Lagetests zum Vergleich 2er Gruppen - Die Qual der Wahl

Christel Weiß
Medizinische Fakultät Mannheim der
Universität Heidelberg
Abteilung für Medizinische Statistik und
Biomathematik
Theodor-Kutzer-Ufer 1-3
68167 Mannheim
christel.weiss@medma.uni-heidelberg.de

Zusammenfassung

In diesem Beitrag werden Lagetests zum Vergleich zweier Stichproben (Student-Test, t-Test nach Welch und Mann-Whitney-U-Test) verglichen. Anhand von Anwendungsbeispielen wird dargelegt, welche Effektgrößen sinnvoll sind, welche Voraussetzungen zu beachten sind und nach welchen Kriterien ein geeignetes Testverfahren auszuwählen ist.

Schlüsselwörter: t-Test für 2 unabhängige Stichproben, Wilcoxon-Test, Mann-Whitney-U-Test, Konfidenzintervall, AUC, Hodges-Lehmann-Schätzer

1 Einleitung

Im Jahre 1908 publizierte der englische Statistiker und Chemiker William Sealy Gosset (1876-1937) einen Beitrag, in dem er die t-Verteilung und den t-Test behandelte. Weil es ihm als Angestellten der Dubliner Brauerei Guinness untersagt war, eigene Arbeiten zu veröffentlichen, erschien diese Abhandlung unter dem Pseudonym „Student“ [1]. Deshalb wird die t-Verteilung häufig auch die „Studentsche Verteilung“ genannt.

Der Student-Test zum Vergleich der Mittelwerte zweier unabhängiger Stichproben erfreut sich nach wie vor großer Beliebtheit, obwohl seine Voraussetzungen formal sehr streng sind: Die Daten sollten normalverteilten Grundgesamtheiten mit gleichen Varianzen entstammen. Einige Jahre später gelang es dem britischen Statistiker Bernhard Lewis Welch (1911-1989), eine Variante des t-Tests zu entwickeln, die die Gleichheit der Varianzen (Homoskedastizität) nicht voraussetzt [2].

Der Grund für die Beliebtheit des t-Tests liegt in seiner Anschaulichkeit und leichten Handhabbarkeit. Insbesondere in der klinischen und der epidemiologischen Forschung treten häufig Fragestellungen auf, bei denen die Mittelwerte zweier Gruppen zu vergleichen sind (etwa bei randomisierten Therapiestudien, in denen die Wirksamkeit einer Therapie mittels einer stetigen Zielgröße erfasst wird). Der Unterschied zwischen den Gruppen lässt sich durch die Differenz der Mittelwerte oder der Effektgröße „Cohens d“

quantifizieren. Wenn allerdings die Voraussetzungen des t-Tests in grober Weise verletzt sind (etwa bei schiefen Verteilungen oder kleinen Stichproben), ist Vorsicht geboten: In diesen Fällen könnte das Testergebnis verzerrt sein und zu unzulässigen Schlussfolgerungen verleiten.

Als Alternative zum t-Test bietet sich der Wilcoxon-Test für zwei unverbundene Stichproben an, benannt nach Frank Wilcoxon (1892-1965) [3]. Dieser Test ist äquivalent zum Mann-Whitney-U-Test, der von den Mathematikern Henry Mann (1905-2000) und Donald Whitney (1915-2007) entwickelt wurde [4]. Dabei handelt es sich um einen nicht-parametrischen Rangsummentest, der keine bestimmte Verteilung voraussetzt. Viele Anwender wenden diesen Test unbesorgt als Lagetest an in der Annahme, dass er an keine Voraussetzungen gebunden sei. Doch stellt sich hier die Frage: Wie ist ein signifikantes Testergebnis des U-Tests zu interpretieren? Wie lässt sich ein Effekt quantifizieren?

Auf diese Fragen wird in diesem Beitrag eingegangen. Anhand von Anwendungsbeispielen werden die Besonderheiten jedes Tests dargelegt und es wird aufgezeigt, was bei deren Anwendung zu beachten ist.

2 Der t-Test für zwei unabhängige Stichproben

2.1 Student-Test

Gegeben seien zwei Stichproben der Umfänge n_1 und n_2 mit den Mittelwerten $\bar{x}$ und $\bar{y}$ sowie den Standardabweichungen s_1 und s_2. Die Prüfgröße des Student-Tests berechnet sich als:

$$t = \frac{\bar{x} - \bar{y}}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Dabei ergibt sich s als Quadratwurzel der „gepoolten“ Varianz, die aus den Stichprobenvarianzen berechnet wird als:

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

Die Anzahl der Freiheitsgrade beträgt $f = n_1 + n_2 - 2$.

Die Überprüfung der Voraussetzungen mag schwierig sein, da sich die Forderungen bezüglich Normalverteilung und Homoskedastizität auf die Grundgesamtheiten beziehen. Anhand der Stichproben kann allenfalls geschlussfolgert werden, dass nichts gegen die Annahme der Normalverteilung oder der Gleichheit der Varianzen spricht. Es stehen

zwar diverse Anpassungstests zur Prüfung der Normalverteilung zur Verfügung (z. B. der Kolmogoroff-Smirnov-Test, der Shapiro-Will-Test oder der Anderson-Darling-Test); die Varianzen lassen sich mit dem F-Test, dem Bartlett-Test oder dem Levene-Test vergleichen. Allerdings ist die Interpretation eines Anpassungstests oder eines Tests zur Überprüfung der Gleichheit der Varianzen problematisch. Bei kleinen Fallzahlen ist die Nullhypothese schwerlich zu verwerfen (was jedoch keinesfalls als Beleg für deren Richtigkeit gedeutet werden sollte). Dagegen tendieren hohe Fallzahlen dazu, dass die Forderungen generell in Frage gestellt werden – was aber nicht unbedingt bedeutet, dass die Voraussetzungen des t-Tests so stark verletzt sind, dass er unbrauchbar ist.

Weitere Anhaltspunkte bezüglich der Verteilungsform liefern graphische Darstellungen in Form eines Histogramms oder eines Box-and-Whisker-Plots. Zur Beurteilung der Verteilungsform eignen sich ferner die Schiefe oder der Vergleich zwischen Mittelwert und empirischem Median.

2.2 Welch-Test

Bei diesem Test [2] ist die Gleichheit der Varianzen keine Bedingung. Die Prüfgröße und die Anzahl der Freiheitsgrade berechnen sich als:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$f = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Die Formel für die Anzahl der Freiheitsgrade wurde von Franklin E. Satterthwaite entwickelt [5]; deshalb ist der Welch-Test auch als der t-Test nach Satterthwaite bekannt.

2.3 Vergleich Student- und Welch-Test

Da der Welch-Test weniger Voraussetzungen beinhaltet als der Student-Test, ist er vielseitiger verwendbar. Die beiden Tests und deren Ergebnisse sollen nun anhand eines Anwendungsbeispiels verglichen werden:

Im Rahmen einer nicht randomisierten Studie mit Adipositaspatienten wurden zwei OP-Techniken verglichen: Bei 129 Patienten wurde eine Bypass-OP durchgeführt; bei 46 Patienten wurde der Magen verkleinert („Sleeve-Gastrektomie“) [6]. Als primärer Endpunkt wurde der „Exzess Weight Loss“ (Reduktion des Übergewichts bezogen auf das Idealgewicht in Prozent) nach 12 Monaten erhoben. Darüber hinaus wurden weitere pa-

tientenspezifische Merkmale, diverse Laborparameter und sekundäre Zielgrößen erfasst. Die Tabelle 1 enthält statistische Kenngrößen sowie die Ergebnisse des Student-Tests und des Welch-Tests für die Parameter „Exzess Weight Loss“ (EWL, in %) und „Kreatinin“ (in *mg/dl*).

Tabelle 1: Vergleich der t-Tests nach Student und Welch. EWL = Exzess Weight Loss in %, Std = Standardabweichung

	EWL in %		Kreatinin im *mg/ml*	
	Bypass	Sleeve	Bypass	Sleeve
Fallzahl	$n_1 = 129$	$n_2 = 46$	$n_1 = 127$	$n_2 = 46$
Mittelwert ± Std	101,1 ± 66,4	72,3 ± 44,2	0,892 ± 0,386	1,124 ± 1,003
Minimum – Maximum	22 – 308	32 – 183	0,46 – 4,56	0,56 – 6,64
p-Wert (F-Test)	$p = 0{,}0024$		$p < 0{,}0001$	
Student-Test:				
Prüfgröße	$t = 2{,}73$		$t = -2{,}20$	
Freiheitsgrade	$f = 173$		$f = 171$	
p-Wert (Student-Test)	$p = 0{,}0069$		$p = 0{,}0290$	
Differenz der Mittelwerte mit Konfidenzintervall	28,8 ± 61,4 [8,0; 49,7]		-0,232 ± 0,612 [-0,440; -0,024]	
Welch-Test:				
Prüfgröße	$t = 3{,}29$		$t = -1{,}53$	
Freiheitsgrade	$f = 119{,}34$		$f = 49{,}907$	
p-Wert (Welch-Test)	$p = 0{,}0013$		$p = 0{,}1328$	
Differenz der Mittelwerte mit Konfidenzintervall	28,8 ± 61,4 [11,5; 46,2]		-0,232 ± 0,612 [-0,537; +0,073]	

Aus Tabelle 1 geht hervor: Bei der primären Zielgröße EWL führen zwar beide Testvarianten zu einem signifikanten Ergebnis. Der p-Wert des Welch-Tests ist jedoch kleiner als der p-Wert des Student-Tests. Das Konfidenzintervall des Welch-Tests ist schmaler; die Präzision der Schätzung präziser. Anders beim Laborparameter Kreatinin: Hier ergibt sich mit dem Welch-Test ein nicht signifikantes Ergebnis und ein breiteres Konfidenzintervall, während der Student-Test zu einem wesentlich kleineren p-Wert mit einem signifikanten Ergebnis führt.

Wie ist das zu erklären? Ein Blick auf Tabelle 1 zeigt, dass die größere Bypass-Gruppe bezüglich der EWL-Zielgröße heterogener ist als die Sleeve-Gruppe ($s_1 > s_2$), während bezüglich der Kreatinin-Werte die Bypass-Gruppe homogener ist ($s_1 < s_2$). Generell gilt:

- Bei gleichen Fallzahlen $n_1 = n_2$ und gleichen Standardabweichungen $s_1 = s_2$ stimmen die Ergebnisse beider Tests (Prüfgröße, Freiheitsgrad und p-Wert) überein.
- Bei gleichen Fallzahlen und ungleichen Standardabweichungen ($n_1 = n_2$, $s_1 \neq s_2$) ergeben sich identische Werte für die Prüfgröße. Jedoch weist der Student-Test die höhere Anzahl von Freiheitsgraden aus. Das führt zu einem geringeren p-Wert beim Student-Test.
- Falls die Standardabweichungen gleich sind, nicht jedoch die Fallzahlen ($n_1 \neq n_2$, $s_1 = s_2$), sind die Anzahl der Freiheitsgrade und die Prüfgröße des Student-Tests höher als die des Welch-Tests. Auch dies führt zu einem geringeren p-Wert beim Student-Test.
- Falls bei $n_1 \neq n_2$ die größere Stichprobe die geringere Variabilität aufweist, erhält man mit dem Student-Test den kleineren p-Wert (siehe Tabelle 1, Kreatinin).
- Falls bei $n_1 \neq n_2$ die größere Stichprobe die höhere Variabilität aufweist, kann keine eindeutige Aussage getroffen werden. Je mehr sich die beiden Stichproben bezüglich ihrer Variabilität unterschieden, desto eher ergibt sich mit dem Welch-Test der kleinere p-Wert (siehe Tabelle 1, EWL).

Selbstverständlich sollte man nicht generell denjenigen Test bevorzugen, der den kleineren p-Wert liefert, sondern die Auswahl der geeigneten Testvariante aufgrund von statistischen Überlegungen vorab treffen. Bei beiden Parametern (EWL und Kreatinin) weichen die Standardabweichungen der Therapiegruppen stark voneinander ab; das Ergebnis des F-Tests ist jeweils signifikant (siehe Tabelle 1). Deshalb bietet sich in beiden Fällen der Welch-Test an. Dennoch wurden aus didaktischen Gründen beide Tests durchgeführt, um die Unterschiede der Ergebnisse darzulegen.

2.4 t-Test nach logarithmischer Transformation

Mit Monte-Carlo-Studien wurde nachgewiesen, dass der t-Test auf geringe Verletzungen seiner Voraussetzungen robust reagiert [7]. Ähnlich große, nicht allzu geringe Fallzahlen (jedes $n \geq 30$), symmetrische empirische Verteilungen und annähernd gleiche Stichprobenvarianzen sind ideale Voraussetzungen für die Anwendung eines t-Tests, auch wenn die Grundgesamtheiten nicht perfekt normalverteilt sind (was sich in der Praxis ohnedies kaum überprüfen lässt). Was ist jedoch zu tun, wenn die Verteilungen ganz offensichtlich schief sind?

Dazu betrachten wir das Merkmal „Triglyzeride“ (in *mg/dl*) in der Adipositasstudie [6]. Ein Blick auf die Box-Plots in Abbildung 1 zeigt die Verteilungsformen. Aus Tabelle 2 geht hervor, dass in beiden Gruppen der Mittelwert wesentlich größer ist als der Median, dass die Daten weit streuen, und dass die Werte für die Schiefe deutlich größer als 0 sind. Dies weist auf eine linksgipfelige (rechtsschiefe) Verteilung hin, was durch einzelne, extrem hohe Werte in beiden Gruppen bedingt ist.

In solchen Fällen kann eine Datentransformation nützlich sein. Bei linksgipfeligen Verteilungen bietet sich die logarithmische Transformation an (wobei die Basis unerheblich ist). Die Triglyzerid-Daten wurden zur Basis e logarithmiert. Statistische Kenngrößen sind in Tabelle 2 aufgelistet. Aufgrund fehlender Werte sind die Stichprobenumfänge geringer als bei den Parametern in Tabelle 1.

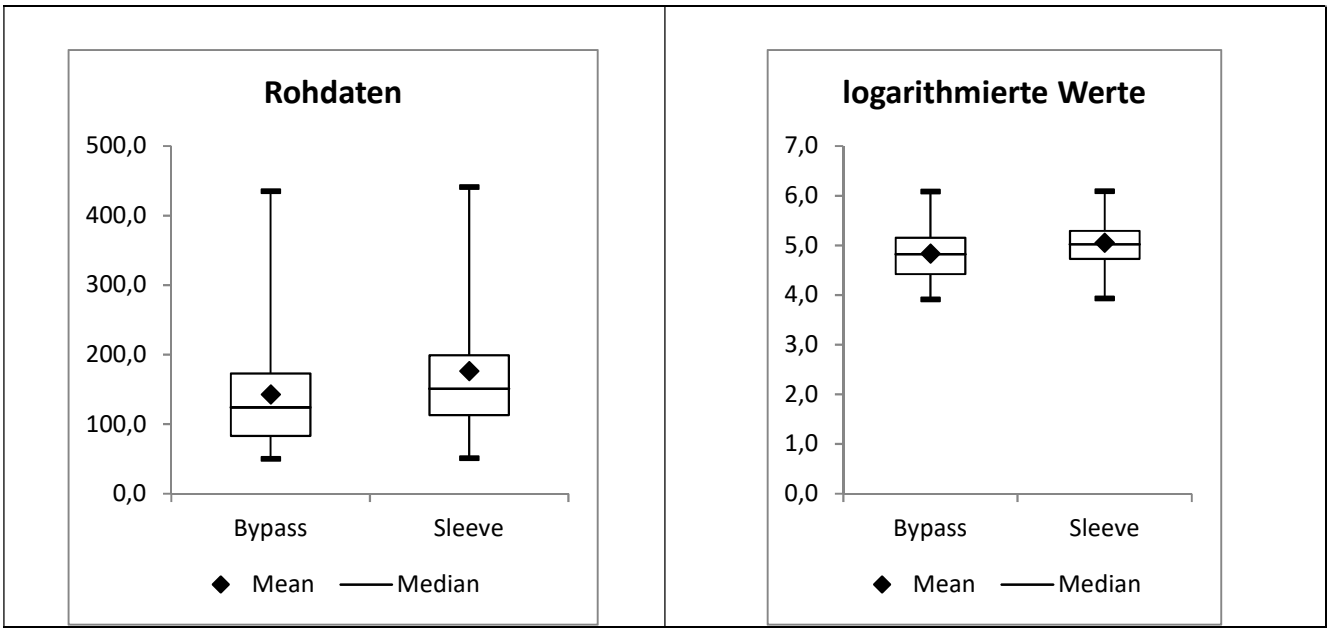

Abbildung 1: Box-Plots der Triglyceride-Werte für die Rohdaten (links) und deren Logarithmen (rechts)

Tabelle 2: Kenngrößen für die Variable „Triglyzeride“ und deren Logarithmen

	Rohdaten		Logarithmen	
	Bypass	Sleeve	Bypass	Sleeve
Fallzahl	$n_1 = 83$	$n_2 = 37$	$n_1 = 83$	$n_2 = 37$
Mittelwert ± Std	142,3 ± 76,9	175,9 ± 93,6	4,83 ± 0,50	5,05 ± 0,49
Median	124	151	4,82	5,02
Minimum – Maximum	50 – 435	51 – 441	3,91 – 6,08	3,93 – 6,09
Variationskoeffizient	54%	53%	10,3%	9,7%
Schiefe	1,62	1,36	0,19	0,24

Aus Tabelle 2 geht hervor:

- Die logarithmierten Daten scheinen in beiden Gruppen annähernd symmetrisch verteilt zu sein. Das wird deutlich anhand der Mediane, die sich nur unwesentlich vom jeweiligen Mittelwert unterscheiden, und anhand der Werte für die Schiefe, die nahe bei 0 liegen.
- In beiden Stichproben ist der Variationskoeffizient der logarithmierten Daten mit Werten um 10% wesentlich geringer als bei den Rohdaten mit Werten über 50%. Ferner sind die Standardabweichungen der logarithmierten Werte sehr ähnlich.

Aus diesen Gründen spricht nichts gegen die Anwendung des t-Tests nach Student. Es ergibt sich ein signifikantes Testergebnis mit $p = 0{,}0286$. Allerdings ist bei dessen Interpretation zu beachten, dass die logarithmierten Werte in die Analyse eingeflossen sind. Die daraus berechneten Mittelwerte (4,83 bzw. 5,05) scheinen wenig anschaulich zu sein. Für die Differenz dieser Mittelwerte ergibt sich -0,2166. Was besagt dieser Wert? Dazu folgende Überlegungen: Der arithmetische Mittelwert der logarithmierten Werte entspricht dem Logarithmus des geometrischen Mittels:

$$\frac{\sum_{i=1}^{n} \log(x_i)}{n} = \log \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

Demnach erhält man durch Exponieren eines Mittelwerts, der aus den logarithmierten Werten berechnet wurde, das geometrische Mittel der Originalwerte. Für die Triglyzerid-Werte ergeben sich:

$$\bar{x}_{geom} = e^{4{,}83} = 125{,}21 \text{ (Bypass-Gruppe)}$$

$$\bar{x}_{geom} = e^{5{,}05} = 156{,}02 \text{ (Sleeve-Gruppe)}$$

Aus der Differenz der Mittelwerte der logarithmierten Werte (4,83 bzw. 5,05) lässt sich also durch Exponieren das Verhältnis der geometrischen Mittelwerte bestimmen:

$$e^{(4{,}83-5{,}05)} = \frac{e^{4{,}83}}{e^{5{,}05}} = \frac{125{,}21}{156{,}02} \approx 0{,}8$$

Analog lassen sich die Konfidenzintervalle interpretieren, die basierend auf den logarithmierten Werten berechnet werden. Nun sind bei rechtsschiefen Verteilungen die geometrischen Mittelwerte ohnedies sinnvoller als die arithmetischen Mittelwerte. Man sollte allerdings – wenn man einen t-Test mit transformierten Daten durchführt – beachten, dass die Ergebnisse adäquat zu interpretieren sind.

2.5 SAS-Procedures

Die SAS-Procedure zur Durchführung eines t-Tests für zwei unverbundene Stichproben für das oben erwähnte Beispiel wird aufgerufen durch

```
PROC TTEST; CLASS OP; VAR EWL; RUN;
```

Mit dieser Procedure werden automatisch der Student-Test („pooled method“) und der Welch-Test („Satterthwaite“) durchgeführt. Neben diversen statistischen Kenngrößen werden standardmäßig 95%-Konfidenzintervalle für die beiden Mittelwerte, deren Differenz und für die Standardabweichungen ermittelt. Außerdem wird der p-Wert des F-Tests angezeigt, mit dem die Stichprobenvarianzen verglichen werden. – Für die An-

wendung eines Anpassungstests steht die SAS-Procedure `PROC UNIVARIATE` zur Verfügung. Die Prüfung auf Normalverteilung muss für beide Stichproben separat erfolgen:

```
PROC SORT; BY OP; RUN;
PROC UNIVARIATE NORMAL PLOT; BY OP; RUN;
```

Die `NORMAL`-Option bewirkt, dass mehrere Anpassungstests durchgeführt werden (Shapiro-Wilk-Test, Kolmogorov-Smirnov-Test, Test nach Cramer-von Mises und der Test nach Anderson-Darling). Die Option `PLOT` veranlasst die Darstellung eines Box-and-Whisker-Plots sowie eines „Normal Probability Plots“. Es wurde bereits oben erwähnt, dass diese Tests zwar Hinweise bezüglich der Verteilungsformen und der Gleichheit der Varianzen liefern. Jedoch sollte man sich bei der Wahl eines geeigneten Tests nicht ausschließlich an diesen Ergebnissen orientieren.

3 Der U-Test nach Mann und Whitney

3.1 Das Prinzip des U-Tests

Der U-Test setzt lediglich voraus, dass sich die Rohdaten in einer sinnvollen Reihenfolge anordnen lassen. Deshalb ist er für alle quantitativen und auch für ordinal skalierte Merkmale geeignet; eine bestimmte Verteilungsform ist (im Gegensatz zum t-Test) nicht erforderlich. Dieser Test bietet sich an, falls die Voraussetzungen des t-Tests nicht erfüllt sind oder (etwa wegen zu kleiner Stichprobenumfänge) nicht überprüft werden können.

Dabei stellt sich allerdings die Frage: Was genau vergleicht dieser Test? In vielen Lehrbüchern wird er als Lagetest zum Vergleich zweier Mediane aufgeführt. Das folgende Beispiel zeigt jedoch, dass dies zu Missverständnissen führen kann. Gegeben seien zwei Datenreihen mit jeweils 10 Werten:

Datenreihe 1:	1 2 3 4 5 5 5 5 5 5
Datenreihe 2:	4 5 5 5 5 5 6 7 8 10

Trotz der Gleichheit der Mediane (beide betragen 5) ergibt sich mit dem U-Test ein signifikantes Ergebnis mit $p = 0{,}0186$. Worin unterscheiden sich also die beiden Stichproben? Dieser Frage soll im Folgenden nachgegangen werden.

Eigentlich vergleicht der U-Test zwei Verteilungsfunktionen. Das Prinzip beruht auf der Auswertung von Rängen. Diese werden gebildet, indem man die Werte beider Stichproben entsprechend ihrer Größe sortiert und Rangzahlen zuweist, beginnend bei 1 für den kleinsten Wert bis zu $(n_1 + n_2)$ für den größten Wert. Falls zwei oder mehrere Stichprobenwerte übereinstimmen, werden mittlere Rangzahlen zugeordnet (in diesem Fall spricht man von „verbundenen Rängen“). Aus diesen Rangzahlen werden nun – separat für jede Stichprobe – Rangsummen gebildet. Frank Wilcoxon, der dieses Verfahren im

Jahre 1945 vorstellte [3], verwendete die Rangsumme der kleineren Stichprobe T als Prüfgröße. Mann und Withney, die zwei Jahre später ebenfalls einen Rangsummentest vorstellten, verwendeten die etwas kompliziertere Prüfgröße

$$U = n_1 \cdot n_2 + \frac{1}{2} \cdot n_1 \cdot (n_1 + 1) - T$$

(wobei angenommen wird, dass n_1 die Fallzahl der kleineren Stichprobe ist). Der Wilcoxon-Test und der U-Test von Mann und Whitney sind also äquivalent. Die Prüfgrößen scheinen auf den ersten Blick wenig anschaulich zu sein. Tatsächlich hat jedoch U eine konkrete Bedeutung. Dazu folgende Überlegungen:

- Seien x_i $(i = 1, \dots, n_1)$ die Werte der ersten und y_j $(j = 1, \dots, n_2)$ die Werte der zweiten Stichprobe. Es lassen sich $n_1 n_2$ Paare bilden mit je einem Partner aus jeder Stichprobe.
- Wenn jeder Wert der ersten Stichprobe kleiner ist als jeder beliebige Wert der zweiten Stichprobe (wenn also $x_i < y_i$ für alle i und j), ist die Rangsumme T die Summe der Zahlen 1 bis n_1, also $T = 1/2 \cdot n_1 \cdot (n_1 + 1)$ und $U = n_1 \cdot n_2$.
- Wenn dagegen $x_i > y_i$ für alle i und j, nimmt T mit $n_1 n_2 + 1/2 \cdot n_1 \cdot (n_1 + 1)$ den maximalen Wert an und $U = 0$.
- Unter der Voraussetzung, dass es keine verbundenen Ränge gibt, quantifiziert die Prüfgröße U die Anzahl der Paare (x_i/y_j), für die gilt: $x_i < y_j$.
- Demnach ist $U/(n_1 \cdot n_2)$ der Anteil der Paare mit $x_i < y_j$ und ist somit ein Schätzer für die Wahrscheinlichkeit, dass ein beliebiger Wert der ersten Stichprobe kleiner ist als ein beliebiger Wert der zweiten (größeren) Stichprobe.

3.2 Anwendungsbeispiele

Das Prinzip des U-Tests wird anhand der Zielgröße EWL bei der Adipositas-Studie [6] erläutert. Das EWL wurde bei 175 Patienten erfasst; demzufolge erstrecken sich die Rangzahlen zwischen 1 und 175. Die kleinere Gruppe umfasst die Patienten mit der Sleeve-OP ($n_1 = 46$). Deren Rangsumme beträgt $T = 3043$. Daraus berechnet sich

$$U = 46 \cdot 129 + \frac{1}{2} \cdot 46 \cdot (46 + 1) - 3043 = 3972$$

Um die Bedeutung von U nachvollziehen zu können, betrachten wir alle theoretisch denkbaren Paare, die sich mit je einem Partner aus der Bypass- und der Sleeve-Gruppe lassen. Deren Anzahl beträgt $n_1 \cdot n_2 = 46 \cdot 129 = 5934$. Es lässt sich nachweisen:

- Bei 3947 Paaren (66,51%) hat der Sleeve-Partner den kleineren EWL-Wert $(x_i < y_j)$.
- Bei 50 Paaren (0,84%) stimmen die EWL-Werte überein $(x_i = y_j)$.

- Bei 1937 Paaren (32,64%) hat der Sleeve-Partner den höheren EWL-Wert ($x_i > y_j$).

Die Prüfgröße $U = 3972$ setzt sich zusammen aus 3947 ($x_i < y_j$) und 25 ($x_i = y_j$). Die Anzahl der Paare mit $x_i = y_j$ fließt also zur Hälfte in die Berechnung von U ein. Dann ist $U/(n_1 \cdot n_2) = 3972/5934 = 66{,}9\%$ ein Schätzer für die Wahrscheinlichkeit, dass ein beliebiger Sleeve-Patient einen geringeren oder gleich hohen EWL hat als ein beliebiger Bypass-Patient. Das Testergebnis des U-Tests ist signifikant ($p = 0{,}0007$) wie auch das Ergebnis des t-Tests nach Welch ($p = 0{,}0013$) – allerdings mit einer anderen Bedeutung. Das Ergebnis des U-Tests besagt, dass bei einem Sleeve-Patienten meist ein geringerer EWL zu beobachten ist als bei einem Bypass-Patienten, und dass dieser Anteil signifikant größer ist als 0,5. Dagegen geht aus dem signifikanten Ergebnis des t-Tests hervor, dass bei einem Sleeve-Patienten im Durchschnitt ein geringerer EWL zu erwarten ist als bei einem Bypass-Patienten. Dieser Unterschied ist signifikant verschieden von 0.

Für das in Abschnitt 3.1 erwähnte Beispiel der beiden Datenreihen lassen sich aus den beiden Stichproben 100 Paare bilden. Die Werte der oberen Datenreihe seien x_i. Man kann leicht ermitteln, dass bei 63 Paaren $x_i < y_j$, bei 31 Paaren ist $x_i = y_j$ und bei 6 Paaren ist $x_i > y_j$. Dann ist $U = 63 + 31/2 = 78{,}5$.
Das bedeutet: Die Wahrscheinlichkeit, dass ein beliebiger Wert der ersten Reihe kleiner oder gleich einem beliebigen Wert der zweiten Datenreihe ist, wird auf 78,5% geschätzt. Dieser Unterschied ist beeindruckend – trotz der Gleichheit der empirischen Mediane.

3.3 Hodges-Lehmann-Schätzer

Es wurde bereits erwähnt, dass der U-Test nicht in jedem Fall zum Vergleich zweier Mediane geeignet ist. Andererseits eignen sich Mediane, um ordinal skalierte oder quantitative Variablen zu charakterisieren. Für den Vergleich zweier Gruppen bietet sich anstelle des direkten Vergleichs der empirischen Mediane die Schätzung eines Medians nach einer Methode von Hodges und Lehmann an [8]. Diese „Location Shift“ wird berechnet aus den $(n_1 \cdot n_2)$ Differenzen $x_i - y_j$.

Für die Abschnitt 3.1 aufgezeigte Datenreihen ergibt sich eine Location Shift -2, basierend auf 100 Paar-Differenzen. Für die Zielgröße EWL der Adipositasstudie ergibt sich der Wert 17%. Er besagt, dass bei einem zufällig ausgewählten Bypass-Patienten ein um 17% höherer EWL-Wert zu erwarten ist als bei einem zufällig ausgewählten Sleeve-Patienten. Nach einem Verfahren von Hollander und Wolfe [9] lassen sich Konfidenzintervalle für die Location Shift berechnen. Deren Grenzen werden in der Literatur auch als Moses-Intervallgrenzen bezeichnet.

3.4 SAS-Procedures für den U-Test

Der U-Test wird mit der SAS-Procedure `PROC NPAR1WAY` aufgerufen:

```
PROC NPAR1WAY WILCOXON HL; CLASS OP; VAR EWL; EXACT; RUN;
```

Die Wilcoxon-Option veranlasst die Durchführung des U-Tests. Im Output erscheinen die Prüfgröße T (nicht jedoch U) und drei p-Werte entsprechend drei unterschiedlicher Approximationen der Prüfgröße (Normalapproximation mit Stetigkeitskorrektur, t-Approximation und Chi^2-Approximation des Kruskal-Wallis-Tests). Die Stetigkeitskorrektur lässt sich mit der Option `CORRECT = NO` unterdrücken. Mit dem `EXACT`-Statement wird ein exakter Test durchgeführt (und ein vierter p-Wert berechnet), was allerdings bei hohen Stichprobenumfängen sehr rechen- und zeitintensiv werden kann.

Die `HL`-Option bewirkt die Berechnung des Location Shifts und des dazugehörenden Konfidenzintervalls. Wegen des `EXACT`-Statements werden zusätzlich exakte Intervallgrenzen berechnet. Diese Konfidenzintervalle sind in der Regel unsymmetrisch. – Um den Anteil $U/(n_1 \cdot n_2)$ zu bestimmen, bietet sich die Procedure `PROC LOGISTIC` an:

```
PROC LOGISTIC; MODELL OP = EWL; ROC EWL; RUN;
```

Die „Area under the curve“ (AUC) entspricht dem Anteil $U/(n_1 n_2)$, falls dieser größer ist als 0,5. Ansonsten ist $U/(n_1 n_2) = 1 - \text{AUC}$. Mit dem `ROC`-Statement lassen sich Konfidenzintervalle für die AUC ermitteln.

4 Zusammenfassung und Schlussfolgerungen

4.1 Voraussetzungen des t-Tests

Der Student-Test, der Welch-Test und der U-Test zählen neben dem Chi^2-Test und Fishers exaktem Test zu den bekanntesten und am häufigsten verwendeten Analysetechniken zum Vergleich zweier unverbundener Stichproben. Die Wahl eines geeigneten Tests sollte aufgrund der Fragestellung, des Studiendesigns und der Eigenschaften der zu analysierenden Daten erfolgen.

Der t-Test stellt nach wie vor den beliebtesten Lagetest dar. Dies ist durch seine Anschaulichkeit begründet: Der Vergleich zweier Mittelwerte und die Bedeutung des Testergebnisses leuchten unmittelbar ein. Die SAS-Procedure `PROC TTEST` stellt alle relevanten Informationen zur Verfügung. Auch in einfacheren Programmen wie beispielsweise in MS-Excel ist der t-Test implementiert, was freilich auch zu dessen Verbreitung beiträgt.

Die Überprüfung der Voraussetzungen des t-Tests ist zuweilen problematisch. Die Vielzahl der verfügbaren Tests zur Prüfung der Voraussetzungen (Normalverteilung, gleiche

Varianzen) macht dies nicht einfacher. Manchen Anwendern mag es verlockend erscheinen, nach dem Blick auf die Ergebnisse des F-Tests (das die TTEST-Procedure automatisch liefert) und eines Anpassungstests zur Prüfung auf Normalverteilung befriedigt festzustellen, dass die p-Werte über 0,05 liegen, um danach bedenkenlos den t-Test anzuwenden. Andere Anwender gehen noch einen Schritt weiter: Sie wenden diverse Tests zur Prüfung der Normalverteilung oder der Homoskedasdizität an in der Hoffnung, wenigstens ein nicht-signifikantes Ergebnis zu erhalten und werten dieses als „Beweis“ dafür, dass die jeweilige Voraussetzung erfüllt ist. Es wurde bereits darauf hingewiesen, dass dieses Vorgehen nicht sinnvoll ist. Da die p-Werte von der Fallzahl abhängen und da andererseits t-Tests bei Fallzahlen ab 30 robust gegenüber Verletzungen ihrer Voraussetzungen reagieren, ist die Aussagekraft dieser Ergebnisse sehr eingeschränkt. Bei kleinen Fallzahlen kann der ß-Fehler eines Anpassungstests oder eines F-Tests sehr hoch sein, weshalb zuweilen das Signifikanzniveau $\alpha = 0{,}10$ zugrunde gelegt wird (womit das Problem jedoch nicht wirklich entschärft wird).

Es sollte darauf verzichtet werden, quasi gewaltsam die Daten zu manipulieren: Unliebsame Ausreißer werden eliminiert, Daten werden logarithmiert oder in anderer Weise transformiert – so lange, bis die Histogramme einer Gauß’schen Glockenkurve nahekommen. Auch dieses Vorgehen ist von höchst zweifelhaftem Nutzen.

Normalverteilte Variablen sind in der Medizin keineswegs die Regel, sondern eher die Ausnahme. Lognormalverteilte Merkmale (wie das in Abschnitt 2.3 behandelte Merkmal „Triglyzeride“) lassen sich transformieren. Die Verteilungsformen dieser Merkmale, die nur positive Werte annehmen können, entstehen, wenn die Einflussfaktoren multiplikativ zusammenwirken. Als Beispiele seien Laborparameter, Blutdruckwerte, Körpergewicht, Lebensdauern oder Inkubationszeiten genannt. Eine Datentransformation sollte jedoch nicht allein dem Zwecke dienen, normalverteilte Daten zur Anwendung des t-Tests zu erhalten. Die logarithmische Transformation ist nur dann sinnvoll, wenn sich der Unterschied zwischen zwei Messwerten besser durch einen Quotienten als durch eine Differenz beschreiben lässt.

4.2 Voraussetzungen des U-Tests

U-Tests sind vielfältig verwendbar: Sie bieten sich an bei ordinal skalierten oder bei quantitativ-diskreten Merkmalen sowie bei quantitativ-stetigen, nicht normalverteilten Variablen, oder auch dann, wenn der quantitative Charakter der Daten zweifelhaft erscheint (etwa bei subjektiven Beurteilungen). In Abschnitt 3 wurde gezeigt, dass das aus der Prüfgröße U berechnete Maß $U/(n_1 n_2)$ oder die Location Shift als Effektgrößen geeignet sind. Allerdings werden diese Kenngrößen nicht standardmäßig von SAS ausgegeben; dafür muss die HL-Option in der PROC NPAR1WAY inkludiert bzw. PROC LOGISTIC gestartet werden. Möglicherweise ist darin ein Grund zu sehen, weshalb diese Maße in Zusammenhang mit dem U-Test bislang wenig Verwendung finden.

Der große Vorteil des U-Tests besteht darin, dass er weit weniger Voraussetzungen beinhaltet als der t-Test und deshalb sorgloser angewandt werden kann. Es ist nicht erforderlich, die Stichprobenvarianzen zu vergleichen oder die Verteilungsformen näher zu untersuchen. Ein einzelner oder ein paar wenige Ausreißer werden das Ergebnis eines U-Tests kaum beeinflussen, weil letztlich nur die Rangzahlen in die Teststatistik einfließen und nicht die Rohdaten.

Allerdings gibt es wesentliche Punkte, die beim U-Test zu beachten sind: Aus einem signifikanten Ergebnis kann nicht ohne Weiteres auf einen Unterschied der Mediane geschlossen werden. Es könnte auch aufgrund unterschiedlicher Streuungen oder Verteilungsformen zustande gekommen sein. Wenn der U-Test als Lagetest anstelle des t-Tests eingesetzt wird, muss deshalb vorab überprüft werden, ob die Verteilungsformen zumindest ähnlich sind. Der U-Test verliert an Power bei zahlreichen übereinstimmenden Werten, was zu verbundenen Rängen führt. Bei ordinal-skalierten Merkmalen mit einem überschaubaren Wertebereich ist möglicherweise der Trendtest nach Cochran-Armitage eine Alternative (implementiert in der SAS-Procedure `PROC FREQ`). Darüber hinaus steht noch der Median-Test zur Verfügung, bei dem die Stichprobenwerte mit dem Gesamtmedian verglichen werden (aufzurufen mit `PROC NPAR1WAY MEDIAN`). Dieser Test hat jedoch eine geringere Power als der U-Test und findet deshalb eher selten Verwendung.

4.3 Ausreißer

Ausreißer beeinflussen Mittelwerte und Standardabweichungen, insbesondere bei kleinen Stichproben. Bei geringen Stichprobenumfängen können sie das Ergebnis eines t-Tests massiv verzerren. U-Tests eignen sich dagegen auch für Datenreihen mit Ausreißern, da deren Einfluss durch die Bildung von Rängen abgeschwächt wird.

Wie sind Ausreißer zu handhaben? In keinem Fall sollte man sie entfernen, nur weil es sich um besonders große oder besonders kleine Werte handelt, die nicht zur Datenreihe zu passen scheinen. Die Elimination muss inhaltlich begründet sein. Messwerte, die offensichtlich nicht korrekt erfasst oder dokumentiert wurden, sind von der Analyse auszuschließen, weil sie das Ergebnis verfälschen würden. Auch wenn Daten von Beobachtungseinheiten stammen, die inhaltlich nicht zur Datenreihe passen, mag es vernünftig sein, sie zu eliminieren, weil ansonsten das Testergebnis an Aussagekraft verlieren würde. Dies betrifft beispielsweise Patienten in einer Therapiestudie, die etwa aufgrund von Komorbiditäten Besonderheiten aufweisen, die untypisch für die zu untersuchende Studienpopulation sind.

Dagegen ergeben sich in anderen Fällen aus Ausreißern mitunter wichtige Erkenntnisse: Sie zeigen den Wertebereich, innerhalb dessen ein Merkmal variiert. Wenn man diese Werte von der Analyse generell ausschließen würde, wäre das Ergebnis des statistischen Tests für die Praxis wenig relevant.

4.4 Multiple Testmethoden

Was macht die Normalverteilung so beliebt? Es geht dabei nicht nur um die Anwendung eines t-Tests, sondern um viel mehr Möglichkeiten. Normalverteilte Zielgrößen lassen sich sehr effizient auswerten: 1-faktorielle Varianzanalysen eignen sich zum Vergleich von mehreren Gruppen, mit 2- und mehr-faktoriellen Varianzanalysen lassen sich mehrere Einflussfaktoren simultan analysieren und Interaktionen (Wechselwirkungen) zwischen ihnen untersuchen. Zudem gibt es Varianzanalysen für Messwiederholungen und multiple Regressionsanalysen, mit denen sich der Einfluss mehrerer quantitativer und qualitativer Variablen auf elegante Weise analysieren lässt. Freilich ist bei der Anwendung dieser komplexen Modelle eine leistungsfähige Statistiksoftware und statistisches Know-How bzw. die Konsultation eines kompetenten Statistikers vonnöten.

Die Zweckmäßigkeit einer multiplen Analyse soll an folgendem Beispiel verdeutlicht werden: Der signifikante Unterschied beim Vergleich der beiden OP-Methoden bezüglich der Zielgröße EWL (siehe Abschnitt 2.3) ist keineswegs auf die Überlegenheit der Bypass-Technik zurückzuführen. Dieser Unterschied ist vielmehr dadurch bedingt, dass die zu vergleichenden Gruppen nicht strukturgleich sind: Die Bypass-Patienten hatten vor der Operation einen durchschnittlichen BMI von $(45{,}7 \pm 5{,}8)$ kg/m^2, während der Mittelwert der Sleeve-Patienten mit $(55{,}9 \pm 7{,}8)$ kg/m^2 wesentlich höher war. Mittels einer Covarianzanalyse konnte bestätigt werden, dass dies die eigentliche Ursache für die Differenz der EWL-Werte war ($p = 0{,}0126$). Der Einfluss der OP-Methode war darüber hinaus nicht mehr signifikant ($p = 0{,}4219$).

Die Analysetechniken, die auf Rängen basieren, sind bei Weitem nicht so flexibel. Es gibt zwar den Kruskal-Wallis-Test als Rangsummentest zum Vergleich von mehreren Stichproben; darüber hinaus wurden 2- und mehr-faktorielle nicht-parametrische Verfahren als Pendant zu Varianzanalysen entwickelt [10]. Eine Covarianzanalyse wie oben beschrieben, bei der ein qualitativer Faktor (OP-Methode) zusammen mit einer stetigen Variablen (BMI) analysiert wird, ist jedoch mit Rangsummenverfahren nicht durchführbar. Dies mag erklären, weshalb die Prüfung der Normalverteilung hin und wieder ein wenig lax gehandhabt oder stillschweigend vorausgesetzt wird. Dennoch zeigt ein Vergleich der p-Werte, welche Einflussgröße kausal mit der Zielgröße assoziiert ist und welche einen Confounder darstellt.

4.5 Studiendesign

Ein adäquates Studiendesign kann dazu beitragen, dass die Datenanalyse effizient durchgeführt wird und dass unzulässige Schlussfolgerungen vermieden werden. Durch Randomisierung entstehen gleich große, strukturgleiche Gruppen. Dadurch wird der Einfluss von Confoundern verhindert. Zudem lassen sich gleich große Gruppen effizienter analysieren als Gruppen mit stark unterschiedlichen Fallzahlen. Präzise Ein- und Ausschlusskriterien tragen zu homogenen Gruppen bei. Mit adäquaten Messverfahren und sorgfältiger Dokumentation werden fehlerhafte Werte (und dadurch bedingte Aus-

reißer) vermieden. Die statistischen Verfahren sollten a priori ausgewählt werden. Freilich sollte auch sichergestellt sein, dass die Fallzahl ausreichend hoch ist, um einen klinisch relevanten Effekt nachzuweisen.

Letzten Endes muss die Wahl eines geeigneten Tests empirisch erfolgen, und der Anwender muss diese Wahl plausibel begründen. In jedem Fall sollte zusätzlich zum p-Wert eine Effektgröße angegeben werden, idealerweise zusammen mit einem Konfidenzintervall. Nur so kann der Leser einer Publikation die klinische Relevanz oder die wissenschaftliche Bedeutung des Testergebnisses beurteilen und die Präzision einer Schätzung erkennen. Der p-Wert ist vom Stichprobenumfang abhängig und daher für die Beurteilung dieser Fragen unzureichend.

Diese Ausführungen belegen, dass bei der Planung und der Durchführung einer Studie Statistiker und Vertreter des jeweiligen Fachgebietes (z. B. Kliniker) zusammenarbeiten sollten. Nur so kann es gelingen, die enorme Datenmenge, die während der Durchführung einer Studie anfällt, sinnvoll zu strukturieren, effizient zu analysieren und so zu neuen Erkenntnissen zu gelangen. Dies erscheint aus wissenschaftlichen, ökonomischen und ethischen Gründen unabdingbar zu sein.

Literatur

[1] Student: The Probable Error of the Mean. *Biometrika* 6(1): 1-25, 1908.

[2] B. L. Welch: The generalization of „Student's" problem when several different population variances are involved. *Biometrika* 34(1-2): 28-35, 1947

[3] F. Wilcoxon: Individual Comparisons by ranking Methods. *Biometrics* 1: 80-83, 1945

[4] H. Mann, D. Whitney: On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* 18: 50-60, 1947

[5] F. E. Satterthwaite: Synthesis of variance. *Psychometrika* 6(5), 309-316, 1941

[6] M. Otto, M. Elrefai, J. Krammer, C. Weiß, P. Kienle, T. Hasenberg: Sleeve Gastrectomy and Roux-en-Y Gastric Bypass Lead to Comparable changes in Body Composition After Adjustment for Initial Body Mass Index. *Obes Surg*, 26 (3), 479-485, 2016

[7] L. L. Havlicek, N. L. Peterson: Robustness of the t-Test: A guide for researchers on effect of violations of assumptions. *Psychol Reports* 34, 1095-1114, 1974

[8] J. L. Hodges, E. L. Lehmann: Estimation of location based on rank tests. *Ann Math Stat* 34(2): 598-611, 1963

[9] M. Hollander, D. A. Wolfe: Nonparametric statistical methods. 2nd ed. New York: John Wiley & Sons, 1999

[10] E. Brunner, U. Munzel: Nicht-parametrische Datenanalyse. Unverbundene Stichproben, 2. Auflage. Springer-Spektrum-Verlag, Berlin Heidelberg, 2013