerbar für den Fall von mehr als zwei Gruppen).

- Hypothesen: Die ungerichteten statistische Hypothesen zur Prüfung der Varianzhomogenität von zwei Gruppen lauten:
  - $H_0: \sigma_1^2 = \sigma_2^2$ (Die Varianzen beider Gruppen sind in der Population gleich)
  - $H_1: \sigma_1^2 \neq \sigma_2^2$ (Die Varianzen beider Gruppen sind in der Population ungleich)

- Voraussetzungen: Normalverteilung in beiden Gruppen.

Prüfung von Voraussetzungen: Levene-Test

- Vorgehen: Bestimmung der absoluten Abweichungen aller Werte vom jeweiligen Gruppenmittelwert $y_{i1} = |x_{i1} - \bar{x}_1|$ bzw. $y_{i2} = |x_{i2} - \bar{x}_2|$ und Bestimmung der Mittelwerte dieser Werte getrennt in den beiden Gruppen $\bar{y}_1$ bzw. $\bar{y}_2$ sowie über alle Werte $\bar{y}$.

- Die folgende Prüfgröße

$$W = (n_1 + n_2 - 2) \cdot \frac{\sum_{j=1}^{2} n_j \cdot (\bar{y}_j - \bar{y})^2}{\sum_{j=1}^{2} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}$$

Im Zähler stehen die Quadratsummen zwischen den beiden Gruppen QS_{zw}.

Im Nenner stehen die Quadratsummen innerhalb der beiden Gruppen QS_{in}.

ist dann F-verteilt mit $df_1 = 1$ und $df_2 = n_1 + n_2 - 2$ Freiheitsgraden.

- Entscheidung: Obwohl die Hypothese bei der Voraussetzungsprüfung ungerichtet formuliert ist, wird sie einseitig getestet, da durch die Quadrierung der Abweichungen der W-Wert mit zunehmendem Unterschied zwischen beiden Varianzen (gleichgültig welche der beiden größer ist) immer größer wird. Es erfolgt also eine Zurückweisung der H_0, falls $W > F_{crit}$ mit $F_{crit} = F_{1;n_1+n_2-2;1-\alpha}$.

Gr. 1	Gr. 2
x_{11}	x_{12}
x_{21}	x_{22}
…	…
…	x_{n2}
x_{n1}	
$\bar{x}_1$	$\bar{x}_2$

Y_1	Y_2
y_{11}	y_{12}
y_{21}	y_{22}
…	…
…	y_{n2}
y_{n1}	
$\bar{y}_1$	$\bar{y}_2$
$\bar{y}$	

Prüfung von Voraussetzungen: Levene-Test

- Beispiel: Datensatz zur Verfälschung von Persönlichkeitstests

$$W = (n_1 + n_2 - 2) \cdot \frac{\sum_{j=1}^{2} n_j \cdot (\bar{y}_j - \bar{y})^2}{\sum_{j=1}^{2}\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2}$$

$$= (10 + 8 - 2) \cdot \frac{10 \cdot (6.00 - 5.06)^2 + 8 \cdot (3.88 - 5.06)^2}{134.25}$$

$$= 2.39$$

- Der F-Verteilung entnehmen wir für $\alpha = 0.10$, $df_1 = 1$ und $df_2 = 10 + 8 - 2 = 16$ den Wert $F_{crit} = F_{1;16;0.90} = 3.048$.

Da $W < F_{crit}$ behalten wir die H_0 bei und gehen von keinem Verstoß gegen die Varianzhomogenität aus.

Standard	Faking good				
Gr. 1	Gr. 2	y_{i1}	y_{i2}	$(y_{i1} - \bar{y}_1)^2$	$(y_{i2} - \bar{y}_2)^2$
46	38	11.50	2.38	30.25	2.25
29	45	5.50	4.63	0.25	0.56
41	36	6.50	4.38	0.25	0.25
29	41	5.50	0.63	0.25	10.56
31	40	3.50	0.38	6.25	12.25
44	32	9.50	8.38	12.25	20.25
37	45	2.50	4.63	12.25	0.56
27	46	7.50	5.63	2.25	3.06
28		6.50		0.25	
33		1.50		20.25	

$n_1 = 10$	$n_2 = 8$
$\bar{x}_1 = 34.50$	$\bar{x}_2 = 40.38$
$s_1^2 = 49.39$	$s_2^2 = 24.27$

$\bar{y}_1 = 6.00$	$\bar{y}_2 = 3.88$
$\bar{y} = 5.06$	

$$\sum_{j=1}^{2}\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 = 134.25$$

Prüfung von Voraussetzungen

- Zur Prüfung der Hypothese homogener Varianzen stehen neben dem Levene-Test auch alternative Verfahren zur Verfügung (vgl. Diehl & Arbinger, 2001, Kap. 15). Die Verfahren unterscheiden sich vor allem darin, wie robust sie auf Verletzung der Normalverteilungsannahme in beiden Gruppen reagieren.

- Der Levene-Test wurde hier dargestellt, da er verallgemeinerbar auf mehr als zwei Gruppen ist (in dem die Zahl 2 in den Summen durch die Zahl g der Gruppen ersetzt wird) und in SPSS implementiert ist.

  - Er ist robuster gegenüber Verletzung der Normalverteilungsannahme als der einfachere, hier nicht behandelte „F-Test zur Prüfung der Varianzhomogenität“.
  - Ein weiteres Verfahren ist der Brown-Forsyth-Test, der wie der Levene-Test abläuft, mit dem einzigen Unterschied, dass die Abweichungswerte nicht vom Mittelwert, sondern vom Median bestimmt werden. Bei schiefen Verteilungen und symmetrischen Verteilungen mit $Exzess \neq 0$ weist er eine größere Robustheit auf als der Levene-Test, die mit wachsendem n noch zunimmt.

Prüfung von Voraussetzungen

- Der Befund, dass Varianzen heterogen sind, ist nicht nur als Problem in Bezug auf die Anwendung von Verfahren zu sehen, die die Homogenität als Voraussetzung haben. Sie ist selbst auch ein empirischer Befund, der inhaltlich interessant sein kann.
- So können Treatments neben der Veränderung der Mittelwerte auch den (Neben-) Effekt haben, die Varianz zu verändern (z.B. eine Induktion positiver Stimmung).

- Zudem können Verfahren zur Prüfung der Gleichheit zweier Varianzen auch direkt zur Prüfung von Forschungshypothesen eingesetzt werden.
- Beispiel: Frauen unterscheiden sich weniger im Persönlichkeitsmerkmal Gewissenhaftigkeit als Männer.
- Zu diesem Zwecke gibt es auch Verfahren, die die Gleichheit von Varianzen in zwei abhängigen Gruppen prüfen (vgl. z.B. Diehl & Arbinger, 2001, Kap. 15).

Levene-Test in SPSS

- Der Levene-Test wird standardmäßig beim t-Test für unabhängige Gruppen unter `Analysieren/Mittelwerte_vergleichen/T-Test_bei_unabhängigen_Stichproben….` ausgegeben (und auch später im erweiterten Fall von Varianzanalysen):

Test bei unabhängigen Stichproben

		Levene-Test der Varianzgleichheit		T-Test für die Mittelwertgleichheit						
									95% Konfidenzintervall der Differenz	
		F	Signifikanz	T	df	Sig. (2-seitig)	Mittlere Differenz	Standardfehler der Differenz	Untere	Obere
gewissen	Varianzen sind gleich	2,392	,142	-1,999	16	,063	-5,875	2,939	-12,106	,356
	Varianzen sind nicht gleich			-2,081	15,79	,054	-5,875	2,824	-11,867	,117

Hier steht die Prüfgröße W [F] und der p-Wert [Signifikanz] für den Levene-Test. Im Beispiel führt der Test bei $\alpha = 0.10$ zu keinem statistisch signifikanten Ergebnis (da $.142$ nicht kleiner ist als $.10$). Die Nullhypothese homogener Varianzen wird also beibehalten. (Entsprechend kann der t-Test für unabhängige Gruppen in der gleichen Zeile zur Testung der Mittelwertsunterschiede herangezogen werden.)

Unterschiede in zentraler Tendenz in einer oder zwei Gruppen: Gliederung

1 Ein-Stichproben-Tests

(z-Test, t-Test, weitere Tests)

2 Parametrische Tests für Mittelwertsunterschiede zweier Gruppen

(Unabhängigkeit und Abhängigkeit von Gruppen, t-Test für unabhängige und abhängige Gruppen, Welch-Test)

3 Prüfung der Voraussetzungen

(Robustheit, Varianzhomogenität, Normalverteilung)

4 Nicht-parametrische Tests für Unterschiede zweier Gruppen in der zentralen Tendenz

(Mann-Whitney U-Test, Vorzeichen-Test, Wilcoxon-Test)

Prüfung auf Normalverteilung

- Für die Prüfung, ob eine empirische Verteilung in der Population einer bestimmten Verteilung folgt (z.B. der Normalverteilung) gibt es eine Reihe von Testverfahren. Einige der hier nicht dargestellten Verfahren prüfen z.B. nur, ob Schiefe und Exzess der empirischen Verteilung statistisch signifikant von 0 verschieden sind. Wir stellen im Folgenden die am häufigsten eingesetzten sog. Anpassungstests (goodness of fit tests), den Kolmogorov-Smirnov und den Lilliefors-Test dar.

- Beispiel: An einer Stichprobe von 20 Schülern wurde die Intelligenz erfasst. Es soll geprüft werden, ob das Merkmal normalverteilt ist. Es wird davon ausgegangen, dass Mittelwert und Standardabweichung der Intelligenztestwerte in der Population der Schüler bekannt sind: $\mu = 100$ und $\sigma = 15$.

„Everyone believes in the normal law of errors, the experimenters because they think it is a mathematical theorem, and the mathematicians because they think it is an experimental fact.“ (Henri Poincaré)

Nr.	X
1	99
2	107
3	119
4	93
5	105
6	91
7	107
8	75
9	130
10	70

Nr.	X
11	94
12	115
13	106
14	90
15	79
16	117
17	97
18	92
19	107
20	121

Prüfung auf Normalverteilung

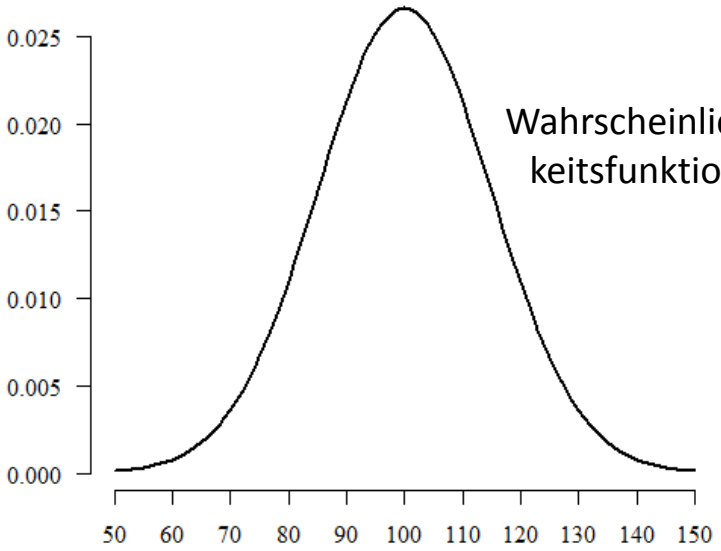
- Bezeichnung: Kolmogorov-Smirnov Test (K-S Test).
- Einsatzbereich: Prüfung auf Normalverteilung (oder eine andere Verteilungsform).
- Hypothesen: In der H_0 wird behauptet, dass die Verteilung eines Merkmals X in der Population $\Phi(x)$ einer bestimmten Verteilungsfunktion $\Phi_0(x)$ entspricht. Im Folgenden betrachten wir für $\Phi_0(x)$ als Spezialfall die Normalverteilung.
  - H_0: $\Phi(x) = \Phi_0(x)$ für alle x (Empirische und theoretische Verteilung sind identisch)
  - H_1: $\Phi(x) \neq \Phi_0(x)$ für mindestens ein x (Beide Verteilungen unterscheiden sich)

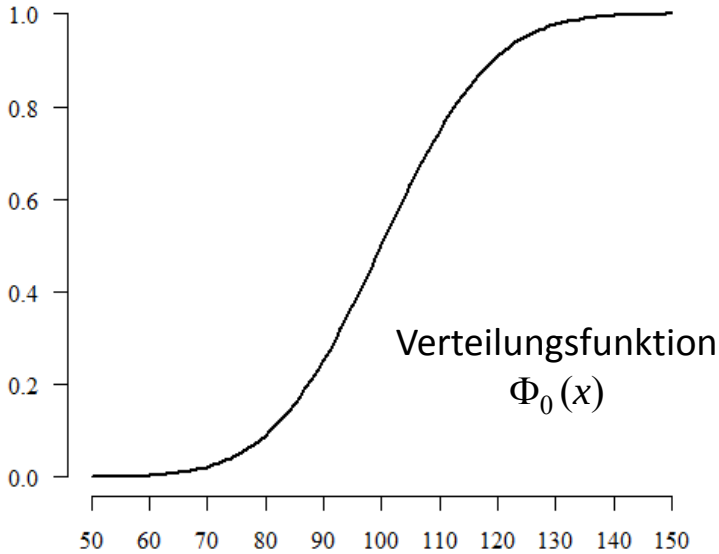
  Die Nullhypothese besagt also, dass es keinen Punkt gibt, an dem die (empirische) Verteilungsfunktion von X in der Population von der (theoretischen) Verteilungsfunktion (z.B. der Normalverteilung) abweicht.
- Voraussetzungen: Die Verteilung des Merkmals X ist stetig und weist in der Population den bekannten Mittelwert μ und die Standardabweichung σ auf.
- Vorgehen: Als Prüfgröße wird die maximale Abweichung der empirischen Verteilungsfunktion $F(x)$ und der theoretischen Verteilungsfunktion $\Phi_0(x)$ herangezogen:

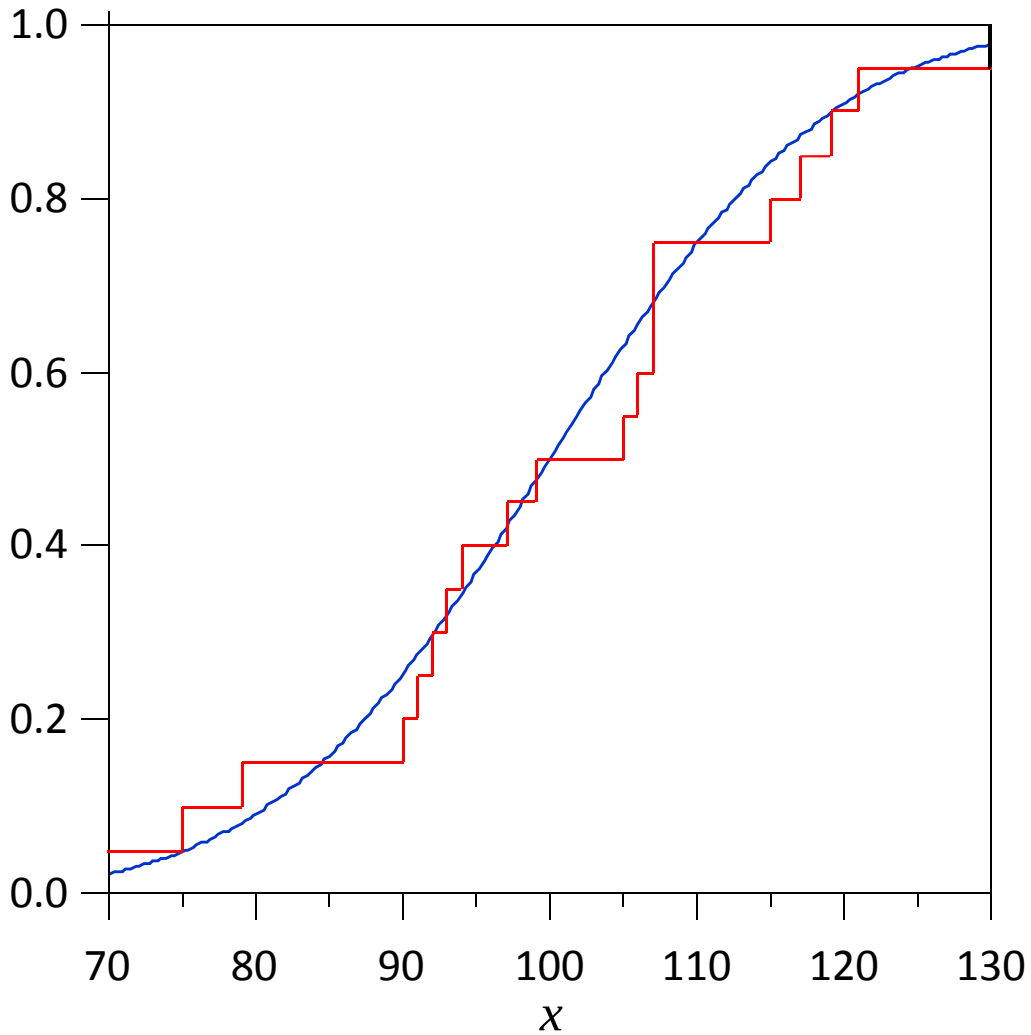
$$D_{\max} = \max_x \left| F(x) - \Phi_0(x) \right|$$

Prüfung auf Normalverteilung

Normalverteilung N(100,15)







Abweichungen der empirischen Verteilungsfunktion $F(x)$
von der theoretischen Verteilungsfunktion $\Phi_0(x)$
der Normalverteilung $\mathrm{N}(100,15)$

Prüfung auf Normalverteilung

- Im ersten Schritt werden die Werte in X aufsteigend sortiert.
- Für jeden Messwert x_i wird der Anteil der Messwerte bestimmt, der kleinergleich x_i ist $F(x_{(i)}) = (i) / n = (i) / 20.$
- Dann wird für jeden Messwert der unter der H_0 angenommene Wert der Verteilungsfunktion geschätzt. Dazu wird jeder Messwert z-transformiert

$$z_{(i)} = \frac{x_{(i)} - \mu}{\sigma} = \frac{x_{(i)} - 100}{15}$$

und die zugehörigen Wahrscheinlichkeiten bestimmt:

$$\hat{\Phi}_0(x_{(i)}) = P(Z \le z_{(i)})$$

z.B. $x_{(16)} = 115 \Rightarrow F(x_{(i)}) = 16/20 = 0.800$

$z_{(16)} = (115 - 100)/15 = 1 \Rightarrow \hat{\Phi}_0(x_{(16)}) = 0.841$

i	(i)	$x_{(i)}$	$F(x_{(i)})$	$z_{(i)}$	$\hat{\Phi}_0(x_{(i)})$
10	1	70	0.050	-2.000	0.023
8	2	75	0.100	-1.667	0.048
15	3	79	0.150	-1.400	0.081
14	4	90	0.200	-0.667	0.252
6	5	91	0.250	-0.600	0.274
18	6	92	0.300	-0.533	0.297
4	7	93	0.350	-0.467	0.320
11	8	94	0.400	-0.400	0.345
17	9	97	0.450	-0.200	0.421
1	10	99	0.500	-0.067	0.473
5	11	105	0.550	0.333	0.631
13	12	106	0.600	0.400	0.655
2	13	107	0.750	0.467	0.680
7	14	107			
19	15	107			
12	16	115	0.800	1.000	0.841
16	17	117	0.850	1.133	0.871
3	18	119	0.900	1.267	0.897
20	19	121	0.950	1.400	0.919
9	20	130	1.000	2.000	0.977

Prüfung auf Normalverteilung

- Für alle Messwerte werden nun die Abweichungen zwischen den beiden Verteilungen gebildet.
- Dabei ist zu beachten, dass sich der Abstand zwischen $F(x_i)$ und $\hat{\Phi}_0(x_{(i)})$ im Intervall $[x_{(i-1)}, x_{(i)}]$ ändert.
- Daher werden sowohl die Abweichungen am oberen und am unteren Ende des Intervalls herangezogen:

$$D_{1i} = F(x_{(i)}) - \hat{\Phi}_0(x_{(i)})$$

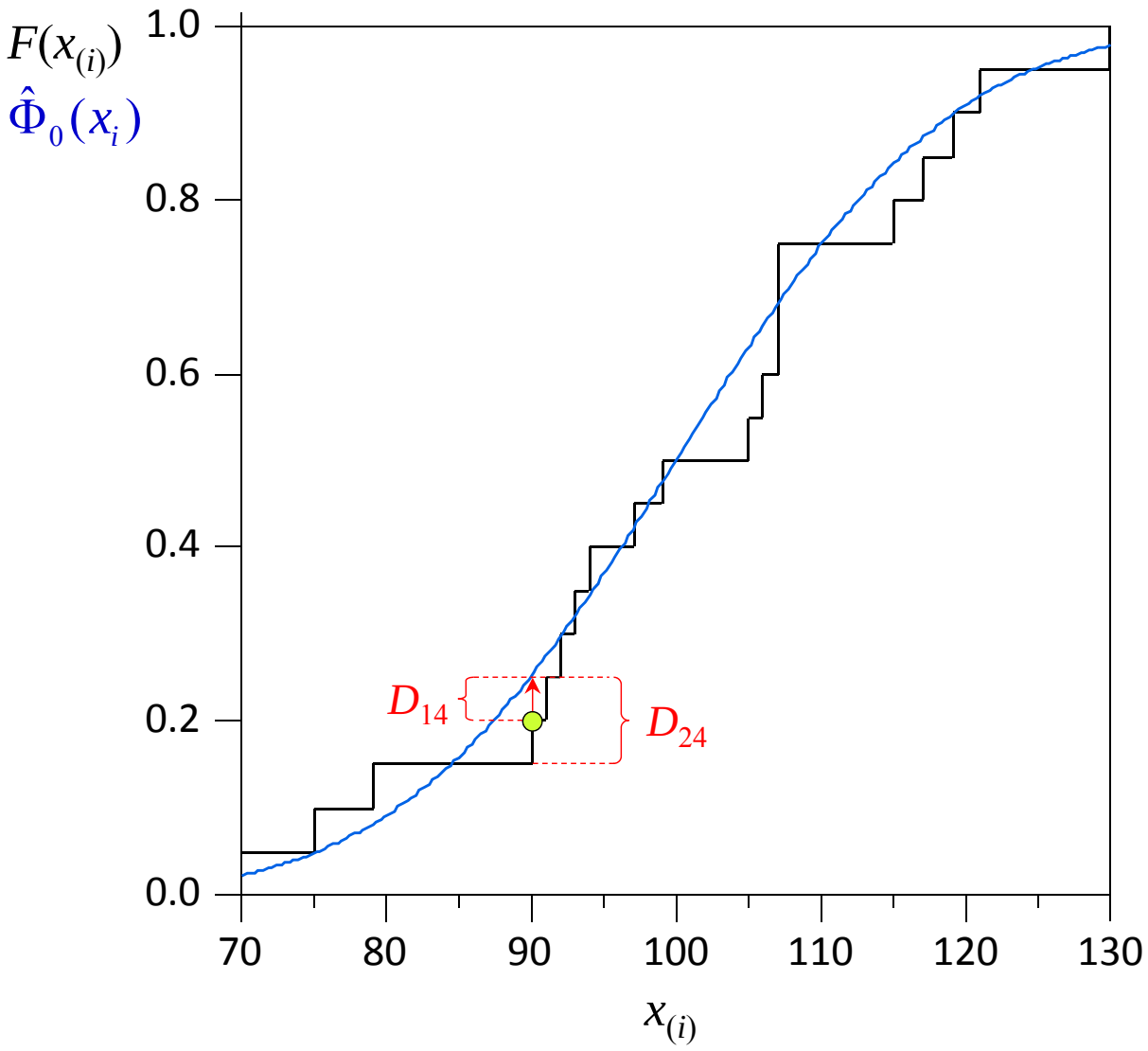
$$D_{2i} = F(x_{(i-1)}) - \hat{\Phi}_0(x_{(i)}); \quad F(x_{(0)}) := 0$$

- Die Prüfgröße D_{max} ergibt sich dann als Maximum aller Abweichungen:

$$D_{\max} = \max_{i=1\ldots n}\left(\left|D_{1i}\right|, \left|D_{2i}\right|\right) = 0.131$$

(i)	$x_{(i)}$	$F(x_{(i)})$	$F(x_{(i-1)})$	$z_{(i)}$	$\hat{\Phi}_0(x_{(i)})$	D_{1i}	D_{2i}
1	70	0.050	0.000	-2.000	0.023	0.027	-0.023
2	75	0.100	0.050	-1.667	0.048	0.052	0.002
3	79	0.150	0.100	-1.400	0.081	0.069	0.019
4	90	0.200	0.150	-0.667	0.252	-0.052	-0.102
5	91	0.250	0.200	-0.600	0.274	-0.024	-0.074
6	92	0.300	0.250	-0.533	0.297	0.003	-0.047
7	93	0.350	0.300	-0.467	0.320	0.030	-0.020
8	94	0.400	0.350	-0.400	0.345	0.055	0.005
9	97	0.450	0.400	-0.200	0.421	0.029	-0.021
10	99	0.500	0.450	-0.067	0.473	0.027	-0.023
11	105	0.550	0.500	0.333	0.631	-0.081	-0.131
12	106	0.600	0.550	0.400	0.655	-0.055	-0.105
13	107	0.750	0.600	0.467	0.680	0.070	-0.080
14	107						
15	107						
16	115	0.800	0.750	1.000	0.841	-0.041	-0.091
17	117	0.850	0.800	1.133	0.871	-0.021	-0.071
18	119	0.900	0.850	1.267	0.897	0.003	-0.047
19	121	0.950	0.900	1.400	0.919	0.031	-0.019
20	130	1.000	0.950	2.000	0.977	0.023	-0.027

Prüfung auf Normalverteilung



Abweichungen der empirischen und der theoretischen Normalverteilung $\mathrm{N}(100,15)$

(i)	$x_{(i)}$	$F(x_{(i)})$	$F(x_{(i-1)})$	$\hat{\Phi}_0(x_{(i)})$	D_{1i}	D_{2i}
1	70	0.050	0.000	0.023	0.027	-0.023
2	75	0.100	0.050	0.048	0.052	0.002
3	79	0.150	0.100	0.081	0.069	0.019
4	90	0.200	0.150	0.252	-0.052	-0.102
5	91	0.250	0.200	0.274	-0.024	-0.074
6	92	0.300	0.250	0.297	0.003	-0.047
7	93	0.350	0.300	0.320	0.030	-0.020
8	94	0.400	0.350	0.345	0.055	0.005
9	97	0.450	0.400	0.421	0.029	-0.021
10	99	0.500	0.450	0.473	0.027	-0.023
11	105	0.550	0.500	0.631	-0.081	-0.131
12	106	0.600	0.550	0.655	-0.055	-0.105
13	107	0.750	0.600	0.680	0.070	-0.080
14	107					
15	107					
16	115	0.800	0.750	0.841	-0.041	-0.091
17	117	0.850	0.800	0.871	-0.021	-0.071
18	119	0.900	0.850	0.897	0.003	-0.047
19	121	0.950	0.900	0.919	0.031	-0.019
20	130	1.000	0.950	0.977	0.023	-0.027

Prüfung auf Normalverteilung

- Entscheidung: Bei der ungerichtet formulierten Hypothese erfolgt eine Zurückweisung der H_0, falls $D_{max} > D_{crit}$. Kritische Werte D_{crit} für den K-S Test lassen sich Tabellen im Anhang von Statistik-Büchern entnehmen (z.B. Tabelle A.6a in Eid, Gollwitzer & Schmitt, 2010) und hängen ab von n und α.

  Im Beispiel ergibt sich für $n = 20$ ein kritischer Wert von $D_{crit} = 0.265$ (bei $\alpha = 0.10$) bzw. $D_{crit} = 0.232$ (bei $\alpha = 0.20$). Da $D_{max} = 0.131 < D_{crit}$ wird die Nullhypothese beibehalten: Die Werte weichen nicht statistisch signifikant von der Normalverteilung ab.

- Sind, wie meist, der Mittelwert und die Standardabweichung des Merkmals X in der Population unbekannt, so sollten nicht die entsprechenden Schätzungen aus der Stichprobe herangezogen werden und dann der K-S Test bestimmt werden, da er dann zu konservativ entscheidet (d.h. entscheidet zu häufig zugunsten der Normalverteilung). Stattdessen sollte dann eine im Folgenden dargestellte Korrektur durchgeführt werden.

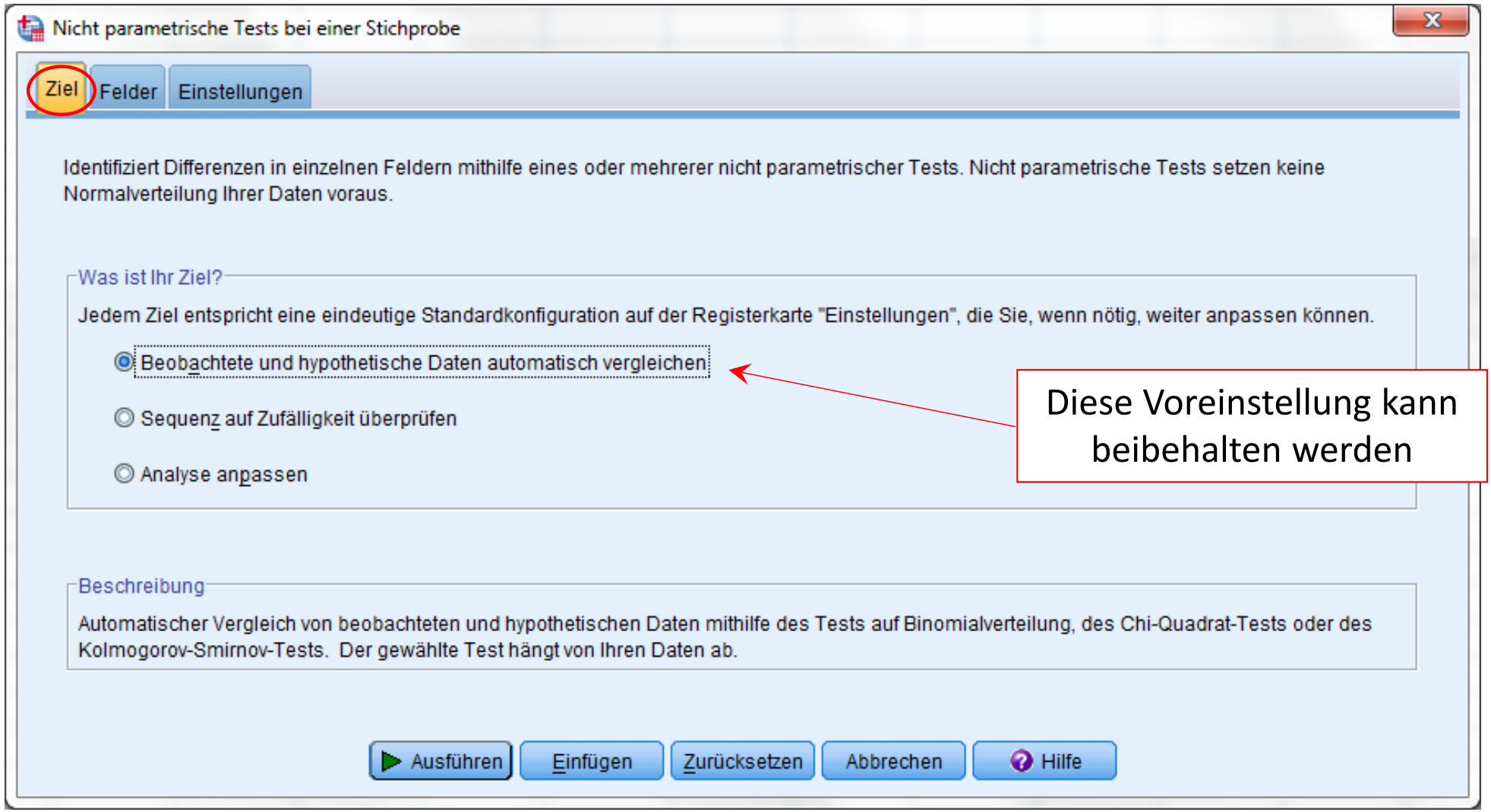
Prüfung auf Normalverteilung

- Bezeichnung: Lilliefors-Korrektur des K-S Tests, Lilliefors-Test.
- Einsatzbereich, Hypothesen: wie beim K-S Test.
- Voraussetzungen: Die Verteilung des Merkmals X ist stetig.
- Vorgehen: Analog zum K-S Test, nur dass jetzt die Parameter μ und σ der Verteilung $\Phi_0(x)$ aus den Daten geschätzt werden und dies bei der Testung adäquat berücksichtigt wird.
- Im Beispiel ist $\hat{\mu} = \overline{x} = 100.70$ und $\hat{\sigma} = s = 15.76$. Entsprechend werden die Abweichungen von der $\mathrm{NV}(100.70,\ 15.76)$ bestimmt. Dabei resultiert $D_{max} = 0.108$.
- Entscheidung: Bei der ungerichtet formulierten Hypothese erfolgt eine Zurückweisung der H_0, falls $D_{max} > D_{crit}$. Kritische Werte D_{crit} für den Lilliefors-Test lassen sich Tabellen im Anhang von Statistik-Büchern entnehmen (z.B. Tabelle A.6b in Eid, Gollwitzer & Schmitt, 2010) und hängen wiederum ab von n und α.

  Im Beispiel ergibt sich für $n = 20$ ein kritischer Wert von $D_{crit} = 0.167$ (bei $\alpha = 0.10$) bzw. $D_{crit} = 0.159$ (bei $\alpha = 0.20$). Da $D_{max} = 0.108 < D_{crit}$ wird auch in diesem Fall die Nullhypothese beibehalten: Die Werte weichen nicht statistisch signifikant von der Normalverteilung ab.

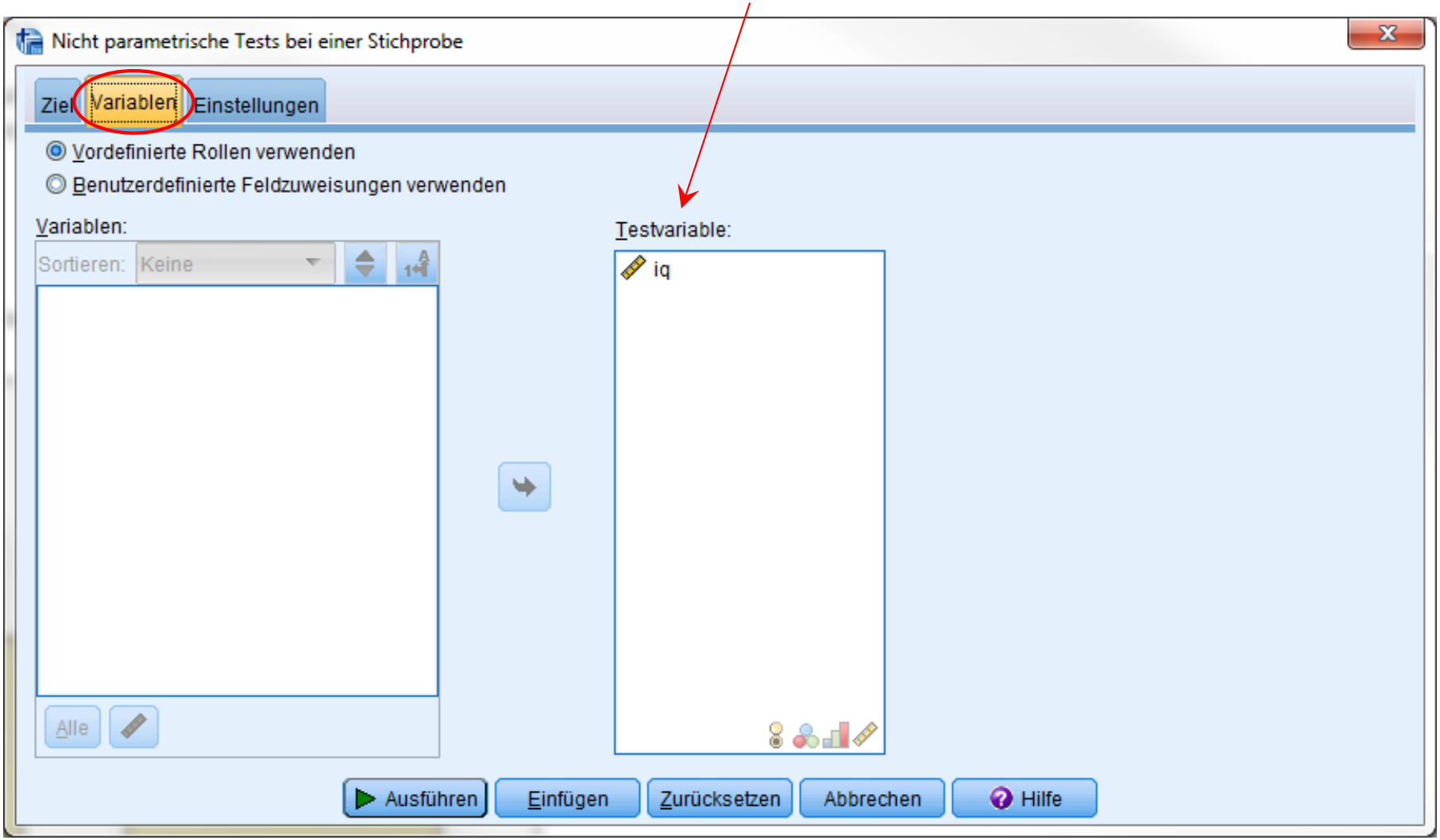
Kolmogorov-Smirnov-Test in SPSS

- Der Kolmogorov-Smirnov Test lässt sich in SPSS anfordern unter `Analysieren/Nicht-parametrische_Tests/Eine_Stichprobe`…



Kolmogorov-Smirnov-Test in SPSS

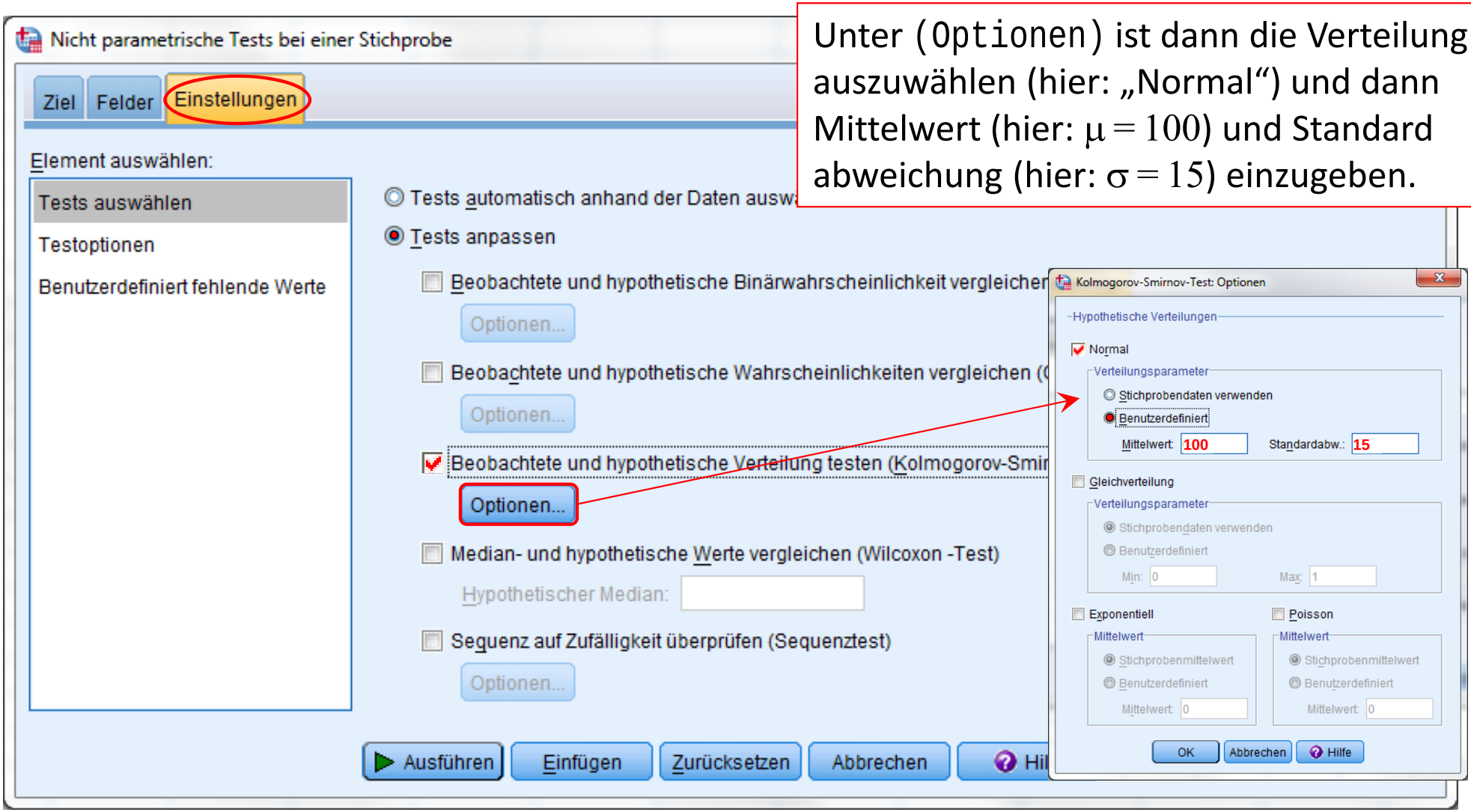
- In dem Reiter ⌈Variablen⌉ ist dann unter „Testfelder“ die Variable anzugeben, die auf Normalverteilung geprüft werden soll (hier: `iq`).



Kolmogorov-Smirnov-Test in SPSS

- Abschließend ist unter ⌈Einstellungen⌉ die Option „Tests anpassen“ und dort „Beobachtete und hypothetische Verteilung testen (Kolmogorov-Smirnov-Test)“ zu aktivieren.

Unter (Optionen) ist dann die Verteilung auszuwählen (hier: „Normal“) und dann Mittelwert (hier: $\mu = 100$) und Standard abweichung (hier: $\sigma = 15$) einzugeben.



Kolmogorov-Smirnov-Test in SPSS

Nichtparametrische Tests

Hypothesentestübersicht

	Nullhypothese	Test	Sig.	Entscheidung
1	Die Verteilung von iq ist normal mit Mittelwert 100 und Standardabweichung 15,000.	Kolmogorov-Smirnov-Test bei einer Stichprobe	885,000	Nullhypothese beibehalten

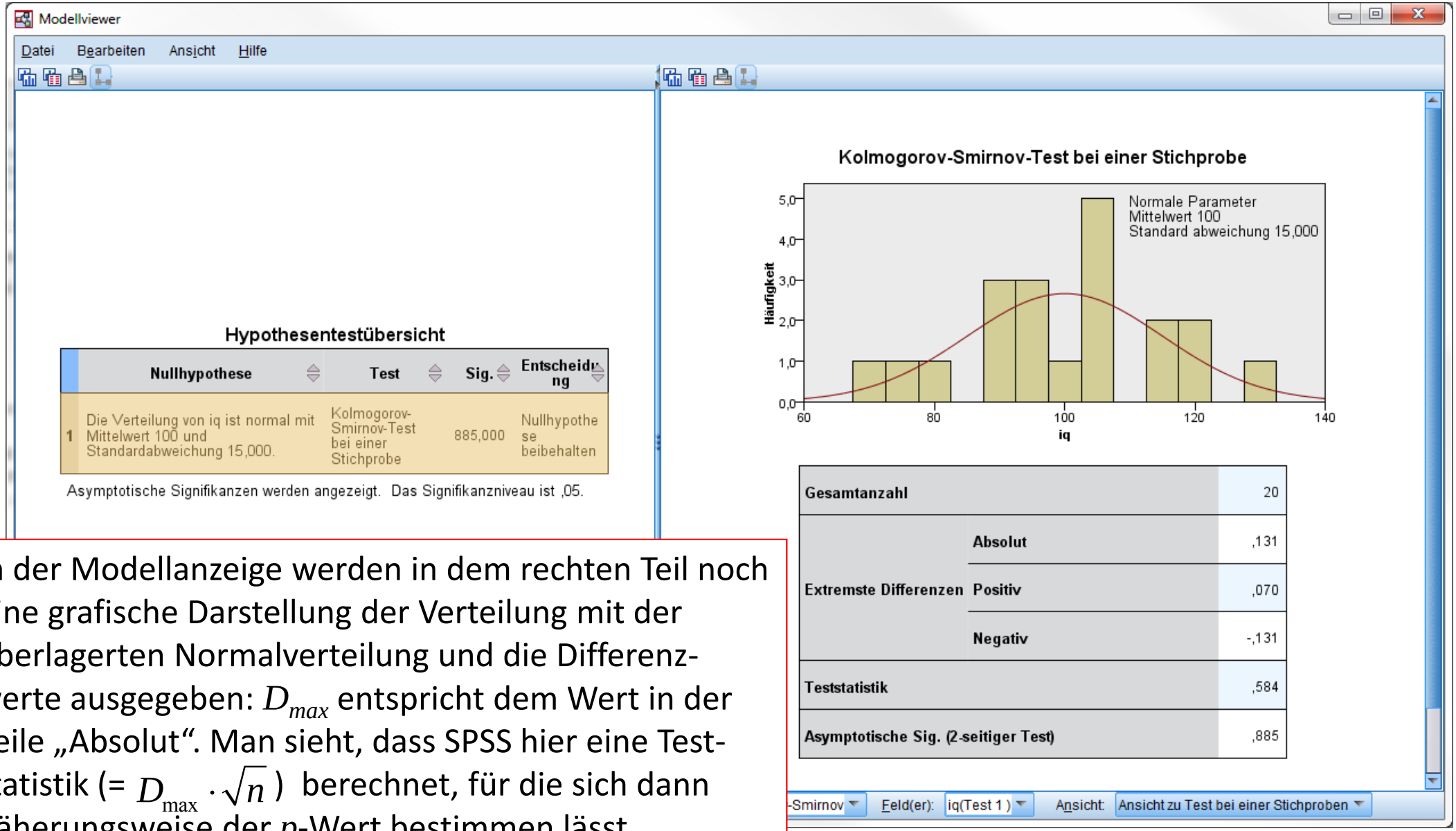
Asymptotische Signifikanzen werden angezeigt. Das Signifikanzniveau ist ,05.

Achtung: Fehler in SPSS Version 24: Statt 0.885 wird in der Spalte Sig der Wert 885 ausgegeben. (In der Modellansicht wird der p-Wert dann korrekt angezeigt.)

Da der ausgegebene p-Wert (Spalte Sig) von 0.885 nicht kleiner als unser gewähltes α von 0.10 ist, wird die Nullhypothese beibehalten.

Durch dass Doppelklicken auf die Tabelle öffnet sich ein auf der folgenden Folie dargestelltes Fenster „Modellansicht“, dem noch weitere Informationen zu entnehmen sind.

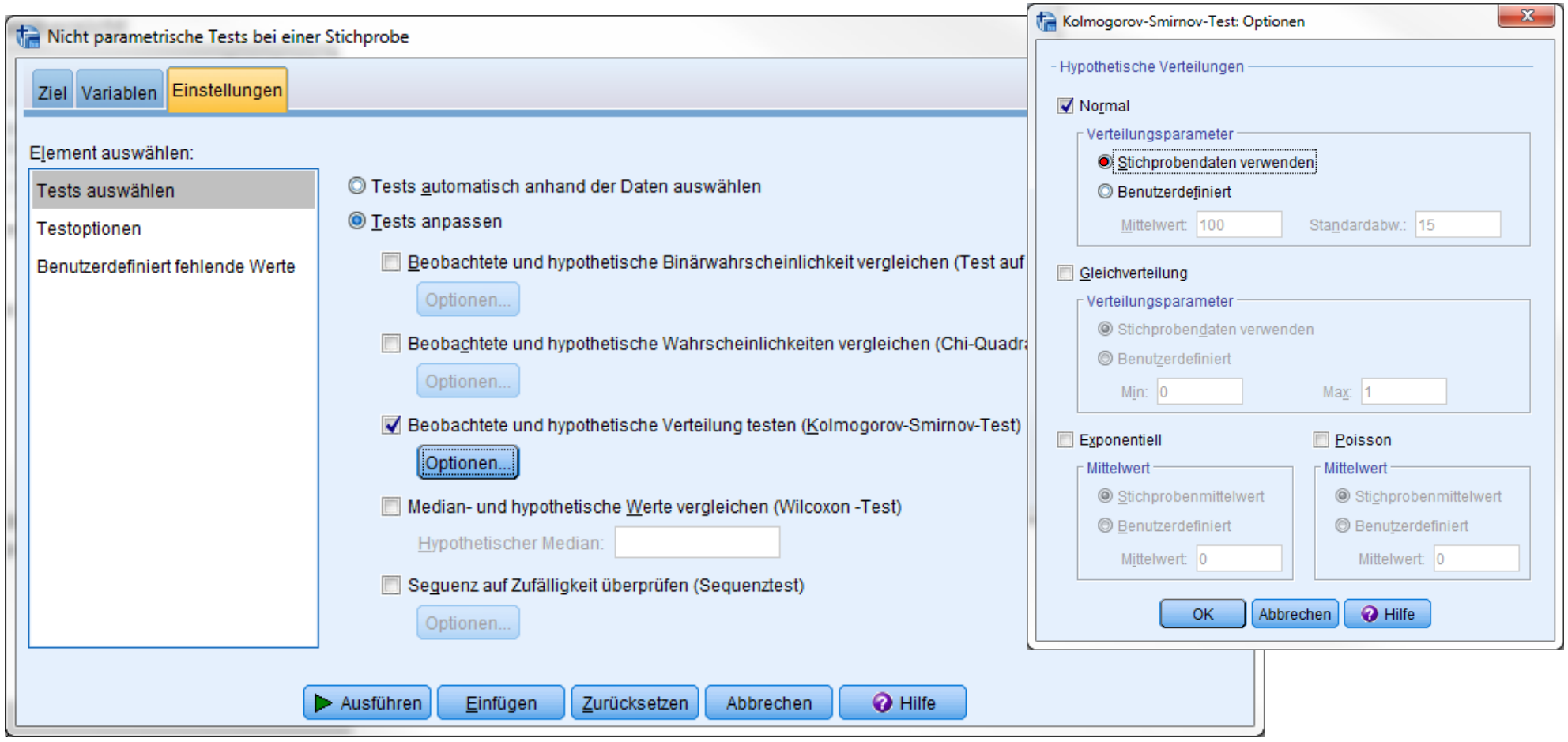
Kolmogorov-Smirnov-Test in SPSS



In der Modellanzeige werden in dem rechten Teil noch eine grafische Darstellung der Verteilung mit der überlagerten Normalverteilung und die Differenzwerte ausgegeben: D_{max} entspricht dem Wert in der Zeile „Absolut“. Man sieht, dass SPSS hier eine Teststatistik (= $D_{\max} \cdot \sqrt{n}$) berechnet, für die sich dann näherungsweise der p-Wert bestimmen lässt.

Lilliefors-Test in SPSS

- Für den Lilliefors-Test muss lediglich abweichend unter (Optionen) unter „Verteilungs-parameter“ die Option „Stichprobendaten verwenden“ aktiviert werden.



Lilliefors-Test in SPSS

Hypothesentestübersicht

	Nullhypothese	Test	Sig.	Entscheidung
1	Die Verteilung von iq ist normal mit Mittelwert 101 und Standardabweichung 15,762.	Kolmogorov-Smirnov-Test bei einer Stichprobe	,200[1,2]	Nullhypothese beibehalten

Asymptotische Signifikanzen werden angezeigt. Das Signifikanzniveau ist ,05.

[1]Lilliefors korrigiert

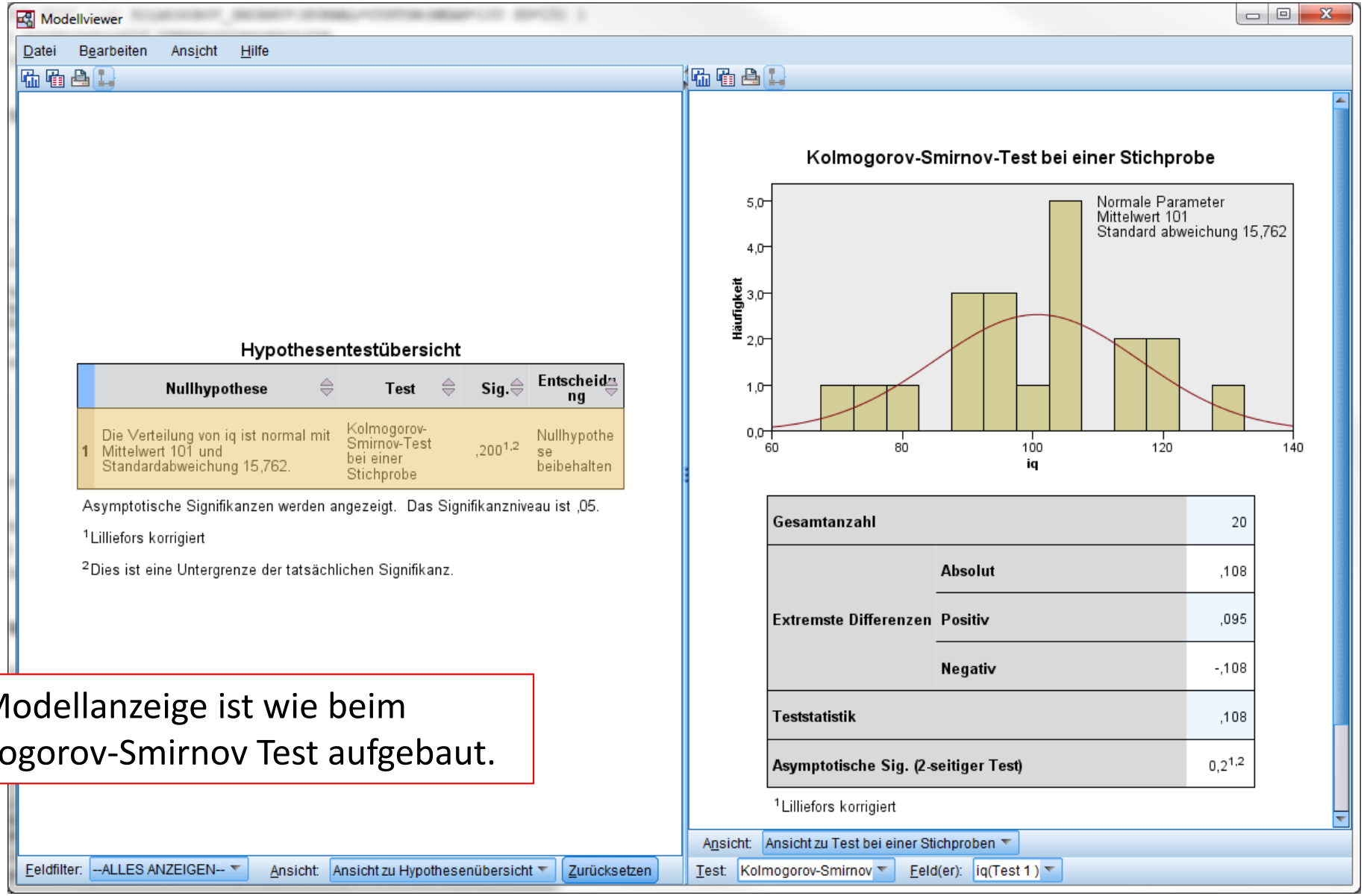
[2]Dies ist eine Untergrenze der tatsächlichen Signifikanz.

Da der ausgegebene p-Wert (Spalte Sig) von 0.200 nicht kleiner als unser gewähltes α von 0.10 ist, wird die Nullhypothese beibehalten.

Durch dass Doppelklicken auf die Tabelle öffnet sich wieder die „Modellansicht“.

Hinweis: Der Lillifors-Test wird in SPSS in dieser Prozedur erst ab Version 24 korrekt berechnet. Alternativ konnte und kann man ihn anfordern unter `Analysieren/Deskriptive_Statistiken/Explorative Datenanalyse`…. Dort muss dann die Variable (hier: `IQ`) unter „Abhängige Variable“ angegeben werden und unter (`Diagramme`) die Option „Normalverteilungsdiagramm mit Tests“ aktiviert werden.

Lilliefors-Test in SPSS



Die Modellanzeige ist wie beim Kolmogorov-Smirnov Test aufgebaut.

Unterschiede in zentraler Tendenz in einer oder zwei Gruppen: Gliederung

1 Ein-Stichproben-Tests

(z-Test, t-Test, weitere Tests)

2 Parametrische Tests für Mittelwertsunterschiede zweier Gruppen

(Unabhängigkeit und Abhängigkeit von Gruppen, t-Test für unabhängige und abhängige Gruppen, Welch-Test)

3 Prüfung der Voraussetzungen

(Robustheit, Varianzhomogenität, Normalverteilung)

4 Nicht-parametrische Tests für Unterschiede zweier Gruppen in der zentralen Tendenz

(Mann-Whitney U-Test, Vorzeichen-Test, Wilcoxon-Test)

Parametrische vs. nonparametrische Verfahren

- Dis bisher betrachteten statistischen Verfahren machen alle Annahmen über die Verteilung der Variablen in der Population – meist die Normalverteilung – und deren Parameter. Statistische Tests, die solche Annahmen beinhalten, bezeichnet man als parametrische Verfahren.
- Demgegenüber benötigen nonparametrische (verteilungsfreie) Testverfahren zur Ableitung der Prüfgröße keine Annahmen über die Verteilungsform der Variablen.
- Wir werden im Folgenden drei Verfahren kennenlernen, die den Unterschied in den zentralen Tendenzen in zwei unabhängigen bzw. abhängigen Gruppen prüfen. Sie können alternativ zu den entsprechenden t-Tests (bzw. dem Welch-Test) herangezogen werden, wenn dort die Voraussetzungen bzgl. der Form der Verteilung nicht erfüllt sind oder auch, wenn die abhängige Variable X nicht intervall- sondern nur ordinalskaliert ist.

- Beispiel (modifiziert aus Sedlmeier & Renkewitz, 2008, S. 582ff): In einer Studie wird geprüft, ob ein bestimmtes neues Medikament als Nebenwirkung eine Verlangsamung der Reaktionszeit zur Folge hat. 19 Personen werden randomisiert zu einer Experimentalbedingung (Medikament) bzw. Kontrollgruppe (Placebo) zugewiesen. Gemessen wird die Reaktionszeit in einer einfachen Reaktionsaufgabe in Millisekunden.

Placebo	Medikament
X_1	X_2
177	178
161	186
163	172
219	185
154	223
159	171
192	184
155	194
158	196
	179

$n_1 = 9$	$n_2 = 10$

- Wie bei Reaktionszeiten zu erwarten weichen die Daten systematisch von der Normalverteilung ab, was man auch z.B. an dem nebenstehenden (back-to-back) stem-and-leaf-Diagramm erkennen kann. Zudem ist die Stichprobe klein.

Placebo		Medikament
	22	3
9	21	
	20	
2	19	46
	18	456
7	17	1289
31	16	
9854	15	

Nonparametrische Verfahren: Mann-Whitney U-Test

- Bezeichnung: Mann-Whitney U-Test, U-Test. (Der Test führt zum gleichen Ergebnis wie der Wilcoxon-Rangsummen-Test, vgl. Eid, Gollwitzer & Schmitt, 2010, Kap. 11.2.)
- Einsatzbereich: Prüfung der Hypothese, dass sich die Verteilungen bzw. die zentralen Tendenzen (Mediane) der Variable X in zwei unabhängigen Gruppen unterscheiden.
- Hypothesen: Die Hypothesen lassen sich unterschiedlich formulieren, im zweiseitigen Fall:
  - Möglichkeit 1: Die beiden Verteilungen sind identisch bzw. unterscheiden sich: $H_0: \Phi(X_1) = \Phi(X_2)$ und $H_1: \Phi(X_1) \neq \Phi(X_2)$
  - Möglichkeit 2: Hier formuliert man die Hypothesen über die Wahrscheinlichkeit, dass ein zufällig aus Gruppe 1 gezogener Messwert kleiner ist als ein zufällig aus Gruppe 2 gezogener Messwert:

    $H_0: \pi(X_1 < X_2) = 0.5$ und $H_1: \pi(X_1 < X_2) \neq 0.5$

    bzw. einseitig z.B. $H_0: \pi(X_1 < X_2) \leq 0.5$ und $H_1: \pi(X_1 < X_2) > 0.5$
  - Möglichkeit 3: Unter bestimmten Voraussetzungen (z.B. bei symmetrischen Verteilungen) kann mit dem Verfahren auch die Hypothese gleicher Mediane in beiden Gruppen getestet werden: $H_0: \eta_1 = \eta_2$ und $H_1: \eta_1 \neq \eta_2$ (Eta = Median in der Population)

Nonparametrische Verfahren: Mann-Whitney U-Test

- Voraussetzungen: Das Merkmal X ist mindestens ordinalskaliert.
- Vorgehen: Man zählt für jeden Wert von Gruppe 1 aus, wie viele Vpn der Gruppe 2 einen höheren Wert in X haben. Im Falle von $x_{11} = 177$ ist das z.B. 8 Mal der Fall. Macht man dies für alle Personen in Gruppe 1 und summiert über alle $n_1 \cdot n_2$ Vergleiche, so erhält man $U_1 = 8 + 10 + 10 + 1 + 10 + 10 + 3 + 10 + 10 = 72$.

  Statt auszuzählen kann man U_1 auch berechnen. Dazu werden alle Messwerte aus beiden Gruppen aufsteigend in eine gemeinsame Rangordnung gebracht (also von Rangplatz 1 bis $n_1 + n_2 := n$, hier $9 + 10 = 19$). Statt auszuzählen berechnet man nun die Summe der Ränge

$$rs_1 = \sum_{i=1}^{n_1} \mathrm{Rg}(x_{i1}) = 9 + 5 + 6 + \ldots + 3 = 63$$

  und erhält dann U_1 durch

$$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - rs_1 = 9 \cdot 10 + \frac{9 \cdot (9+1)}{2} - 63 = 72$$

X_1	$\mathrm{Rg}(X_1)$
177	9
161	5
163	6
219	18
154	1
159	4
192	15
155	2
158	3

X_2	$\mathrm{Rg}(X_2)$
178	10
186	14
172	8
185	13
223	19
171	7
184	12
194	16
196	17
179	11

$n_1 = 9$	$rs_1 = 63$	$n_1 = 10$	$rs_2 = 127$
	$rm_1 = 7.0$		$rm_2 = 12.7$

Mittlere Ränge $rm_j = rs_j / n_j$

Nonparametrische Verfahren: Mann-Whitney U-Test

- Führt man dies analog in die andere Richtung aus (bestimmt also, wie häufig jeder Wert in Gruppe 1 größer ist als ein Messwert in Gruppe 2), dann erhält man

$$rs_2 = \sum_{i=1}^{n_2} \mathrm{Rg}(x_{i2}) \quad \text{und} \quad U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2+1)}{2} - rs_2 = n_1 \cdot n_2 - U_1 = 9 \cdot 10 - 72 = 18$$

- Offensichtlich gilt $U_1 + U_2 = n_1 \cdot n_2$ (wenn keine Ties existieren). Je mehr die beiden U-Werte von $n_1 \cdot n_2 / 2$ abweichen, desto stärker unterscheiden sich die Gruppen. [Man kann sowohl $U_1/(n_1 \cdot n_2)$ als auch $1 - U_2/(n_1 \cdot n_2)$ als Schätzungen von $P(X_1 < X_2)$ heranziehen, die umso mehr für Unterschiede beider Gruppen sprechen, je stärker sie von 0.5 abweichen.]

- Die kritischen Werte der exakten Verteilung der Prüfgröße $U = \mathrm{Min}(U_1, U_2)$ lassen sich für kleine Stichprobenumfänge (meist n_1 und $n_2 < 20$) Tabellen in Statistik-Büchern entnehmen.

- Bei größeren Stichproben verteilt sich unter der Annahme, dass beide Verteilungen sich in ihrer Form nicht unterscheiden (sondern nur in ihrer Lage, sich also bei einer Verschiebung perfekt zur Deckung bringen lassen) die Prüfgröße U annähernd normal:

$$z = \frac{U - \mu_U}{\sigma_U} \quad \text{mit} \quad \mu_U = \frac{n_1 \cdot n_2}{2} \quad \text{und dem Standardfehler} \quad \sigma_U = \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}$$

Nonparametrische Verfahren: Mann-Whitney U-Test

- Entscheidung: Bei kleinem n_1 und n_2 lassen sich die kritischen Werte U_{crit} entsprechenden Tabellen in Abhängigkeit von α, ein- bzw. zweiseitiger Fragestellung sowie n_1 und n_2 entnehmen. Die H_0 wird zurückgewiesen, wenn $U < U_{crit}$!

Im Beispiel lässt sich für $\alpha = 0.05$, zweiseitiger Fragestellung und $n_1 = 9$ und $n_2 = 10$ (oder umgekehrt) den Tabellen ein kritischer Wert von $U_{crit} = 20$ entnehmen. Da $U = \text{Min}(U_1, U_2) = \text{Min}(72, 18) = 18 < U_{crit}$ wird in diesem Fall die Nullhypothese zurückgewiesen: Die Reaktionszeiten sind unter der Medikation statistisch signifikant länger.

Kritische U-Werte für Mann-Whitney U-Test
($\alpha = 0.05$, zweiseitige Testung)

	n_2																		
n_1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2							0	0	0	0	1	1	1	1	1	2	2	2	2
3				0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4			0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
5		0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6		1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7		1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	69	74	78	83
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	2	6	11	17	22	28	34	39	45	51	57	63	69	75	81	87	93	99	105
18	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

Nonparametrische Verfahren: Mann-Whitney U-Test

- Bei größerem n_1 und n_2 sind die exakten Werte nicht mehr tabelliert und man verwendet die Normalverteilungsapproximation. Zur Illustration ergäbe sich für unsere Daten:

$$z = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}} = \frac{18 - \frac{9 \cdot 10}{2}}{\sqrt{\frac{9 \cdot 10 \cdot (9 + 10 + 1)}{12}}} = \frac{-27}{12.247} = -2.205$$

- Die Zurückweisung der H_0 erfolgt, falls $|z| > z_{crit}$ mit $z_{crit} = z_{1-\alpha/2}$ bei zweiseitiger Prüfung, also $z_{crit} = z_{0.975} = 1.96$. Dies ist hier der Fall; die Entscheidung ist also identisch mit der obigen.

- Bei einseitiger Testung muss zunächst (z.B. an den Rangplatzmittelwerten) geprüft werden, ob der Effekt in die postulierte Richtung liegt. Falls nicht, wird wie immer die H_0 beibehalten. Andernfalls wird U berechnet und in den Tabellen der kritische U-Wert für gegebenes α und einseitiges Testen (bzw. $2 \cdot \alpha$ und zweiseitigem Testen) aufgesucht bzw. die z-Prüfgröße mit der kritischen $z_{crit} = z_{1-\alpha}$ verglichen.

Nonparametrische Verfahren: Mann-Whitney U-Test

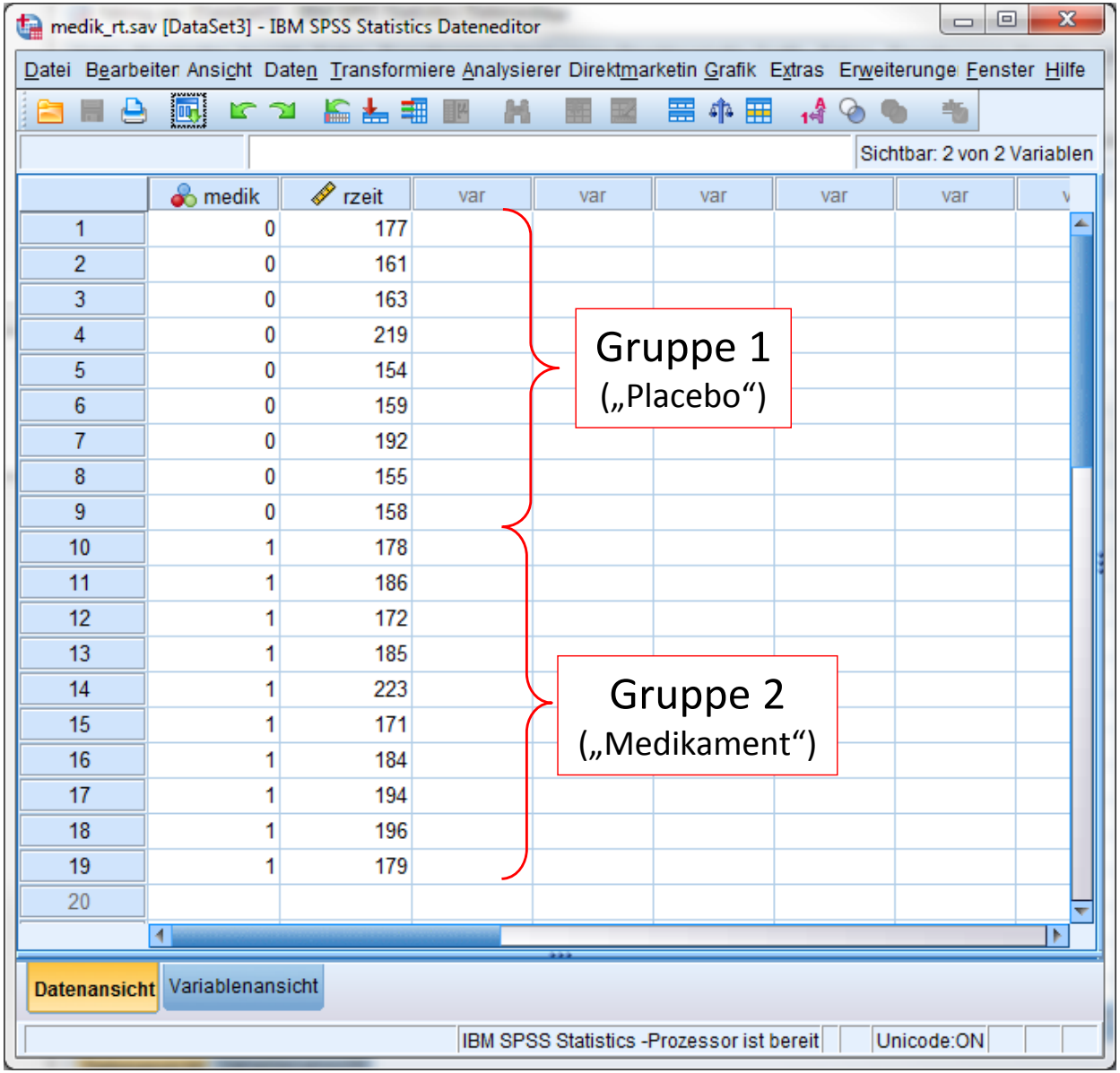
- Verbleibendes Problem: Bei gleichen Messwerten (=Ties) in den Daten können (wie bei der Bestimmung der Rangkorrelation nach Spearman gezeigt) mittlere Rangplätze vergeben werden. Unter diesen Bedingungen führt die Anwendung des U-Tests aber zu fehlerhaften Entscheidungen, die zu konservativ ausfallen.
- In diesem Fall liegen verschiedene Optionen für Korrekturformel vor (vgl. Wilcox, 2005). SPSS etwa verwendet den folgenden korrigierten Standardfehler

$$\sigma_{U_T} = \sqrt{\frac{n_1 \cdot n_2}{n \cdot (n-1)} \cdot \left(\frac{n^3 - n}{12} - \sum_{i=1}^{k} \frac{t_i^3 - t_i}{12} \right)}$$

mit t_i = Zahl der Personen, die sich Rangplatz i teilen und k = Zahl der Tie-Blöcke. (Für ein Beispiel siehe Bortz & Schuster, 2010, S. 132ff.)

Nonparametrische Tests für unabhängige Gruppen in SPSS

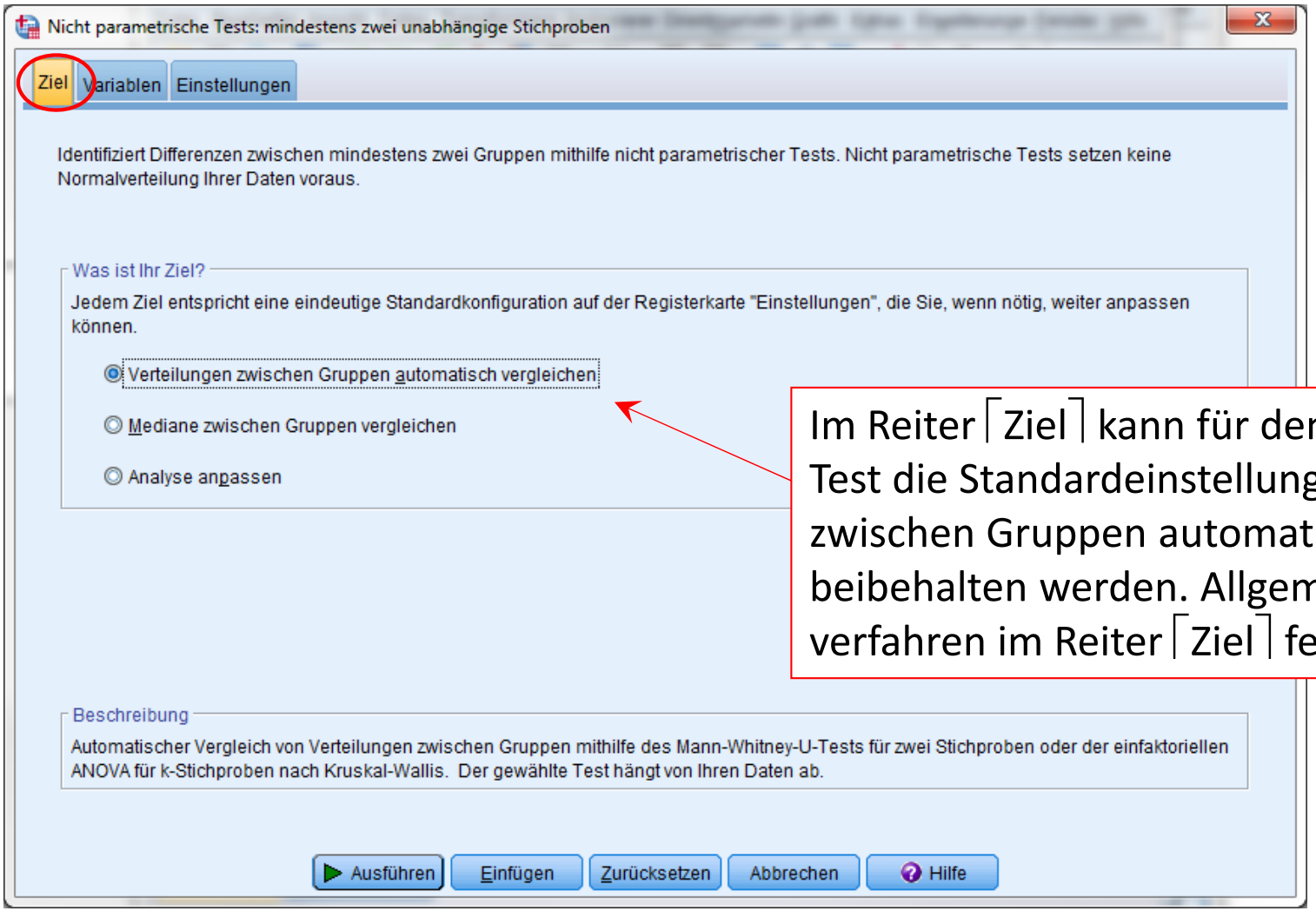
- Im Falle unabhängiger Gruppen sind (wie schon beim t-Test für unabhängige Gruppen) die Daten in SPSS so einzugeben, dass die Gruppenzugehörigkeit in einer Variable codiert wird (hier: MEDIK) und die abhängige Variable X in einer zweiten Variablen (hier: RZEIT).



	medik	rzeit
1	0	177
2	0	161
3	0	163
4	0	219
5	0	154
6	0	159
7	0	192
8	0	155
9	0	158
10	1	178
11	1	186
12	1	172
13	1	185
14	1	223
15	1	171
16	1	184
17	1	194
18	1	196
19	1	179
20		

Mann-Whitney U-Test in SPSS

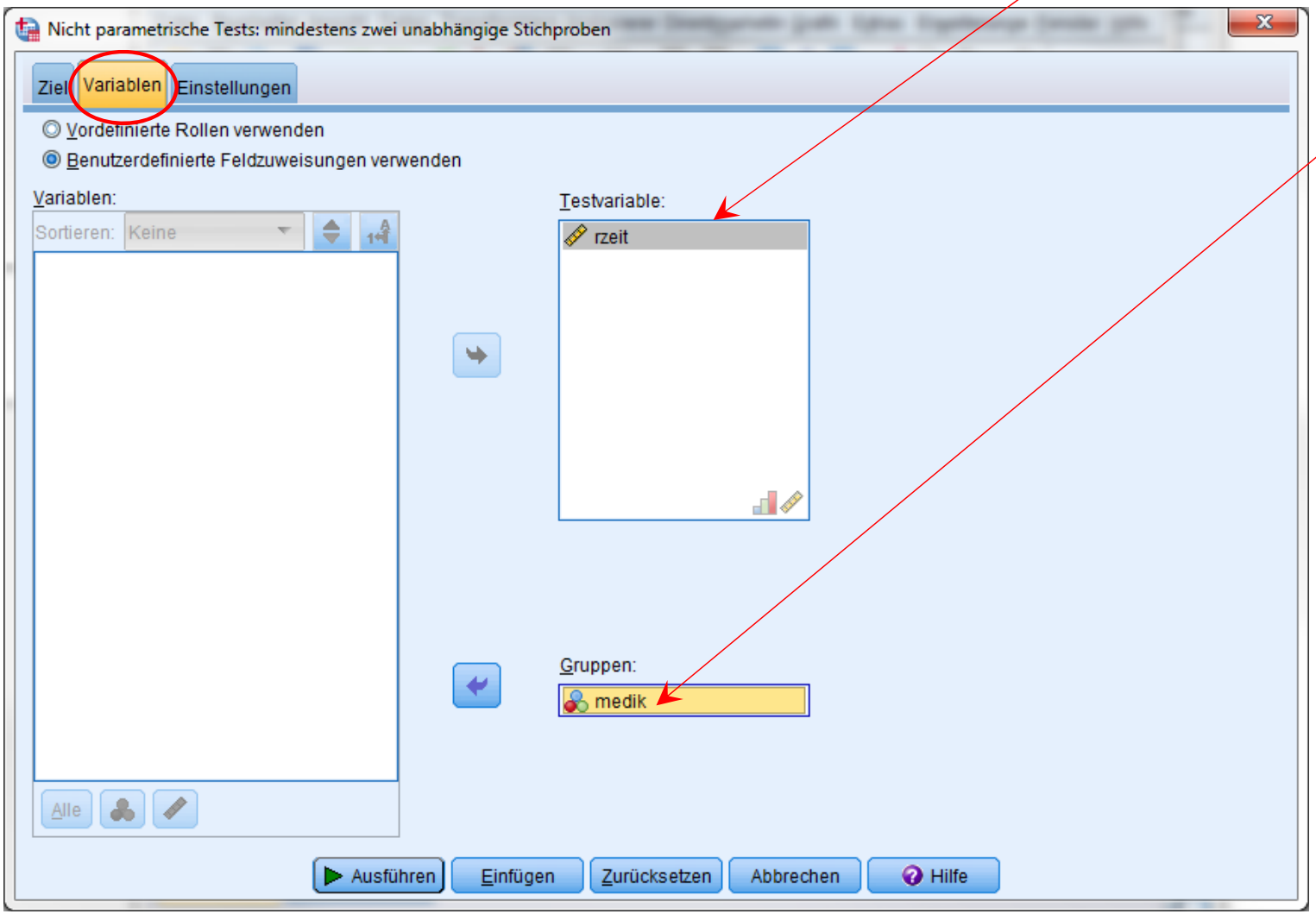
- Der Mann-Whitney U-Test lässt sich in SPSS anfordern unter `Analysieren/Nicht-parametrische_Tests/Unabhängige_Stichproben…`



Im Reiter ⌈Ziel⌉ kann für den Mann-Whitney U-Test die Standardeinstellung „Verteilungen zwischen Gruppen automatisch vergleichen“ beibehalten werden. Allgemein kann das Test-verfahren im Reiter ⌈Ziel⌉ festgelegt werden.

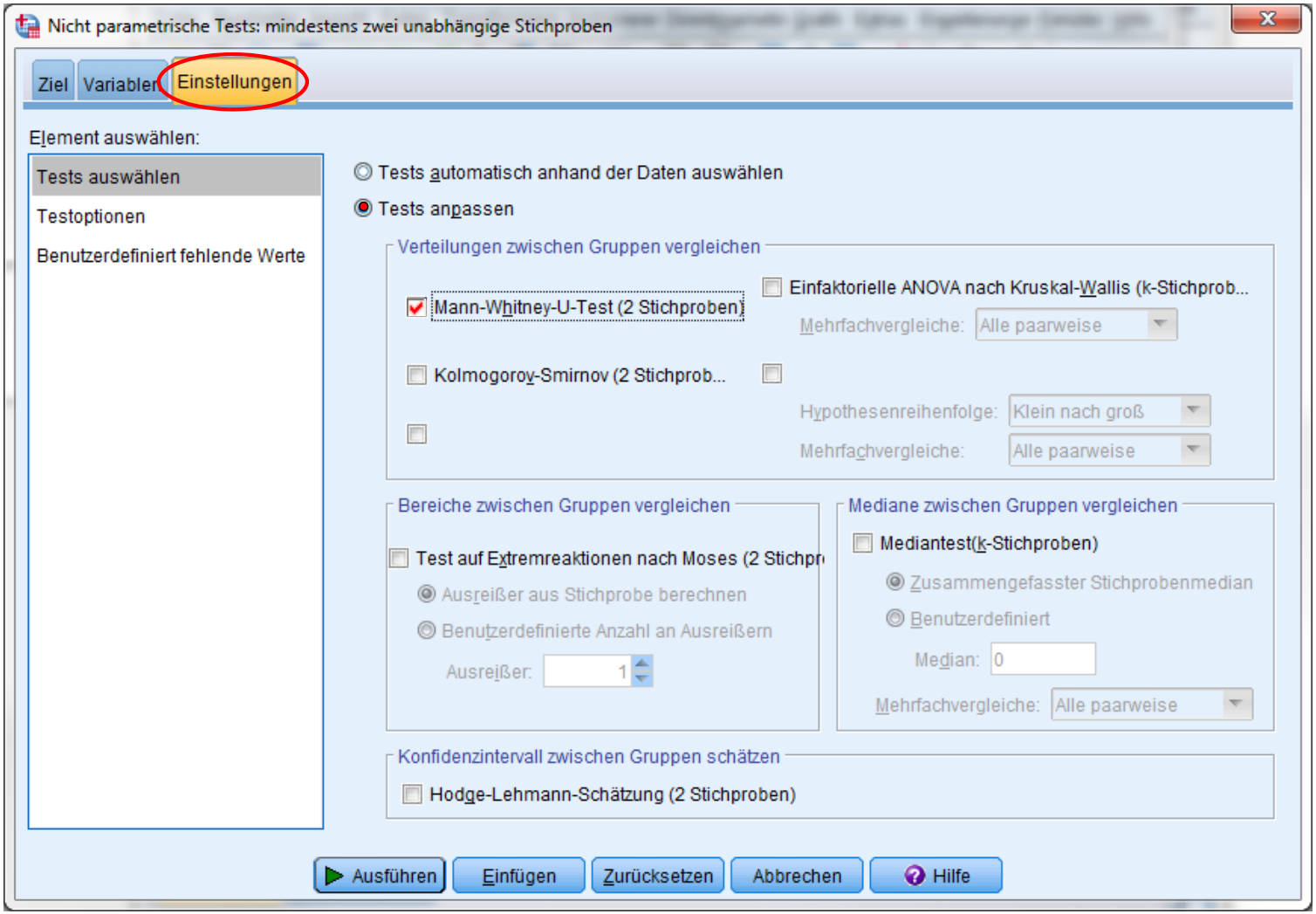
Mann-Whitney U-Test in SPSS

- Im Reiter ⌈Variablen⌉ muss unter „Testvariable“ die AV (hier: RZEIT) und unter „Gruppen“ die Variable angegeben werden, die die Gruppenzugehörigkeit enthält (UV; hier: MEDIK).



Mann-Whitney U-Test in SPSS

- Im Reiter ⌈Einstellungen⌉ könnte man die Option „Tests anpassen“ anwählen und dort auch explizit den „Mann-Whitney-U-Test (2 Stichproben)“ aktivieren. (Nicht erforderlich, da dies die Standardeinstellung ist.)



Mann-Whitney U-Test in SPSS

Nichtparametrische Tests

Hypothesentestübersicht

	Nullhypothese	Test	Sig.	Entscheidung
1	Die Verteilung von rzeit ist über die Kategorien von medik identisch.	Mann-Whitney-U-Test bei unabhängigen Stichproben	28,000[1]	Nullhypothese ablehnen

Asymptotische Signifikanzen werden angezeigt. Das Signifikanzniveau ist , 05.

[1]Für diesen Test wird die exakte Signifikanz angezeigt.

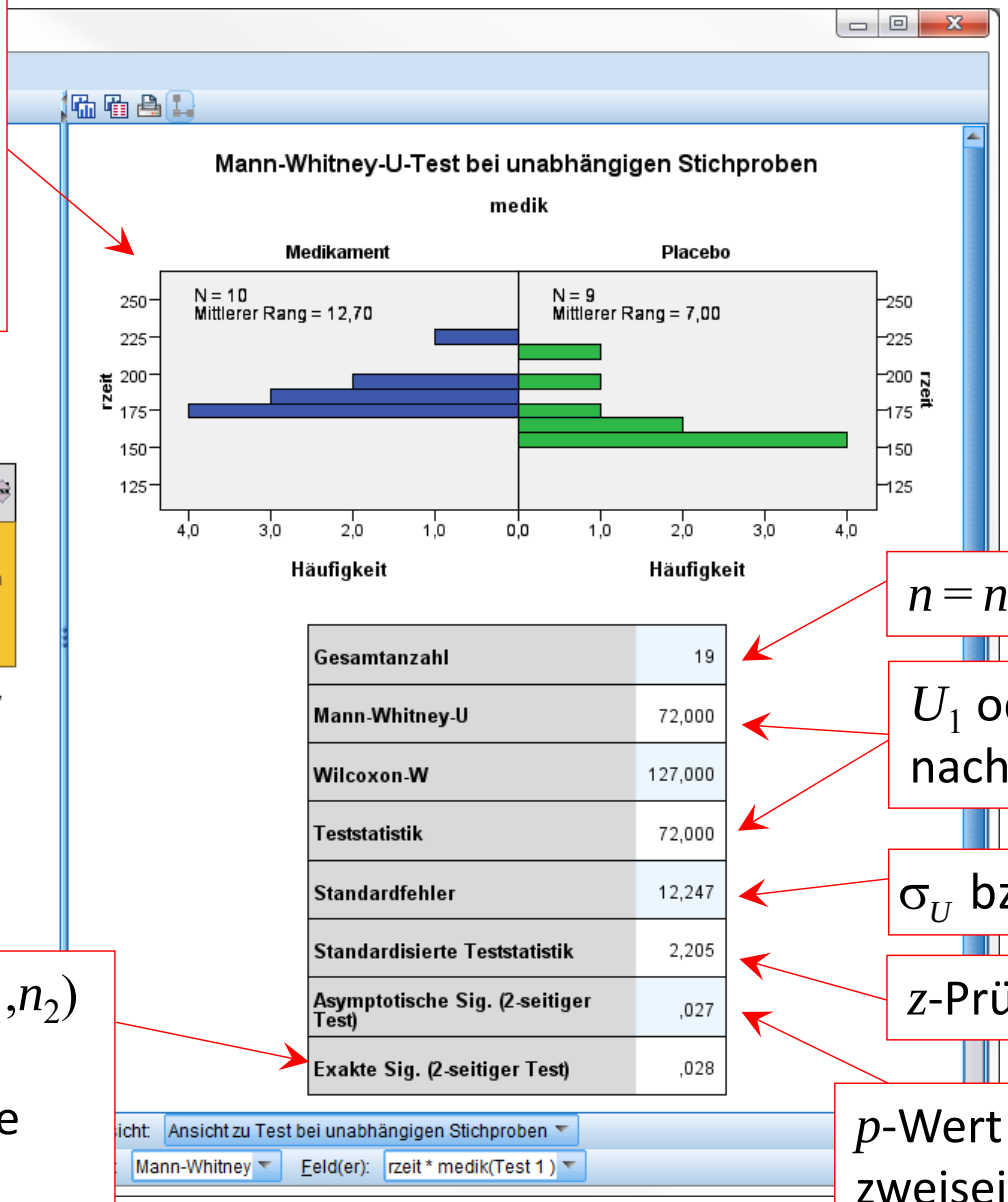
In der Spalte Sig taucht in Version 24 derselbe Fehler wie oben bereits dargestellt auf: Es muss 0.028 statt 28 heißen.

Da der ausgegebene p-Wert (Spalte Sig) von 0.028 kleiner als unser gewähltes α von 0.05 ist, wird die Nullhypothese zurückgewiesen.

Durch dass Doppelklicken auf die Tabelle öffnet sich ein auf der folgenden Folie dargestelltes Fenster „Modellansicht“, dem noch weitere Informationen zu entnehmen sind.

Mann-Whitney U-Test in SPSS

In der Modellanzeige wird rechts eine grafische Darstellung der Verteilungen des Merkmals in beiden Gruppen sowie n_1 und n_2 und die mittleren Ränge rm_1 sowie rm_2 ausgegeben.



Hypothesentestübersicht

	Nullhypothese	Test	Sig.	Entscheidung
1	Die Verteilung von rzeit ist über die Kategorien von medik identisch.	Mann-Whitney-U-Test bei unabhängigen Stichproben	28,000[1]	Nullhypothese ablehnen

Asymptotische Signifikanzen werden angezeigt. Das Signifikanzniveau ist ,05.

[1]Für diesen Test wird die exakte Signifikanz angezeigt.

Gesamtanzahl	19
Mann-Whitney-U	72,000
Wilcoxon-W	127,000
Teststatistik	72,000
Standardfehler	12,247
Standardisierte Teststatistik	2,205
Asymptotische Sig. (2-seitiger Test)	,027
Exakte Sig. (2-seitiger Test)	,028

$n = n_1 + n_2$

U_1 oder U_2 (je nach Codierung)

σ_U bzw. σ_{U_T} (bei Ties)

z-Prüfgröße

p-Wert für z-Wert bei zweiseitiger Testung

Falls $n_1 \cdot n_2 \leq 400$ und $n_1 \cdot n_2 / 2 + \min(n_1, n_2) \leq 200$ wird hier der exakte p-Wert für zweiseitige Testung ausgegeben (ohne Tie-Korrektur).

- Beispiel (fiktiv aus Bortz & Schuster, 2010, S. 133f): Es soll die Wirksamkeit einer Schulung zur Unfallverhütung überprüft werden. In zehn Betrieben werden dazu alle Mitarbeiter geschult und die monatliche Zahl der Unfälle vor und nach der Schulung erfasst.

- Hier kann nicht davon ausgegangen werden, dass die Unfallzahlen normalverteilt sind; zudem ist die Stichprobe klein. Daher entscheiden wir uns für ein nichtparametrisches Verfahren.

vorher	nachher
X_1	X_2
8	4
23	16
7	6
11	12
5	6
9	7
12	10
6	10
18	13
9	6

$n = 10$

Nonparametrische Verfahren: Vorzeichen-Test

- Bezeichnung: Vorzeichen-Test (sign test).
- Einsatzbereich: Prüfung der Hypothese, dass sich die Verteilungen bzw. die zentralen Tendenzen (Mediane) der Variable X in zwei abhängigen Gruppen unterscheiden.
- Hypothesen: Können wieder analog wie beim Mann-Whitney U-Test unterschiedlich formuliert werden, z.B. zweiseitig
  - $H_0: \Phi(X_1) = \Phi(X_2)$ und $H_1: \Phi(X_1) \neq \Phi(X_2)$
  - $H_0: \pi(X_1 < X_2) = 0.5$ und $H_1: \pi(X_1 < X_2) \neq 0.5$ bzw. entsprechend einseitig
  - $H_0: \eta_D = 0$ und $H_1: \eta_D \neq 0$ (η_D = Median der Differenzen der Messwerte) bzw. entsprechend einseitig. Im Falle symmetrischer Verteilungen kann mit dem Verfahren auch die Hypothese gleicher Mediane in beiden Gruppen zweiseitig wie folgt getestet werden: $H_0: \eta_1 = \eta_2$ und $H_1: \eta_1 \neq \eta_2$ bzw. entsprechend einseitig.
- Voraussetzungen: Das Merkmal X ist mindestens ordinalskaliert.
- Vorgehen: Wir zählen aus, wie häufig ein Messwert in Gruppe 2 größer als der korrespondierende Wert in Gruppe 1 ist, bzw. wie häufig die Differenz der Messwerte $d_i = x_{i1} - x_{i2}$ negativ ist.

Nonparametrische Verfahren: Vorzeichen-Test

- Im Beispiel sind dies $V^+ = 7$ positive (und $V^- = 3$ negative) Differenzen d_i, wobei $V^+ + V^- = n$ (wenn keine Ties existieren). Je mehr die beiden V-Werte von $n/2$ abweichen, desto stärker unterscheiden sich die Gruppen.
- Die Zahl der negativen (und positiven) Differenzen folgt einer Binomialverteilung mit n Wiederholungen und einer Wahrscheinlichkeit von $p = 0.5$. Deren Verteilungsfunktion liefert uns dann die exakte Wahrscheinlichkeit, dass $\mathrm{Min}(V^+, V^-)$ mal oder weniger die minimale (positive oder negative) Differenz auftritt.
- Bei größerem n (meist über 20 bis 40) sind die Wahrscheinlichkeiten nicht mehr tabelliert und man kann wieder die Approximation durch die Normalverteilung heranziehen und einsetzen in:

$$z = \frac{V^- - \mu_V}{\sigma_V} = -\frac{V^+ - \mu_V}{\sigma_V} \quad \text{mit} \quad \mu_V = \frac{n}{2} \quad \text{und dem Standardfehler} \quad \sigma_V = \sqrt{\frac{n}{4}}$$

$n = 10$

vorher x_{i1}	nachher x_{i2}	$d_i = x_{i1} - x_{i2}$
8	4	4
23	16	7
7	6	1
11	12	-1
5	6	-1
9	7	2
12	10	2
6	10	-4
18	13	5
9	6	3

Nonparametrische Verfahren: Vorzeichen-Test

$n = 10$

vorher	nachher	
x_{i1}	x_{i2}	d_i
8	4	4
23	16	7
7	6	1
11	12	-1
5	6	-1
9	7	2
12	10	2
6	10	-4
18	13	5
9	6	3

- Entscheidung: Im Beispiel entnehmen wir einer tabellierten Binomialverteilung für $n = 10$, $p = 0.5$ sowie für $\text{Min}(V^+, V^-) = \text{Min}(7, 3) = 3$, dass die Summe der Auftretenswahrscheinlichkeiten von $3, 2, 1$ und 0 mal des Ereignisses „negative Differenz“ $0.1172 + 0.0439 + 0.0098 + 0.0010 = 0.1719$ beträgt. (Der gleiche Wert resultiert für $7, 8, 9$ und 10 mal die positive Differenz.)

  Bei $\alpha = 0.05$ und zweiseitiger Fragestellung müsste dieser Wert kleiner als $.025$ sein, was nicht der Fall ist. Wir behalten also die Nullhypothese bei und schließen, dass die Intervention nicht zu einer statistisch signifikanten Veränderung der Unfallhäufigkeit geführt hat.

$$P(X = k = 3) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \binom{10}{3} \cdot 0.5^3 \cdot (1-0.5)^{10-3} = 0.1172$$

- Die Normalverteilungsapproximation ergibt

$$z = \frac{V^- - n/2}{\sqrt{n/4}} = \frac{3 - 10/2}{\sqrt{10/4}} = -1.26$$

Die Zurückweisung der H_0 erfolgt, falls $|z| > z_{crit}$ mit $z_{crit} = z_{1-\alpha/2}$ bei zweiseitiger Prüfung, also $z_{crit} = z_{0.975} = 1.96 > |-1.26| = |z|$. Dies ist nicht der Fall; die Entscheidung erfolgt wie oben.

- Manchmal (auch von SPSS) wird bei der Approximation der diskreten Binomialverteilung durch die stetige Normalverteilung noch eine sog. Stetigkeitskorrektur (Kontinuitäts-korrektur) vorgenommen (vgl. Bortz, Lienert & Boehnke, 1990, S. 90f), indem im Betrag des Zählers von z noch 0.5 subtrahiert wird. In unserem Beispiel:

$$z' = \frac{\left|V^- - n/2\right| - 0.5}{\sqrt{n/4}} = \frac{\left|3 - 10/2\right| - 0.5}{\sqrt{10/4}} = -0.949$$

Dies führt zu einer konservativeren Entscheidung (und wirkt sich vor allem bei kleinem n, wie hier, stärker aus).

- Liegen Ties in den Daten vor (also Differenzen von $d_i = 0$), so spricht dies zunächst prinzipiell für die H_0. Es gibt verschiedene Methoden damit umzugehen (vgl. Bortz, Lienert & Boehnke, 1990, S. 256f). Das einfachste - auch von SPSS verwendete - Vorgehen besteht darin, alle Null-Differenzen zu ignorieren und mit dem reduzierten $n' := V^+ + V^- < n$ wie oben dargestellt vorzugehen.

Nonparametrische Verfahren: Wilcoxon-Test

- Bezeichnung: Wilcoxon Vorzeichen Rangtest (signed rank test), Wilcoxon-Test.
- Einsatzbereich, Hypothesen: wie beim Vorzeichentest.
- Voraussetzungen: Das Merkmal X ist mindestens ordinalskaliert. Zusätzlich müssen die Differenzen der Messwerte ordinalskaliert sein. (Dies ist etwa der Fall, wenn eine Gleichabständigkeit der Messwerte angenommen werden kann.)
- Vorgehen: Es wird wieder von den Differenzen der Messwerte $d_i = x_{i1} - x_{i2}$ ausgegangen. Diesmal zählen wir nicht nur aus, wie viele dieser Differenzen positiv bzw. negativ sind sondern bringen die Absolutbeträge der Differenzen zunächst in eine Rangreihe (ggf. gemittelt bei Ties, d.h. Differenzen mit gleichem Absolutbetrag) und summieren die Ränge dann getrennt für positive und negative Differenzen. Im Beispiel:

$$W^+ = \sum_{d_i>0} \mathrm{Rg}(|d_i|) = 7.5+10+2+4.5+4.5+9+6 = 43.5$$

$$W^- = \sum_{d_i<0} \mathrm{Rg}(|d_i|) = 2+2+7.5 = 11.5$$

x_{i1}	x_{i2}	d_i	$\mathrm{Rg}(\lvert d_i \rvert)$
8	4	4	7.5
23	16	7	10
7	6	1	2
11	12	-1	2
5	6	-1	2
9	7	2	4.5
12	10	2	4.5
6	10	-4	7.5
18	13	5	9
9	6	3	6

Nonparametrische Verfahren: Wilcoxon-Test

- Es gilt: $W^+ + W^- = n \cdot (n+1)/2$ (bei fehlenden Ties, also keinen d_i Werten $= 0$). Je mehr die beiden W-Werte von $n \cdot (n+1)/4$ abweichen, desto stärker unterscheiden sich die Gruppen.
- Die kritischen Werte der exakten Verteilung der Prüfgröße $W = \mathrm{Min}(W^+, W^-)$ lassen sich für kleine Stichprobenumfänge (meist $n \leq 25$) Tabellen in Statistik-Büchern entnehmen.
- Bei größeren Stichproben verteilt sich die Prüfgröße W approximativ normal:

$$z = \frac{W - \mu_W}{\sigma_W} \quad \text{mit} \quad \mu_W = \frac{n \cdot (n+1)}{4} \quad \text{und dem Standardfehler} \quad \sigma_W = \sqrt{\frac{n \cdot (n+1) \cdot (2 \cdot n+1)}{24}}$$

- Entscheidung: Bei kleinem n lassen sich die kritischen Werte W_{crit} entsprechenden Tabellen in Abhängigkeit von α, ein- bzw. zweiseitiger Fragestellung sowie n entnehmen. Die H_0 wird zurückgewiesen, wenn $W < W_{crit}$!

  Im Beispiel lässt sich für $\alpha = 0.05$, zweiseitiger Fragestellung und $n = 10$ den Tabellen ein kritischer Wert von $W_{crit} = 8$ entnehmen. Da $W = \mathrm{Min}(W^+, W^-) = \mathrm{Min}(43.5, 11.5) = 11.5 > W_{crit}$ wird in diesem Fall die Nullhypothese beibehalten: Die Intervention hat zu keiner statistisch signifikanten Veränderung der Unfallhäufigkeit geführt.

x_{i1}	x_{i2}	d_i	$\mathrm{Rg}(\lvert d_i \rvert)$
8	4	4	7.5
23	16	7	10
7	6	1	2
11	12	-1	2
5	6	-1	2
9	7	2	4.5
12	10	2	4.5
6	10	-4	7.5
18	13	5	9
9	6	3	6

- Die Normalverteilungsapproximation ergibt:

$$z = \frac{W - \frac{n\cdot(n+1)}{4}}{\sqrt{\frac{n\cdot(n+1)\cdot(2\cdot n+1)}{24}}} = \frac{11.5 - \frac{10\cdot(10+1)}{4}}{\sqrt{\frac{10\cdot(10+1)\cdot(2\cdot 10+1)}{24}}} = -1.631$$

- Liegen Ties in den Rangsummen vor (wie im Beispiel), so kann wiederum entsprechend korrigiert werden:

$$z^* = \frac{W-\mu_W}{\sigma_{W_T}} = \frac{W - \frac{n\cdot(n+1)}{4}}{\sqrt{\frac{n\cdot(n+1)\cdot(2\cdot n+1) - \sum_{i=1}^{k}\frac{t_i^3 - t_i}{2}}{24}}} = -1.673$$

$$\sum_{i=1}^{k}\frac{t_i^3 - t_i}{2} = \frac{3^3-3}{2} + 2\cdot\frac{2^3-2}{2} = 12 + 2\cdot 3 = 18$$

mit t_i = Zahl der Differenzen, die sich Rangplatz i teilen und k = Zahl der Tie-Blöcke.

Die Zurückweisung der H_0 erfolgt, falls $|z| > z_{crit}$ bzw. $|z^*| > z_{crit}$ mit $z_{crit} = z_{1-\alpha/2}$ bei zweiseitiger Prüfung, also hier $z_{crit} = z_{0.975} = 1.96$. Dies ist hier nicht der Fall; die Entscheidung ist also identisch mit der obigen.

Nonparametrische Verfahren: Wilcoxon-Test

- Verbleibendes Problem: Bezüglich der Ties in den Daten (also $d_i = 0$) gilt das beim Vorzeichentest Gesagte. Meist werden die Differenzen ignoriert und, obwohl $W^+ + W^- \leq n \cdot (n+1)/2$, in obige Formeln eingesetzt. Da 0-Differenzen für die H_0 sprechen, führt dieses Vorgehen zu progressiven Entscheidungen.

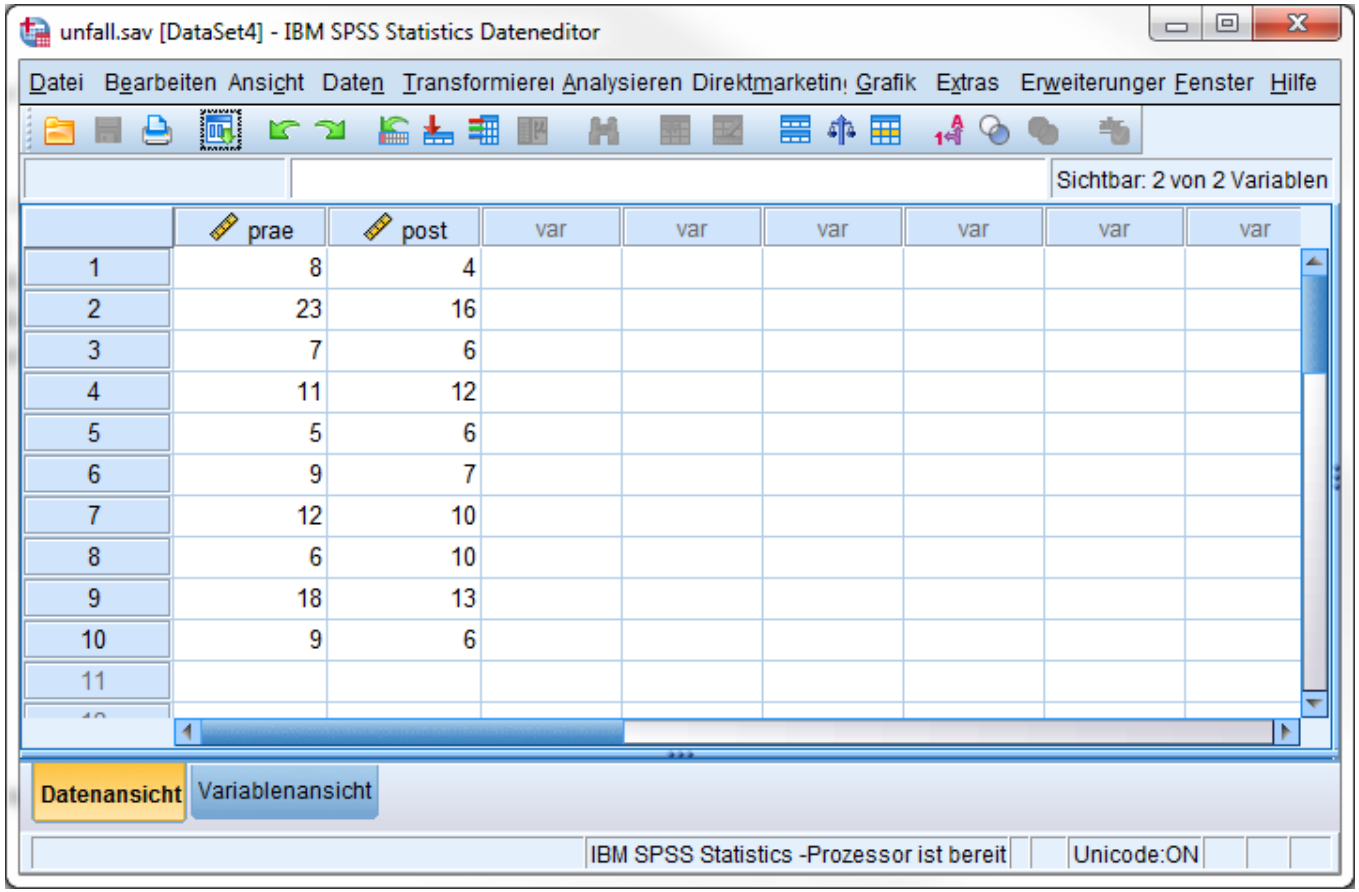
- Verglichen mit dem Vorzeichentest weist der Wilcoxon-Test unter den meisten Bedingungen eine höhere Power auf (vgl. Bortz, Lienert & Boehnke, 1990, S. 265). Dafür macht er die strengere Annahme der Ordinalskaliertheit der Differenzen. Diese ist bei kardinalen abhängigen Variablen (d.h. Anzahlen, wie im Beispiel) unproblematischer, aber bei bestimmten ordinalen Variablen schwieriger zu rechtfertigen. (Ist z.B. der Unterschied zwischen den Schulnoten 1 und 2 sicher kleiner als zwischen den Schulnoten 3 und 5?)

- Immer dann, wenn empirisch nur Informationen über die Richtung des Unterschieds existieren (z.B. bei der Frage an Schüler, ob ihnen Lehrer A oder B sympathischer ist) stellt der Vorzeichentest die einzige Alternative dar.

- Es existieren noch weitere nonparametrische Tests zum Vergleich zweier Gruppen, etwa auch im Hinblick auf Variabilitätsunterschiede, die wir hier nicht behandeln werden (vgl. Bortz, Lienert & Boehnke, 1990, Kap. 6 & 7; Diehl & Arbinger, 2001, Kap. 23 & 25)

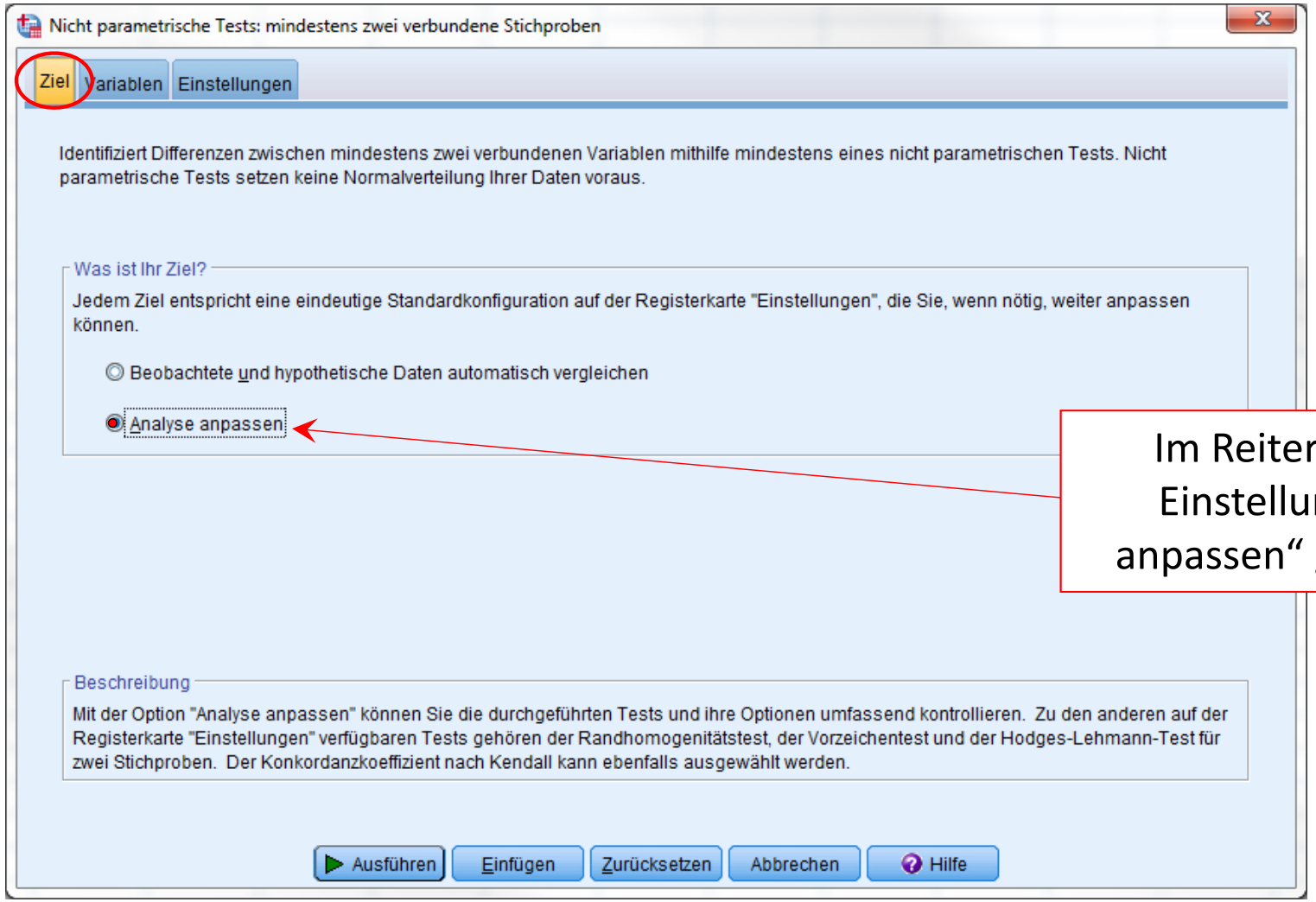
Nonparametrische Tests für abhängige Gruppen in SPSS

- Im Falle abhängiger Gruppen sind die Daten (wie schon beim t-Test für abhängige Gruppen) in SPSS so einzugeben, dass die zusammengehörigen Messwertpaare immer in einer Zeile von zwei Variablen stehen (hier SCHUL1 und SCHUL2).



Vorzeichen- & Wilcoxon-Test in SPSS

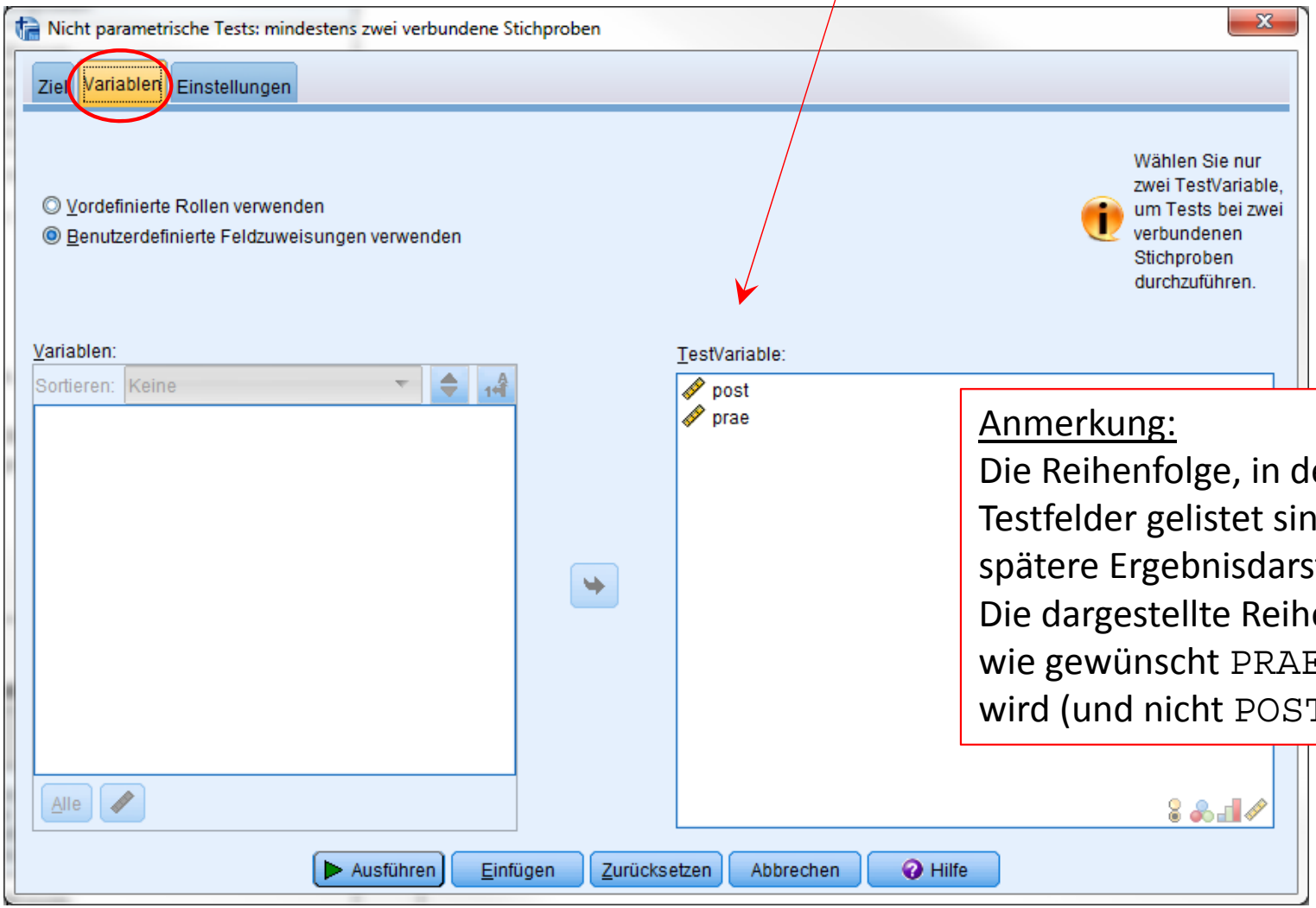
- Vorzeichen- und Wilcoxon-Test lassen sich in SPSS anfordern unter `Analysieren/ Nicht_parametrische_Tests/Verbundene_Stichproben…`



Im Reiter ⌈Ziel⌉ muss die Einstellung auf „Analyse anpassen“ geändert werden.

Vorzeichen- & Wilcoxon-Test in SPSS

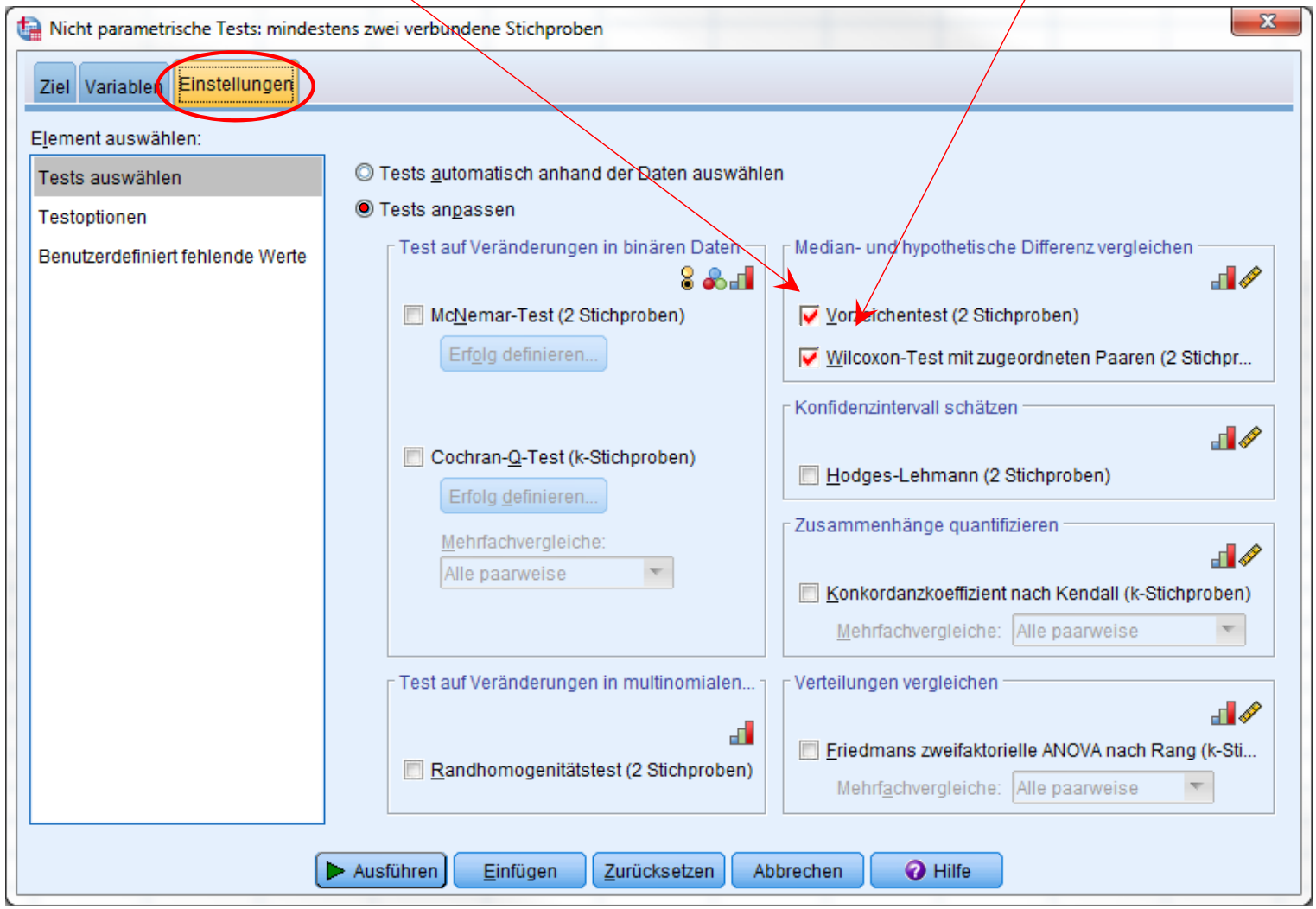
- Im Reiter ⌈Variablen⌉ müssen unter „Testvariable“ die beiden Variablen (hier: PRAE und POST) angegeben werden, die gegeneinander getestet werden sollen.



Anmerkung:
Die Reihenfolge, in der die Variablen unter Testfelder gelistet sind, wirkt sich auf die spätere Ergebnisdarstellung im Viewer aus. Die dargestellte Reihenfolge bewirkt, dass wie gewünscht `PRAE-POST` berechnet wird (und nicht `POST-PRAE`).

Vorzeichen- & Wilcoxon-Test in SPSS

- Im Reiter ⌈Einstellungen⌉ müssen die Option „Tests anpassen“ aktiviert und dann die Verfahren „Vorzeichentest (2 Stichproben)“ bzw. „Wilcoxon-Test mit …“ ausgewählt werden.



Nichtparametrische Tests

Hypothesentestübersicht

	Nullhypothese	Test	Sig.	Entscheidung
1	Der Median der Differenzen zwischen prae und post ist gleich 0.	Vorzeichentest bei verbundenen Stichproben	344,000[1]	Nullhypothese beibehalten
2	Der Median der Differenzen zwischen prae und post ist gleich 0.	Wilcoxon-Vorzeichenrangtest bei verbundenen Stichproben	102,000	Nullhypothese beibehalten

Asymptotische Signifikanzen werden angezeigt. Das Signifikanzniveau ist ,05.

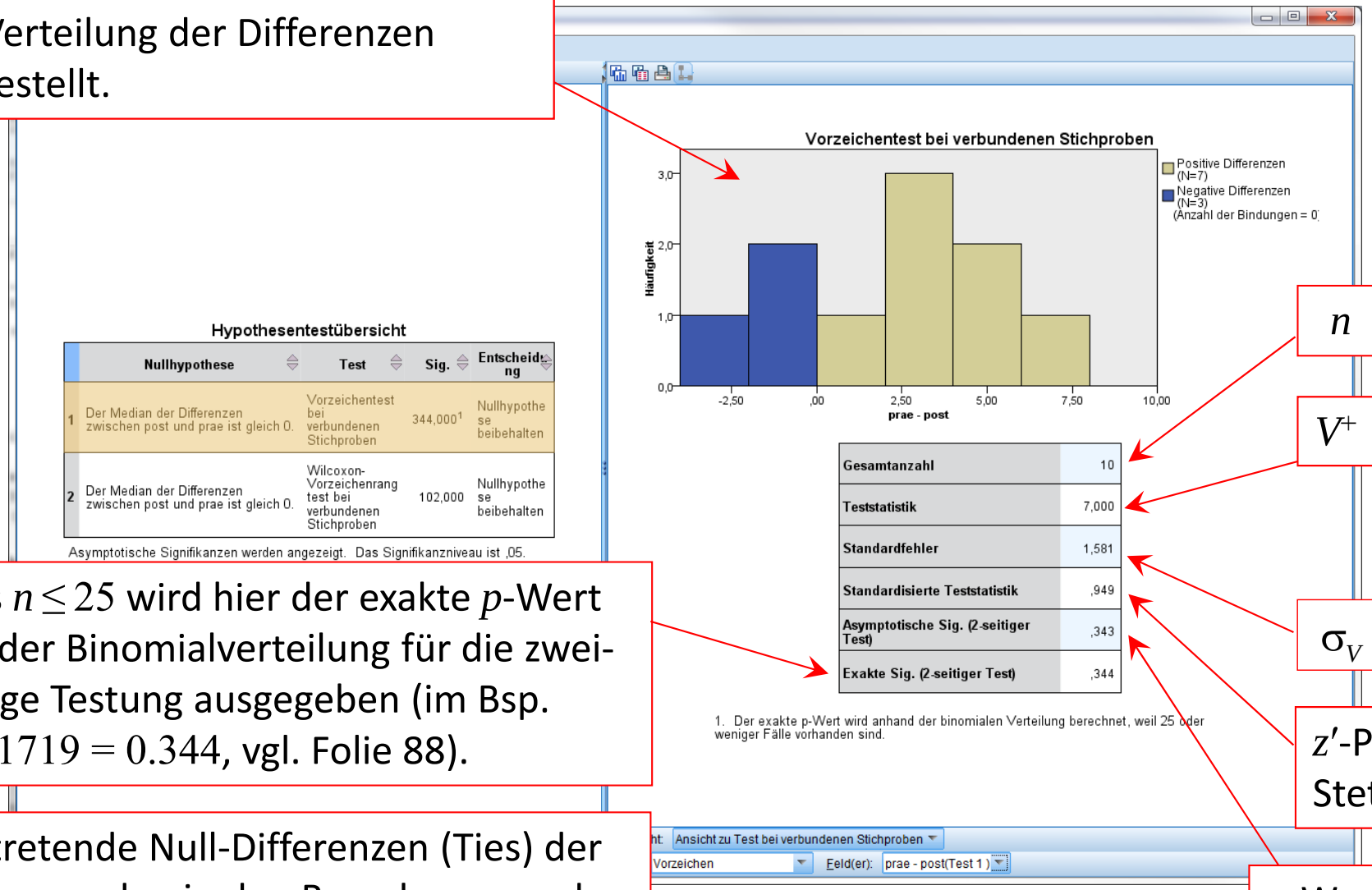
[1]Für diesen Test wird die exakte Signifikanz angezeigt.

Die ausgegebenen p-Werte (Spalte Sig) sind sowohl für den Vorzeichentest (.344) als auch für den Wilcoxon-Test (.102) bei zweiseitigem Testen größer als das vorab gewählte $\alpha = 0.05$ (Fehler in der Anzeige wie oben). Damit führen beide Tests zur Beibehaltung der H_0.

Durch dass Doppelklicken auf die betreffende Zeile der Tabelle öffnet sich die Modellansicht für das betreffende Verfahren.

Vorzeichen- & Wilcoxon-Test in SPSS

Beim Vorzeichentest wird rechts die Verteilung der Differenzen dargestellt.



Hypothesentestübersicht

	Nullhypothese	Test	Sig.	Entscheidung
1	Der Median der Differenzen zwischen post und prae ist gleich 0.	Vorzeichentest bei verbundenen Stichproben	344,000[1]	Nullhypothese beibehalten
2	Der Median der Differenzen zwischen post und prae ist gleich 0.	Wilcoxon-Vorzeichenrang test bei verbundenen Stichproben	102,000	Nullhypothese beibehalten

Asymptotische Signifikanzen werden angezeigt. Das Signifikanzniveau ist ,05.

Gesamtanzahl	10
Teststatistik	7,000
Standardfehler	1,581
Standardisierte Teststatistik	,949
Asymptotische Sig. (2-seitiger Test)	,343
Exakte Sig. (2-seitiger Test)	,344

1. Der exakte p-Wert wird anhand der binomialen Verteilung berechnet, weil 25 oder weniger Fälle vorhanden sind.

n

V^+

σ_V

z'-Prüfgröße (mit Stetigkeitskorrektur)

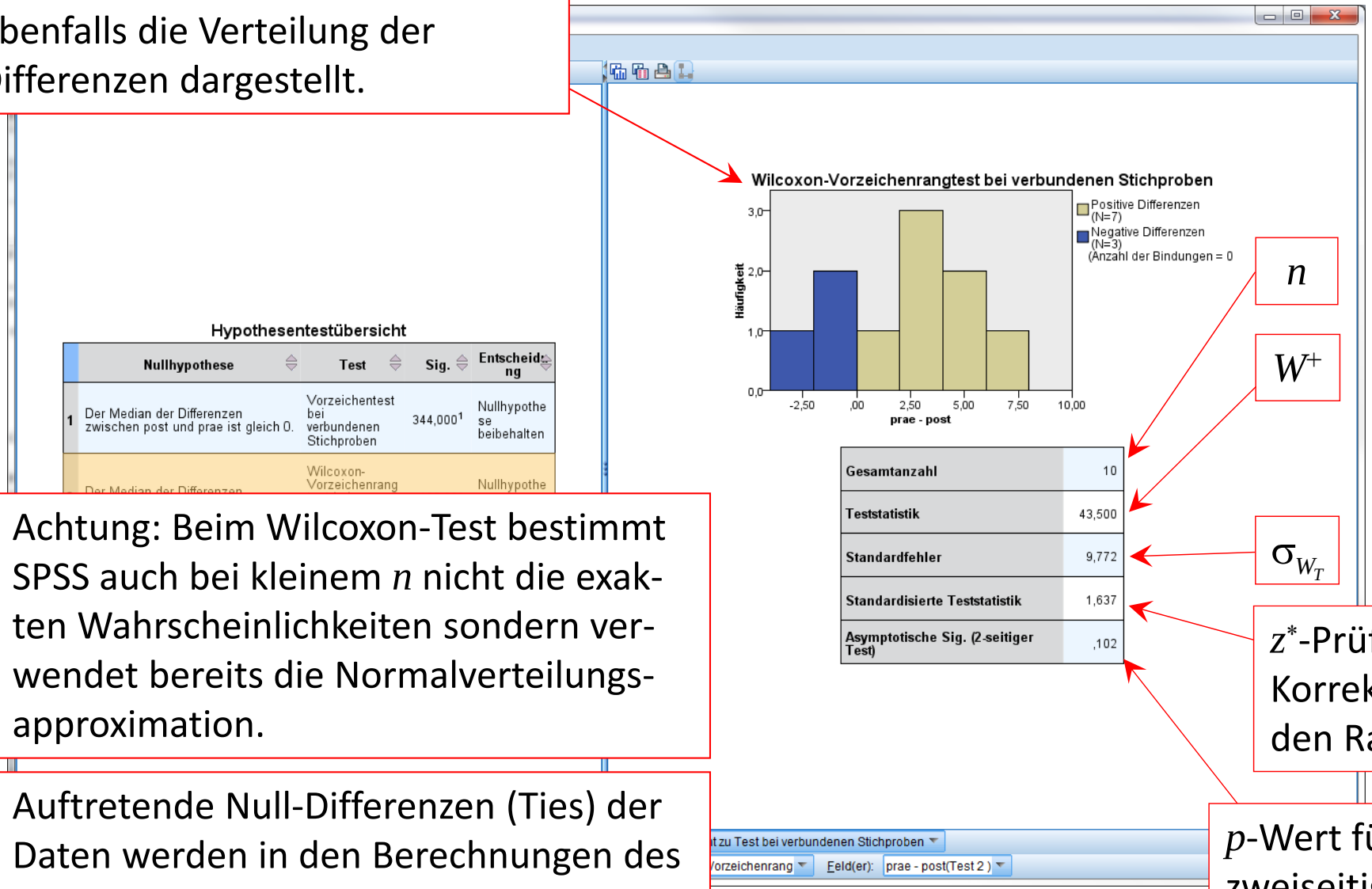
Falls $n \leq 25$ wird hier der exakte p-Wert aus der Binomialverteilung für die zweiseitige Testung ausgegeben (im Bsp. $2 \cdot 0.1719 = 0.344$, vgl. Folie 88).

Auftretende Null-Differenzen (Ties) der Daten werden in den Berechnungen des Vorzeichentests von SPSS ignoriert.

p-Wert für z'-Wert bei zweiseitiger Testung

Vorzeichen- & Wilcoxon-Test in SPSS

Beim Wilcoxon-Test wird rechts ebenfalls die Verteilung der Differenzen dargestellt.



n

W^+

σ_{W_T}

z^*-Prüfgröße (mit Korrektur für Ties in den Randsummen)

p-Wert für z^*-Wert bei zweiseitiger Testung

Achtung: Beim Wilcoxon-Test bestimmt SPSS auch bei kleinem n nicht die exakten Wahrscheinlichkeiten sondern verwendet bereits die Normalverteilungsapproximation.

Auftretende Null-Differenzen (Ties) der Daten werden in den Berechnungen des Wilcoxon-Tests von SPSS ignoriert.

Unterschiede in zentraler Tendenz in einer oder zwei Gruppen: Zitierte Quellen

Zitierte Quellen

- 🕮 Bortz, J., Lienert, G. A. & Boehnke, K. (1990). *Verteilungsfreie Methoden in der Biostatistik*. Berlin: Springer.
- 🕮 Howell, D. C. (2010). *Statistical methods for psychology* (7th ed.). Belmont, CA: Cengage Wadsworth.
- 🕮 Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego: Academic Press.

Unterschiede in zentraler Tendenz in einer oder zwei Gruppen: Errata

Änderungen am 10.01.2017

- Folie 66: Oben links fehlerhaftes Outputfenster von SPSS ausgetauscht.